UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

FACTORES QUE INFLUYEN EN EL DESEMPEÑO DE UN EQUIPO DE LA NBA: UNA APLICACIÓN DE REGRESIÓN LOGÍSTICA Y COMPONENTES PRINCIPALES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A :

CARLOS IVAN MADRIGAL NAVARRO

DIRECTOR DE TESIS:

DR. RICARDO RAMÍREZ ALDANA

Ciudad Universitaria, Cd. De México, 2019

Datos del Jurado

1. Datos del Alumno
   Madrigal
   Navarro
   Carlos Ivan
   33 19 42 92 25
   Universidad Nacional
   Autónoma de México
   Facultad de Ciencias
   Actuaría
   414061161

2. Datos del tutor
   Dr.
   Ricardo
   Ramírez
   Aldana

3. Datos del sinodal 1
   Dra.
   Lizbeth
   Naranjo
   Albarrán

4. Datos del sinodal 2
   M. en C.
   José Salvador
   Zamora
   Muñoz

5. Datos del sinodal 3
   M. en C.
   Graciela
   Martínez
   Sánchez

6. Datos del sinodal 4
   Act.
   Miguel Ángel
   Chong
   Rodríguez

7. Datos del trabajo escrito
   Factores que influyen en el desempeño de un equipo de la NBA: una aplicación de regresión logística y componentes principales
   241 p
   2019

Agradecimientos y motivación

La motivación por realizar el presente trabajo tiene sus raíces desde la infancia del autor. Siendo un apasionado del deporte del baloncesto y fanático de uno de los mejores equipos en esa época, los Phoenix Suns, el autor ha sido testigo de ver a su equipo caer desde el tercer mejor equipo de la liga, hasta los puestos más bajos, con cada intento de regresar a la élite de la liga o simplemente al rango de ser competitivos, fracasado durante ocho años consecutivos.

Dicha larga sequía, aunado con el hecho de que el actual dueño del equipo, Robert Sarver, se ha rehusado frecuentemente a invertir fuertes cantidades monetarias para mejorar el equipo, han llevado al autor del documento a formularse la pregunta de cómo, dados los problemas anteriormente mencionados, poder mejorar al equipo de una manera que no implique derrochar dinero por jugadores populares. Dicha pregunta ha sido la causa de la motivación del autor por realizar el presente estudio, pues él piensa que podría existir una solución matemática que si bien podría no llevar al equipo a un campeonato de la NBA (dado que podrían influir mil factores para esto), sí podría hacerlo regresar a la élite de la liga, todo esto contemplando las restricciones económicas del equipo. Con el fin de encontrar dicha solución matemática, se ha motivado el autor a realizar el presente documento.

Antes de iniciar con el presente documento, el autor del mismo quisiera plasmar un agradecimiento a un grupo de personas, ya que si por algo se logró realizar satisfactoriamente el presente documento, fue por ellos. El autor del documento agradece profundamente a su familia (papá, mamá y Emma) por todo el sacrificio y apoyo realizado a lo largo de estos cinco años, para que el autor pudiera estudiar la carrera que el decidió en la universidad que el decidió. El autor sabe que no fue sencillo para ellos, tanto económicamente como sentimentalmente, y es por eso que agradece profundamente por ello. Además, el autor agradece profundamente a su pareja, Chiara, por ser su pilar, apoyo y mano derecha en todo el tiempo que han estado juntos. Por la fiel compañía en los momentos de frustración, enojo, alegría, tristeza y festejo. Por estar ahí en todos ellos. Y por último, pero nunca menos importante, el autor agradece grandemente a su

asesor y amigo, Ricardo Ramírez Aldana. Gracias por tanto tiempo dedicado a este proyecto, tanta paciencia y tanta entrega de su parte al mismo. Sin la ayuda del mismo, este trabajo nunca hubiera tenido la calidad que posee, y el autor del trabajo nunca hubiera logrado tener la misma cantidad de aprendizajes que logró obtener. El autor del documento estará por siempre agradecido con él.

Finalmente, hubo una gran cantidad de personas que han ayudado y motivado al autor a realizar el presente trabajo a pesar de todas las adversidades que se presentaron a lo largo del camino. A todas ellas, el autor agradece de todo corazón su apoyo y sabe que de no ser por ellos, el camino para realizar dicho trabajo hubiera sido más complicado. Gracias sinceras.

Introducción

El presente trabajo es orientado a la aplicación práctica de dos de las técnicas estadísticas de mayor importancia en la actualidad, como es el caso de la regresión logística multinomial y el análisis de componentes principales, para encontrar información relevante respecto al comportamiento de los equipos de la NBA (*National Basketball Association*), liga profesional estadounidense de baloncesto y la de mayor popularidad respecto a ese deporte en el mundo entero.

Dicho trabajo tiene por objetivos encontrar un modelo estadístico que sea capaz de realizar dos funciones a la vez de manera satisfactoria, que son el predecir correctamente el nivel de desempeño de un equipo de la NBA, y el encontrar el tipo de influencia que tienen las variables respecto a dicho nivel de desempeño, con el fin de localizar a las que pudieran tener el mayor impacto positivo para algún equipo de la NBA. En otras palabras, se busca un modelo que pueda desempeñarse de excelente manera tanto en aspectos de ajuste, como en poder predictivo.

Dicho objetivo parecería a simple vista ser ambicioso, dado que la mayoría de los modelos son creados enfocándose en sólo uno de los dos aspectos mencionados anteriormente. En esto radica la importancia y dificultad de este estudio.

Además de los aspectos prácticos de los modelos, también se tendrá importancia en presentar a minucioso detalle la teoría detrás de las dos herramientas estadísticas mencionadas anteriormente, ya que para poder aplicar cualquier técnica estadística a un conjunto de datos, es imprescindible el conocimiento de la derivación de dicha técnica, así como sus supuestos y limitaciones.

En los primeros cuatro capítulos se presentará la teoría detrás del modelo de regresión logística multinomial, así como algunos consejos para intentar obtener el mayor provecho de dicho modelo. Posteriormente, el capítulo cinco presentará la teoría necesaria para el entendimiento del análisis de componentes principales. En el sexto capítulo se dará una breve introducción a las estadísticas de la NBA que se utilizarán en el estudio en cuestión. En el séptimo capítulo se procede al análisis que permitirá lograr los dos objetivos del presente trabajo de una manera satisfactoria. Finalmente, en el octavo capítulo se presentarán las conclusiones obtenidas.

Índice general

Capítulo 1

Modelo de Regresión Logística Multinomial

1.1. Introducción

La regresión logística multinomial es una herramienta ampliamente utilizada como método para explicar el comportamiento de una variable categórica a través de un conjunto variables.

En vanas palabras, lo que dicha herramienta realiza es estimar probabilidades que dicten qué tan propensa es una observación, dados los valores que posee en las variables explicativas, a pertenecer a cada una de las categorías de la variable dependiente.

Algunos ejemplos de aplicaciones de dicha técnica engloban la propensión de los pacientes a padecer cierta enfermedad en el área de ciencias de la salud, el conocer las características de las personas que tienen preferencia sobre algún candidato o partido en el área de ciencias políticas, o detección de posibles operaciones fraudulentas en áreas industriales y financieras.

Antes de iniciar con el estudio del modelo de regresión logística multinomial, es importante entender el objetivo de este análisis: Encontrar el modelo que mejor describa la relación entre una variable dependiente categórica (también llamada de respuesta) y una combinación lineal de variables independientes (también llamadas explicativas o predictoras). A la vez, dicho modelo debe de cumplir con los requerimientos de acuerdo a los objetivos específicos del estudio, como ajuste, poder predictivo, principio de parsimonia, etcétera. Cabe mencionar que de los requerimientos mencionados anteriormente, algunos son fundamentales en cualquier modelo.

1.2. El modelo

La regresión logística forma parte de los llamados modelos lineales generalizados. Esto significa que el modelo, es decir, la ecuación donde se relaciona a la variable dependiente (en este caso la categoría) con el conjunto de variables independientes es de la forma:

$$f(\boldsymbol{E}[y]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_P x_P \tag{1.1}$$

Con $f(\boldsymbol{E}[y])$ una función de la esperanza de la variable dependiente y, la cual tiene una distribución asociada a la familia exponencial, β_p un coeficiente asociado a la p-ésima variable, P el número de variables independientes y x_p la p-ésima variable independiente.

La función de la esperanza de la variable dependiente que es igualada a la combinación lineal de variables independientes varía de acuerdo al tipo de modelo lineal generalizado, pero para el caso del modelo de regresión logística, dicha función recibe el nombre de "función *logit*", y corresponderá al logaritmo natural del cociente de dos probabilidades. No debe de causar preocupación el no tener claridad sobre las definiciones mencionadas anteriormente, pues más adelante se explicarán a detalle.

Cuando se tienen dos posibles valores para la variable dependiente y se utilizara a la función *logit*, la ecuación 1.1 corresponde al modelo de regresión logística. Pero si se llegaran a presentar más de dos categorías en la variable dependiente, entonces el modelo de regresión logística a utilizar tendría la siguiente forma:

$$f(\boldsymbol{E}[y_j]) = \beta_{j0} + \beta_{j1} x_1 + \cdots + \beta_{jP} x_P \tag{1.2}$$

Con $f(\boldsymbol{E}[y_j])$ la función *logit* descrita anteriormente pero asociada a la categoría j y β_{jp} el coeficiente asociado a la p-ésima variable y a la j-ésima categoría. Es decir, se tendrá una ecuación por cada categoría de la variable de respuesta, menos uno.

A este tipo de modelo, con más de dos categorías, se le llama modelo de regresión logística multinomial, y será de ahora en adelante el mayor enfoque en que se analizará a la regresión logística en este trabajo.

Una de las características del modelo de regresión logística (sea multinomial o no) es que el modelo se basa, entre otras cosas, en las probabilidades de que cierta observación pertenezca a cada una de las posibles categorías de la variable dependiente. Así pues, se define a π_{ij} como la probabilidad de que la i-ésima

observación tome la categoría j dados sus valores en las variables independientes. Por lo tanto, se tendrá una probabilidad por cada categoría de la variable de respuesta.

Con esto puesto sobre la mesa, sea N el número de observaciones, J el número de categorías que puede tomar la variable de respuesta y π_{ij} definida como en el párrafo anterior, entonces se define a la función *logit* como:

$$f(\boldsymbol{E}[y_{ij}]) = \log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) \tag{1.3}$$

Con $i = 1, 2, \ldots, N$ y $j = 1, 2, \ldots, J-1$. Es decir, la función *logit* para la observación i y la categoría j corresponde al logaritmo natural del cociente entre la probabilidad de pertenecer a esa categoría (π_{ij}) y la probabilidad de pertenecer a una categoría fija, misma que recibirá el nombre de categoría de referencia. A partir de ahora, la categoría J fungirá como categoría de referencia en el presente documento.

Nótese también, que la esperanza de que la i-ésima observación pertenezca a la categoría j, $\boldsymbol{E}[y_{ij}]$, corresponde a la probabilidad de que la observación pertenezca a la categoría j, π_{ij}, multiplicada por una constante N.

Ahora, asumiendo que se tienen P variables explicativas, el modelo de regresión logística multinomial corresponde a:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \beta_{0j} + x_{i1}\beta_{1j} + x_{i2}\beta_{2j} + \ldots + x_{iP}\beta_{Pj} \tag{1.4}$$

Donde x_{ip} corresponde al valor de la variable p para la observación i y β_{pj} corresponde al coeficiente asociado a la variable p para la categoría j, con $p = 0, 1, \ldots, P$ y $j = 1, 2, \ldots, J-1$. El valor β_{0j} aparece sin variable porque este coeficiente será el asociado al intercepto, por lo que $x_{i0} = 1$ para toda $i = 1, 2, \ldots, N$.

Por simples cuestiones de practicidad, se denotará de ahora en adelante a la función *logit* (también llamada únicamente "el *logit*") correspondiente a la i-ésima observación y la j-ésima categoría como:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = g(x_{i\cdot})_j \tag{1.5}$$

con $i = 1, \ldots, N$ y $j = 1, \ldots, J$.

Ahora, a partir del modelo logístico multinomial definido en la ecuación 1.4, puede obtenerse que la probabilidad de que la observación i pertenezca a la categoría j, con $j = 1, \ldots, J-1$ es:

$$\pi_{ij} = \frac{e^{g(x_{i\cdot})_j}}{1 + \sum_{v=1}^{J-1} e^{g(x_{i\cdot})_v}}$$

Y que la probabilidad de que la observación pertenezca a la categoría de referencia, denotada por J, es:

$$\pi_{iJ} = \frac{1}{1 + \sum_{v=1}^{J-1} e^{g(x_{i\cdot})_v}}$$

Las ecuaciones anteriores se demostrarán en la siguiente sección del capítulo.

Una de las grandes ventajas de la regresión logística multinomial, es que permite utilizar como variables predictoras (o explicativas) a cualquier tipo de variables, ya sean continuas, categóricas, etc. Pero si alguna variable fuese categórica, al igual que en el caso del modelo de regresión lineal, se tratará a dicha variable de manera distinta a las demás dentro del modelo.

El tratamiento que se les aplicará a dichas variables será el utilizar "variables de diseño", mejor conocidas como "variables *dummy*". De manera general, si la variable categórica a utilizar contara con L posibles niveles y únicamente pudiera tomar un sólo valor, entonces se crearán $L - 1$ variables de diseño. A continuación se procede a explicar dicho tratamiento.

Supóngase que se tuviera interés en introducir dentro de un modelo como variable predictora a la variable continente, y dicha variable únicamente pudiera tomar los valores "África", "Asia" y "Europa". Entonces, el siguiente paso sería elegir una de esas 3 categorías y tomarla como la categoría de referencia.

Supóngase que, además, se eligió como categoría de referencia al continente asiático. Entonces se le destinará una variable de diseño a cada una de las categorías (también llamadas "niveles") que no fueron elegidas como categoría de referencia, en este caso a África y Europa. Ahora, si una observación tuviese el valor África, entonces la variable de diseño asignada a África tendría valor igual a uno, mientras que la variable de diseño Europa tendrá valor cero. Si la observación tuviese el valor Europa, entonces se revertirían los valores de las variables de diseño; es decir, a la variable de diseño correspondiente a África se le asignaría un cero y a la de Europa un uno. Si en cambio, el valor de la observación fuese la categoría de referencia, Asia, entonces todas las variables de diseño tendrán valor cero.

Ahora se generalizará lo explicado anteriormente. Supóngase ahora un modelo con P variables independientes y con la variable de respuesta de J niveles. También supóngase que la k-ésima variable independiente es categórica, y que ésta pudiera tomar uno de entre L niveles. Al tener la variable k un total de $L - 1$ variables de diseño, se denotará por D_{ikl} al valor de la l-ésima variable de diseño correspondiente a la variable independiente k, para la observación i, mientras que los coeficientes de dichas variables de diseño se denotarán por β_{klj}, con los subíndices de esta última expresión indicando que el coeficiente pertenece a la

l-ésima variable de diseño asociada a la k-ésima variable explicativa, y que dicho coeficiente pertenece al grupo de coeficientes asociados al j-ésimo nivel de la variable de respuesta (recuerde que la variable dependiente puede tomar una de entre J categorías).

Así pues, el *logit* asociado a este modelo para la variable de respuesta j y la observación i se expresaría de la forma:

$$g(x_{i\cdot})_j = \sum_{p\neq k}^{P} x_{ip}\beta_{pj} + \sum_{l=1}^{L-1} D_{ikl}\beta_{klj}$$

Con $i = 1, \ldots, N$ y $j = 1, \ldots, J-1$, donde N representa el número de observaciones a utilizar en el modelo.

De manera general, se suprimirá la suma de las variables de diseño, esto con los fines de expresar a la función *logit* como una única suma, contemplando que las variables asociadas a la suma pudieran ser continuas o categóricas.

1.3. Estimación de parámetros vía Máxima Verosimilitud

Antes de iniciar con el desarrollo del método para calcular los coeficientes del modelo, es de urgencia presentar motivos que expliquen la gran importancia del modelo de regresión logística multinomial al tener una variable categórica como variable de respuesta.

A alguien se le podría haber ocurrido la idea de, en vez de utilizar el modelo de regresión logística multinomial para el análisis, convertir los niveles de la variable de respuesta a números enteros y utilizar el método de regresión lineal, esto con el fin de evitar tener que recurrir a una herramienta diferente. Sin embargo, tal decisión sería un grave error.

Algunos de los motivos por los cuales no se debe de realizar dicha práctica son:

- En modelos de regresión lineal, los parámetros son estimados mediante el método de mínimos cuadrados. Sin embargo, si se quisiera ajustar un modelo para estimar una variable categórica o discreta, el método de mínimos cuadrados no sería capaz de producir estimadores insesgados (es decir, que la esperanza del estimador sea el parámetro que se quiere estimar) con

varianza mínima para los parámetros de dicho modelo, cualidad buscada al momento de calcular estimadores y que posee cuando la variable de respuesta del modelo es continua.

- Dado que el valor ajustado que se busca es una probabilidad (probabilidad de que el individuo i sea categorizado el nivel j), se buscaría que la respuesta del modelo estuviera acotada inferiormente por cero y superiormente por uno; sin embargo, la respuesta que brinda el modelo de regresión lineal no respeta dicho supuesto y podría ser mayor a uno o incluso negativa.
- Si la variable respuesta constara de J niveles y fuera no ordinal, los resultados del modelo de regresión lineal dependerían totalmente del número asignado a cada categoría.

En cambio, los puntos a favor de utilizar el modelo de regresión logística multinomial al tener a una variable categórica como variable de respuesta son:

- En vez de utilizar el método de mínimos cuadrados para la estimación de coeficientes, el modelo de regresión logística multinomial utiliza el método de máxima verosimilitud, mismo que consiste en proporcionar a los parámetros los valores que hagan más probable (o más verosímil) la ocurrencia de las observaciones de los datos. Los estimadores obtenidos bajo máxima verosimilitud, tienen la propiedad de ser asintóticamente eficientes, por lo que también son asintóticamente insesgados y de varianza mínima.
- Las estimaciones realizadas por el modelo de regresión logística multinomial, como se verá a continuación, sí se encuentran acotadas por los valores cero y uno.
- Para variables de respuesta con J niveles, del modelo de regresión logística multinomial se pueden obtener J probabilidades, una por cada categoría.

Ahora que se han dado argumentos sólidos para favorecer el uso del modelo de regresión logística multinomial, se procede a presentar el método utilizado para la estimación de los coeficientes: el método de máxima verosimilitud.

1.3.1. Descripción del modelo

Considérense los puntos siguientes:

- Una base de datos compuesta de N observaciones, todas ellas independientes entre sí.
- El número de categorías que puede tomar la variable dependiente de cada observación es J, con $J \geq 2$.

- Es conveniente tanto por trabajo computacional como para dar mayor robustez al ajuste del modelo, agrupar las N observaciones de acuerdo a su combinación de variables independientes. Así, si dos observaciones tuvieran los mismos valores en cada variable explicativa, entonces formarán parte del mismo grupo sin importar sus respectivos valores asociados a la variable independiente. Estos grupos serán llamados de ahora en adelante "patrones de covariables". Sea M el número total de patrones de covariables en la base de datos y m el vector compuesto por todos los valores de m_i, con m_i el número total de observaciones dentro del patrón de covariables i .

  Se cumple que $\sum_{i=1}^{M} m_i = N$.

- Se define a $\boldsymbol{y}$ como una matriz de dimensión $M \times J$, con y_{ij} el número de observaciones dentro del patrón de covariables i que cumplan con la característica que el valor de su variable dependiente sea j.

- De manera similar a $\boldsymbol{y}$, sea $\boldsymbol{\pi}$ una matriz de las mismas dimensiones que $\boldsymbol{y}$, con π_{ij} representando la probabilidad de que una observación dentro del patrón de covariables i sea categorizada en j.

- La matriz $\boldsymbol{X}$ será llamada la matriz de diseño. Dicha matriz será de dimensión $M \times (P+1)$, y contendrá dentro de ella a cada patrón de covariables con sus respectivos valores para cada variable. La primer columna de dicha matriz estará compuesta de números uno, es decir, $x_{i0} = 1$ para todo $i = 1, 2, \ldots, M$. Esta columna será la asociada al intercepto.

- $\boldsymbol{\beta}$ será una matriz de dimensión $(P+1) \times (J-1)$, de tal manera que sus valores β_{pj} representarán al coeficiente asignado a la variable p para la categoría j, con $p = 0, 1, \ldots, P$ (cero porque recuérdese que se toma en cuenta al intercepto) y $j = 1, 2, \ldots, J-1$ (recuérdese también que la categoría J será la categoría de referencia).

Para la generalización de la regresión logística multinomial se igualará, como se había visto antes, la componente lineal a la función *logit*, que corresponde al logaritmo natural del cociente entre la probabilidad de que el patrón de covariables i tenga categoría j, y la probabilidad de que el mismo patrón pertenezca a la categoría de referencia, J.

Es decir, la ecuación anterior se representa como:

$$g(x_{i\cdot})_j = \log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \log\left(\frac{\pi_{ij}}{1 - \sum_{j=1}^{J-1} \pi_{ij}}\right) = \sum_{p=0}^{P} x_{ip}\beta_{pj} \tag{1.6}$$

con $i = 1, 2, \ldots, M$ y $j = 1, 2, \ldots, J-1$.

De la ecuación anterior, al intentar despejar las probabilidades π_{ij} y π_{iJ} se llega a las siguientes expresiones:

$$\pi_{ij} = \frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}}{1 + \sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}} \qquad \pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}} \tag{1.7}$$

recordando siempre que $j < J$ y J es la categoría de referencia.

Las ecuaciones anteriores se demostrarán a continuación:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \sum_{p=0}^{P} x_{ip}\beta_{pj}$$
$$\Rightarrow$$
$$\pi_{ij} = e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}\pi_{iJ}$$

Al mismo tiempo, se sabe que $\pi_{iJ} = 1 - \sum_{v=1}^{J-1} \pi_{iv}$. Sustituyendo el valor de π_{ij} en esta última ecuación se obtiene:

$$\pi_{iJ} = 1 - \sum_{v=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pv}}\pi_{iJ}$$
$$\Rightarrow$$
$$\pi_{iJ}\left(1 + \sum_{v=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pv}}\right) = 1$$
$$\Rightarrow$$
$$\pi_{iJ} = \frac{1}{1 + \sum_{v=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pv}}}$$

Y sustituyendo a π_{iJ} en π_{ij} se obtiene:

$$\pi_{ij} = e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}\pi_{iJ}$$
$$= \frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}}{1 + \sum_{v=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pv}}}$$

1.3.2. Máxima verosimilitud

No se deben de olvidar los valores de las probabilidades π_{ij} y π_{iJ} ya que se utilizarán a la brevedad. Ahora pues, para cada patrón de covariables, se sigue que la variable de respuesta se distribuye de manera multinomial con J posibles categorías. Por lo tanto, la función de densidad de probabilidad conjunta de la variable de respuesta de todos los patrones de covariables se calcula como:

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{M} \left[\frac{m_i!}{\prod_{j=1}^{J} y_{ij}!} \cdot \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \right] \tag{1.8}$$

Recuérdese que en la ecuación anterior, $\boldsymbol{\beta}$ se presenta de manera implícita en π_{ij}.

La función de verosimilitud es, algebraicamente, idéntica a la ecuación (1.3). La única diferencia con respecto a esta ecuación es que, inversamente a la función de densidad conjunta, la función de verosimilitud expresa los valores desconocidos de $\boldsymbol{\beta}$ en función de los valores conocidos $\boldsymbol{y}$.

Lo que ahora se buscará será maximizar la ecuación 1.8 con respecto a $\boldsymbol{\beta}$.

Gracias a la ecuación 1.8 se puede observar que la función de verosimilitud buscada es proporcional a:

$$\boldsymbol{L}(\boldsymbol{\beta}|\boldsymbol{y}) \propto \prod_{i=1}^{M} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \tag{1.9}$$

Ahora, sustituyendo a y_{iJ} por $m_i - \sum_{j=1}^{J-1} y_{ij}$, la ecuación es igual a :

$$\prod_{i=1}^{M} \left[\left(\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \right) \cdot \pi_{iJ}^{m_i - \sum_{j=1}^{J-1} y_{ij}} \right] \tag{1.10}$$

$$= \prod_{i=1}^{M} \left[\left(\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \right) \cdot \frac{\pi_{iJ}^{m_i}}{\pi_{iJ}^{\sum_{j=1}^{J-1} y_{ij}}} \right] \tag{1.11}$$

$$= \prod_{i=1}^{M} \left[\left(\prod_{j=1}^{J-1} \pi_{ij}^{y_{ij}} \right) \cdot \frac{\pi_{iJ}^{m_i}}{\prod_{j=1}^{J-1} \pi_{iJ}^{y_{ij}}} \right] \tag{1.12}$$

$$= \prod_{i=1}^{M} \left[\left(\prod_{j=1}^{J-1} \left(\frac{\pi_{ij}}{\pi_{iJ}} \right)^{y_{ij}} \right) \cdot \pi_{iJ}^{m_i} \right] \tag{1.13}$$

Sustituyendo los valores de π_{ij} y π_{iJ} en la última ecuación, se obtiene:

$$\prod_{i=1}^{M}\left[\left(\prod_{j=1}^{J-1}\left(e^{\sum_{p=0}^{P}x_{ip}\beta_{pj}}\right)^{y_{ij}}\right)\cdot\left(\frac{1}{1+\sum\limits_{j=1}^{J-1}e^{\sum_{p=0}^{P}x_{ip}\beta_{pj}}}\right)^{m_i}\right] \tag{1.14}$$

$$=\prod_{i=1}^{M}\left[\left(\prod_{j=1}^{J-1}e^{y_{ij}\sum_{p=0}^{P}x_{ip}\beta_{pj}}\right)\cdot\left(1+\sum_{j=1}^{J-1}e^{\sum_{p=0}^{P}x_{ip}\beta_{pj}}\right)^{-m_i}\right] \tag{1.15}$$

Y al tomar el logaritmo natural de la última ecuación (que hace que se faciliten los cálculos y que preserva a los puntos máximos de la verosimilitud), se obtiene la función "log-verosimilitud" para el modelo de regresión logística multinomial. La ecuación de la log-verosimilitud para el modelo mencionado queda expresada como:

$$\boldsymbol{l}(\boldsymbol{\beta})=\sum_{i=1}^{M}\sum_{j=1}^{J-1}\left(y_{ij}\sum_{p=0}^{P}x_{ip}\beta_{pj}\right)-\sum_{i=1}^{M}\left(m_i\log\left(1+\sum_{j=1}^{J-1}e^{\sum_{p=0}^{P}x_{ip}\beta_{pj}}\right)\right) \tag{1.16}$$

El siguiente objetivo es encontrar los valores de $\boldsymbol{\beta}$ que maximicen la ecuación previa. Para estimar dichos valores se hará uso del método iterativo "Newton-Raphson", dado que no se puede utilizar un procedimiento algebraico para encontrarlos.

El método de Newton-Raphson involucra a las primeras y segundas derivadas de la ecuación 1.16 con respecto a sus valores β_{pj} para todo $p = 0, 1, \ldots, P$ y $j = 1, 2, \ldots, J-1$.

Las primeras derivadas de la ecuación 1.16 son calculadas a continuación:

$$\begin{aligned}\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj}} &= \sum_{i=1}^{M} y_{ij}x_{ip} - \sum_{i=1}^{M}\left(\frac{m_i}{1+\sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}} \cdot \frac{\partial}{\partial \beta_{pj}}\left(1+\sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}\right)\right) \\ &= \sum_{i=1}^{M} y_{ij}x_{ip} - \sum_{i=1}^{M}\left(\frac{m_i}{1+\sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}} \cdot e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \cdot \frac{\partial}{\partial \beta_{pj}}\left(\sum_{p=0}^{P} x_{ip}\beta_{pj}\right)\right) \\ &= \sum_{i=1}^{M} y_{ij}x_{ip} - \sum_{i=1}^{M}\left(\frac{m_i}{1+\sum_{j=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}} \cdot e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \cdot x_{ip}\right)\end{aligned}$$

Y al sustituir la ecuación 1.7 de π_{ij} en la ecuación anterior se obtiene que:

$$\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj}} = \sum_{i=1}^{M} [y_{ij}x_{ip} - m_i\pi_{ij}x_{ip}] \tag{1.17}$$

Aunque tradicionalmente las primeras derivadas de $\boldsymbol{l}(\boldsymbol{\beta})$ tienen forma de una matriz de dimensión $(P+1)\times(J-1)$, de ahora en adelante se analizarán tanto a $\boldsymbol{\beta}$ como a las primeras derivadas $\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_{pj}}}$, como a vectores de dimensión $(P+1)\cdot(J-1)$, acomodando la j-ésima columna de la correspondiente matriz debajo de la columna número $j-1$. De esta manera, se podrán ver a las segundas derivadas de $\boldsymbol{l}(\boldsymbol{\beta})$ en forma matricial.

Ahora se calcularán las segundas derivadas de $\boldsymbol{l}(\boldsymbol{\beta})$, que como se había mencionado en el párrafo anterior, se interpretarán como una matriz cuya dimensión será de $((P+1)\cdot(J-1))\times((P+1)\cdot(J-1))$. Cada término de la nueva matriz cuadrada se puede ver de la siguiente forma:

$$\frac{\partial^2 \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj}\partial \beta_{p'j'}} = \frac{\partial}{\partial \beta_{p'j'}} \sum_{i=1}^{M} [y_{ij}x_{ip} - m_i\pi_{ij}x_{ip}] \tag{1.18}$$

$$= -\sum_{i=1}^{M}\left[m_i x_{ip} \cdot \frac{\partial}{\partial \beta_{p'j'}}\left(\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}}\right)\right] \tag{1.19}$$

Ahora, si $j \neq j'$ entonces se sigue la ecuación anterior como

$$\begin{aligned}\frac{\partial^2 \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj}\partial \beta_{p'j'}} &= \sum_{i=1}^{M}\left[m_i x_{ip}\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}}{\left(1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}\right)^2} \cdot \frac{\partial}{\partial \beta_{p'j'}}\left(1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}\right)\right] \\ &= \sum_{i=1}^{M}\left[m_i x_{ip}\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}}}{\left(1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}\right)^2} \cdot \frac{\partial}{\partial \beta_{p'j'}}\left(e^{\sum_{p=0}^{P} x_{ip}\beta_{pj'}}\right)\right]\end{aligned}$$

$$= \sum_{i=1}^{M} \left[m_i x_{ip} \pi_{ij} \pi_{iJ} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj'}} \cdot \frac{\partial}{\partial \beta_{p'j'}} \left(\sum_{p=0}^{P} x_{ip}\beta_{pj'} \right) \right]$$
$$= \sum_{i=1}^{M} \left[m_i x_{ip} \pi_{ij} \pi_{ij'} x_{ip'} \right]$$

En cambio, si $j = j'$ la ecuación 1.19 se desarrolla de la forma:

$$\frac{\partial^2 \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj} \partial \beta_{p'j}} = -\sum_{i=1}^{M} \left[m_i x_{ip} \cdot \left(\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \cdot \frac{\partial}{\partial \beta_{p'j}} \left(\sum_{p=0}^{P} x_{ip}\beta_{pj} \right)}{1 + \sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}} \right) \right]$$
$$- \sum_{i=1}^{M} \left[-m_i x_{ip} \cdot \left(\frac{e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \cdot \frac{\partial}{\partial \beta_{p'j}} \left(e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \right)}{\left(1 + \sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}} \right)^2} \right) \right]$$
$$= -\sum_{i=1}^{M} \left[m_i x_{ip} \left(\pi_{ij} x_{ip'} - \frac{\pi_{ij} e^{\sum_{p=0}^{P} x_{ip}\beta_{pj}} \cdot \frac{\partial}{\partial \beta_{p'j}} \left(\sum_{p=0}^{P} x_{ip}\beta_{pj} \right)}{1 + \sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\beta_{pk}}} \right) \right]$$
$$= -\sum_{i=1}^{M} \left[m_i x_{ip} (\pi_{ij} x_{ip'} - \pi_{ij} \pi_{ij} x_{ip'}) \right]$$
$$= -\sum_{i=1}^{M} \left[m_i x_{ip} x_{ip'} \pi_{ij} (1 - \pi_{ij}) \right]$$

En resumen, los valores de la matriz de segundas derivadas se pueden calcular como:

$$\frac{\partial^2 \boldsymbol{l}(\boldsymbol{\beta})}{\partial \beta_{pj} \partial \beta_{p'j'}} = \begin{cases} -\sum_{i=1}^{M} \left[m_i x_{ip} x_{ip'} \pi_{ij} (1 - \pi_{ij}) \right] & \text{si } j = j' \\ \sum_{i=1}^{M} \left[m_i x_{ip} \pi_{ij} \pi_{ij'} x_{ip'} \right] & \text{si } j \neq j' \end{cases} \tag{1.20}$$

Esta matriz de segundas derivadas tiene otro papel importante además de ser utilizada para el cálculo de los estimadores mediante el método Newton-Raphson: la teoría de máxima verosimilitud señala que la matriz de segundas derivadas se utiliza también para la obtención de la matriz de varianzas y covarianzas de los parámetros estimados. El cálculo de esta matriz se realizará en la sección 5 del presente capítulo, ya que las varianzas de los estimadores serán necesarias tanto para calcular los intervalos de confianza de los mismos, como para algunas pruebas de hipótesis.

1.3.3. Método de Newton-Raphson

Al igualar cada una de las entradas del vector de primeras derivadas a cero se obtiene un sistema de $(P+1)\cdot(J-1)$ ecuaciones, cada una con $(P+1)\cdot(J-1)$ variables que corresponden a cada β_{pj}. Se procederá ahora a encontrar la solución para cada β_{pj}, corroborando primero que dicha solución corresponderá a un máximo y no un mínimo .

Después de verificar que la matriz de segundas derivadas es negativa definida (es decir, que todos los elementos de su diagonal sean negativos), se asegura que la solución es un máximo y no un mínimo, por lo que se comprueba que al resolver el sistema de $(P+1)\cdot(J-1)$ ecuaciones dadas al igualar a cero el vector de primeras derivadas, la solución a dicho sistema contendrá a las estimaciones de los parámetros tales que las observaciones tendrían la mayor probabilidad de ocurrencia.

Los elementos de la diagonal de la matriz de segundas derivadas tienen la forma:

$$-\sum_{i=1}^{M} m_i x_{ip}^2 \pi_{ij}(1-\pi_{ij})$$

Y como m_i es siempre mayor a cero y π_{ij} una probabilidad, entonces todos los términos de la suma son positivos, lo que al agregar el signo negativo se convertiría en negativo.

Sin embargo, la solución a un sistema de ecuaciones no lineales, como es ahora el caso, no se puede encontrar algebraicamente, por lo que tendrá que ser calculada de manera numérica a través del método iterativo antes mencionado: el método Newton-Raphson.

El método Newton-Raphson inicia con una propuesta de la raíz de una función, para después utilizar los primeros dos términos de la expansión de Taylor evaluada en la solución o raíz propuesta y así obtener otra estimación que se encontrará más cercana de la verdadera raíz. Este proceso se repite hasta que las estimaciones converjan.

Cabe recordar que la expansión de Taylor de grado "n" para alguna función "f" en el punto x_0 está dada por la fórmula:

$$f(x) = \sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x-x_0)^i$$

Claro está, suponiendo que las primeras n derivadas de f en x_0 existen.

La expansión de Taylor utilizando únicamente los primeros dos términos de la expansión en el punto x_0 se representa como:

$$f(x) \approx f(x_0) + f'(x_0)\cdot(x-x_0)$$

Ahora, en palabras coloquiales se busca igualar las primeras derivadas a cero porque esto indicaría que los puntos a encontrar causarían una pendiente igual a cero en la función de log-verosimilitud, y esto sólo sucede cuando el punto es máximo o mínimo. Nótese que por lo tanto, se utilizará como f en el método Newton-Raphson a las primeras derivadas, y no a la función de log-verosimilitud.

Al igualar a cero la expresión anterior y despejar a x (la solución), se tiene que las raíces de f se aproximan al valor:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

El método comenzará aquí. Sea x_0 la propuesta inicial a la solución y f la función a maximizar. Al valor calculado x en la ecuación anterior se le llamará x_1 y será la nueva aproximación de f a la verdadera solución. Sucesivamente, se vuelve a calcular x bajo la misma ecuación, pero ahora utilizando x_1 en el lugar de x_0, y al nuevo resultado se le llamará x_2. El proceso se repite iterativamente hasta que no haya diferencia significativa entre x_i y x_{i+1} para algún i.

Nótese que el método presentado anteriormente corresponde a funciones $\mathbb{R} \to \mathbb{R}$. El caso $\mathbb{R}^n \to \mathbb{R}^n$, que es el que se necesita para estimar los parámetros del modelo logístico multinomial, se presenta a continuación.

Sea $f(x)$ la función a la que se quisiera encontrar sus raíces y sea
$x = (x_1, x_2, \ldots, x_n)$ el vector de raíces de f. Cabe mencionar que tanto x como $f(x)$ deben ser de la misma dimensión. Entonces la aproximación de $f(x)$ alrededor del vector $x^{(0)}$ utilizando los dos primeros términos de la expansión de Taylor para el caso vectorial se representan como:

$$f\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} = f\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times 1} + f'\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times n} \begin{bmatrix} x_1 - x_1^{(0)} \\ x_2 - x_2^{(0)} \\ \vdots \\ x_n - x_n^{(0)} \end{bmatrix}_{n\times 1}$$

Con f' una matriz de $n \times n$. Dado que las raíces de f deben de cumplir que

$$f\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}$$

Entonces

$$\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1} = f\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times 1} + f'\begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times n} \begin{bmatrix} x_1 - x_1^{(0)} \\ x_2 - x_2^{(0)} \\ \vdots \\ x_n - x_n^{(0)} \end{bmatrix}_{n\times 1}$$

Y el vector de raíces x se calcularía como:

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times 1} - \left(f' \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix} \right)^{-1}_{n\times n} f \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}_{n\times 1}$$

A partir de este momento, el método procede de igual manera que en el caso real.

Ahora, las ecuaciones a las que se les desea encontrar su solución, corresponden en este caso a las $(P+1)\cdot(J-1)$ ecuaciones de las primeras derivadas de la función log-verosimilitud. Se denotará entonces a la matriz que contenga todas estas ecuaciones por $\boldsymbol{l'(\beta)}$.

También se denota a β^0 como la solución propuesta en primera instancia para las ecuaciones. Entonces, la fórmula iterativa de Newton-Raphson aplicado a este caso y con la notación propuesta tiene la forma:

$$\beta^{(1)} = \beta^{(0)} + [-\boldsymbol{l''(\beta^{(0)})}]^{-1}\boldsymbol{l'(\beta^{(0)})} \tag{1.21}$$

Con $\beta^{(1)}$, $\beta^{(0)}$ y $\boldsymbol{l'(\beta^{(0)})}$ vectores de dimensión $(P+1)\cdot(J-1)$, y $[-\boldsymbol{l''(\beta^{(0)})}]^{-1}$ una matriz de dimensión $((P+1)(J-1))\times((P+1)(J-1))$.

Dado que las soluciones a las ecuaciones bajo este método se obtienen de manera simultánea y no de manera individual, es conveniente utilizar de ahora en adelante notación matricial.

Aunque ya se había calculado el valor exacto de cada elemento de la matriz de segundas derivadas y el vector de primeras derivadas, se hará otra notación con el único propósito de referenciarlos de una manera sencilla.

Sea $1D_{ij}$ el valor referente a la primera derivada de la log-verosimilitud con respecto a β_{ij}, entonces se tiene que:

$$\big[\boldsymbol{l'(\beta)}\big] = \begin{bmatrix} \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{01}}} \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{02}}} \\ \vdots \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{0J-1}}} \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{11}}} \\ \vdots \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{1J-1}}} \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{21}}} \\ \vdots \\ \frac{\partial \boldsymbol{l(\beta)}}{\partial \boldsymbol{\beta_{PJ-1}}} \end{bmatrix} = \begin{bmatrix} 1D_{01} \\ 1D_{02} \\ \vdots \\ 1D_{0J-1} \\ 1D_{11} \\ \vdots \\ 1D_{1J-1} \\ 1D_{21} \\ \vdots \\ 1D_{PJ-1} \end{bmatrix}$$

con

$$1D_{ij} = \frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_{ij}}} = \sum_{v=1}^{M} (y_{vj}x_{vi} - m_v\pi_{vj}x_{vi})$$

Ahora bien, sea $2D_{ij,ab}$ el valor correspondiente a $\frac{\partial \boldsymbol{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_{ij}} \partial \boldsymbol{\beta_{ab}}}$ con

$$2D_{ij,ab} = \begin{cases} -\sum_{v=1}^{M} m_v x_{vi} x_{va} \pi_{vj}(1-\pi_{vj}) & \text{si } j = b \\ \sum_{v=1}^{M} m_v x_{vi} x_{va} \pi_{vj} \pi_{vb} & \text{si } j \neq b \end{cases}$$

Entonces la expresión $\beta^{(1)} = \beta^{(0)} + [-\boldsymbol{l''(\beta^{(0)})}]^{-1}\boldsymbol{l'(\beta^{(0)})}$ queda descrita por:

$$
\begin{bmatrix} \beta_{01}^{(1)} \\ \beta_{02}^{(1)} \\ \vdots \\ \beta_{0J-1}^{(1)} \\ \beta_{11}^{(1)} \\ \vdots \\ \beta_{1J-1}^{(1)} \\ \beta_{21}^{(1)} \\ \vdots \\ \beta_{PJ-1}^{(1)} \end{bmatrix} = \begin{bmatrix} \beta_{01}^{(0)} \\ \beta_{02}^{(0)} \\ \vdots \\ \beta_{0J-1}^{(0)} \\ \beta_{11}^{(0)} \\ \vdots \\ \beta_{1J-1}^{(0)} \\ \beta_{21}^{(0)} \\ \vdots \\ \beta_{PJ-1}^{(0)} \end{bmatrix} - \begin{bmatrix} 2D_{01,01} & \cdots & 2D_{01,0J-1} & 2D_{01,11} & \cdots & 2D_{01,PJ-1} \\ 2D_{02,01} & \cdots & 2D_{02,0J-1} & 2D_{02,11} & \cdots & 2D_{02,PJ-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 2D_{0J-1,01} & \cdots & 2D_{0J-1,0J-1} & 2D_{0J-1,11} & \cdots & 2D_{0J-1,PJ-1} \\ 2D_{11,01} & \cdots & 2D_{11,0J-1} & 2D_{11,11} & \cdots & 2D_{11,PJ-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 2D_{1J-1,01} & \cdots & 2D_{1J-1,0J-1} & 2D_{1J-1,11} & \cdots & 2D_{1J-1,PJ-1} \\ 2D_{21,01} & \cdots & 2D_{21,0J-1} & 2D_{21,11} & \cdots & 2D_{21,PJ-1} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 2D_{PJ-1,01} & \cdots & 2D_{PJ-1,0J-1} & 2D_{PJ-1,11} & \cdots & 2D_{PJ-1,PJ-1} \end{bmatrix}^{-1} \begin{bmatrix} 1D_{01} \\ 1D_{02} \\ \vdots \\ 1D_{0J-1} \\ 1D_{11} \\ \vdots \\ 1D_{1J-1} \\ 1D_{21} \\ \vdots \\ 1D_{PJ-1} \end{bmatrix}
$$

Utilizando iterativamente la ecuación matricial anterior se llegará a la solución del sistema de ecuaciones deseado.

La soluciones encontradas para el vector β tendrán el nombre de "Estimadores máximos verosímiles de β" , y se denotarán de forma vectorial como "$\hat{\beta}$".

Una observación interesante es, que la estimación de parámetros vía máxima verosimilitud aplicada a regresión lineal arroja como resultado el método de mínimos cuadrados; es decir, mínimos cuadrados es un caso particular de máxima verosimilitud.

1.3.4. Advertencias

Hay dos problemas que podrían surgir al utilizar el método de Newton-Raphson:

1. Es posible que algún parámetro llegara a tender a infinito. Esto normalmente es señal de que el modelo está ajustado pobremente o bien que se comporta mal debido a escasez de datos en algún patrón de covariables.

   Esto es un grave problema, pues el parámetro que presente esta particularidad nunca convergerá, independientemente del número de iteraciones que se realicen. Sin embargo, existe la posibilidad de que si aún conociendo el problema que presenta el parámetro se quisieran encontrar los parámetros para ese modelo, se puede arreglar el algoritmo para fijar al parámetro que tienda a infinito y que siga iterando los demás parámetros hasta que converjan. Esta modificación al algoritmo se encuentra disponible en diversos softwares estadísticos.

2. Dadas ciertas condiciones, es posible que algún parámetro encuentre su verdadera raíz, y en la siguiente iteración cambie de forma drástica alejándose de ésta y cayendo en un ciclo que se acerca y aleja de la solución verdadera.

   En caso de tener problemas con los resultados del algoritmo y se sospeche que una posible causa pueda ser que las iteraciones caigan en este tipo de ciclo, revisar en cada iteración el valor de la función de verosimilitud utilizando los parámetros dados y verificar que, en cada nueva iteración, la función de verosimilitud aumente su valor. En caso de que llegara a decrecer para alguna iteración, se confirma el problema del ciclo. Dado este problema, quedarse con los parámetros de la iteración con mayor valor en su función de verosimilitud.

1.4. Evaluación de la significancia del modelo

Una vez estimados los coeficientes, se recomienda evaluar la significancia de éstos mediante pruebas de hipótesis. El objetivo de estas pruebas son, en otras palabras, corroborar si al enfocarse en una variable explicativa en específico, ésta contribuye a la explicación de la variable respuesta o no. Existen varios métodos para calcular la significancia del coeficiente de una variable explicativa, aunque los más utilizados y que se presentarán en este trabajo son dos: La prueba de Wald y la prueba del cociente de verosimilitud, mejor conocida como *Likelihood Ratio Test*.

Dichas pruebas, tienen por hipótesis nula que el valor del coeficiente estimado es igual a cero, mientras que la hipótesis alternativa enuncia que el valor de dicho coeficiente es distinto de cero. Por lo tanto, para poder realizar inferencias respecto a algún coeficiente en particular, lo ideal sería rechazar la hipótesis nula al evaluar estas pruebas en el coeficiente.

La prueba de Wald, consiste en dividir un estimador máximo verosímil de un parámetro, dígase $\hat{\beta}_{pj}$, entre su respectivo error estándar. Dado que una propiedad de los estimadores máximos verosímiles es que éstos se distribuyen de manera asintóticamente normal, entonces la división mencionada tendría una distribución asintóticamente normal estándar suponiendo que el valor real del parámetro asociado al estimador fuese cero. Por lo tanto, a un nivel de significancia buscado, se podría comparar el valor de este cociente con el cuantil de una distribución normal estándar para decidir si rechazar o no rechazar la hipótesis nula que dice que el verdadero valor del parámetro es cero.

De manera explícita, el estadístico de Wald asociado al coeficiente $\hat{\beta}_{pj}$ para la variable p en la categoría j es:

$$W_{pj} = \frac{\hat{\beta}_{pj}}{\hat{\text{SE}}(\hat{\beta}_{pj})}$$

Donde $\hat{\text{SE}}(\hat{\beta}_{pj})$ representa el error estándar estimado para el coeficiente estimado $\hat{\beta}_{pj}$. En la siguiente sección se mencionará cómo calcular $\hat{\text{SE}}(\hat{\beta}_{pj})$.

Si el valor W_{pj} fuera mayor al cuantil referente a la significancia deseada, entonces se podrá rechazar la hipótesis nula y se concluirá que el valor del parámetro evaluado es distinto de cero al nivel de significancia elegida. En cualquier otro caso, no se podrá rechazar la hipótesis nula.

Antes de explicar la prueba del cociente de verosimilitud, se introducirá un concepto importante para su entendimiento: se define a un "modelo saturado" como a un modelo que contiene el mismo número de variables que de patrones de covariables.

La propiedad más importante de un modelo saturado es que el ajuste es perfecto, es decir $y_{ij} = m_i\pi_{ij}$ para todo patrón de covariables i y categoría j. Por lo tanto, lo que el investigador buscaría sería encontrar un modelo con características similares a las de un modelo saturado, pero con un número de variables mucho menor al de patrones de covariables, pues el tener un gran número de variables ocasiona muchos problemas que podrían ser de gravedad. Dichos problemas se presentarán en los siguientes capítulos.

Una función utilizada con frecuencia para comparar al modelo a analizar con respecto al modelo saturado, es la llamada "Devianza" cuya expresión es:

$$D = -2\log\left[\frac{\text{Verosimilitud del modelo en cuestión}}{\text{Verosimilitud del modelo saturado}}\right] \tag{1.22}$$

El cociente dentro del logaritmo natural de la Devianza recibe el nombre de "cociente de verosimilitud". Al aplicarle al cociente de verosimilitud la función logaritmo natural y multiplicarlo por menos dos, se obtiene un valor que se sabe se distribuye Ji-cuadrada con grados de libertad iguales al número de categorías menos uno, multiplicado por la diferencia entre el número de patrones de covariables y el número de variables del modelo en cuestión más el intercepto. En otras palabras, los grados de libertad de la devianza serán $(J-1)(M-(P+1))$. Una manera sencilla de recordar esta cantidad, es que los grados de libertad son iguales al número de patrones de covariables multiplicado por el número de categorías menos uno, menos el número total de coeficientes calculados por el modelo.

Una vez obtenido el valor de D y sus grados de libertad, se puede proceder entonces a realizar pruebas de hipótesis con ella.

En dichas pruebas de hipótesis, la hipótesis nula consiste en la suposición de que ambos modelos (el saturado y el actual) se ajustan de igual manera a los datos y, por lo tanto, los coeficientes de las variables en el modelo saturado que no están en el modelo ajustado tienen un valor igual a cero. Si al comparar el valor de la devianza con un cuantil de la distribución teórica a un nivel de significancia específico α, el valor de la devianza es mayor al del cuantil, entonces se rechaza la hipótesis nula y por lo tanto el modelo ajustado no se ajusta de igual manera que el saturado. En cambio, si la devianza resulta ser menor que el cuantil, entonces no se rechaza la hipótesis nula y se supone que el modelo ajustado se ajusta de igual manera a los datos que el saturado.

Mientras más cercana se encuentre la verosimilitud del modelo ajustado de la verosimilitud del modelo saturado, más se parecerán ambos modelos, que es lo que se quisiera encontrar. Por lo tanto, mientras menor sea la devianza, mejor será el modelo.

Ahora bien se desarrollará la devianza de manera algebraica, esto con el propósi-

to de encontrar una fórmula más amigable y que facilite el trabajo computacional requerido. Utilizando la ecuación de la verosimilitud:

$$f(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{M} \left[\frac{m_i!}{\prod_{j=1}^{J} y_{ij}!} \cdot \prod_{j=1}^{J} \pi_{ij}^{y_{ij}} \right] \tag{1.23}$$

dentro de la ecuación de la devianza y utilizando el supuesto de que bajo el modelo saturado, $y_{ij} = m_i \pi_{ij_s}$ con π_{ij_s} la probabilidad correspondiente al modelo saturado y $\hat{\pi}_{ij}$ la probabilidad correspondiente al modelo ajustado, se llega a:

$$\begin{aligned} D =& -2 \log \left[\frac{\text{(Verosimilitud del modelo ajustado)}}{\text{(Verosimilitud del modelo saturado)}} \right] \\ =& -2 \log \left(\frac{\prod_{i=1}^{M} \left(\frac{m_i!}{\prod_{j=1}^{J} y_{ij}} \prod_{j=1}^{J} \hat{\pi}_{ij}^{y_{ij}} \right)}{\prod_{i=1}^{M} \left(\frac{m_i!}{\prod_{j=1}^{J} y_{ij}} \prod_{j=1}^{J} \pi_{ij_s}^{y_{ij}} \right)} \right) \\ =& -2 \log \left(\prod_{i=1}^{M} \prod_{j=1}^{J} \left(\frac{\hat{\pi}_{ij}}{\pi_{ij_s}} \right)^{y_{ij}} \right) \end{aligned}$$

Y dado que $y_{ij} = m_i \pi_{ij_s}$ para el modelo saturado, entonces se sigue que:

$$\begin{aligned} D =& -2 \log \left(\prod_{i=1}^{M} \prod_{j=1}^{J} \left(\frac{m_i \hat{\pi}_{ij}}{y_{ij}} \right)^{y_{ij}} \right) \\ =& -2 \sum_{i=1}^{M} \sum_{j=1}^{J} y_{ij} \log \left(\frac{m_i \hat{\pi}_{ij}}{y_{ij}} \right) \end{aligned}$$

Nótese que en la última igualdad, se tiene que el numerador del logaritmo es el número esperado de observaciones en cada categoría bajo el modelo en cuestión, mientras que en el denominador se encuentra el número observado.

Ahora, dado que $\hat{\pi}_{iJ} = \left(1 - \sum_{k=1}^{J-1} \hat{\pi}_{ik}\right)$ y $y_{iJ} = \left(m_i - \sum_{k=1}^{J-1} y_{ik}\right)$, entonces

$$D = -2 \sum_{i=1}^{M} \left[\sum_{j=1}^{J-1} y_{ij} \log \left(\frac{m_i \hat{\pi}_{ij}}{y_{ij}} \right) + \left(m_i - \sum_{k=1}^{J-1} y_{ik} \right) \log \left(\frac{m_i \left(1 - \sum_{k=1}^{J-1} \hat{\pi}_{ik}\right)}{m_i - \sum_{k=1}^{J-1} y_{ik}} \right) \right]$$

Más aún, cuando m_i toma el valor de 1 para todo $i = 1, 2, \ldots, M$ o cuando $y_{ij} = m_i$ para algún j, la verosimilitud del modelo saturado es 1, pues:

$$\begin{aligned}\boldsymbol{L}(\boldsymbol{\beta}|\boldsymbol{y}) =& \prod_{i=1}^{M}\left[\frac{m_i!}{\prod_{j=1}^{J} y_{ij}!}\cdot\prod_{j=1}^{J}\pi_{ij_s}^{y_{ij}}\right]\\ =& \prod_{i=1}^{M}\prod_{j=1}^{J}\pi_{ij_s}^{y_{ij}}\\ =& 1\end{aligned}$$

y además se cumpliría que

$$D = -2\log\left(\text{Verosimilitud del modelo ajustado}\right)$$

Es importante hacer mención en que esto sólo se cumple cuando $M = N$ (es decir, cuando cada patrón de covariables se encuentra compuesto por una sola observación) o cuando $y_{ij} = m_i$ para todo $i = 1, 2, \ldots, M$ y para algún $j = 1, 2, \ldots, J-1$; de otro modo, la verosimilitud del modelo saturado no será uno.

Ahora, si se quisiera verificar si una variable es significativa o no para el modelo en cuestión, se implementaría una estadística similar a la devianza, con la diferencia de que en vez de comparar al modelo ajustado con el modelo saturado, se compararía al modelo ajustado con el modelo ajustado sin la variable a evaluar.

Esta nueva estadística se deriva de la diferencia entre la devianza del modelo con la variable a evaluar y la devianza del modelo sin la variable a evaluar:

$$\begin{aligned}G =& D(\text{Modelo sin la variable}) - D(\text{Modelo con la variable})\\ =& -2\log\left(\frac{\text{Verosimilitud sin la variable}}{\text{Verosimilitud modelo saturado}}\right)\\ &+2\log\left(\frac{\text{Verosimilitud con la variable}}{\text{Verosimilitud modelo saturado}}\right)\\ =& -2\left(\log\left(\text{Verosimilitud sin la variable}\right) - \log\left(\text{Verosimilitud modelo saturado}\right)\right) +\\ &-2\left(-\log\left(\text{Verosimilitud con la variable}\right) + \log\left(\text{Verosimilitud modelo saturado}\right)\right)\\ =& -2\left(\log\left(\text{Verosimilitud sin la variable}\right) - \log\left(\text{Verosimilitud con la variable}\right)\right)\end{aligned}$$

Dando al final como resultado

$$G = -2 \log \left[\frac{\text{(Verosimilitud sin la variable)}}{\text{(Verosimilitud con la variable)}} \right] \tag{1.24}$$

Bajo la hipótesis de que el coeficiente de la variable a evaluar, β_{pj}, es igual a cero, el estadístico G sigue una distribución Ji-cuadrada con $J - 1$ grados de libertad, pues es la diferencia del número de coeficientes a estimar para cada modelo.

Esta prueba, bajo la hipótesis nula referente a si los modelos a evaluar son similares, y la hipótesis alternativa referente a que no lo son, forma parte de las llamadas pruebas de cocientes de verosimilitud. Las pruebas de cocientes de verosimilitud están conformadas por todas las pruebas donde se comparan a modelos anidados. La definición de dichos modelos se presentará en los puntos siguientes.

Cabe mencionar que la devianza es un caso particular de una prueba de cocientes de verosimilitud, y de manera general, la distribución que tendrá dicha prueba será Ji-cuadrada con grados de libertad iguales a la diferencia en el número de coeficientes a calcular en cada uno de los dos modelos.

Algunas observaciones importantes respecto a las pruebas de hipótesis antes presentadas se mencionan a continuación:

- Hauck y Donner (1977) analizaron el desempeño de la prueba de Wald y encontraron que no se comportaba de una manera adecuada, pues había fallado constantemente en rechazar la hipótesis nula aún cuando el coeficiente era significativo. Por estas razones, ellos han recomendado ampliamente utilizar el cociente de verosimilitud en vez de esta prueba.
- La log-verosimilitud de cualquier modelo se calcula sustituyendo en la ecuación 1.16 a los parámetros estimados bajo el modelo, o bien si se cumplen los casos donde la verosimilitud del modelo saturado es uno, se puede despejar la log-verosimilitud del modelo en cuestión utilizando la ecuación 1.22.
- La devianza en regresión logística multinomial juega el mismo papel que la suma residual de cuadrados en regresión lineal.
- Para poder comparar dos modelos utilizando la prueba del cociente de verosimilitud, se necesita que los modelos en cuestión sean anidados, es decir, que un modelo tenga más variables que otro, y que además el modelo con mayor número de variables contenga a todas las variables utilizadas por el modelo con menor número de variables.

  En caso de que dichos modelos no sean anidados, no se podría asegurar si una variable es o no significativa para el estudio pues el comportamiento de cada variable depende también de las demás variables en el modelo.

- Existen muchas otras consideraciones para la evaluación de un modelo además de las mencionadas anteriormente, por lo que sería un error elegir al modelo definitivo únicamente basándose en estas pruebas. En el capítulo tres, referente al ajuste del modelo, se cubrirán otros métodos que son de vital importancia para la elección del modelo final.

- Para el caso de las variables independientes categóricas, si algunas variables de diseño fueran significativas mientras otras no, se debe tomar la decisión de incluir todas las variables de diseño o ninguna. Por ningún motivo se pueden considerar sólo algunas para el modelo. Otra opción que también se puede realizar en este caso es combinar dos o más categorías en una sola para intentar que todas las categorías sean significativas, siempre y cuando dicha combinación tenga sentido.

- Si L fuera el número de niveles de una variable independiente categórica, entonces la contribución en grados de libertad que tendría ésta para motivos de la prueba del cociente de verosimilitud sería $L-1$ multiplicado por el número de categorías de la variable de respuesta menos uno, es decir, $(L-1)\cdot(J-1)$.

1.5. Estimación de intervalos de confianza

Otra técnica importante para el análisis del modelo, además de la verificación de la significancia de los estimadores máximos verosímiles, es el cálculo e interpretación de los intervalos de confianza de los mismos.

Dado que los estimadores máximos verosímiles siguen una distribución asintóticamente normal, a un nivel de confianza deseado 100(1-α) %, los puntos extremos del intervalo de confianza de un coeficiente $\hat{\beta}_{pj}$ se calculan como:

$$\hat{\beta}_{pj} \pm z_{1-\alpha/2}\hat{\mathrm{SE}}(\hat{\beta}_{pj}) \tag{1.25}$$

Donde $z_{1-\alpha/2}$ es el cuantil $100(1-\alpha/2)\,\%$ de una distribución normal estándar y $\hat{\mathrm{SE}}(\cdot)$ representa el error estándar estimado del respectivo parámetro.

Hasta ahora no se ha estimado el error estándar de ningún parámetro, únicamente se mencionó en la sección 3.2 del presente capítulo que dicho error se obtendría a través de la matriz de varianzas y covarianzas estimada, misma obtenida a partir de la matriz de segundas derivadas del logaritmo natural de la verosimilitud de $\boldsymbol{\beta}$.

Se trabajará entonces con esta matriz de segundas derivadas. Denótese como $\boldsymbol{I(\beta)}$ a dicha matriz pero con los signos opuestos en cada valor de ella. A esta nueva matriz $\boldsymbol{I(\beta)}$ se le llamará "Matriz observada de información". La matriz de varianzas y covarianzas estimada se obtiene mediante la inversa de esta matriz, es decir:

$$\hat{\text{Var}}(\hat{\boldsymbol{\beta}}) = [\boldsymbol{I}(\hat{\boldsymbol{\beta}})]^{-1}$$

Se utilizará la notación "$\hat{\text{Var}}(\hat{\beta}_{pj})$" para denotar al pj-ésimo elemento de la diagonal de esta matriz, que corresponderá a la varianza estimada de $\hat{\beta}_{pj}$, y $\text{Cov}(\hat{\beta}_{pj}, \hat{\beta}_{qr})$ para denotar a cualquier elemento arbitrario de la matriz fuera de la diagonal, que será la covarianza estimada entre $\hat{\beta}_{pj}$ y $\hat{\beta}_{qr}$, con $p, q = 0, 1, \ldots, P$ y $j, r = 1, 2, \ldots, (J-1)$.

Ahora que ya se tienen las varianzas estimadas de los parámetros estimados, se procede a calcular los errores estándar estimados como:

$$\hat{\text{SE}}(\hat{\beta}_{pj}) = \left[\hat{\text{Var}}(\hat{\beta}_{pj})\right]^{1/2}$$

para $p = 0, 1, \ldots, P$ y $j = 1, 2, \ldots, (J-1)$.

Además de los estimadores de los parámetros, existen otros estimadores a los que es de interés calcular sus intervalos de confianza, como es el caso de la función *logit*.

Para poder hacer el cálculo de su respectivo intervalo de confianza, se debe de conocer el error estándar del mismo y, por ende, su varianza.

El estimador de la varianza de la función *logit* para el patrón de covariables i con la categoría j es denotada por:

$$\hat{\text{Var}}[\hat{g}(x_{i\cdot})_j] = \sum_{p=0}^{P} x_i^2 \hat{\text{Var}}(\hat{\beta}_{pj}) + 2 \sum_{p=0}^{P-1} \sum_{k=p+1}^{P} x_{ip} x_{ik} \hat{\text{Cov}}(\hat{\beta}_{pj}, \hat{\beta}_{kj}) \tag{1.26}$$

Y dado que $\hat{g}(x_{i\cdot})_j$ es una combinación lineal de estimadores distribuidos asintóticamente normal, $\hat{g}(x_{i\cdot})_j$ también se distribuye asintóticamente normal, dando como consecuencia que el intervalo de confianza de la función estimada *logit* se encuentre delimitado por los puntos:

$$\hat{g}(x_{i\cdot})_j \pm z_{1-\alpha/2} \hat{\text{SE}}[\hat{g}(x_{i\cdot})_j] \tag{1.27}$$

donde $\hat{\text{SE}}[\hat{g}(x_{i\cdot})_j]$ es la raíz cuadrada positiva del estimador de la varianza de $\hat{g}(x_{i\cdot})_j$.

El estimador del *logit*, así como su intervalo de confianza, son de gran importancia debido a que con ellos se pueden obtener los intervalos de confianza de

las probabilidades estimadas $\hat{\pi}_{ij}$.

Los puntos delimitantes del intervalo al $100(1-\alpha)\,\%$ de confianza para la probabilidad de que el sujeto i se encuentre en la categoría j, π_{ij} son:

$$\frac{e^{\hat{g}(x_{i\cdot})_j \pm z_{1-\alpha/2}\hat{\mathrm{SE}}[\hat{g}(x_{i\cdot})_j]}}{1+e^{\hat{g}(x_{i\cdot})_j \pm z_{1-\alpha/2}\hat{\mathrm{SE}}[\hat{g}(x_{i\cdot})_j]}} \tag{1.28}$$

Capítulo 2

Búsqueda del Mejor Modelo

2.1. Introducción

Este capítulo está enfocado a la selección del modelo a utilizar. En la práctica, es común encontrar que uno de los objetivos principales de un estudio sea encontrar las variables que logren categorizar de manera acertada a las observaciones con respecto a alguna variable categórica. Uno de los problemas a solucionar en este caso es qué variables utilizar, dado que a medida que el volumen de información aumenta, es más sencillo contar con una mayor cantidad de variables explicativas. Si se tuviera un conjunto de variables explicativas grande, dígase por ejemplo diez o más, y a primera instancia no se tuviera idea alguna sobre cual de ellas podría brindar el mejor rendimiento en el modelo, entonces el reto de encontrar a las variables indicadas para el modelo a implementar sería desafiante.

En esta sección se brindarán algunas herramientas que podrían ayudar a facilitar el trabajo de búsqueda de variables adecuadas, así como la búsqueda del mejor modelo a encontrar dadas dichas variables. Cabe mencionar que las herramientas presentadas a continuación no son las únicas que existen, y tal vez podrían no ser las mejores para algún análisis según los objetivos de éste, pero son los métodos más populares y sencillos de implementar y que al mismo tiempo brindan una efectividad confiable.

Los métodos señalados a continuación son únicamente sugerencias, y ninguno de ellos asegura completamente al investigador que encontrará un modelo que satisfaga sus necesidades; sin embargo pueden ser de gran utilidad dependiendo del objetivo del estudio, sobre todo si éste fuera ajustar los datos lo mejor posible.

Algo importante a mencionar antes de mostrar dichos métodos es que, dado que la buena selección de las variables predictoras por parte del investigador

corresponde a uno de los factores de mayor importancia para el éxito de un modelo, se debe de tener siempre una idea clara de la naturaleza e interpretación de cada una de las variables a utilizar.

A continuación, se tomará el tema de la selección de las variables con mayor significancia para el modelo, para posteriormente tratar sobre la selección del mejor o posibles mejores modelos.

Aunque las técnicas explicadas a continuación pueden auxiliar al investigador en la búsqueda del modelo, debe de ser prioridad el realizar el análisis de éste en compañía de expertos del tema en cuestión, ya que como se ha mencionado anteriormente, para lograr un buen ajuste del modelo se necesita tanto de buenas herramientas teóricas y prácticas, como de experiencia profesional tanto en el área de aplicación como en la elaboración de modelos. Así también, se debe de dar prioridad al conocimiento del área de aplicación para decidir si el modelo en cuestión es coherente o no.

2.2. Algoritmos de selección de variables y modelos

Una de las convicciones respecto a la selección de modelos estadísticos es encontrar el modelo con la menor cantidad de variables posibles que aún pueda explicar de manera acertada a la variable de respuesta; esto es en otras palabras el principio de parsimonia, que dicta que casi siempre la mejor solución es la más sencilla.

Existen otros argumentos para buscar siempre el modelo con menor cantidad de variables. Uno de ellos es que mientras menor sea el número de variables, mayor es la tendencia a que el modelo sea estable; esto es, que los estimadores de los coeficientes y errores estándar no sean muy grandes, que tengan sentido, y que el modelo no se sobreajuste a los datos observados.

Por el otro lado, otra convicción sugiere incluir en el modelo a todas las variables cuyas naturalezas sean relevantes para la variable de respuesta, sin importar su significancia estadística en el modelo. En parte, esto lo hacen debido al pensamiento de que aunque no sean significativas en sí mismas dichas variables, tal vez tengan efectos sobre otras variables que sí lo sean (es decir, son variables tipo confusoras o *confounders*) y se podrá tener un mejor control de las variables significativas. Sin embargo, se podrían presentar los problemas de

sobreajuste ya mencionados si se ingresaran demasiadas variables al modelo. Si se encontrara con dichos problemas, se diría que el modelo se encuentra "sobre ajustado". Es por esto que es común encontrarse con casos los que al ajustar un modelo con una gran cantidad de variables explicativas, aunque las pruebas del modelo indican que éste se ajusta bien a los datos, al momento de utilizarlo con nuevas observaciones los resultados del modelo distan mucho de la realidad.

Además de las causas de sobreajuste mencionadas anteriormente, también se puede presentar este problema cuando aunque el número de variables sea pequeño en cantidad, en proporción con respecto al número de observaciones sea muy grande, o el número de individuos en cada categoría de respuesta sea demasiado desproporcional.

Desde un punto de vista personal, lo que el autor de este documento sugiere es primero incluir en el modelo a todas las variables significativas y que sean de importancia dentro del área de aplicación, y si se encontraran problemas como errores estándar altos, intentar reducir el número de variables hasta estabilizar al modelo.

Ahora que se conocen los motivos por los cuales reducir el número de variables del modelo, se procede a explicar un algoritmo enfocado en encontrar a las posibles variables indicadas para el modelo.

1. El primer paso consiste en realizar análisis univariados para cada variable explicativa candidata a ser incluida en el modelo; es decir si se tuviera un conjunto de P variables explicativas candidatas a usarse en el modelo, ajustar P regresiones logísticas multinomiales univariadas, una por cada variable. En dicho análisis se deben de tomar en cuenta diversas estadísticas, como el estadístico de Wald y la prueba del cociente de verosimilitud. Con base en los *p-values* obtenidos se tomará la decisión de seguir considerando a la variable como posible candidata al modelo o no. Se debe hacer énfasis en que el hecho de que el *p-value* de alguna variable haya sido significativo no significa que la variable será incluida en el modelo, si no que seguirá siendo candidata a pertenecer al modelo, mientras las variables que no hayan sido significativas ya no se tomarán en cuenta sino hasta el paso número cuatro del algoritmo.

   Dado que en un modelo los coeficientes de las variables y su significancia son calculados de manera conjunta, es posible que alguna variable no aparezca como significativa en el modelo univariado, pero dentro de un modelo múltiple sí lo sea. Es por eso que en esta primera etapa del algoritmo, se recomienda tomar un nivel de significancia α no tan rígido, esto con el fin de incluir posibles variables que si bien en un primer momento no cumplan con los estándares necesarios para pertenecer al modelo, tal

vez lo hagan de manera conjunta con otras variables incluidas en el mismo.

Autores como Hosmer y Lemeshow (2000, p.95) recomiendan en primer lugar utilizar un nivel de significancia $\alpha = 0.25$ para esta primera etapa. Además de las variables que cumplan con el requisito de significancia, también seguirán siendo candidatas a pertenecer al modelo final las variables que se consideren importantes dada el área de aplicación de los datos.

2. Una vez realizado el filtro de variables candidatas al modelo mediante análisis univariados, se procede a crear un modelo que contenga a todas las variables candidatas.

3. Seguido de la elaboración del modelo con todas las variables candidatas, se prosigue a evaluar el impacto de cada variable dentro del modelo.
Para realizar dicha evaluación, se tomarán en cuenta los estadísticos de Wald de cada variable. Las variables cuyo estadístico de Wald indiquen no ser significativas para el modelo se eliminarán de una por una iniciando con la que tenga menor nivel de significancia, y se elaborará un nuevo modelo únicamente con las variables no eliminadas. Posteriormente, el nuevo modelo se compara con el modelo con todas las variables explicativas mediante pruebas de cocientes de verosimilitud. Se recomienda fijar atención en las variables cuyos coeficientes cambien drásticamente de magnitud, pues es señal de que alguna variable que funcionaba como confusora fue eliminada, y se podría buscar dicha variable para reingresarla.

   Si la prueba de cocientes de verosimilitud indica que el ajuste de ambos modelos es igual de bueno, entonces se procede nuevamente a eliminar la variable del modelo con menor significancia y se repiten los pasos anteriores. El proceso iterativo termina una vez que se haya encontrado al modelo con menor cantidad de variables explicativas que siga cumpliendo la hipótesis nula bajo la prueba del cociente de verosimilitud, misma que consiste en que el modelo en cuestión explica de igual manera a la variable de respuesta que el modelo saturado.

4. Terminado el paso anterior, se procede a incluir en el modelo a las variables desechadas en el paso uno, esto para tener mayor seguridad en no olvidar alguna variable que por sí sola no sea significativa, pero en conjunto con otras sea de importancia. Si alguna de estas variables resultó ser significativa al nivel de significancia comúnmente utilizado (0.05 o 0.10), entonces se incluye al modelo; si no, se elimina.

5. Se procede a realizar un análisis más severo para las variables que resultaron ser incluidas en el modelo. Para el caso de las variables categóricas se procede al siguiente paso del algoritmo, pero para las variables continuas, se debe de revisar el supuesto de que toda función *logit* sea lineal con respecto a cada variable. Esto se realiza mediante una gráfica como la expuesta en la figura 2.1. Dicha figura fue obtenida utilizando los datos

de la base de datos de entrenamiento expuesta en el apéndice, con la variable "Categoría" como variable de respuesta y la variable "Número de jugadores con experiencia en semifinales" como única variable explicativa. El *logit* corresponde a la categoría dos con respecto a la categoría de referencia, la categoría uno.

Esta gráfica se realiza para todas las variables continuas. Si alguna variable no cumpliera con linealidad en el logit, se contemplará su eliminación del modelo o bien el aplicarle alguna transformación con el objetivo de buscar dicha linealidad.

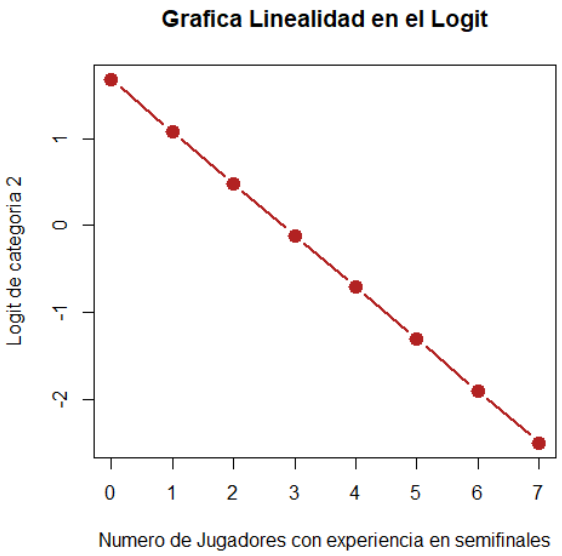

Figura 2.1: Gráfica correspondiente al paso cinco del algoritmo: linealidad en el logit.

6. Como último paso, se prosigue a buscar interacciones dentro de las variables del modelo. Si se tuvieran preguntas sobre como calcularlas, acudir al capítulo 4, sección 3, pues ahí se explicará detalladamente tanto su función como su cálculo.

   Las interacciones se deben de incluir de una por una, esto para tener certeza del efecto que brinda cada una de ellas en específico, y así poder tomar la decisión de incluirla en el modelo o desecharla utilizando pruebas de cocientes de verosimilitud. Se recomienda que la significancia para decidir si incluirlas o desecharlas sea la usual, 0.05 o 0.10.

   Es importante mencionar que aunque la decisión de incluir o no una inter-

acción recae en gran parte en consideraciones estadísticas, siempre se debe de cumplir que una interacción tenga sentido desde un punto de vista del área de aplicación de los datos. Si la interpretación de la interacción no fuera coherente, no se propondrá su inclusión en el modelo. Así también antes de señalar al modelo como el mejor modelo ajustado encontrado, se deben de revisar estadísticas de bondad de ajuste y de diagnóstico.

Como se había mencionado anteriormente, ésta no es la única manera propuesta para elegir algún modelo. Existen otras alternativas a los pasos dos y tres del algoritmo anterior. En particular, a continuación se hablará del método *stepwise*, teniendo en cuenta siempre que éste no puede fungir como alternativa a todo el algoritmo anterior, sino como una alternativa a los pasos dos y tres de dicho algoritmo.

2.2.1. Método *stepwise*

Una de las muchas alternativas a los pasos dos y tres del algoritmo anterior es el llamado método *stepwise* para regresión logística multinomial. En pocas palabras, éste es un método iterativo en el cual se empiezan a incluir las variables con mayor significancia en el modelo y se eliminan las de menor significancia. Este proceso se repite hasta haber encontrado un modelo en el cual todas sus variables sean estadísticamente significativas. El nivel de significancia, y por lo tanto la decisión de incluir una variable o no, son determinados por la prueba de cocientes de verosimilitud (*Likelihood Ratio Test*).

Aunque en general este algoritmo puede ser utilizado para cualquier tipo de análisis, el caso en el que se le obtiene el mayor provecho es cuando es incierta la relación entre la variable de respuesta y las posibles variables explicativas, por lo que se utilizan en primera instancia un gran número de variables como candidatas a pertenecer al modelo.

Existen 2 maneras de realizar dicho método:

- Introducción progresiva (*Forward selection* en inglés).
- Eliminación progresiva (*Backward elimination* en inglés).

En la primera, se inicia con un modelo que contiene únicamente al intercepto, y en cada iteración se agregan variables que tengan un nivel de significancia importante, mientras que en el segundo, se inicia con un modelo que incluye a todas las posibles variables, y en cada iteración se eliminan las variables no significativas hasta encontrar al modelo en el que todas sus variables sean significativas.

Se mostrará a continuación el proceso para la primera manera, que es la introducción progresiva o *método forward*.

1. Se ajusta un modelo únicamente con el intercepto y se obtiene su log-verosimilitud. Posteriormente, se calculan todos los posibles modelos univariados (con una variable más el intercepto) y se obtienen sus respectivas log-verosimilitudes.

2. Se aplican pruebas de cocientes de verosimilitud entre cada modelo univariado y el modelo únicamente con intercepto. Se busca a la variable correspondiente al *p-value* más pequeño y que cumpla que sea menor a α_E, con α_E el nivel de significancia que se utilizará para incluir variables (en total se utilizarán dos niveles de significancia, que se presentarán a continuación), y se tomará al modelo con esa variable como base para seguir buscando al mejor modelo.

   Podría existir la posibilidad de que ningún *p-value* cumpliera las condiciones especificadas. Si llegara a suceder esto, existen dos maneras de interpretarlo: La primera es que ninguna variable serviría para explicar a la variable de respuesta, por lo que es necesario buscar otras alternativas de variables o métodos estadísticos; la segunda consiste en aumentar el nivel de significancia asociado a la inclusión de variables, α_E, para que los criterios no sean tan estrictos y puedan permitir la inclusión de más variables.

   Como se había mencionado, a lo largo del desarrollo del método se utilizarán dos valores diferentes de significancia, definidas por α_E y α_S (alfa de entrada y alfa de salida). La primera corresponde al nivel de significancia que se utilizará para decidir si incluir o no variables nuevas al modelo, mientras la segunda se utilizará para decidir si remover o no alguna variable que ya se encuentre en dicho modelo.

   Se debe de ser cuidadoso al elegir los valores α_E y α_S. Para el caso de α_E, autores como Hosmer y Lemeshow (2000, p.118) han sugerido sea entre 0.15 y 0.20.

3. Una vez elegida la primera variable del modelo, se procede a realizar el mismo procedimiento de elaborar modelos pero que ahora contengan dos variables, la elegida en el paso anterior más alguna otra. Si en un inicio se tuvieran P posibles variables, entonces se elaboran un total de $P-1$ modelos. Para cada modelo, se obtiene su log-verosimilitud y se realiza la prueba de cocientes de verosimilitud entre dicho modelo y el obtenido en el paso anterior (el que contiene a una variable más el intercepto).

   Se elige al modelo que cumpla con las dos condiciones mencionadas anteriormente (que el *p-value* de su prueba sea el menor de los $P-1$ modelos

y que al mismo tiempo el *p-value* sea menor a α_E) como nueva base para encontrar el mejor modelo. Si ningún *p-value* cumpliera con las dos condiciones, entonces se termina el método y se elige como mejor modelo al modelo anterior.

4. Este paso inicia con un procedimiento similar al realizado en el paso anterior. Se calculan $P-2$ modelos, cada uno con las dos variables elegidas anteriormente más una de las restantes. Se calculan sus log-verosimilitudes y se realizan $P-2$ pruebas de cocientes de verosimilitud entre los nuevos modelos y el modelo que sólo contiene a dos variables más el intercepto. Si existiera algún modelo al que su *p-value* de la prueba cumpliera con las dos condiciones ya mencionadas, entonces se elegiría a dicho modelo como base para el mejor modelo. Si ningún *p-value* cumpliera con las dos condiciones, entonces se terminaría el método y se elegiría como mejor modelo al modelo anterior.

   Aquí sucede algo diferente a los demás pasos. Dado que se corrió la prueba de cocientes de verosimilitud se sabe que bajo las hipótesis que se propusieron (si el *p-value* es menor a α_E entonces el nuevo modelo es mejor que el anterior), el modelo con tres variables sería mejor al modelo anterior con dos variables. Pero también podría suceder el caso en que ahora que se utilizaron tres variables, dígase por ejemplo, la primer variable en ser introducida podría parecer ya no ser significativa en el modelo. Esto podría suceder, dado que podría ser que la combinación de la segunda y tercera variable en ser incluidas fuera muy significativa pero individualmente no. Si este llegara a ser el caso, entonces se deberían de realizar pruebas de cocientes de verosimilitud entre el modelo con tres variables, y los posibles modelos que contengan dos de las tres variables.

   El *p-value* de cada prueba será comparado con el nivel de significancia elegido α_S, y si el *p-value* obtenido fuera el mayor de todos los *p-values*, y además fuera mayor a α_S, entonces se elegiría al modelo de dos variables correspondiente a ese *p-value*, se desechará a la variable no incluida en este modelo, y se tomará como base para el mejor modelo a este modelo de dos variables. Si ninguno de los *p-values* cumpliera con los dos requisitos de ser el mayor de los *p-values* y además ser mayor que α_S, entonces no se elimina ninguna variable y se sigue tomando como base para el mejor modelo al modelo con tres variables.

   Es de vital importancia no confundir las funciones de α_E y α_S, así como sus valores. Normalmente, α_S tiende a ser más grande que α_E, aunque en realidad los valores de dichos niveles de significancia dependerán tanto de la experiencia del investigador, como de la severidad que le quiera dar a los criterios de inclusión o exclusión de variables.

5. Se repite el paso anterior únicamente aumentando el número de variables. Es decir, ahora se comparará un modelo con cuatro variables con un mo-

delo de tres utilizando α_E. Si el modelo con cuatro variables resulta ser igual de bueno que el modelo con tres variables termina el proceso, mientras que si resultara ser mejor, se toma como base. Luego se asegura que todas las variables dentro de este modelo sean significativas utilizando a α_S y se toma el mejor modelo.

Después se repite para cinco variables contra cuatro, y así sucesivamente hasta que la significancia no aumente al incluir o eliminar variables del modelo, o se hayan introducido todas las posibles variables al modelo. En ese momento se habrá encontrado el mejor modelo posible según el método forward.

2.3. Advertencias

Antes de continuar con la siguiente sección, es importante remarcar algunos detalles respecto a la elaboración de modelos, ya que éstos podrían marcar una gran diferencia en el momento de aplicar las técnicas y conocimientos de este capítulo.

Con respecto al método *stepwise*:

- En vez de pruebas de cocientes de verosimilitud, se podrían utilizar otras estadísticas para decidir si incluir o excluir variables (como el estadístico de Wald, Test de *Score*, etcétera.). También se podrían utilizar otros criterios como el criterio de información de Akaike o criterio de información bayesiano para la elección de las variables del modelo.

- Una desventaja de este método es, que al ser iterativo y tener que calcular tantos modelos y pruebas de cocientes de verosimilitud, este proceso puede llegar a ser muy lento si el número de observaciones o de variables explicativas fuera muy grande, y algunos casos, podría necesitarse un poder computacional que no cualquier computadora podría poseer.

- Puede llegar a ocurrir que el método *stepwise* dé como resultado un modelo incoherente. Es por eso que es de gran importancia que el investigador o analista revise, evalúe y comente los resultados del modelo con los expertos del área, pues en estos casos, el dar por hecho que el modelo resultante por el método *stepwise* es el adecuado podría ser un grave error para el estudio. Los métodos como el *stepwise* son herramientas que ayudan al proceso de encontrar el mejor modelo, más no son el proceso para encontrarlo en sí.

- El método de *stepwise* para regresión logística multinomial también puede ser utilizado en el paso seis del algoritmo general dedicado a encontrar interacciones. Simplemente se debe de forzar al método a que inicie con las variables explicativas seleccionadas y que elija como posibles variables a ser introducidas a las interacciones que tengan sentido según el experto del área a la que pertenecen los datos.

- Se debe recordar que si se quisiera incluir en el modelo alguna interacción de orden mayor a dos, se deben de incluir también a las interacciones de orden menor entre las variables involucradas, esto con el fin de conocer el efecto absoluto de las interacciones en cada orden y no atribuirle el efecto de un orden inferior a otro superior. En otras palabras, el modelo de regresión logística multinomial es un modelo jerárquico.

- Existe una gran cantidad de métodos utilizados para selección de variables además del método *stepwise*, por ejemplo la regresión tipo *lasso*, entre otros.

- Por último cabe recordar que este método calcula modelos utilizando como condiciones únicamente significancias estadísticas, por lo que es necesario una vez obtenido algún posible buen modelo mediante este método, revisarlo con detenimiento para asegurar su coherencia y plausibilidad, así como evaluar su respectivo ajuste mediante pruebas formales.

Con respecto a la linealidad en el *logit*:

- Uno de los problemas con los que se pueden encontrar tanto el investigador como el experto del área es que los resultados no sean coherentes debido a alguna variable explicativa continua que tenga un efecto no lineal en el *logit*. Este es un problema, pues uno de los supuestos del modelo era que el riesgo aumentaba de manera lineal y con la misma pendiente para cada variable continua, independientemente del resultado de las demás variables explicativas.

  Si se llegase a topar con este problema, existen varias maneras de intentar resolverlo. Una de ellas es convertir la variable continua a categórica, tomando diferentes niveles que vayan a concordar con las diferentes magnitudes del efecto de dicha variable en los *logits*.

  Así mismo, se podrían incluir términos cuadráticos o de otro orden, o algún tipo de función como la función logaritmo natural al modelo, todo dependiendo del comportamiento que se crea tenga alguna o algunas de las variables explicativas respecto a los *logits*.

  Otra posible solución consiste en, si alguna variable llegara a aumentar su efecto en el *logit* de diferente manera dependiendo del valor que tome

alguna otra variable explicativa; es decir, si la pendiente del *logit* con respecto a una sola variable dependiera del valor que tomara otra variable, entonces se podrían introducir al modelo alguna o algunas interacciones entre las variables involucradas. Este caso se verá en futuros capítulos.

Existen algunos problemas que más que con el modelo en si, surgen con base en la naturaleza de los datos. Algunas advertencias respecto a ellos son:

- El mejor indicador de que hay problemas de ajuste, son los errores estándar asociados a los coeficientes de las variables. Si éstos fueran irreales o muy grandes en proporción al coeficiente estimado, podría significar existencia de problemas respecto a la estructura de los datos.

- Pueden existir casos donde un grupo de variables pudiera separar perfectamente a la variable de respuesta. Si esto llegara a suceder, traería como consecuencia que los estimadores del conjunto de variables no existieran o sus valores y los de sus errores estándar fueran extremadamente grandes. Esto se debe a que a medida que aumentan o disminuyen las variables involucradas en la separación, la función de verosimilitud aumenta, por lo que a medida que aumentaran las iteraciones del método Newton-Raphson dichos coeficientes tenderán a infinito o menos infinito, y por lo tanto nunca convergerán. Es por ello que para poder calcular los estimadores vía máxima verosimilitud, se tiene que garantizar que no exista separación completa entre algún conjunto de variables.

  Cabe mencionar que una desventaja de utilizar muchas variables explicativas es que se tendrá mayor propensión a que se presente este tipo de separación.

- Como en el caso de otras herramientas de análisis, es necesario que las variables explicativas no muestren signos de multicolinealidad para fiarse de las estimaciones calculadas y que el modelo no confunda el efecto real de cada variable. Para esto también se pueden realizar análisis de colinealidad como los realizados en regresión lineal.

- El método de estimación usa como supuesto, al igual que la independencia de variables mencionada en el punto anterior, la independencia entre observaciones (se supuso independencia entre observaciones multiplicarse las funciones de densidad de probabilidad de los patrones de covariables para la elaboración de la teoría de regresión logística multinomial, como el cálculo de los estimadores de máxima verosimilitud, su matriz de varianzas y covarianzas, la función log-verosimilitud, etc.), por lo que se tiene que realizar un análisis de éstas para corroborar que no se encuentren muy correlacionadas.

En resumen, la principal prueba o consecuencia de que existen problemas en el modelo son los errores estándar grandes asociados a los coeficientes estimados.

Una vez detectado esto, el reto estará en detectar el problema en sí, ya que se deben de analizar profundamente todos los puntos mencionados anteriormente.

Capítulo 3

Evaluando la Bondad de Ajuste del Modelo

3.1. Introducción

Antes de hablar sobre el ajuste del modelo, primero es necesario entender su definición. El término "ajuste del modelo" se refiere a evaluar qué tan acertadamente describe el modelo a la variable dependiente.

Se iniciará la evaluación del ajuste del modelo una vez que el analista o investigador se encuentre satisfecho a primera instancia con el modelo encontrado. Con el fin de ser auxiliados en este capítulo, se hará uso de la siguiente notación.

Antes se había denotado a $\boldsymbol{y}$ como la matriz de dimensión $(M) \times (J-1)$ que contenía a los elementos y_{ij}, es decir al número de observaciones en el patrón de covariables i que fueron clasificados en la categoría j . Ahora se denota a $\hat{\boldsymbol{y}}$ como la matriz asociada a los valores ajustados de la variable dependiente; es decir, para cada elemento $\hat{y}_{ij}$ de la matriz $\hat{\boldsymbol{y}}$, se cumple que $\hat{y}_{ij} = m_i \hat{\pi}_{ij}$.

Para que un modelo se ajuste correctamente, se buscaría que:

- La distancia entre y_{ij} y $\hat{y}_{ij}$ fuera pequeña para toda $i = 1, 2, \ldots, M$ y $j = 1, 2, \ldots, J-1$.
- La contribución a la suma total de toda pareja $(y_{ij}, \hat{y}_{ij})$ fuera pequeña en comparación al valor total de la suma.

Para la evaluación de dichos puntos se recurrirán a dos tipos de estadísticas. El primer punto se evaluará con las llamadas estadísticas de "resumen", mientras el segundo, con las estadísticas de "diagnóstico". Se procede a estudiar las estadísticas y herramientas de cada uno de estos dos tipos.

3.2. Estadísticas resumen para bondad de ajuste

Las estadísticas resumen, por más útiles que sean, no proveen información alguna acerca de las contribuciones individuales del modelo, por lo que serán de gran ayuda únicamente para el primero de los dos puntos mencionados anteriormente; es decir, para el ajuste general del modelo y no para el individual, ya que un valor pequeño de estas estadísticas no descarta la posibilidad de una gran desviación del ajuste del modelo por parte de algunas observaciones. En otras palabras, aunque las distancias entre cada valor observado y_{ij} y ajustado $\hat{y}_{ij}$ parecieran ser pequeñas a simple vista, esto no asegura que no existan parejas en las que aunque su distancia sea pequeña, en proporción con las demás distancias sea muy grande.

3.2.1. Distribución de las estadísticas resumen

Para esta sección, se dará énfasis en el término "patrón de covariables". Los patrones de covariables son de gran importancia en este tema dado que, el número de patrones de covariables de un modelo o el número de observaciones dentro de cada patrón de covariables podrían ocasionar problemas durante la evaluación y ajuste del mismo.

Es de gran importancia mencionar, que las pruebas de bondad de ajuste son evaluadas con base en los patrones de covariables que se presentaron en el modelo, no con base en el total de posibles valores de los patrones de covariables, es por eso que el número de patrones de covariables que se utilizaría para un análisis donde se incluyera una variable explicativa continua sería a lo más el número total de observaciones, y no el valor infinito al utilizar una variable continua.

Supóngase la misma definición utilizada en capítulos anteriores para los valores M, P, J, m_i, y_{ij} y $\hat{\pi}_{ij}$. La distribución de todas las estadísticas de bondad de ajuste se basan en que N, el número total de observaciones, sea suficientemente grande. Aquí es donde juegan su papel los patrones de covariables. Si al momento de incrementar el valor de N, también se incrementara el número de patrones de covariables (M), entonces los valores m_i tenderán a ser pequeños para toda i. Bajo estas condiciones, la distribución de las estadísticas de resumen será de la forma *N-asintótica*.

Si en cambio, al incrementar N el número de patrones de covariables no llegara a crecer, entonces m_i incrementará al igual que N para toda i. Bajo estas condiciones, la distribución de las estadísticas de resumen será de la forma *M-*

asintótica.

La diferencia entre ambas distribuciones, así como la importancia de saberlas distinguir quedarán claras conforme se vaya avanzando en el capítulo. Sin embargo, cabe mencionar que de los dos casos, el más complicado es cuando $M \approx N$ (distribución *N-asintótica*), situación común cuando alguna de las variables explicativas es continua.

3.2.2. Estadísticas de resumen: Ji-cuadrada de Pearson y Devianza

- Ji-cuadrada de Pearson

  Recordando que $\hat{y}_{ij} = m_i\hat{\pi}_{ij}$, se define al residuo de Pearson asaociado al patrón de covariables i y la categoría j como:

$$r(y_{ij}, \hat{\pi}_{ij}) = \frac{(y_{ij} - m_i\hat{\pi}_{ij})}{\sqrt{m_i\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} \tag{3.1}$$

  Ahora, la estadística resumen Ji-cuadrada de Pearson, denotada como χ^2, se calculará como:

$$\chi^2 = \sum_{i=1}^{M}\sum_{j=1}^{J-1} r(y_{ij}, \hat{\pi}_{ij})^2 \tag{3.2}$$

- Devianza

  Esta estadística es la misma que la calculada en el capítulo 1 sección 4, y es de gran utilidad para evaluar el ajuste del modelo. Se recuerda que la ecuación para calcular dicha estadística es:

$$\begin{aligned}
D =& -2\log\left[\frac{\text{(Verosimilitud del modelo ajustado)}}{\text{(Verosimilitud del modelo saturado)}}\right] \\
=& -2\sum_{i=1}^{M}\sum_{j=1}^{J} y_{ij}\log\left(\frac{m_i\hat{\pi}_{ij}}{y_{ij}}\right) \\
=& -2\sum_{i=1}^{M}\left[\sum_{j=1}^{J-1} y_{ij}\log\left(\frac{m_i\hat{\pi}_{ij}}{y_{ij}}\right)\right] \\
& -2\sum_{i=1}^{M}\left[\left(m_i - \sum_{k=1}^{J-1} y_{ik}\right)\log\left(\frac{m_i\left(1-\sum_{k=1}^{J-1}\hat{\pi}_{ik}\right)}{\left(m_i - \sum_{k=1}^{J-1} y_{ik}\right)}\right)\right]
\end{aligned}$$

A partir de ella, se obtienen otras estadísticas llamadas residuos de devianza, que también serán de ayuda cuando se tome el tema de estadísticas de diagnóstico. Supóngase que se quisiera ver a D como una suma de cuadrados, de tal forma que

$$D = \sum_{i=1}^{M} d_i^2$$

Entonces d_i sería calculado como

$$d_i = \pm \left[2 \sum_{j=1}^{J} y_{ij} \log \left(\frac{y_{ij}}{m_i \hat{\pi}_{ij}} \right) \right]^{\frac{1}{2}}$$

$$= \pm \left[2 \left(\sum_{j=1}^{J-1} y_{ij} \log \left(\frac{y_{ij}}{m_i \hat{\pi}_{ij}} \right) + \left(m_i - \sum_{k=1}^{J-1} y_{ik} \right) \log \left(\frac{m_i - \sum_{k=1}^{J-1} y_{ik}}{m_i \left(1 - \sum_{k=1}^{J-1} \hat{\pi}_{ik} \right)} \right) \right) \right]^{\frac{1}{2}}$$

Estos valores d_i serán los llamados residuos de devianza, y por lo pronto no se profundizará más sobre éstos ya que se retomarán cuando se hable de las estadísticas de diagnóstico.

Tanto la estadística de resumen χ^2 como D siguen una distribución Ji-cuadrada con $(J-1)(M-(P+1))$ grados de libertad, con P el número de variables y $(P+1)$ el número total de coeficientes estimados en el modelo por cada nivel de la variable de respuesta, incluyendo la constante. Los grados de libertad de dichas distribuciones se siguen del hecho que $(J-1)(M)$ es el número total de coeficientes que se estimarían para el modelo saturado (M por cada categoría a excepción de la de referencia), mientras que $(J-1)(P+1)$ corresponde al número total de coeficientes estimados por el modelo en cuestión.

La hipótesis nula para ambas estadísticas (dado que la Ji-cuadrada de Pearson es una aproximación de la Devianza) es que el modelo en cuestión es "similar" al modelo saturado (el mejor modelo posible, que consiste en un modelo hipotético con M variables predictoras). Por lo tanto, se busca no rechazar la hipótesis nula. Típicamente, esto se lograría si el *p-value* fuera mayor a 0.05.

Existe un problema que es de vital importancia hacer referencia. Como se había hablado antes, cuando $M \approx N$ la distribución es de la clase *N-asintótica*, por lo tanto al aumentar la muestra de observaciones, se incrementa también el número de patrones de covariables y por ende también lo hacen los grados de libertad asociados a la distribución. En consecuencia, cuando $M \approx N$ los *p-values* obtenidos mediante χ^2 y D utilizando la distribución Ji-cuadrada con $(J-1)(M-(P+1))$ grados de libertad serán incorrectos.

Una manera de solucionar el problema de distribución cuando $M \approx N$, es agrupar los datos de tal manera que se pueda utilizar una distribución de clase

M-asintótica. Esto se hace convirtiendo las variables explicativas continuas a categóricas. De esta manera, el número de patrones de covariables no aumentará al aumentar el número de observaciones.

El autor del documento recomienda que antes de aplicar el tratamiento de categorización a las variables continuas, el investigador o analista se cerciore que cuente con una gran cantidad de observaciones a su disposición para la elaboración del modelo, ya que de ser así, se podrían realizar una gran cantidad de particiones en las variables continuas, las cuales tendrían contenidas a una gran cantidad de observaciones en cada una de ellas. La consecuencia directa de esto sería que, la pérdida de información que se produciría al categorizar dichas variables se vería disminuida considerablemente.

3.2.3. Prueba de Hosmer-Lemeshow

Esta prueba puede ser utilizada como alternativa al problema que presentan las pruebas χ^2 y D cuando la distribución es *N-asintótica*, es decir, cuando $M \approx N$. Consiste en agrupar los datos de acuerdo a sus probabilidades estimadas. Supóngase que $M = N$ y supóngase también que los datos se encuentran ordenados de forma descendente de acuerdo a las probabilidades estimadas del nivel de la variable de respuesta al que se le tenga el mayor interés en el estudio. Existen tradicionalmente 2 maneras de realizar el agrupamiento de datos, mismas que se presentan a continuación.

1. Dividir los datos en deciles, cuantiles o cualquier medida que haga que el número de observaciones por cada grupo sea el mismo o casi el mismo. El autor Hosmer, recomienda dividir la muestra en 10 grupos de $N/10$ observaciones cada uno aproximadamente.

2. Dividir los datos de acuerdo a los puntos de corte. En este caso, el número de observaciones en cada grupo puede llegar a variar bastante.

Una vez realizado el agrupamiento de datos en base a la probabilidad del nivel elegido, supóngase se hicieron G grupos distintos. Se procede ahora a calcular los valores estimados de cada grupo por cada categoría de la variable de respuesta (denotados por E_{gj} con g denotando al grupo y j a la categoría) y observados (denotados por O_{gj}) de cada grupo de la siguiente manera:

$$O_{gj} = \sum_{w=1}^{M_g} y_{wj}$$

$$E_{gj} = \sum_{w=1}^{M_g} \hat{\pi}_{wj}$$

Con $y_{wj} = 1$ si la observación w del grupo g tiene el valor j por variable de respuesta y $y_{wj} = 0$ en otro caso, $\hat{\pi}_{wj}$ la probabilidad estimada de que la observación w del grupo g sea categorizada en j y M_g el número total de observaciones dentro del grupo g.

La estadística de Hosmer-Lemeshow (C) se calcula como:

$$C = \sum_{g=1}^{G} \sum_{j=1}^{J-1} \frac{(O_{gj} - E_{gj})^2}{E_{gj}} \tag{3.3}$$

El *p-value* de C bajo la hipótesis nula de que el modelo ajustado es el correcto se obtiene mediante una distribución Ji-cuadrada con $(G-2)(J-1)$ grados de libertad.

Algunas recomendaciones que son de importancia recalcar acerca de esta estadística se mencionan a continuación:

- Diversos investigadores (Hosmer y Lemeshow (2000, p.149) , Klar (1988)) han mostrado que el tipo de agrupamiento de datos número uno (es decir el referente a agrupamientos con aproximadamente el mismo número de observaciones por grupo), es preferido sobre el agrupamiento por puntos de corte, ya que se ajusta mejor a la distribución $\chi^2_{(g-2)(J-1)}$, especialmente cuando las probabilidades estimadas son pequeñas (menores a 0.2).
- Cuando se utilizan muy pocos grupos para calcular C, se corre el riesgo de no tener la sensibilidad necesaria para distinguir entre valores observados y esperados. Cuando C es calculada utilizando menos de seis grupos, es muy probable que indique que el modelo ajusta bien a los datos, aún cuando no lo haga. El número diez para el total de grupos es ampliamente recomendado según Hosmer y Lemeshow.
- Al mismo tiempo, se recomienda que no tengan muy pocas observaciones los grupos, dado que podría ser incorrecto el resultado de la prueba.
- Se recomienda que siempre que se realice esta prueba y se vaya a aceptar la hipótesis nula (que el modelo ajusta bien a los datos), realizar posteriormente un análisis de cada uno de los residuos de esta estadística de manera individual.
- Es importante hacer énfasis en que C no puede ser utilizada para comparar diferentes modelos que cumplan con la hipótesis de que se ajusten bien a los datos. Esta estadística únicamente será de ayuda para decidir

si un modelo ajusta bien a los datos o no; sin embargo, si un modelo que ajustara bien a los datos tuviera un *p-value* mayor al de otro modelo que también se ajustara bien a los datos, no se podría concluir que uno es mejor que otro.

3.2.4. Tablas de clasificación

Una manera aparentemente atractiva de resumir los resultados de un modelo de regresión logística multinomial es por medio de tablas de clasificación; sin embargo, la clasificación no está relacionada con el ajuste del modelo (como se verá más adelante). Es por ello que no deben de utilizarse como medida de bondad de ajuste de ningún modelo dichas tablas, aunque sí serán de ayuda para plasmar en ellas los resultados de una herramienta bastante útil en temas de clasificación para detectar posibles puntos de corte: la curva ROC.

Las tablas de clasificación se crean contrastando los valores observados con los valores predichos derivados de las probabilidades ajustadas por el modelo.

Un ejemplo de dicha herramienta es la tabla 3.1. Para la creación de la tabla de clasificación mostrada, se utilizaron 23 observaciones con $Y = 0$, de las cuales 21 fueron predichas correctamente. Con respecto a las observaciones con $Y = 1$ que suman 20, 16 fueron predichas correctamente por el modelo.

Tabla 3.1: Ejemplo de tabla de clasificación

		Observados	
		$Y = 0$	$Y = 1$
Predichos	$Y = 0$	21	4
	$Y = 1$	2	16

Con ayuda del ejemplo mostrado anteriormente, se definirá a la "especificidad" como el porcentaje de datos observados con valor $Y = 0$ que fueron predichos correctamente, mientras que el término "sensibilidad" se refiere al porcentaje de datos observados con valor $Y = 1$ que fueron predichos correctamente. En dicho ejemplo, la especificidad tiene un valor de $(21/23)\cdot 100$ que corresponde a 91.3 % y la sensibilidad $(16/20)\cdot 100$ que corresponde a 80 %.

Estos dos términos no se pueden generalizar para J con $J > 2$, pero de igual manera, se puede proceder a calcular el porcentaje de las observaciones con variable de respuesta igual a la categoría j que fueron pronosticados correctamente

por el modelo, para todo $j = 1, 2, \ldots, J$.

Cabe recordar que el modelo no obtiene en si a valores pronosticados de las categorías, sino que obtiene probabilidades. Ahora se explicará cómo a partir de estas probabilidades obtener los valores predichos.

Si se quisiera encontrar el número de observaciones dentro del patrón de covariables i que se espera tengan la categoría j, simplemente se hace el producto del número de individuos dentro del patrón de covariables i por la probabilidad de ser catalogado en la categoría j; es decir se obtiene a $\hat{y}_{ij}$.

Así también, el porcentaje de observaciones que se espera sean catalogadas correctamente para la categoría j se calcularía como:

$$\frac{\sum_{i=1}^{M} \text{mín}(m_i \hat{\pi}_{ij}, y_{ij})}{\sum_{i=1}^{M} y_{ij}} \cdot 100$$

Una situación común es que $m_i = 1$ para todo $i = 1, 2, \ldots, M$. Si esto llegara a suceder, los valores ajustados $\hat{y}_{ij} = m_i \hat{\pi}_{ij}$ serían todos menores a uno. En estos casos, la categoría en la que el modelo pronosticará a la observación i podría ser la categoría con mayor probabilidad de ocurrir; es decir, bajo esta perspectiva y únicamente para el caso $M = N$:

- $\hat{y}_{ij} = 1$, si $\hat{\pi}_{ij} = \text{máx}(\hat{\pi}_{ik})$, con $k = 1, 2, \ldots, J$.
- $\hat{y}_{ij} = 0$ en cualquier otro caso.

Asimismo, dependiendo del tipo del estudio y sus objetivos, se podría elegir algún punto de corte (denotado por c), a partir del cual las observaciones con probabilidad ajustada mayor a c, pronosticarán el valor uno, mientras que las observaciones con probabilidad ajustada menor a c, pronosticarán cero.

Es importante mencionar que ninguno de estos dos acercamientos para clasificar es mejor que el otro, pues cada uno será de mayor utilidad dependiendo de la naturaleza de los datos. A continuación se mencionan algunas deficiencias para cada uno de los dos métodos de clasificación mencionados.

Respecto a la clasificación vía máxima probabilidad:

- Aunque la interpretación de este método es sencilla y su implementación no requiere mayor esfuerzo, la clasificación vía máxima probabilidad comúnmente sólo se utiliza cuando la muestra o universo de observaciones que se utilizará para elaborar el modelo está balanceada; es decir, el número de observaciones asignado a cada categoría de la variable de respuesta es similar.

De no ser este el caso, es muy probable que todas o casi todas las observaciones tengan como probabilidad mayor a la categoría con el mayor número de observaciones asociadas.

Respecto a la clasificación vía punto de corte:

- Se debe tener extremo cuidado si se quisiera optar por dicha forma de pronosticar, dado que se deben establecer medidas para asegurar que una observación sólo pueda tener una única categoría y no más ni menos. Este problema se presenta sobre todo cuando $J > 2$. Cuando $J = 2$, este método tiende a ser poderoso, ya que a diferencia del método de máxima probabilidad, no depende de que las observaciones se encuentren balanceadas.
- Mientras mayor sea el número de observaciones con probabilidades ajustadas cercanas al punto de corte, mayor será el número de clasificaciones incorrectas que tendrá el modelo.

Sería intuitivo pensar que si el modelo predijera acertadamente los valores, esto sería una justificación suficiente para asegurar que el modelo ajusta bien a los datos. Sin embargo, existen ocasiones en las que este no es el caso. El hecho de que la clasificación sea precisa o imprecisa no tiene relación con los criterios de bondad de ajuste: que las distancias entre valores observados y las probabilidades ajustadas sean pequeñas y dentro de los estándares de variación del modelo.

A continuación se lista un conjunto de problemas que se pudieran presentar al utilizar tablas de clasificación para decidir si el ajuste de un modelo es bueno o no:

- Al utilizar tablas de clasificación, la clasificación resulta ser sensible a los tamaños de los grupos de observaciones por categoría, por lo que siempre favorecerá en términos predictivos al grupo con el mayor número de observaciones, hecho totalmente independiente del ajuste del modelo.

- Haciendo mención al concepto de "probabilidad", de entre N sujetos todos con la misma estimación $\hat{\pi}$ de presentar algún evento de interés, entonces el número de sujetos que se espera presenten dicho evento es $N\hat{\pi}$ y el número de sujetos que se espera no lo presenten es $N(1 - \hat{\pi})$. Ahora, supóngase para la clasificación de 100 sujetos que todos poseen la misma probabilidad de pertenecer a la categoría j; en este ejemplo $\hat{\pi}_{ij} = 0.51$ para todo i perteneciente a dichas observaciones.

  Entonces, bajo el criterio de clasificación, todos los sujetos serían predichos como si presentaran el evento, pero al mismo tiempo, únicamente

$N\hat{\pi} = 51$ sujetos se esperan que lo presenten. Por lo tanto, un total de 49 sujetos serían clasificados incorrectamente.

En resumen, las tablas de clasificación son más apropiadas cuando el objetivo principal del análisis sea clasificar; en cualquier otro caso, podrán ser utilizadas únicamente como complemento de otras pruebas de ajuste más robustas y no se les dará mucha importancia.

Asimismo, se debe de recordar que la probabilidad calculada es una estimación de la media, o proporción, de los sujetos con patrón de covariables i que "podrían" categorizarse en j.

3.2.5. Curva ROC

Esta herramienta es útil comúnmente cuando $M = N$ y se quisieran realizar predicciones con base en puntos de corte. Sin embargo, las curvas ROC están elaboradas para modelos de regresión logística con únicamente dos niveles en la variable de respuesta, por lo que si se quisiera utilizar dicha herramienta para modelos multinomiales, se tendría que tomar al nivel de referencia como el mismo, pero a los demás niveles como a uno sólo, con la probabilidad de ocurrencia como la suma de las probabilidades de todas las categorías indexadas. De igual forma se puede realizar la mezcla de variables tomando como nivel individual a cualquier otro nivel, no necesariamente el de referencia.

Una vez sustituyendo los J niveles de la variable de respuesta por los nuevos dos niveles, se procede a calcular la sensibilidad y especificidad del modelo.

Tanto la sensibilidad como la especificidad dependen de un sólo punto de corte para su clasificación, aunque normalmente se calculan para varios puntos de corte para poder así obtener de entre ellos al más adecuado de acuerdo a los objetivos del estudio, que usualmente consisten en maximizar la sensibilidad, la especificidad o ambas. La curva ROC entonces, funge como auxiliar para encontrar este mejor punto de corte.

La curva ROC grafica la probabilidad de detectar verdaderos positivos y falsos positivos para cualquier rango de puntos de corte. El porcentaje de verdaderos positivos (o *true positive rate* en inglés) se obtiene mediante la sensibilidad, mientras que el porcentaje de falsos positivos (*false positive rate*) se obtiene mediante uno menos la especificidad.

Si el objetivo del análisis fuera encontrar el punto de corte óptimo para propósi-

tos de clasificación, con ayuda de dicha curva se elegiría el punto de corte que maximizara tanto sensibilidad como especificidad.

En la figura 3.1 se puede apreciar un ejemplo de una curva ROC, así como el punto de corte que maximiza tanto sensibilidad como especificidad.

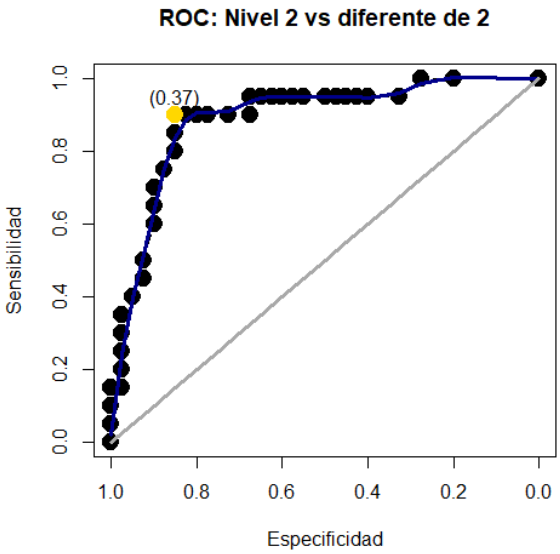

Figura 3.1: Ejemplo de curva ROC. Punto de corte óptimo de color amarillo.

Además de encontrar el punto de corte óptimo (dado por el punto amarillo en el ejemplo), la curva ROC también ayuda a saber qué tan bien clasifica dicho modelo, al calcular el área debajo de la curva, que de ahora en adelante se denotará por AUC (*Area Under the Curve* en inglés).

Como regla general, Hosmer y Lemeshow (2000, p.162) sugieren categorizar la calidad del AUC de acuerdo a los siguientes puntos de corte:

- AUC=0.50 $\rightarrow$ Clasificación mala, aleatoria.
- 0.70 $\leq$AUC$\leq$ 0.80 $\rightarrow$ Clasificación aceptable.
- 0.80 $\leq$AUC$\leq$ 0.90 $\rightarrow$ Clasificación excelente.
- AUC$\geq$ 0.90 $\rightarrow$ Clasificación excepcional.

Cabe mencionar que dichos autores afirman que es extremadamente inusual encontrar AUC con valores mayores a 0.90.

Si se tuviera en el modelo original un total de J niveles para la variable de respuesta, y fuera de interés realizar predicciones con base en puntos de corte, es recomendable generar J gráficas de curvas ROC, donde en cada una se grafique al nivel j contra el nivel que agrupa todos los valores distintos de j, con $j = 1, 2, \ldots, J$.

Como sugerencia personal, si el objetivo del estudio fuera clasificar a las observaciones en diversas categorías y se quisiera emplear algún modelo de regresión logística multinomial para ello, el autor del documento recomienda utilizar el método de clasificación por máxima probabilidad si la densidad de observaciones en cada categoría fuera similar, y utilizar el método de clasificación por puntos de corte si las densidades entre categorías llegaran a ser desiguales, ya que de utilizar el método de máxima probabilidad para este caso, todas las observaciones serían clasificadas en la categoría con mayor densidad. Un ejemplo de este caso se presenta a continuación.

Supóngase se quisiera utilizar un modelo de regresión logística multinomial para predecir la categoría de 100,000 observaciones. Supóngase también que el número de categorías es 3, y dentro de éstas, las que son de mayor importancia predecir correctamente son la categoría uno y dos, con 5,000 y 20,000 observaciones en cada una de ellas, respectivamente. En este caso, dado que son de mayor importancia las categorías con menor densidad, y la diferencia entre las densidades de las tres categorías es muy grande, el autor del documento recomendaría fijar puntos de corte apoyados en las tres curvas ROC que se obtendrían al agrupar las categorías dos a uno, dado que por ejemplo, realizar dos análisis diferentes no sería una buena opción, al ser el propósito del estudio la predicción de categorías.

3.2.6. Medidas *pseudo-*R^2

En regresión lineal, la R^2 es una estadística utilizada para medir la bondad de ajuste del modelo. Dado que esta estadística es calculada utilizando mínimos cuadrados, no puede ser utilizada para modelos de regresión logística multinomial, pues en dichos modelos los coeficientes son calculados vía máxima verosimilitud. Sin embargo, se han desarrollado otras estadísticas cuya interpretación es muy similar a la R^2 y pueden ser utilizadas en modelos de regresión logística multinomial. Dichas estadísticas reciben el nombre de "*pseudo-*R^2".

En general, las medidas *pseudo-*R^2 para regresión logística multinomial se basan en comparaciones entre los valores ajustados del modelo actual y los ajustados por el modelo nulo, es decir, el modelo que no utiliza variables explicativas y

cuyo único coeficiente es una constante, y por ende, realmente no evalúan bondad de ajuste. Sin embargo, las medidas *pseudo*-R^2 pueden ser de gran utilidad para comparar diferentes modelos ajustados a los mismos datos.

Actualmente existe una gran cantidad de maneras para calcular la *pseudo*-R^2 para regresión logística multinomial, por lo que Mittlböck y Schemper (1996) las estudiaron con mayor profundidad con el objetivo de encontrar las medidas *pseudo*-R^2 que cumplieran con las siguientes características:

1. La medida debe de tener una interpretación sencilla.

2. El valor mínimo de la medida puede ser cero y el valor máximo uno.

De entre la gran variedad de maneras de calcular la *pseudo*-R^2 para regresión logística multinomial, dichos investigadores únicamente encontraron dos que cumplían con las dos características mencionadas: la *pseudo*-R^2 de correlación de Pearson de valores observados con la probabilidad ajustada y la *pseudo*-R^2 como suma de cuadrados de una regresión lineal.

Así, el valor de la *pseudo*-R^2 del coeficiente de correlación de Pearson se calcula como:

$$pseudo\text{-}R^2 = \frac{\left[\sum_{j=1}^{J-1}\sum_{i=1}^{M}(y_{ij}-\bar{y}_j)(m_i\hat{\pi}_{ij}-m_i\bar{\hat{\pi}}_j)\right]^2}{\left[\sum_{j=1}^{J-1}\sum_{i=1}^{M}(y_{ij}-\bar{y}_j)^2\right]\cdot\left[\sum_{j=1}^{J-1}\sum_{i=1}^{M}(m_i\hat{\pi}_{ij}-m_i\bar{\hat{\pi}}_j)^2\right]} \tag{3.4}$$

Sin embargo, la *pseudo*-R^2 denotada por la ecuación 3.4 tiene el gran problema de que calcularía a la correlación tanto negativa como positiva de y y $\hat{\pi}$ como cerca de 1, siendo este un grave problema dado que si $\hat{\pi}$ clasificara exactamente de manera opuesta a la debida, es decir si por ejemplo $m_i\hat{\pi}_{ij} = m_i - y_{ij}$ o en un caso más concreto si $y_{ij} = m_i$ y $\hat{\pi}_{ij} = 0$, el modelo no clasificaría correctamente a ninguna de las observaciones; sin embargo, la *pseudo*-R^2 propuesta en la ecuación 3.4 tendría valor 1, que indicaría ajuste perfecto.

Este es un grave problema incluso si se utilizara la *pseudo*-R^2 propuesta exclusivamente para comparar modelos, dado que se podrían preferir modelos con errores graves en vez de modelos con errores moderados. Dado este problema, si se quisiera utilizar la *pseudo*-R^2 descrita en la ecuación 3.4, antes se debe de comprobar que el modelo se ajuste lo suficientemente bien como para evitar lo más posible las complicaciones mencionadas.

La otra manera de calcular la *pseudo*-R^2 que sugieren Mittlböck y Schemper (1996), es la R^2 de regresión lineal cuya fórmula modificada para el caso de

regresión logística multinomial es:

$$pseudo\text{-}R^2 = 1 - \left(\frac{\sum_{j=1}^{J-1} \sum_{i=1}^{M} (y_{ij} - m_i \hat{\pi}_{ij})^2}{\sum_{j=1}^{J-1} \sum_{i=1}^{M} \left(y_{ij} - m_i \frac{\sum_{k=1}^{M} y_{kj}}{\sum_{k=1}^{M} m_k} \right)^2} \right) \tag{3.5}$$

Otra versión de *pseudo*-R^2 no mencionada por los autores debido a que su interpretación no es tan sencilla como las otras dos, ya que se calcula utilizando log-verosimilitudes, es presentada a continuación.

Sean L_0 la log-verosimilitud del modelo que contiene únicamente el intercepto y L_P la del modelo con las P variables más el intercepto, entonces la *pseudo*-R^2 basada en log-verosimilitudes se calcula de la forma:

$$pseudo\text{-}R^2 = \frac{L_0 - L_P}{L_0} = 1 - \frac{L_P}{L_0} \tag{3.6}$$

El máximo valor de esta *pseudo*-R^2 se obtiene cuando L_P se aproxima a la log-versosimilitud del modelo saturado, denotado por L_s, pues L_s=0 y la *pseudo*-R^2 tendría valor igual a uno.

Sin embargo, existe un caso donde el máximo valor posible para la ecuación 3.6 ya no sería uno. Si el número de observaciones fuera diferente al número de patrones de covariables, la log-verosimilitud del modelo saturado no sería uno, por lo que para preservar las propiedades de la *pseudo*-R^2 la fórmula se modificaría a:

$$pseudo\text{-}R^2 = \frac{L_0 - L_P}{L_0 - L_s} \tag{3.7}$$

donde el valor máximo sí puede ser uno.

En estos casos, se puede calcular el valor de la log-verosimilitud para el modelo saturado de manera algebraica teniendo presente que para dicho modelo

$$\pi_{ij_s} = \frac{y_{ij}}{m_i}$$

O bien se puede aproximar mediante la ecuación:

$$L_s = L_P + 0.5 D_P$$

Esta última expresión proviene del hecho que:

$$\begin{aligned} D_p &= -2\log\left[\frac{\text{(Verosimilitud del modelo ajustado)}}{\text{(Verosimilitud del modelo saturado)}}\right] \\ &= -2\left(L_p - L_s\right) \\ &\Rightarrow \\ -0.5D_P &= L_P - L_s \\ &\Rightarrow \\ L_s &= L_P + 0.5D_P \end{aligned}$$

Con D_P la devianza del modelo con P variables y L_P, L_s las log-verosimilitudes de los modelos con P variables y el modelo saturado, respectivamente.

Antes de terminar con las *pseudo*-R^2, es importante mencionar que desafortunadamente, los valores que normalmente toman las *pseudo*-R^2 para regresión logística multinomial tienden a ser muy pequeños, tanto así que un buen modelo podría llegar a tener incluso valores de *pseudo*-R^2 cercanos a 0.30, por lo que es necesario mencionar dicha particularidad de los modelos logísticos multinomiales si se presentaran resultados de dichos modelos frente a audiencias familiarizadas únicamente con el modelo de regresión lineal, donde una R^2 con valor cercano a 0.30 indicaría que el modelo en cuestión es ineficiente.

3.3. Estadísticas de diagnóstico para bondad de ajuste

Las estadísticas resumen proveen un único número que resume el ajuste del modelo. Aunque esto es realmente útil, sería erróneo justificar el ajuste del modelo únicamente con este valor, dado que podrían existir problemas de ajuste con específicas variables o patrones de covariables, los cuales el investigador o analista nunca se percataría si terminara su análisis de ajuste en este momento. Para complementar el análisis de las estadísticas resumen, se utilizan las llamadas "estadísticas de diagnóstico".

3.3.1. *Leverage* y residuos

Recordando a las estadísticas de ajuste

$$\chi^2 = \sum_{i=1}^{M} \sum_{j=1}^{J-1} r(y_{ij}, \hat{\pi}_{ij})^2$$

y

$$D = \sum_{i=1}^{M} d_i^2$$

Algunas de las medidas de diagnóstico de la regresión logística multinomial corresponden en utilizar a los componentes individuales de la suma de cada una de estas dos estadísticas, es decir, a $r(y_{ij}, \hat{\pi}_{ij})^2$ y d_i^2. A estos valores se les llamarán residuos de Pearson y residuos de devianza, respectivamente, y servirán para detectar patrones de covariables que podrían ser *outliers*, o bien que el modelo tuvo problemas en ajustar. Aunque en general, la decisión de juzgar a un patrón como atípico proviene del especialista del área, valores grandes de estos residuos tienden a ser una fuerte indicación de que el patrón de covariables seleccionado pudiera ser un *outlier*.

Además de los residuos mencionados anteriormente, otros conceptos importantes que contribuyen al desarrollo de las estadísticas de diagnóstico corresponden a los términos "matriz sombrero" y "valores de apalancamiento" (*leverage values*), donde los segundos se obtienen a partir de la matriz sombrero correspondiente al modelo de regresión logística multinomial y fungirán como otra estadística de diagnóstico además de los residuos. A continuación se presentan dichas herramientas.

Sea $\mathbf{X}$ la matriz de dimensión $M \times (P+1)$ cuyos contenidos sean los valores de todos los patrones de covariables para cada variable más el intercepto. $\mathbf{X}$ es usualmente llamada "matriz diseño". En regresión lineal, la matriz sombrero (denotada por $\mathbf{H}$) es calculada como $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Esta matriz sombrero es muy útil, dado que cumple con la característica de que $\hat{y} = \mathbf{H}y$; es decir, que la matriz $\mathbf{H}$ multiplicada por los valores observados y es igual a los valores ajustados $\hat{y}$. En consecuencia, los residuos de la regresión lineal $(y - \hat{y})$ expresados en términos de la matriz sombrero son de la forma $(\mathbf{I} - \mathbf{H})y$, donde $\mathbf{I}$ es la matriz identidad de dimensión $M \times M$.

Una aproximación a la matriz sombrero para el caso del modelo de regresión logística multinomial, calculada por Pregibon (1981) es:

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}_B(\mathbf{X}_B'\mathbf{V}\mathbf{X}_B)^{-1}\mathbf{X}_B'\mathbf{V}^{1/2} \tag{3.8}$$

Con $\mathbf{X}_B$ una matriz de bloques de dimensión $M \times (P+1)$ con cada bloque igual

a la matriz de dimensión $(J-1)\times(J-1)$ con valores:

$$X_{B_{ip}} = \begin{bmatrix} X_{ip} & 0 & \cdots & 0 \\ 0 & X_{ip} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & X_{ip} \end{bmatrix}$$

Donde X_{ip} representa el valor de la p-ésima variable asociada al patrón de covariables i.

Esta modificación a la matriz de diseño $\mathbf{X}$ fue hecha para que fuera compatible la multiplicación de dicha matriz con las matrices $\mathbf{V}$ y $\mathbf{V}^{1/2}$.

También, $\mathbf{V}$ corresponde a una matriz de bloques diagonal de tamaño $M\times M$, donde cada $\mathbf{V}_{ii}$ se compone de una matriz cuadrada de dimensión $(J-1)\times(J-1)$ y para cada valor $g,h=1,2,\ldots,J-1$ se tiene que el elemento perteneciente al g-ésimo renglón y a la h-ésima columna de V_{ii}, es decir $V_{ii_{gh}}$, tendrá el valor:

$$V_{ii_{gh}} = m_i\hat{\pi}_{ig}(\delta_{gh}-\hat{\pi}_{ih}) \tag{3.9}$$

Con δ_{gh} definida como la delta de Kronecker donde $\delta_{gh}=1$ cuando $g=h$, y $\delta_{gh}=0$ en cualquier otro caso.

Cabe destacar que $\mathbf{V}^{1/2}$ se compone de una matriz de bloques nuevamente diagonal, donde cada bloque de la diagnal se compone de $V_{ii}^{1/2}$ y donde $V_{ii}^{1/2}=QD^{1/2}Q^{-1}$, con Q la matriz de eigenvectores y D la matriz diagonal de eigenvalores de V_{ii} (recuérdese que únicamente una matriz elevada a una potencia es igual a la matriz con entradas elevadas a dicha potencia si la matriz es diagonal, como en el caso de D).

Esta medida de utilizar una matriz de bloques tiene el propósito de definir a la matriz sombrero en una sola ecuación, ya que si no se hiciera uso de esta notación, se tendrían que definir un número de ecuaciones equivalentes al número de patrones de covariables y no se podría generalizar a la matriz sombrero de una manera tan sencilla.

Ahora se define otro concepto para el diagnóstico de la regresión logística multinomial, que es el de los valores de apalancamiento, mejor conocidos como *leverage values*.

Los *leverage values* se utilizan en modelos de regresión lineal para medir la distancia entre las observaciones con respecto a su media general. El *leverage value* asociado a la observación i (h_i) es el i-ésimo elemento de la diagonal de la matriz sombrero $\mathbf{H}$ (bajo el modelo lineal $\mathbf{H}$ no es matriz por bloques). Generalmente se utilizan como herramienta de búsqueda de observaciones que sean influyentes, ya que si alguna de ellas se llegara a encontrar muy alejada

de la media, ésta podría influir de manera significativa en los coeficientes de los parámetros estimados.

La propiedad más importante y en la cual recae la relevancia de los *leverage values* para el modelo de regresión lineal es que los *leverage values* son monótonos; es decir, mientras más alejado se encuentre la observación de la media, mayor será el valor de su *leverage value* y más probable será que dicha observación sea un dato influyente. Para el caso de los modelos de regresión logística multinomial, esta propiedad de monotonía no se cumple, es por eso que se debe de tener cuidado con dicha estadística y no confiarse únicamente en ella para hacer conclusiones.

Para el modelo logístico considerando únicamente dos niveles dentro de la variable de respuesta ($J = 2$), la interpretación que se le pudiera dar al *leverage value* asociado a un patrón de covariables i dependerá de la probabilidad estimada asociada a i, $\hat{\pi}_{i1}$. En este caso, la interpretación del *leverage value* del patrón de covariables i será la misma a la de los modelos de regresión lineales cuando el valor de $\hat{\pi}_{i1}$ se encuentre entre 0.1 y 0.9, mientras que si $\hat{\pi}_{i1}$ se encuentra fuera de ese rango, no se puede concluir nada al respecto.

Para el caso multinomial ($J > 2$), el autor se abstiene de describir el comportamiento de dicha estadística por ausencia de dominio del tema; sin embargo, algunos autores sugieren calcular al *leverage value* del patrón de covariables i como el determinante del bloque H_{ii} de la matriz $\mathbf{H}$, esto es $h_i = |H_{ii}|$ (recuérdese que $\mathbf{H}$ también es una matriz por bloques).

Otros autores como Hosmer y Lemeshow (2000, p.172), han sugerido también el uso de la estadística:

$$b_i = |x_{B_{i\cdot}}(\mathbf{X}'_B\mathbf{V}\mathbf{X}_B)^{-1}x'_{B_{i\cdot}}|$$

Con $x_{B_{i\cdot}}$ el i-ésimo renglón de la matriz por bloques $\mathbf{X}_B$, para encontrar una medida de distancia entre el patrón de covariables i y la media de los patrones de covariables (nótese que $|x_{B_{i\cdot}}(\mathbf{X}'_B\mathbf{V}\mathbf{X}_B)^{-1}x'_{B_{i\cdot}}|$ es el determinante de una matriz de $(J-1)\times(J-1)$).

Aunque no se profundizará más sobre estas dos estadísticas en el aspecto teórico (h_i y b_i), ambas se calcularán en el estudio realizado en el capítulo siete y se analizarán sus resultados.

Podrían existir casos en los que $\mathbf{H}$ sea calculada por observaciones en vez de por patrones de covariables. De ser este el caso, entonces $\mathbf{H}$ tendría una dimensión de $N \times N$ y no de $M \times M$. Es importante saber si el paquete estadístico que se desea utilizar para el análisis calcula la matriz sombrero, y por ende las estadísticas h_i y b_i por patrones de covariables o por observaciones. Se recomienda ampliamente que se trabaje con *leverage values* y valores b_i calculados por patrones de covariables. Esto es de gran importancia sobre todo cuando el número

de patrones de covariables M es mucho menor a N, o si algunos valores m_i son muy grandes. Esta misma recomendación aplicará para todas las estadísticas de diagnóstico. Un ejemplo de cómo influye el calcular las estadísticas por patrón de covariable en vez de por observación es el siguiente, utilizando como estadística de diagnóstico a los residuos de Pearson $r(y_{ij}, \hat{\pi}_{ij})$.

Cuando el número de patrones de covariables es similar a N, existe el riesgo de que se falle en identificar patrones de covariables atípicos o con un ajuste pobre. Por ejemplo, considérese un patrón de covariables compuesto por m_i observaciones, con $y_{ij} = 0$ para alguna $j = 1, 2, \ldots, J-1$ y con su respectiva probabilidad estimada $\hat{\pi}_{ij}$. Entonces, los residuos de Pearson definidos en la ecuación 3.1 y calculados individualmente para cada observación de este patrón de covariables tendrán el valor:

$$r(y_{ij}, \hat{\pi}_{ij}) = \frac{(y_{ij} - m_i\hat{\pi}_{ij})}{\sqrt{m_i\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} = \frac{(0 - \hat{\pi}_{ij})}{\sqrt{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} = -\sqrt{\frac{\hat{\pi}_{ij}}{(1-\hat{\pi}_{ij})}}$$

Mientras que los residuos de Pearson calculados por patrón de covariables estarían dados por:

$$r(y_{ij}, \hat{\pi}_{ij}) = \frac{(y_{ij} - m_i\hat{\pi}_{ij})}{\sqrt{m_i\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} = \frac{(0 - m_i\hat{\pi}_{ij})}{\sqrt{m_i\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}} = -\sqrt{m_i}\sqrt{\frac{\hat{\pi}_{ij}}{(1-\hat{\pi}_{ij})}}$$

que aumenta negativamente a manera que m_i aumenta.

Si por ejemplo, $m_i = 1$ y $\hat{\pi}_{ij} = 0.5$, entonces $r(y_{ij}, \hat{\pi}_{ij}) = -1$, que no es un residuo grande. En cambio, si $m_i = 16$, entonces $r(y_{ij}, \hat{\pi}_{ij}) = -4$, que indicaría un grave problema en el ajuste del modelo para dicho patrón de covariables.

Si se supusiera $m_i = 16$, $\hat{\pi}_{ij} = 0.5$ y $y_{ij} = 0$, intuitivamente se sabría que el modelo ajustó mal al patrón de covariables seleccionado, por lo que se buscaría que los residuos también sugirieran eso. Sin embargo, sólo el residuo calculado por patrón de covariables pudo corroborar el problema.

3.3.2. Deltas

Hasta ahora se han definido cuatro posibles estadísticas de diagnóstico: los residuos de Pearson ($r(y_{ij}, \hat{\pi}_{ij})$) y los residuos de devianza (d_i), que sirven para buscar datos atípicos, y los *leverage values* (h_i) y la estadística b_i, que dan una intuición de posibles datos influyentes.

Otras estadísticas de diagnóstico que son de gran utilidad para el análisis de

bondad de ajuste consisten en evaluar el efecto que tendría el eliminar del modelo a todas las observaciones de cierto o ciertos patrones de covariables que se piensan podrían ser datos atípicos o influyentes. Lo que es de mayor interés en esta práctica es observar el cambio en los valores de los coeficientes estimados $\hat{\beta}_{pj}$, así como de las estadísticas resumen χ^2 y D al remover cierto patrón, ya que mientras cambios grandes de $\hat{\beta}_{pj}$ indicarán gran influencia del patrón removido en el modelo, cambios grandes en χ^2 y D podrían evidenciar la presencia tanto de *outliers* como de patrones influyentes.

El cambio mencionado en el párrafo anterior se obtiene, para el caso de los coeficientes estimados por el modelo, mediante una resta estandarizada entre $\hat{\beta}_{pj}$ y $\hat{\beta}_{pj}^{(-i)}$, que representan los estimadores por máxima verosimilitud para la variable p y la categoría j calculados mediante todos los patrones de covariables y mediante todos los patrones de covariables excepto el i-ésimo, respectivamente. Una vez realizada esta diferencia, se procede a dividirla entre el error estándar del coeficiente estimado $\hat{\beta}_{pj}$, multiplicado por el valor asociado a algún cuantil $1-\alpha/2$ de una distribución normal estándar.

Dicha estandarización se utiliza para quitar la importancia de la escala de medición y poder notar el cambio verdadero entre los valores de ambos coeficientes, en proporción a su correspondiente intervalo de confianza a algún nivel $(1-\alpha)100\,\%$ elegido. A esta estadística se le denotará por $\Delta\hat{\beta}_{pj}^{(-i)}$.

A manera de ecuación, la estadística anterior se calcula de la siguiente manera:

$$\Delta\hat{\beta}_{pj}^{(-i)} = \frac{\hat{\beta}_{pj} - \hat{\beta}_{pj}^{(-i)}}{z_{1-\alpha/2} \cdot \hat{\mathrm{SE}}(\hat{\beta}_{pj})} \tag{3.10}$$

El valor de $\Delta\hat{\beta}_{pj}^{(-i)}$ se puede pues, interpretar como la proporción de la distancia entre $\hat{\beta}_{pj}$ y $\hat{\beta}_{pj}^{(-i)}$ con respecto a la distancia máxima que se pudiera alejar el verdadero valor β_{pj} de su valor estimado, $\hat{\beta}_{pj}$, y aún así seguir perteneciendo al intervalo de confianza a un nivel $100(1-\alpha)\,\%$.

Si por ejemplo, $\Delta\hat{\beta}_{pj}^{(-i)} = 2$, entonces el coeficiente $\hat{\beta}_{pj}^{(-i)}$ se encontraría a una distancia equivalente a dos veces lo que podría alejarse el verdadero valor β_{pj} de su estimador $\hat{\beta}_{pj}$ según el intervalo de confianza a un nivel de $100(1-\alpha)\,\%$. Para patrones de covariables no influyentes se esperaría entonces, que todos los valores $\Delta\hat{\beta}_{pj}^{(-i)}$ fueran menores a uno bajo algún nivel de confianza alto, pues esto significaría que el cambio en los coeficientes se sigue encontrando dentro del intervalo de confianza permitido y no sería señal de influencia por parte del patron de covariables i hacia el modelo.

Cabe recordar que el error estándar estimado del coeficiente $\hat{\beta}_{pj}$, $\hat{\mathrm{SE}}(\hat{\beta}_{pj})$, se obtiene a partir de la matriz de varianzas y covarianzas estimada calculada en el capítulo uno.

De manera similar a $\Delta\hat{\beta}_{pj}^{(-i)}$, se definen a $\chi^{2(-i)}$ y $D^{(-i)}$ como la Ji-cuadrada de Pearson y la devianza calculadas mediante todos los patrones de covariables a excepción del i-ésimo patrón de covariables, respectivamente. Las estadísticas de diagnóstico antes presentadas se expresan a continuación:

$$\Delta\chi^{2(-i)} = \chi^2 - \chi^{2(-i)} \tag{3.11}$$

$$\Delta D^{(-i)} = D - D^{(-i)} \tag{3.12}$$

Estas estadísticas son muy útiles, pues permiten encontrar aquellos patrones de covariables en los que el ajuste es pobre (mediante $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$) y al mismo tiempo aquellos patrones que tienen una gran influencia en los parámetros del modelo (mediante $\Delta\hat{\beta}_{pj}^{(-i)}$, $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$).

Los resultados que se pueden obtener de dichas estadísticas, así como su interpretación, serán los siguientes:

- $\Delta\chi^{2(-i)}$

  Para $\Delta\chi^{2(-i)} = \chi^2 - \chi^{2(-i)}$, el mejor hipotético caso sería que la estadística χ^2 fuera pequeña y además que $\chi^{2(-i)}$ tuviera un valor muy cercano a χ^2 para toda $i = 1, 2, \ldots, M$, haciendo a $\Delta\chi^{2(-i)}$ muy cercana a cero.

  Las interpretaciones de cada caso que podría tomar $\Delta\chi^{2(-i)}$ para algún i en específico se mencionan a continuación:

  - Caso $\Delta\chi^{2(-i)} = \chi^2 - \chi^{2(-i)}$ muy grande

    Este caso implicaría que $\chi^{2(-i)}$ fuera mucho menor a χ^2, lo que implica que el remover al patrón de covariables i del estudio resultaría en un modelo con un mejor ajuste a los datos restantes. Por mejor ajuste puede referirse a dos situaciones. La primera, que al remover dicho patrón las probabilidades de todos los demás patrones que no fueron eliminadas se aproximen más a la realidad; es decir, las probabilidades asociadas a la categoría observada aumentan, mientras las asociadas a las categorías no observadas disminuyen. La segunda, que el residuo del patrón eliminado fuera demasiado grande, y aunque las probabilidades de los demás patrones no sean modificadas, la sustracción del residuo causaría una disminución significativa en la estadística χ^2.

    Mientras que la primer situación sucede cuando el patrón eliminado es influyente que empeoraba el ajuste del modelo, la segunda sucede cuando el patrón eliminado es un patrón que el modelo falla en ajustar, y que tal vez pudiera ser un *outlier*. También se podría presentar una combinación de ambas.

Como comentario final, el que $\Delta\chi^{2(-i)}$ fuera muy grande sería una señal de que el modelo carece de habilidad para clasificar patrones de covariables similares al patrón i, característica negativa del modelo.

- Caso $\Delta\chi^{2(-i)} = \chi^2 - \chi^{2(-i)}$ es cercano a cero

  El que suceda este caso para algún patrón de covariables i implica que la estadística χ^2 no se ve afectada por la sustracción de dicho patrón, lo que a la vez significa que muy probablemente no es ni influyente ni atípico.

  El que sucediera este caso para todos los patrones o la mayoría de ellos sería muy positivo para el modelo, ya que sería una señal de robustez; sin embargo sería complicado lograrlo si se contara con pocas observaciones, pues cada patrón tendría un peso mayor en el modelo, y en consecuencia, en la estadística χ^2.

  Nuevamente se menciona que este caso no asegura la calidad de influyente o *outlier* del patrón de covariables; la primera se asegura con la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$, mientras la segunda con los expertos del área.

- Caso $\Delta\chi^{2(-i)} = \chi^2 - \chi^{2(-i)}$ es mucho menor a cero

  El que un patrón tuviera valores $\Delta\chi^{2(-i)}$ negativos y muy alejados a cero implica que dicho patrón es considerado para el ajuste del modelo de vital importancia y, sin él, dicho ajuste se empobrecería y perdería parte de su efectividad (es decir, la el patrón es influyente).

  Este caso tampoco refleja una buena señal para el modelo, ya que demuestra falta de robustez hacia el mismo al hacer una gran parte de la efectividad de su ajuste dependiente de un único patrón de covariables.

- $\Delta D^{(-i)}$

  Para la gráfica de esta estadística se siguen los mismos casos y descripciones que de la estadística anterior, $\Delta\chi^{2(-i)}$.

- $\Delta\hat{\beta}_{pj}^{(-i)}$

  Similarmente a $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$, en la gráfica de $\Delta\hat{\beta}_{pj}^{(-i)}$ se buscan valores cercanos a cero, pues esto implicaría que en proporción respecto al intervalo de confianza elegido asociado al coeficiente $\hat{\beta}_{pj}$, dicho coeficiente no cambiara o cambiara poco con la sustracción del patrón de covariables i, es decir el patrón no sería influyente.

Lo ideal sería encontrar valores entre menos uno y uno para todos los patrones en esta estadística, pues esto indicaría que aún con la sustracción de cada patrón i, el cambio en los coeficientes en promedio, se encuentra dentro de sus respectivos intervalos de confianza aceptados.

Existen casos en los que, para algún patrón i, el valor $\Delta\hat{\beta}_{pj}^{(-i)}$ fuera muy grande mientras que $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$ tuvieran valores muy pequeños. Esto podría suceder de manera común y se debe a que la modificación de los coeficientes del modelo no necesariamente implica una reducción en la capacidad de ajuste del modelo, ajuste medido mediante $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$. Es decir, el modelo podría seguir teniendo el mismo ajuste o incluso un ajuste mejor al sustraer un patrón i que ocasione valores grandes para la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$.

Entonces, una pregunta que se podría hacer es, ¿por qué preocuparse por $\Delta\hat{\beta}_{pj}^{(-i)}$ si no tiene relación alguna con el efectividad del ajuste del modelo?

El motivo por el cual se le da importancia a la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$ es, que si dichas estadísticas fueran pequeñas para todas o para la gran mayoría de los patrones sustraídos i, y para todos los coeficientes $\hat{\beta}_{pj}$, esto sería una gran señal de estabilidad por parte del modelo; es decir, que al adherir o sustraer algún patrón de covariables, se podría esperar un modelo con coeficientes similares a los actuales, y no uno completamente diferente como sería el caso si se tuvieran valores $\Delta\hat{\beta}_{pj}^{(-i)}$ grandes para la mayoría de las observaciones y coeficientes.

Esta estabilidad, al mismo tiempo daría mayor seguridad al analista o investigador para la realización de inferencias, interpretaciones y conclusiones basadas en los coeficientes del modelo. Este tipo de interpretaciones y conclusiones, normalmente constituyen uno de los objetivos principales en la mayoría de los análisis de regresión logística multinomial.

Si se llegaran a obtener valores de $\Delta\hat{\beta}_{pj}^{(-i)}$ grandes en la mayoría de los coeficientes para algún modelo, ¿cuánta confianza le daría al investigador las inferencias obtenidas por los coeficientes de dicho modelo, si al remover uno de los patrones de covariables, el modelo arrojara inferencias completamente diferentes a las anteriores?

Por lo tanto, se buscarán modelos con valores $\Delta\hat{\beta}_{pj}^{(-i)}$ pequeños.

En regresión lineal existen dos maneras de interpretar el valor de las estadísticas de diagnóstico: de manera gráfica o de manera teórica, desarrollando la distribución de los diagnósticos bajo el supuesto de que el modelo ajustado es el correcto. Para el caso de la regresión logística multinomial, se utilizará únicamente la manera gráfica.

En resumen, lo que se hará será identificar a los valores de los diagnósticos que se encuentren muy separados de los demás. Los patrones de covariables asociados a estos valores serán analizados detalladamente y podrán ser juzgados ya sea como *outliers* o como influyentes dependiendo el punto de vista tanto del investigador como de los expertos en el área de la cual se realiza el estudio.

Para algunas estadísticas de diagnóstico que poseen una distribución conocida (como los residuos de Pearson, cuya distribución es asintóticamente normal), se podría considerar como patrón de covariables atípico si su respectiva estadística excediera algún cuantil específico denotado por su respectiva distribución.

Teóricamente, si el modelo ajustara correctamente a los datos entonces no existiría ningún valor extremo para los diagnósticos. Sin embargo, no siempre será tan sencillo este procedimiento, pues cuando se cumpla que $M = N$ un supuesto de gran importancia sería violado. Lo anterior será expuesto mediante un ejemplo para dar mayor claridad al problema.

Considérense los residuos de Pearson, $r(y_{ij}, \hat{\pi}_{ij})$. A menudo se asegura que la distribución de $r(y_{ij}, \hat{\pi}_{ij})$ es aproximadamente $N(0, 1)$ cuando el modelo ajustado es el correcto. Realmente esto sólo se satisface cuando m_i es lo suficientemente grande como para justificar que la distribución normal provee una aproximación adecuada para la distribución binomial, condición que se satisface bajo la distribución de naturaleza *M-asintótica*. Pero si $m_i = 1$, entonces $r(y_{ij}, \hat{\pi}_{ij})$ tiene únicamente dos posibles valores, por lo que difícilmente se esperaría se distribuyera normal.

Por lo tanto, cuando algún modelo contenga variables continuas y se decidiera por no categorizarlas, el poder juzgar mediante estadísticas de diagnóstico a alguna observación como atípica provendrá prácticamente de la experiencia del experto del área.

Hasta ahora se han mencionado siete posibles estadísticas de diagnóstico: $r(y_{ij}, \hat{\pi}_{ij})$, d_i, h_i, b_i , $\Delta\chi^{2(-i)}$, $\Delta D^{(-i)}$ y $\Delta\hat{\beta}_{pj}^{(-i)}$. El siguiente paso consistiría en realizar gráficas entre ellas con el fin de localizar comportamientos y patrones de covariables candidatos a ser *outliers* o influyentes. Se puede también hacer uso de los valores $\hat{\pi}_{ij}$ para la realización de dichas gráficas.

Una gráfica que se considera bastante útil es la gráfica de $\Delta\chi^{2(-i)}$ versus $\hat{\pi}_{ij}$ donde el tamaño de cada punto es proporcional al tamaño de su respectivo $\Delta\hat{\beta}_{pj}^{(-i)}$, ya que al mismo tiempo se pueden detectar observaciones con una gran influencia tanto en el ajuste del modelo, como en los coeficientes estimados.

Otros puntos importantes a considerar son:

- Se piensa que $\Delta\chi^{2(-i)}$ y $\Delta D^{(-i)}$ están correlacionados positivamente, por lo que es probable que al eliminar algún patrón, los valores de ambas

estadísticas también sean modificados.

- Antes de tomar la decisión de eliminar algún patrón de covariables debido a que se tomará como dato influyente o atípico, es necesario consultar dicha decisión con un especialista del área.

Después de haber realizado el procedimiento antes mencionado, se puede proceder a la presentación e interpretación del modelo ajustado.

En resumen, antes de presentar los resultados de algún modelo, es necesario que se analice el ajuste del modelo tanto con estadísticas resumen como con estadísticas de diagnóstico.

3.4. Ajuste del modelo por validación externa

Habrán situaciones en las que será posible obtener nuevas observaciones diferentes a las utilizadas en el modelo para poder evaluar dicho modelo en ellas. También habrá ocasiones en las que será posible dividir las observaciones en dos conjuntos, realizar el modelo con uno y probar el modelo con el conjunto restante. A este tipo de prácticas (probar el modelo con datos diferentes a los utilizados en la elaboración del mismo) se les llama validación del modelo, y es de extrema importancia cuando el objetivo del modelo en cuestión es pronosticar la categoría de futuras observaciones.

Una de las razones más importantes para realizar la validación de un modelo es que éste siempre se ajustará mejor a los datos con los que fue elaborado que con nuevos datos.

Para realizar la validación del modelo, lo que se tiene que hacer es crear un nuevo modelo con los nuevos datos, pero con los mismos coeficientes de los parámetros del modelo con los datos originales.

Los métodos para validación del modelo son, salvo a unas cuantas excepciones, iguales a los utilizados para el ajuste del modelo con los datos originales. Supóngase que la base de datos de validación contiene un total de N_v observaciones, distribuidos en M_v patrones de covariables. Se mantendrá la misma notación utilizada con anterioridad para y_{ij} y $\hat{\pi}_{ij}$, pero ahora todas estas notaciones harán referencia a los datos de validación y no a los originales. Se procede a calcular las estadísticas χ^2, D y $\hat{C}$ para los nuevos datos.

Para los casos de χ^2 y D, el cálculo se realiza de la misma manera a excepción de que la suma ahora será sobre M_v términos. Si dentro de cada patrón de covariables, $m_i\hat{\pi}_{ij}$ es lo suficientemente grande como para utilizar una aproximación normal a la distribución binomial de los residuos, entonces la estadística χ^2 se distribuirá como una $\chi^2_{(J-1)(M_v-(P+1))}$ bajo la hipótesis de que el modelo evaluado es el correcto. Desafortunadamente, el número de observaciones dentro de la base de datos de validación suele ser pequeño, y podría suceder que $m_i = 1$ para todos o para la mayoría de los patrones de covariables; por lo tanto la distribución binomial no se podría aproximar por una distribución normal y por ende no se puede utilizar la teoría de aproximación *M-asintótica* de la distribución.

Si este fuera el caso, se recomienda utilizar la prueba de Hosmer-Lemeshow, ya que ésta no tiene ningún inconveniente al utilizarse en modelos donde $m_i = 1$.

Las tablas de clasificación serán de gran utilidad para la muestra de validación sobre todo cuando el propósito del modelo sea pronosticar. La tabla resultante puede utilizarse para calcular algunas medidas como la sensibilidad, especificidad, poder predictivo positivo o negativo, o bien para propósitos específicos del análisis.

En general, únicamente se tiende a utilizar tablas de clasificación como herramientas de validación del modelo, dejando las demás estadísticas únicamente para analizar los datos con los que se elaboró dicho modelo, también llamados datos de entrenamiento.

Por último, si lo que se desea es pronosticar con nuevos datos, se debe de tener cuidado con que los valores de las variables explicativas se encuentren dentro del rango de los valores tomados por la base de entrenamiento; es decir, no tomar valores que sean mayores ni menores a todos los valores de la base de entrenamiento, pues de hacerse, podrían llegar a presentarse problemas en la validación del modelo.

Capítulo 4

Interpretación del Modelo

4.1. Introducción

El objetivo de este capítulo es proveer métodos para obtener el mayor provecho a los resultados obtenidos por el modelo de regresión logística multinomial. La mayoría de estos métodos recaen en obtener valores e intervalos de confianza para cocientes de momios, cocientes de riesgos relativos, etc.

Una vez que se haya elegido al que el investigador considere como el "mejor modelo encontrado" para los objetivos del estudio y se haya ajustado de manera apropiada, se puede proceder a interpretarlo. Para entender mejor lo que se presentará a continuación, se mencionará un ejemplo análogo al modelo de regresión logística multinomial, pero con un poco más de sencillez en su interpretación.

Supóngase un modelo de regresión lineal con la ecuación $y_x = \beta_0 + \beta_1 x$, con el subíndice de y_x denotando que y es función de x. Sea también $y_{(x+1)} = \beta_0 + \beta_1(x+1)$, entonces se sigue de ambas ecuaciones que:

$$\begin{aligned} y_{x+1} - y_x &= (\beta_0 + \beta_1 x) - (\beta_0 + \beta_1(x+1)) \\ &= \beta_1 \end{aligned}$$

Por lo que se podría interpretar al coeficiente β_1 como el número de unidades que aumenta la variable dependiente y, por cada unidad añadida a la variable x.

En el caso del modelo de regresión logística multinomial, recuérdese que:

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = g(x_{i\cdot})_j = \sum_{p=0}^{P} x_{ip}\beta_{pj}$$

Ahora, si $x_{v\cdot}$ fuera una observación con valores idénticos a $x_{i\cdot}$ en todas las variables salvo en la variable q, donde $x_{iq} = x_{vq} + 1$ para algún $q = 1, 2, \ldots, P$,

entonces:

$$g(x_{i\cdot})_j - g(x_{v\cdot})_j = \beta_{qj}$$

Es decir, en el caso de regresión logística multinomial el coeficiente β_{qj} se puede interpretar como el número de unidades que aumenta la función *logit* al aumentar en una unidad a la variable q, y dejando a las demás variables fijas.

Supóngase ahora un modelo logístico con la q-ésima variable de tipo categórica con K niveles. Entonces, las funciones *logit* para dicho modelo serían de la forma:

$$g(x_{i\cdot})_j = \sum_{p \neq q}^{P} x_{ip}\beta_{pj} + \sum_{k=1}^{K-1} D_{qk}\beta_{qkj}$$

Ahora, sean $x_{i\cdot}$ y $x_{v\cdot}$ dos observaciones idénticas a excepción de la variable categórica antes mencionada, donde la observación $x_{v\cdot}$ toma el nivel u mientras que $x_{i\cdot}$ toma el valor K, definido en este modelo como el nivel de referencia de dicha variable.

En el caso anterior, se tiene que utilizando como base a la ecuación anterior:

$$g(x_{v\cdot})_j = \sum_{p \neq q}^{P} x_{vp}\beta_{pj} + \sum_{k=1}^{K-1} D_{qk}\beta_{qkj} = \sum_{p \neq q}^{P} x_{vp}\beta_{pj} + D_{qu}\beta_{quj} = \sum_{p \neq q}^{P} x_{vp}\beta_{pj} + \beta_{quj}$$

y

$$g(x_{i\cdot})_j = \sum_{p \neq q}^{P} x_{ip}\beta_{pj} + \sum_{k=1}^{K-1} D_{qk}\beta_{qkj} = \sum_{p \neq q}^{P} x_{ip}\beta_{pj}$$

Entonces, al obtener la diferencia de ambos *logits* se llegaría a:

$$\begin{aligned} g(x_{v\cdot})_j - g(x_{i\cdot})_j &= \sum_{p \neq q}^{P} x_{vp}\beta_{pj} + D_{qu}\beta_{quj} - \sum_{p \neq q}^{P} x_{ip}\beta_{pj} \\ &= D_{qu}\beta_{quj} \\ &= \beta_{quj} \end{aligned}$$

Es decir, el coeficiente β_{quj} se puede interpretar como la diferencia entre el *logit* con el nivel u en la variable q y el logit con el nivel de referencia en la variable q, teniendo todos los valores de las demás variables idénticos y para la categoría de respuesta j.

Cabe mencionar que los casos mostrados anteriormente, y en consecuencia la interpretación de los coeficientes en los mismos, sólo se cumplirá cuando los *logits* evaluados sean de la misma categoría de la variable dependiente. Si se quisiera evaluar sobre dos categorías diferentes, por ejemplo a las categorías j y w el resultado de la ecuación anterior sería:

$$g(x_{v\cdot})_w - g(x_{i\cdot})_j = \sum_{p=1}^{P} x_{vp}\beta_{pw} - \sum_{p=1}^{P} x_{ip}\beta_{pj}$$

Que no tendría ninguna interpretación sencilla.

4.2. Cociente de momios y cociente de riesgos relativos

Se introducen a continuación dos términos de gran importancia en modelos de regresión logística: el cociente de momios y el cociente de riesgos relativos. Éstos, serán de gran ayuda para dar otra interpretación a los coeficientes del modelo y simplificará la realización de inferencias.

4.2.1. Cociente de momios

Antes de definir a los "cocientes de momios" (*odds ratios* en inglés), primero se debe de definir lo que es un "momio".

Los momios son una herramienta útil para medir cuantas veces es más probable que suceda algún evento a comparación de que no suceda, y se definen como:

$$\frac{p}{1-p} \tag{4.1}$$

con p la probabilidad de que suceda algún evento en específico. Es decir, se definen como el cociente entre la probabilidad de ocurrencia de algún evento, y su respectivo complemento. Así pues, si el momio de algún evento en específico tomara el valor dos, entonces su interpretación sería que es dos veces más probable que suceda el evento del momio a que no suceda. Si el valor del momio fuese 0.5, entonces sería dos veces más probable que no suceda dicho evento a que suceda.

Ahora, el cociente de momios consiste, como su nombre lo dice, en la división de dos momios, y en el caso de la regresión logística sirve para comparar qué tan probable es que suceda uno con respecto al otro.

Uno de los usos principales del cociente de momios consiste en relacionarlo con una función sencilla de los coeficientes estimados, de tal manera que permita tanto una interpretación sencilla de dichos coeficientes, como un método rápido de encontrar el valor de algún cociente de momios en específico. Dicho uso se presenta a continuación.

Supóngase un modelo de regresión logística con P variables independientes y cuya variable de respuesta se componga únicamente de dos categorías; la categoría cero, que será la de referencia, y la uno. Supóngase también a $x_{i\cdot}$ como una observación con las mismas variables del modelo, y a $x_{v\cdot}$ como otra observación con idénticos valores que $x_{i\cdot}$ salvo en la q-ésima variable, donde $x_{iq} = x_{vq} + 1$.

Entonces, el cociente de momios entre el momio de que $x_{i\cdot}$ pertenezca a la categoría uno, y el momio de que $x_{v\cdot}$ pertenezca a la categoría uno se expresaría como:

$$\frac{\frac{\hat{\pi}_{i1}}{1-\hat{\pi}_{i1}}}{\frac{\hat{\pi}_{v1}}{1-\hat{\pi}_{v1}}} = \frac{\frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{p1}}}{1+e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{p1}}}}{\frac{1}{1+e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{p1}}}}}{\frac{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{p1}}}{1+e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{p1}}}}{\frac{1}{1+e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{p1}}}}} = \frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{p1}}}{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{p1}}} = e^{\sum_{p=0}^{P} \hat{\beta}_{p1}(x_{ip}-x_{vp})} = e^{\hat{\beta}_{q1}} \qquad (4.2)$$

Por lo que $e^{\hat{\beta}_{q1}}$ se podría interpretar como el efecto que tiene en el momio de $\hat{\pi}_{v1}$ el aumentar en una unidad el valor de la variable q; es decir, si $e^{\hat{\beta}_{q1}} =$ "c", entonces el momio de $\hat{\pi}_{i1}$ es "c" veces más grande que el de $\hat{\pi}_{v1}$. Esto, en una interpretación para un público no muy relacionado al término de momios, sería lo mismo a decir que es "c" veces más probable que al aumentar en una unidad a la variable q, una observación perteneciera a la categoría uno y no a su complemento, a que perteneciera a esa misma categoría y no a su complemento si no se aumentara el valor de dicha variable.

De igual manera, si se quisiera conocer el efecto que tendría en el momio el aumentar la q-ésima variable en 5 unidades, simplemente se calcularía el valor de $e^{5\hat{\beta}_{q1}}$, y si dicho valor fuera "c", entonces se sabría que el aumentar la q-ésima variable en 5 unidades ocasiona un momio "c" veces mayor al momio original; o bien, que es "c" veces más probable que la observación perteneciera a la categoría uno y no a su complemento al aumentar la q-ésima variable en 5 unidades, a que dicha observación perteneciera a la categoría uno y no a su complemento si el aumento en unidades hubiera sido nulo.

La interpretación anterior es de bastante utilidad, al permitir medir el efec-

to que tendría en los momios, y por ende, en las probabilidades, el aumentar o disminuir el valor de alguna variable en específico, utilizando únicamente la función exponencial del coeficiente correspondiente a la variable.

4.2.2. Cociente de riesgos relativos

Si bien los cocientes de momios son una herramienta poderosa tanto en la interpretación de coeficientes como en la medición del efecto en los momios al aumentar o disminuir el valor de las variables explicativas, éstos tienen una desventaja, y es que al ser aplicados en modelos de regresión logística multinomial (más de dos categorías en la variable dependiente), la interpretación tan sencilla de los coeficientes se pierde. La prueba de ello se muestra a continuación.

Sea J el número de categorías que posee la variable dependiente en un modelo de regresión logística multinomial con J la categoría de referencia y $J > 2$, P el número de variables, y $x_{i\cdot}$ una observación idéntica a $x_{v\cdot}$ salvo en la q-ésima variable, donde $x_{i\cdot} = x_{v\cdot} + 1$. Entonces, el cociente de momios entre estas dos observaciones para la categoría j se calcula como:

$$\frac{\frac{\hat{\pi}_{ij}}{1-\hat{\pi}_{ij}}}{\frac{\hat{\pi}_{vj}}{1-\hat{\pi}_{vj}}} = \frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{1+\sum_{k\neq j} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}} = \frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k\neq j} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pj}}}{1+\sum_{k\neq j} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}} = e^{\hat{\beta}_{qj}} . \frac{\left(1+\sum_{k\neq j} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}\right)}{\left(1+\sum_{k\neq j} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}\right)} \tag{4.3}$$

haciendo imposible simplificar más esta ecuación, y con ello, fracasando en encontrar una interpretación sencilla para los coeficientes.

Dada esta limitación del cociente de momios, se tuvo que recurrir a otra herramienta que intentara brindar resultados similares a los que brinda el cociente de momios en regresión logística simple (con dos categorías), pero aplicado a modelos de regresión logística multinomial. Se introduce entonces el cociente de riesgos relativos.

Sea J el número de categorías de la variable dependiente de un modelo de regresión logística multinomial, sea J también la categoría de referencia y P el número de variables explicativas del modelo. Entonces se define al cociente de

riesgos relativos (*relative risk ratio* en inglés) entre la observación i con categoría j y la observación v con categoría w como:

$$RRR_{(i,j),(v,w)} = \frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vw}}{\hat{\pi}_{vJ}}} \tag{4.4}$$

Esto es, en vez de utilizar como denominador de los cocientes al complemento de la probabilidad, se utiliza la probabilidad asociada a la categoría de referencia.

La gran ventaja de esta herramienta, es que de esta manera sí se puede llegar a tener una interpretación sencilla de los coeficientes, pues si $x_{i\cdot}$ fuera una observación idéntica a $x_{v\cdot}$ salvo en la variable q, donde $x_{iq} = x_{vq} + 1$, entonces el cociente de riesgos relativos para ambas observaciones con la categoría j sería igual a:

$$RRR_{(i,j),(v,j)} = \frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vj}}{\hat{\pi}_{vJ}}} = \frac{\frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}}{\frac{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}} = \frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pj}}} = e^{\hat{\beta}_{pj}} \tag{4.5}$$

Así pues, si $e^{\hat{\beta}_{pj}} =$ "c" se podría concluir que el riesgo relativo de $\hat{\pi}_{vj}$ es "c" veces más grande al aumentar en una unidad a la variable q, o bien, que es "c" veces más probable que al añadir una unidad a la variable q la observación pertenezca a la categoría j y no a la J, a que dicha observación perteneciera a la categoría j y no a la J si no se hubiera incrementado la variable.

A continuación se presentará la fórmula general para el cálculo de los cocientes de riesgos relativos, misma que es de bastante utilidad si se quisieran realizar inferencias respecto a diferentes combinaciones de categorías o variables explicativas. Por lo tanto, el cociente de riesgos relativos entre la observación i con categoría j y la observación v con categoría w es equivalente a:

$$RRR_{(i,j),(v,w)} = \frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vw}}{\hat{\pi}_{vJ}}} = \frac{\frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}}{\frac{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}} = \frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}}} = e^{\left(\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}\right)} \tag{4.6}$$

y si esta fórmula fuera igual a "c", la interpretación sería que es "c" veces más grande el riesgo relativo de $\hat{\pi}_{ij}$ que el de $\hat{\pi}_{vw}$, o bien que es "c" veces más probable que la observación i pertenezca a la categoría j y no a la J, a que la observación v pertenezca a la categoría w y no a la J. Cabe resaltar que el caso más utilizado es cuando j=w.

4.2.3. Interpretación de cocientes: precauciones

Esta es una de las secciones de más importancia del capítulo, ya que si no se sabe interpretar de manera correcta a los cocientes presentados anteriormente, las conclusiones erróneas realizadas podrían contradecir gravemente a la teoría que las respalda.

Supóngase como ejemplo al cociente de riesgos relativos entre la observación i con categoría j y la observación v con categoría w ($RRR_{(i,j),(v,w)}$). Se había visto que si el valor de dicho cociente fuera algún número "c", una de sus interpretaciones sería: "es c veces más probable que la observación i pertenezca a j y no a J, a que la observación v pertenezca a w y no a J".

Es muy importante hacer énfasis en que este cociente no compara directamente a la categoría j y a la w, sino que cada una se compara con respecto a la categoría de referencia (en este caso la J), y estos riesgos relativos se comparan entre sí. Esta explicación puede llegar a ser confusa, por lo que se expondrá un ejemplo para que se intente esclarecer lo más posible las diferencias de ambas comparaciones.

Supóngase que un estudio quisiera predecir el país de procedencia de las personas basado en sus pesos y alturas. La base de datos utilizada es la siguiente:

	País	Altura	Peso
1	Italia	1.85	80
2	Italia	1.90	85
3	Italia	1.70	65
4	Italia	1.75	80
5	E.U.A.	1.90	105
6	E.U.A.	1.80	78
7	E.U.A.	1.70	80
8	E.U.A.	1.85	85
9	Japón	1.70	64
10	Japón	1.55	48
11	Japón	1.65	60
12	Japón	1.75	68

Con Italia como la categoría de referencia para la variable de respuesta. Ahora supóngase que el ajuste del modelo fue bueno y los coeficientes asociados a ambas variables fueron significativos. Los valores de los coeficientes obtenidos por el modelo fueron los siguientes:

	Intercepto	Altura	Peso
Japón	-11.5247	33.0407	-0.6749
E.U.A.	10.5667	-19.7568	0.3094

Entonces, el valor del cociente de riesgos relativos entre la observación siete para la categoría de Estados Unidos y la observación cinco para la categoría de Estados Unidos sería:

$$RRR_{(7,\text{E.U.A.}),(5,\text{E.U.A.})} = \frac{\frac{\hat{\pi}_{7E.U.A.}}{\hat{\pi}_{7It}}}{\frac{\hat{\pi}_{5E.U.A.}}{\hat{\pi}_{5It}}} = \frac{\frac{0.8499}{0.1501}}{\frac{0.9960}{0.0040}} = 0.0227$$

Lo que quiere decir que es 44 veces más probable (1/0.0227=44.05) que la observación cinco sea nativa de Estados Unidos de América y no de Italia, a que la observación siete sea de Estados Unidos de América y no de Italia.

Es de gran importancia hacer notar que la interpretación anterior es totalmente diferente a decir que la observación cinco es 44 veces más probable que sea de Estados Unidos de América a que la observación siete lo sea, pues el cociente que calcula esta aseveración es el siguiente:

$$\frac{\hat{\pi}_{7E.U.A.}}{\hat{\pi}_{5E.U.A.}} = \frac{0.8499}{0.9960} = 0.8533$$

Lo que diría que es 1.17 veces más probable (1/0.8533=1.17) que la observación cinco sea estadounidense a que la observación siete lo sea.

La diferencia entre los resultados de ambas herramientas puede ser abismal

como en el ejemplo anterior, por lo que hay que tener precaución y pensar muy bien qué herramientas utilizar dependiendo de lo que se quiera obtener en el estudio. Ninguna de las dos herramientas es mejor que otra, simplemente sirven para interpretar diferentes situaciones.

Cabe mencionar que, en el caso en que se quisieran realizar comparaciones entre una misma observación (misma observación, diferentes categorías), entonces tanto el cociente de riesgos relativos, como el cociente de momios darán el mismo resultado que el momio, pues utilizando nuevamente el ejemplo anterior:

$$RRR_{(5,\text{Japón}),(5,\text{E.U.A.})} = \frac{\frac{\hat{\pi}_{5Jap}}{\hat{\pi}_{5It}}}{\frac{\hat{\pi}_{5E.U.A.}}{\hat{\pi}_{5It}}} = \frac{\hat{\pi}_{5Jap}}{\hat{\pi}_{5E.U.A.}}$$

Como nota adicional, cabe mencionar que aunque no se presentaron las probabilidades de manera explícita, se tiene todo lo necesario para calcularlas dado que se presentaron los valores de los coeficientes y las observaciones.

Otra observación que se debe de tomar en cuenta en el momento de interpretar los datos es revisar con un experto del área que los resultados tengan sentido.

4.2.4. Intervalos de confianza

Al igual que los coeficientes del modelo y la función *logit*, también se pueden obtener intervalos de confianza para el cociente de momios cuando el número de categorías es dos, y para el cociente de riesgos relativos para cualquier número de categorías. Esto sería de gran ayuda si lo que se deseara en el estudio fuera encontrar un intervalo y no un número como estimación. En esta sección se encontrarán los intervalos de confianza para el cociente de riesgos relativos.

El motivo por el cual solo se calcularán los intervalos para el cociente de riesgos relativos y no para el cociente de momios es, que mientras para modelos multinomiales ($J > 2$) se desconoce la distribución asintótica de los cocientes de momios y por lo tanto no se pueden calcular sus intervalos de confianza, para el caso simple ($J = 2$) el cociente de momios es idéntico al cociente de riesgos relativos.

Para obtener entonces el intervalo de confianza de un cociente de riesgos relativos se necesita conocer primero la distribución del mismo, para entonces poder estimar sus errores estándar y obtener los puntos delimitantes de los intervalos.

Dado que el cociente de riesgos relativos es igual a la función exponencial de

una combinación lineal de coeficientes cuya distribución se conoce es asintóticamente normal, una alternativa para el cálculo de los intervalos sería obtener la distribución del logaritmo natural del cociente de riesgos relativos, obtener los puntos delimitantes del intervalo por medio de la distribución asintótica del logaritmo del cociente y posteriormente aplicar la función exponencial a dichos puntos.

Así pues, cuando el cociente de riesgos relativos sea entre cocientes de la misma categoría, dígase por ejemplo la categoría j, y cuando todas las variables explicativas de las dos observaciones utilizadas en cociente ($x_{i\cdot}$ y $x_{v\cdot}$) sean iguales entre ellas a excepción de una única variable, llámese la variable k ($x_{ik} \neq x_{vk}$), entonces se cumplirá que

$$\log\left(RRR_{(i,j),(v,j)}\right) = (x_{ik} - x_{vk})\hat{\beta}_{kj}$$

Esta identidad se cumple únicamente bajo las condiciones especificadas anteriormente y se cumple tanto para variables k continuas como categóricas. Bajo estas condiciones, el logaritmo natural del cociente de riesgos relativos sigue una distribución también asintóticamente normal. Por lo tanto, los puntos extremos del intervalo de confianza para el cociente de riesgos relativos, bajo las características mencionadas anteriormente, se calcularían como:

$$\exp\left[(x_{ik} - x_{vk})\hat{\beta}_{kj} \pm z_{1-\alpha/2} \cdot (x_{ik} - x_{vk})\hat{SE}(\hat{\beta}_{kj})\right]$$

donde 100(1-α) % es el nivel de confianza asociado al intervalo, $\hat{\beta}_{kj}$ es el coeficiente estimado correspondiente a la variable k para la categoría de respuesta j, $\hat{SE}(\hat{\beta}_{kj})$ es el error estándar estimado del coeficiente estimado y $z_{1-\alpha/2}$ es el valor asociado al cuantil (1-α/2) % de una distribución normal estándar.

Ahora, para el caso general, se sabe que se cumple la siguiente ecuación:

$$RRR_{(i,j),(v,w)} = \frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vw}}{\hat{\pi}_{vJ}}} = \frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}}} = e^{\left(\sum\limits_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum\limits_{p=0}^{P} x_{vp}\hat{\beta}_{pw}\right)}$$

Lo que implica que:

$$\log\left(RRR_{(i,j),(v,w)}\right) = \sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}$$

Enfocando la atención hacia la parte derecha de la ecuación anterior, se recuerda que los estimadores $\hat{\beta}_{pj}$, al ser calculados por el método de máxima verosimilitud, tienen una distribución asintóticamente normal, y como la combinación lineal de variables aleatorias con distribución asintóticamente normal también se distribuye de manera asintóticamente normal, entonces $\log\left(RRR_{(i,j),(v,w)}\right)$ también tendrá esta distribución.

Se procede ahora a estimar la varianza de $\log\left(RRR_{(i,j),(v,w)}\right)$, esto para poder estimar el error estándar del logaritmo del cociente de riesgos relativos y así utilizarlo para encontrar los puntos extremos del intervalo de confianza:

$$\begin{aligned}\hat{\mathrm{Var}}\left(\log\left(RRR_{(i,j),(v,w)}\right)\right) &= \hat{\mathrm{Var}}\left(\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}\right)\\ &= \sum_{p=0}^{P} x_{ip}^2\hat{\mathrm{Var}}(\hat{\beta}_{pj}) + \sum_{p=0}^{P} x_{vp}^2\hat{\mathrm{Var}}(\hat{\beta}_{pw})\\ &+ 2\sum_{p=0}^{P-1}\sum_{k=p+1}^{P} x_{ip}x_{ik}\hat{\mathrm{Cov}}(\hat{\beta}_{pj},\hat{\beta}_{kj})\\ &+ 2\sum_{p=0}^{P-1}\sum_{k=p+1}^{P} x_{vp}x_{vk}\hat{\mathrm{Cov}}(\hat{\beta}_{pw},\hat{\beta}_{kw})\\ &- 2\sum_{p=0}^{P}\sum_{k=0}^{P} x_{ip}x_{vk}\hat{\mathrm{Cov}}(\hat{\beta}_{pj},\hat{\beta}_{kw})\end{aligned}$$

Aunque parezca desafiante la ecuación anterior, realmente ya se tiene todo lo necesario para estimar dicha varianza, pues estos valores se encuentran en la matriz de varianzas y covarianzas estimada de los coeficientes.

Ahora, el valor del error estándar que se utilizará para calcular los intervalos de confianza de $\log\left(RRR_{(i,j),(v,w)}\right)$ se calculará como:

$$\hat{\mathrm{SE}}\left(\log\left(RRR_{(i,j),(v,w)}\right)\right) = \left[\hat{\mathrm{Var}}\left(\log\left(RRR_{(i,j),(v,w)}\right)\right)\right]^{\frac{1}{2}}$$

Por lo tanto, los puntos extremos del intervalo de confianza para el cociente de riesgos relativos $\hat{RRR}_{(i,j),(v,w)}$ se calcularán como:

$$\exp\left[\left(\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pw}\right) \pm z_{1-\alpha/2}\cdot\hat{\mathrm{SE}}\left(\log\left(\hat{RRR}_{(i,j),(v,w)}\right)\right)\right] \quad (4.7)$$

con 100(1-α) % el nivel de confianza asignado al intervalo.

Otra ventaja de utilizar al cociente de riesgos relativos con respecto a otras herramientas más sencillas como los momios ordinarios, es que a los momios ordinarios no se les pueden obtener intervalos de confianza, pues no siguen una distribución asintótica conocida. La prueba de ello está en que:

$$\frac{\hat{\pi}_{ij}}{\hat{\pi}_{vq}} = \frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=0}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq}}}{1+\sum_{k=0}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}$$

Y a partir de este paso, la ecuación anterior no se puede simplificar más.

4.2.5. Interpretación intervalos de confianza en cocientes de riesgos relativos

La interpretación de los intervalos de confianza para el cociente de riesgos relativos es muy sencilla. Simplemente es, que con un nivel 100(1-α) % de confianza, el verdadero valor del cociente de riesgos relativos se encontrará en el intervalo calculado. El nivel de confianza mencionado puede tomar cualquier valor entre cero y 100 por ciento. Si por ejemplo, el nivel de confianza elegido fuera 95 %, entonces la interpretación sería que en 95 de cada 100 ocasiones, el verdadero valor del cociente de riesgos relativos se encontrará en el intervalo de confianza calculado.

4.2.6. Advertencias intervalo de confianza en cocientes de riesgos relativos

El método propuesto para el cálculo del intervalo de confianza del cociente de riesgos relativos, que consiste en calcular el intervalo del logaritmo natural de dicho cociente y posteriormente exponenciarlo, únicamente es válido cuando la varianza del cociente de riesgos relativos es pequeña, puesto que la exponencial de una variable distribuida asintóticamente normal no se distribuye de la misma manera, sino asintóticamente log-normal, y esta distribución únicamente se aproxima a una distribución asintóticamente normal cuando la varianza de la variable es pequeña.

4.3. Interacciones

4.3.1. Introducción a la interacción entre variables

El objetivo de esta sección es introducir y explicar el papel de las interacciones en el modelo de regresión logística multinomial, así como clarificar la

interpretación que podrían llegar a tener.

En vanas palabras, el que dos o más variables explicativas tengan una interacción significa que la función *logit* se comporta diferente para diferentes combinaciones de estas variables. Un ejemplo muy sencillo de entender es el caso de interacción entre una variable categórica y una continua. Si ambas variables no tuvieran interacción, se esperaría entonces que el efecto de cada una de las dos variables hacia la función *logit* fuera el mismo independientemente del valor de la otra variable.

De manera gráfica, esta situación de no interacción entre variables se podría ilustrar en la figura 4.1 por las líneas uno y dos (l_1 y l_2), suponiendo únicamente dos niveles en la variable explicativa categórica (cada una de estas dos líneas corresponden a cada nivel de la variable categórica). De manera general para el caso de una variable continua y otra categórica, la no interacción se podría comprobar al ver paralelismo entre las líneas correspondientes a cada nivel de la variable categórica.

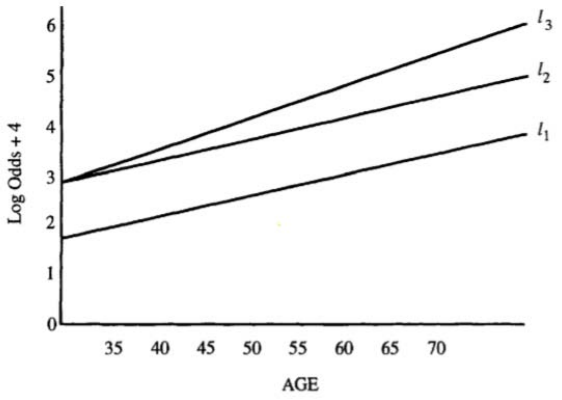

Figura 4.1: Representación gráfica del impacto entre interacción y no interacción de variables respecto al *logit*. Gráfica tomada del libro *Applied Logistic Regression, segunda edición, de D.W. Hosmer y S. Lemeshow.*

Ahora, si las variables explicativas mencionadas con anterioridad tuvieran interacción entre ellas, entonces la gráfica de la función *logit* correspondiente a cada combinación de variables podría ser similar a la figura anterior, pero ahora tomando en cuenta a las líneas uno y tres (l_1 y l_3) como resultado de cada combinación de las variables. Se puede notar que ambas líneas no son paralelas, esto es producto del efecto que produce la combinación de las variables con respecto al *logit*.

Otra manera de asimilarlo sería que bajo supuestos de no interacción, el efecto de la variable continua sobre el *logit* es el mismo independientemente del nivel en que se encuentre la variable categórica, mientras que bajo el supuesto de interacción, el efecto de la variable continua sobre el *logit* dependerá del valor de la variable categórica.

Siguiendo la interpretación de la gráfica anterior, se puede obtener el logaritmo del cociente de riesgos relativos (log *relative risk ratio*) entre los dos niveles de la variable explicativa categórica mediante la diferencia o distancia entre las líneas graficadas en los puntos que se quiera obtener, dejando a la variable continua constante, es decir, la distancia vertical entre ambas líneas. Bajo el caso de no interacción, este logaritmo será constante, mientras que para el caso en que se tenga interacción, el logaritmo de este cociente dependerá del valor de la variable continua. También para el caso de interacción entre dos variables continuas, o cuando se tenga una interacción que involucre a un mayor número de variables, se seguirá el mismo razonamiento.

El ejemplo mencionado anteriormente representa sólo un caso particular de interacción, que es cuando la relación sigue siendo lineal, pero a una pendiente diferente dependiendo del nivel de la variable categórica. Sin embargo otros casos de interacciones pueden ser vistos comúnmente, los cuales se podrían modelar con variables de segundo o mayor orden, o bajo la inclusión de alguna función como el logaritmo natural. Por motivos de practicidad, en este trabajo únicamente se analizará el caso en donde la interacción es lineal con pendiente diferente, como es el caso del ejemplo ilustrado.

Ahora que ya se ha explicado lo que son las interacciones y cómo se manifiestan, se procederá a realizar el cálculo de ellas. Una interacción en términos del modelo se manifiesta como una nueva variable, calculada como el producto de la o las variables a interactuar. Recuérdese que este proceso únicamente aplica si se creyera que la interacción tuviera un comportamiento lineal pero con pendiente diferente.

4.3.2. Interpretación cociente de riesgos relativos con interacción

Se ha mencionado que en modelos con interacción, los cocientes de riesgos relativos no se pueden generalizar, pues dependen del valor de la o las variables que forman parte de la interacción.

Ahora se probará algebraicamente este supuesto. De manera general, sea $\hat{g}(x_{i\cdot})_j$ la función *logit* del patrón de covariables i para la categoría j. Ahora, supóngase

que el modelo cuenta con cierto número de interacciones. Se denotará a I por este número. Entonces el *logit* mencionado queda descrito por:

$$\hat{g}(x_{i\cdot})_j = \sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} + \sum_{w=1}^{I} (\text{Interac}_w)_i \hat{\beta}_{P+w,j}$$

con $(\text{Interac}_w)_i$ la multiplicación de las variables de la interacción número w aplicada a los datos del patrón de covariables i, y $\hat{\beta}_{P+w,j}$ el coeficiente asociado a dicha interacción para la categoría j. Entonces, como se podrá recordar, el cociente de riesgos relativos $RRR_{(i,j),(v,q)}$ puede ser calculado como la exponencial de la resta entre $\hat{g}(x_{i\cdot})_j$ y $\hat{g}(x_{v\cdot})_q$. Pero sucede que:

$$\begin{aligned}\hat{g}(x_{i\cdot})_j - \hat{g}(x_{v\cdot})_q = &\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} + \sum_{w=1}^{I} (\text{Interac}_w)_i \hat{\beta}_{P+w,j} \\ &- \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq} - \sum_{w=1}^{I} (\text{Interac}_w)_v \hat{\beta}_{P+w,q}\end{aligned}$$

Y al aplicar la función exponencial a la ecuación anterior se obtiene $RRR_{(i,j),(v,q)}$.

Ahora, de una manera particular, supóngase j=q, $I = 1$ y $(\text{Interac}_1)_i = x_{i1}x_{i2}x_{i4}$ para toda i. También supóngase $x_{ik} = x_{vk}$ para todo k a excepción de la variable número dos, donde $x_{i2} = x_{v2} + 1$. Entonces para este caso particular, la diferencia de *logits* quedará calculada como:

$$\begin{aligned}\hat{g}(x_{i\cdot})_j - \hat{g}(x_{v\cdot})_j = &\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} + x_{i1}x_{i2}x_{i4}\hat{\beta}_{p+1,j} \\ &- \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pj} - x_{v1}x_{v2}x_{v4}\hat{\beta}_{p+1,j}\end{aligned}$$

Implicando que:

$$\hat{g}(x_{i\cdot})_j - \hat{g}(x_{v\cdot})_j = \hat{\beta}_{2j} + x_{i1}x_{i4}\hat{\beta}_{p+1,j}$$

Y por lo tanto:

$$RRR_{(i,j),(v,j)} = e^{\hat{\beta}_{2j} + x_{i1}x_{i4}\hat{\beta}_{p+1,j}}$$

Es decir, el cociente de riesgos relativos dependerá del valor que tengan las variables con las que interactúa la variable con valores diferentes en los patrones de covariables, en este caso x_{i1} y x_{i4}, por lo que no se puede generalizar la fórmula de cocientes de riesgos relativos con interacciones como en el caso en que no se supone interacción alguna.

Nótese también que ya no hay un único coeficiente, sino el número de coeficientes totales donde interactúa la variable con valores diferentes (en este caso

la segunda variable). Si por ejemplo en el modelo anterior, se tuviera además de la interacción $x_{i1}x_{i2}x_{i4}$ a la interacción $x_{i2}x_{i6}$, entonces el cociente de riesgos relativos hubiera resultado ser:

$$RRR_{(i,j),(v,j)} = e^{\hat{\beta}_{2j}+x_{i1}x_{i4}\hat{\beta}_{p+1,j}+x_{i6}\hat{\beta}_{p+2,j}}$$

Y se tendrían tres coeficientes dentro del cociente de riesgos relativos, número de veces que aparece la variable con valores diferentes dentro del modelo.

4.3.3. Intervalos de confianza para cocientes de riesgos relativos con interacción

Para calcular los intervalos de confianza de un cociente de riesgos relativos con interacción, se seguirá el mismo principio utilizado para el caso en que no existe interacción; es decir, se calculará el intervalo de confianza del logaritmo natural de dicho cociente y posteriormente se le aplicará la función exponencial a éste.

Recuérdese que de forma general el logaritmo del cociente de riesgos relativos con interacción tiene la estructura:

$$\begin{aligned}\log(RRR_{(i,j),(v,q)}) = &\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} + \sum_{w=1}^{I}(\text{Interac}_w)_i\hat{\beta}_{P+w,j} \\ &- \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq} - \sum_{w=1}^{I}(\text{Interac}_w)_v\hat{\beta}_{P+w,q}\end{aligned}$$

Por lo que la varianza de este logaritmo se estimará como:

$$\begin{aligned}
\hat{\mathrm{Var}}\left(\log\left(RRR_{(i,j),(v,q)}\right)\right) = &\sum_{p=0}^{P} x_{ip}^2 \hat{\mathrm{Var}}(\hat{\beta}_{pj}) + \sum_{w=1}^{I} (\mathrm{Interac}_w)_i^2 \hat{\mathrm{Var}}(\hat{\beta}_{P+w,j}) \\
&+ \sum_{p=0}^{P} x_{vp}^2 \hat{\mathrm{Var}}(\hat{\beta}_{pq}) + \sum_{w=1}^{I} (\mathrm{Interac}_w)_v^2 \hat{\mathrm{Var}}(\hat{\beta}_{P+w,q}) \\
&+ 2 \sum_{p_1=0}^{P-1} \sum_{p_2=p_1+1}^{P} x_{ip_1} x_{ip_2} \hat{\mathrm{Cov}}(\hat{\beta}_{p_1 j}, \hat{\beta}_{p_2 j}) \\
&+ 2 \sum_{p_1=0}^{P-1} \sum_{p_2=p_1+1}^{P} x_{vp_1} x_{vp_2} \hat{\mathrm{Cov}}(\hat{\beta}_{p_1 q}, \hat{\beta}_{p_2 q}) \\
&+ 2 \sum_{w_1=1}^{I-1} \sum_{w_2=w_1+1}^{I} (\mathrm{Interac}_{w_1})_i (\mathrm{Interac}_{w_2})_i \hat{\mathrm{Cov}}(\hat{\beta}_{P+w_1,j}, \hat{\beta}_{P+w_2,j}) \\
&+ 2 \sum_{w_1=1}^{I-1} \sum_{w_2=w_1+1}^{I} (\mathrm{Interac}_{w_1})_v (\mathrm{Interac}_{w_2})_v \hat{\mathrm{Cov}}(\hat{\beta}_{P+w_1,q}, \hat{\beta}_{P+w_2,q}) \\
&- 2 \sum_{p_1=0}^{P} \sum_{p_2=0}^{P} x_{ip_1} x_{vp_2} \hat{\mathrm{Cov}}(\hat{\beta}_{p_1 j}, \hat{\beta}_{p_2 q}) \\
&- 2 \sum_{w_1=1}^{I} \sum_{w_2=1}^{I} (\mathrm{Interac}_{w_1})_i (\mathrm{Interac}_{w_2})_v \hat{\mathrm{Cov}}(\hat{\beta}_{P+w_1,j}, \hat{\beta}_{P+w_2,q}) \\
&+ 2 \sum_{p=0}^{P} \sum_{w=1}^{I} x_{ip} (\mathrm{Interac}_w)_i \hat{\mathrm{Cov}}(\hat{\beta}_{pj}, \hat{\beta}_{P+w,j}) \\
&+ 2 \sum_{p=0}^{P} \sum_{w=1}^{I} x_{vp} (\mathrm{Interac}_w)_v \hat{\mathrm{Cov}}(\hat{\beta}_{pq}, \hat{\beta}_{P+w,q}) \\
&- 2 \sum_{p=0}^{P} \sum_{w=1}^{I} x_{ip} (\mathrm{Interac}_w)_v \hat{\mathrm{Cov}}(\hat{\beta}_{pj}, \hat{\beta}_{P+w,q}) \\
&- 2 \sum_{p=0}^{P} \sum_{w=1}^{I} x_{vp} (\mathrm{Interac}_w)_i \hat{\mathrm{Cov}}(\hat{\beta}_{pq}, \hat{\beta}_{P+w,j})
\end{aligned}$$

Todos los valores de varianzas y covarianzas de la ecuación anterior se pueden encontrar en la matriz de varianzas y covarianzas antes estimada.

Aunque pareciera que esta fórmula no es nada amigable, recuérdese que este es el caso general. Un caso particular con mayor interpretabilidad se expondrá posteriomente, pero primero se presentará el intervalo de confianza para la forma general del logaritmo de cocientes de riesgos relativos.

Una vez calculada la varianza de dicho logaritmo de cocientes, su error estándar se estimará como:

$$\hat{\text{SE}}\left[\log(RRR_{(i,j),(v,q)})\right] = \left[\hat{\text{Var}}(\log(RRR_{(i,j),(v,q)}))\right]^{\frac{1}{2}}$$

Estimado el error estándar, se obtiene el intervalo de confianza del cociente de riesgos relativos como:

$$\exp\left[\log(RRR_{(i,j),(v,q)}) \pm z_{1-\alpha/2} \cdot \hat{\text{SE}}\left[\log(RRR_{(i,j),(v,q)})\right]\right]$$

Con 100(1-α) % el nivel de confianza deseado y $z_{1-\alpha/2}$ el valor asociado al cuantil $(1-\alpha/2)$ de una distribución normal estándar.

Ahora, para el caso particular, se seguirá tomando el ejemplo anterior en donde el modelo únicamente tenía una interacción, dada por las variables 1, 2 y 4. Además, $x_{ik} = x_{vk}$ salvo en la variable dos, donde $x_{i2} = x_{v2} + 1$. También se recuerda que $j = q$.

El logaritmo del cociente de riesgos relativos es en este caso:

$$\begin{aligned}\log(RRR_{(i,j),(v,j)}) =& \sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} + x_{i1}x_{i2}x_{i4}\hat{\beta}_{P+1,j} \\ & - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pj} - x_{v1}x_{v2}x_{v4}\hat{\beta}_{P+1,j} \\ =& \hat{\beta}_{2j} + x_{i1}x_{i4}\hat{\beta}_{P+1,j}\end{aligned}$$

La varianza de dicho logaritmo queda estimada por:

$$\begin{aligned}\hat{\text{Var}}(\log(RRR_{(i,j),(v,j)})) =& \hat{\text{Var}}(\hat{\beta}_{2j} + x_{i1}x_{i4}\hat{\beta}_{P+1,j}) \\ =& \hat{\text{Var}}(\hat{\beta}_{2j}) + (x_{i1}x_{i4})^2\hat{\text{Var}}(\hat{\beta}_{P+1,j}) \\ & + 2x_{i1}x_{i4}\hat{\text{Cov}}(\hat{\beta}_{2j}, \hat{\beta}_{P+1,j})\end{aligned}$$

Posteriomente se calcula el error estándar como la raíz cuadrada de la ecuación anterior, para después calcular los puntos extremos del intervalo de confianza a un 100(1-α) % como:

$$\exp\left[\hat{\beta}_{2j} + x_{i1}x_{i4}\hat{\beta}_{P+1,j} \pm z_{1-\alpha/2} \cdot \hat{\text{SE}}\left[\log(RRR_{(i,j),(v,j)})\right]\right]$$

4.3.4. Notas adicionales

Una manera rápida y efectiva para encontrar posibles interacciones es mediante una gráfica del *logit* con respecto a las variables que pudieran poseer alguna interacción.

Si la interacción a comprobar fuera entre una variable continua y otra categórica de K niveles, entonces en la gráfica se apreciarían K líneas representando a cada nivel de la variable categórica, mientras que el eje horizontal representará a la variable continua en cuestión y el eje vertical al valor del *logit*.

En el caso de evaluar una interacción entre dos variables continuas, una de ellas se deberá de categorizar en K niveles y se procede como en el caso anterior referente a una variable continua y una categórica. Si se apreciara carencia de paralelismo entre las diferentes líneas de la gráfica, entonces es muy probable que exista una interacción entre las variables.

Para el caso de interacción entre variables categóricas, el eje horizontal corresponderá a una de ellas y el proceso continuará de igual manera que en los otros casos, sólo que ahora las líneas se trazarán manualmente entre los dos puntos de las categorías asociadas al eje horizontal y se tendrá una línea por cada nivel de la variable no asociada a dicho eje.

Si se quisieran calcular interacciones entre tres o más variables, primero se deberán categorizar a todas las variables continuas. A continuación se elige una variable que será representada en el eje horizontal de la gráfica, y se hará producto cruz entre todos los niveles de las demás variables, para así obtener una línea en la gráfica para cada combinación posible de las demás variables. Cabe mencionar que para realizar el caso anterior, primero se deben de hacer estas gráficas para todos los órdenes de interacciones menores entre las variables a interactuar.

Capítulo 5

Análisis de Componentes Principales

En este capítulo se presentará la técnica estadística de componentes principales. Dado que el análisis de esta técnica puede llegar a ser muy profundo debido al gran número de aplicaciones que tiene, no será posible presentar la técnica detalladamente. Por fines de practicidad, únicamente se presentarán en este trabajo las propiedades y usos que serán de utilidad, pues el uso de esta técnica es meramente un complemento que se utilizará en el modelo de importancia, el modelo de regresión logística multinomial.

Si el lector estuviera interesado en profundizar más respecto a la técnica de componentes principales, puede consultar el libro *Principal Component Analysis*, de *Jolliffe* (2002).

5.1. Introducción y desarrollo de componentes principales

En el área de análisis multivariado, uno de los grandes problemas que enfrenta el investigador es lidiar con una gran cantidad de variables y obtener información relevante respecto a ellas. Componentes principales es un método que puede ayudar, entre otras cosas, a reducir la dimensionalidad de los datos (el número de variables) preservando aún la mayor variabilidad posible de éstos.

No se vaya a pensar que componentes principales elige a las variables con ma-

yor variabilidad y desecha a las demás de manera directa. Lo que realiza esta técnica es crear nuevas variables a partir de las originales, de tal modo que cada una de las nuevas variables se obtenga mediante una combinación lineal de las variables originales. Este conjunto de nuevas variables serán llamadas las componentes principales de los datos.

Una de las propiedades más importantes de las componentes principales, es que no están correlacionadas entre ellas. Esta propiedad se obtiene por construcción como se verá próximamente en el cálculo de las componentes mediante el método de *Lagrange*.

Supóngase $\boldsymbol{X}$ es la matriz de datos de dimensión $N \times P$ que consta de N observaciones y P variables por cada observación. En casos donde P fuera un valor muy grande, sería complicado hacer inferencias tanto de las P variables como de la respectiva matriz de varianzas-covarianzas o de correlaciones de $\boldsymbol{X}$. Como se había mencionado antes, la técnica de componentes principales puede ayudar a disminuir la cantidad de variables a analizar conservando una buena proporción de la variabilidad de los datos, y así simplificar un poco el análisis de dichos datos.

Hay ocasiones en las que no se entiende muy bien el por qué se busca mantener la variabilidad de los datos. El motivo de buscar esta aseveración es, que mientras mayor variación tengan los datos más sencillo será obtener información relevante respecto a ellos (un conjunto de observaciones con casi los mismos valores en sus variables para cada observación no sería de mucha utilidad para hacer inferencias estadísticas, por ejemplo).

Es importante hacer mención que por "conservar variabilidad", el método se refiere a conservar la varianza de cada una de las variables originales. Componentes principales no toma en cuenta a las covarianzas entre variables dentro de este concepto, por lo que el principal punto de atención serán los valores de la diagonal de la matriz de varianzas y covarianzas de $\boldsymbol{X}$.

Se procede ahora al cálculo de dichas componentes principales. Se había mencionado que las componentes principales eran una combinación lineal de las variables originales de los datos. Entonces pues, se espera que la primera componente principal de la i-ésima observación sea de la forma:

$$z_{i1} = a_{11}x_{i1} + a_{12}x_{i2} + \ldots + a_{1p}x_{ip} = \sum_{j=1}^{P} a_{1j}x_{ij} = x_{i\cdot}a_{1\cdot} \tag{5.1}$$

Donde x_{ip} representa el valor de la p-ésima variable para la observación i, a_{1p} representa el coeficiente por el cual multiplicar la p-ésima variable para la primera componente principal, y $x_{i\cdot}$, $a_{1\cdot}$ representan el vector de variables correspondientes a la i-ésima observación y al vector de coeficientes a multiplicar por cada variable para la primera componente, respectivamente.

De manera similar, el vector de tamaño N correspondiente a la primera componente principal para todas las observaciones (denotado por $z_{\cdot 1}$) se expresa como:

$$z_{\cdot 1} = a_{11}x_{\cdot 1} + a_{12}x_{\cdot 2} + \ldots + a_{1p}x_{\cdot p} = \sum_{j=1}^{P} a_{1j}x_{\cdot j} = \boldsymbol{X}a_{1\cdot} \tag{5.2}$$

Donde $x_{\cdot p}$ representa el vector de tamaño N correspondiente a los valores de la p-ésima variable de cada observación. Además de esto, se buscaría que $z_{\cdot 1}$ fuera la función de tipo combinación lineal de $\boldsymbol{X}$ con la mayor varianza posible.

De una manera similar, se busca que la segunda componente principal sea la segunda combinación lineal de las variables con mayor varianza posible, pero además se busca que no se encuentre correlacionada con la primer componente. De manera algebraica, el vector de segundas componentes para las N observaciones se expresaría como:

$$z_{\cdot 2} = a_{21}x_{\cdot 1} + a_{22}x_{\cdot 2} + \ldots + a_{2p}x_{\cdot p} = \sum_{j=1}^{P} a_{2j}x_{\cdot j} = \boldsymbol{X}a_{2\cdot} \tag{5.3}$$

Y al generalizar para toda componente principal se obtiene que la j-ésima componente principal debe de ser la combinación lineal de las variables originales que tenga la j-ésima mayor varianza, además cumpliendo que esta componente no se encuentre correlacionada con ninguna de las anteriores $j-1$ componentes principales. Este proceso se continúa hasta haber calculado P componentes principales, dado que el número máximo de componentes a calcular es equivalente al número de variables. El por qué de esta aseveración también se esclarecerá en el momento de calcular las componentes.

Se procede ahora al cálculo de dichas componentes:

Sea $\boldsymbol{X}$ descrita anteriormente, y sea $\boldsymbol{X^c}$ la matriz centrada por variables de $\boldsymbol{X}$. Esto quiere decir que:

$$x_{ip}^{c} = x_{ip} - \frac{1}{N}\sum_{v=1}^{N} x_{vp} = x_{ip} - \bar{x}_{\cdot p} \tag{5.4}$$

para toda $i = 1, 2, \ldots, N$ y $p = 1, 2, \ldots, P$.

De ahora en adelante se trabajará con $\boldsymbol{X^c}$ y no con $\boldsymbol{X}$, ya que el cálculo de las componentes se simplificará mucho más bajo la matriz centrada. Por ende, todos los resultados, interpretaciones y conclusiones de las componentes principales serán con respecto a la matriz centrada $\boldsymbol{X^c}$, y no respecto a la matriz de datos originales $\boldsymbol{X}$.

El motivo por el cual se recomienda usar $\boldsymbol{X^c}$ es que su varianza viene dada por el valor:

$$\text{Var}(\boldsymbol{X^c}) = \boldsymbol{S^c} = \frac{1}{M-1}(\boldsymbol{X^c})'\boldsymbol{X^c} \tag{5.5}$$

Mientras que:

$$\text{Var}(\boldsymbol{X}) = \boldsymbol{S} = \frac{1}{M-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}}) \tag{5.6}$$

Esta última ecuación sería la que complicaría los cálculos si no se centrara la matriz de datos.

Ahora que se tiene a la matriz centrada, se había mencionado que se buscaba que la primer componente principal fuera de la forma $z_{\cdot 1} = \sum_{p=1}^{P} a_{1p}x^c_{\cdot p} = \boldsymbol{X^c}a_{1\cdot}$, se procede a calcular la varianza de la primera componente, para posteriormente maximizarla utilizando el método de multiplicadores de *Lagrange*. Se obtiene entonces que:

$$\text{Var}(z_{\cdot 1}) = \text{Var}(\boldsymbol{X^c}a_{1\cdot}) = a'_{1\cdot}\text{Var}(\boldsymbol{X^c})a_{1\cdot} = a'_{1\cdot}\boldsymbol{S^c}a_{1\cdot} \tag{5.7}$$

El siguiente paso es maximizar dicha varianza. Para poder utilizar el método de multiplicadores de Lagrange, es necesario fijar una condición. Dicha condición será que $a_{1\cdot}$ sea un vector unitario, esto es, que $a'_{1\cdot}a_{1\cdot} = 1$. Aunque bien se podrían haber elegido otras condiciones respecto a la norma de $a_{1\cdot}$, esta condición es la que da mayor sencillez al cálculo de las componentes, además de ser sencilla de interpretar.

La ecuación para maximizar a $a'_{1\cdot}\boldsymbol{S^c}a_{1\cdot}$ sujeto a la restricción $a'_{1\cdot}a_{1\cdot} = 1$ es:

$$a'_{1\cdot}\boldsymbol{S^c}a_{1\cdot} - \lambda_1(a'_{1\cdot}a_{1\cdot} - 1) \tag{5.8}$$

donde λ_1 es el multiplicador de Lagrange correspondiente a la restricción. Ahora, al diferenciar con respecto a $a_{1\cdot}$ e igualar a cero se obtiene que:

$$\frac{\partial}{\partial a_{1\cdot}}\left(a'_{1\cdot}\boldsymbol{S^c}a_{1\cdot} - \lambda_1(a'_{1\cdot}a_{1\cdot} - 1)\right) = 2\boldsymbol{S^c}a_{1\cdot} - 2\lambda_1 a_{1\cdot} = 0$$

$\Rightarrow$

$$\boldsymbol{S^c}a_{1\cdot} = \lambda_1 a_{1\cdot}$$

En este momento se concluye que la igualdad anterior es la definición de "eigenvalores" y "eigenvectores" asociados a una matriz, en este caso la matriz $\boldsymbol{S^c}$, por lo que se deduce que λ_1 es un eigenvalor asociado a la matriz de varianzas y covarianzas centrada $\boldsymbol{S^c}$, y $a_{1\cdot}$ es el eigenvector asociado al eigenvalor λ_1.

Ahora, al multiplicar la igualdad anterior por $a'_{1\cdot}$ por el lado izquierdo, se obtiene la varianza de la primer componente principal:

$$a'_{1\cdot}\boldsymbol{S^c}a_{1\cdot} = a'_{1\cdot}\lambda_1 a_{1\cdot} = \lambda_1 a'_{1\cdot}a_{1\cdot} = \lambda_1(1) = \lambda_1$$

Por lo que se obtiene que λ_1 es la varianza de la primer componente principal, y dado que se buscaba que la varianza de la primer componente principal fuera la mayor varianza posible para una combinación lineal de los datos, entonces λ_1 tomará el valor del eigenvalor más grande asociado a la matriz $\boldsymbol{S^c}$.

El motivo por el cuál el número de componentes principales de un conjunto de datos es a lo mucho el número de variables de dichos datos viene del hecho de que las componentes se obtienen por medio de los eigenvectores, y los eigenvectores asociados a una matriz de varianzas y covarianzas asociada al conjunto de datos son a lo mucho el número de variables de los datos.

Ya que se ha calculado el valor y varianza de la primer componente principal, se procede a calcular la segunda. La segunda componente principal se calculará de una manera muy similar a la primer componente, con las diferencias de que en vez de buscar la mayor varianza para la componente se buscará la segunda mayor varianza, y esta nueva componente no tiene que estar correlacionada con la anterior, es decir, $\text{Cov}(z_{\cdot 1}, z_{\cdot 2}) = 0$. Pero:

$$\text{Cov}(z_{\cdot 1}, z_{\cdot 2}) = \text{Cov}(\boldsymbol{X^c} a_{1\cdot}, \boldsymbol{X^c} a_{2\cdot}) = a'_{2\cdot}\text{Cov}(\boldsymbol{X^c}, \boldsymbol{X^c})a_{1\cdot} = a'_{2\cdot}\boldsymbol{S^c}a_{1\cdot} = a'_{2\cdot}\lambda_1 a_{1\cdot}$$

Por lo que

$$\text{Cov}(z_{\cdot 1}, z_{\cdot 2}) = a'_{2\cdot}\lambda_1 a_{1\cdot} = \lambda_1 a'_{2\cdot}a_{1\cdot} = 0$$

Por lo tanto el que $z_{\cdot 1}$ y $z_{\cdot 2}$ no sean correlacionados implica que $a'_{2\cdot}a_{1\cdot} = 0$.

Esta última igualdad también se utilizará como restricción para el método de multiplicadores de Lagrange.

La ecuación a maximizar para encontrar la segunda componente principal es, entonces:

$$a'_{2\cdot}\boldsymbol{S^c}a_{2\cdot} - \lambda_2(a'_{2\cdot}a_{2\cdot} - 1) - \delta a'_{2\cdot}a_{1\cdot}$$

donde λ_2 y δ son multiplicadores de Lagrange. Diferenciando nuevamente con respecto a $a_{2\cdot}$ e igualando a cero para encontrar el máximo se obtiene:

$$\frac{\partial}{\partial a_{2\cdot}}(a'_{2\cdot}\boldsymbol{S^c}a_{2\cdot} - \lambda_2(a'_{2\cdot}a_{2\cdot} - 1) - \delta a'_{2\cdot}a_{1\cdot}) = 2\boldsymbol{S^c}a_{2\cdot} - 2\lambda_2 a_{2\cdot} - \delta a_{1\cdot} = 0$$

Ahora, multiplicando la expresión anterior por $a'_{1\cdot}$ por la izquierda se obtiene:

$$\begin{aligned}
2a'_{1\cdot}\boldsymbol{S^c}a_{2\cdot} - 2\lambda_2 a'_{1\cdot}a_{2\cdot} - \delta a'_{1\cdot}a_{1\cdot} =& 2(a'_{2\cdot}(\boldsymbol{S^c})'a_{1\cdot})' - 2\lambda_2(a'_{2\cdot}a_{1\cdot})' - \delta a'_{1\cdot}a_{1\cdot} \\
=& 2(a'_{2\cdot}\boldsymbol{S^c}a_{1\cdot})' - 2\lambda_2(a'_{2\cdot}a_{1\cdot})' - \delta a'_{1\cdot}a_{1\cdot} \\
=& 2\lambda_1(a'_{2\cdot}a_{1\cdot})' - 2\lambda_2(a'_{2\cdot}a_{1\cdot})' - \delta a'_{1\cdot}a_{1\cdot} \\
=& -\delta \\
=& 0
\end{aligned}$$

Lo anterior debido a las restricciones $a'_{2\cdot}a_{1\cdot} = 0$ y $a'_{1\cdot}a_{1\cdot} = 1$. Dado que $\delta = 0$, entonces se puede omitir el último término de la derivada, por lo que la derivada

anterior sería igual a:

$$\frac{\partial}{\partial a_{2\cdot}}(a'_{2\cdot}\boldsymbol{S^c}a_{2\cdot} - \lambda_2(a'_{2\cdot}a_{2\cdot} - 1) - \delta a'_{2\cdot}a_{1\cdot}) = 2\boldsymbol{S^c}a_{2\cdot} - 2\lambda_2 a_{2\cdot}$$

dando como resultado al igualar a cero que $\boldsymbol{S^c}a_{2\cdot} = \lambda_2 a_{2\cdot}$. Esto indicaría nuevamente que la componente a calcular se obtiene mediante un eigenvector, mismo que correspondería al eigenvalor λ_2. De manera similar a la primera componente, al calcular la varianza de la segunda componente principal se obtendría:

$$\text{Var}(z_{\cdot 2}) = a'_{2\cdot}\boldsymbol{S^c}a_{2\cdot} = \lambda_2 a'_{2\cdot}a_{2\cdot} = \lambda_2$$

Por lo que bajo el supuesto de máxima varianza para las componentes, el eigenvector que se utilizará para calcular a la segunda componente $z_{\cdot 2}$ corresponderá al eigenvector asociado al segundo eigenvalor más grande de la matriz de varianzas y covarianzas $\boldsymbol{S^c}$.

Este procedimiento para calcular componentes principales se puede generalizar para las demás $P-2$ componentes, dando siempre como restricciones que la norma de $a_{j\cdot}$ es igual a uno, $a_{j\cdot}$ y $a_{k\cdot}$ son no correlacionados para $k = 1, 2, \ldots, j-1$, y la varianza de la j-ésima componente principal es la varianza más grande posible excluyendo a los valores de las varianzas de las primeras $j-1$ componentes. También se obtiene para toda $j = 1, 2, \ldots, P$ que la varianza de la j-ésima componente principal viene denotada por λ_j, el j-ésimo mayor eigenvalor.

De ahora en adelante, a los valores que conforman al vector $a_{j\cdot}$ que corresponden al eigenvector que será de utilidad para calcular el valor de la j-ésima componente principal $z_{\cdot j}$, serán llamados los *loadings* de la j-ésima componente principal.

5.2. Propiedades importantes

A continuación, se listarán algunas de las propiedades de mayor importancia para el análisis de componentes principales. La notación que se utilizará en esta sección coincide con la utilizada en secciones anteriores.

Propiedad 1: Dada una matriz cuadrada, sus eigenvalores pueden ser tanto positivos como negativos. Dado que en este caso los eigenvalores representan varianzas, dichos eigenvalores deben ser no negativos. La pregunta es, ¿cómo asegurar que los eigenvalores de $\boldsymbol{S^c}$ cumplen con esta propiedad? La justificación de dicha aseveración proviene del siguiente resultado de álgebra lineal.

Sea $\boldsymbol{S}$ una matriz de dimensión $P \times P$ simétrica en la que todos los elementos de su diagonal son mayores o iguales a cero. Entonces, para cualquier vector x de dimensión $P \times 1$ diferente del vector cero, se cumple que $x'\boldsymbol{S}x \geq 0$. A este tipo de matrices se les llaman semidefinidas positivas.

Dado que la matriz utilizada $\boldsymbol{S^c}$ es una matriz de varianzas y covarianzas, entonces dicha matriz es simétrica y a la vez todos sus elementos de la diagonal son mayores o iguales a cero, por lo que para cualquier vector diferente del vector nulo se cumple que $x'\boldsymbol{S^c}x \geq 0$. Pero a la misma vez, por definición se conoce que para cualquier eigenvector de $\boldsymbol{S^c}$ se cumple que $\boldsymbol{S^c}x = \lambda x$. Por lo tanto, para todo eigenvector x de $\boldsymbol{S^c}$ se cumple que,

$$\begin{aligned} x'\boldsymbol{S^c}x \geq 0 &\implies x'\lambda x \geq 0 \\ &\implies \lambda x'x \geq 0 \end{aligned}$$

Pero dado que $x'x = \sum_{i=1}^{P} x_i^2 > 0$ entonces en orden para seguir cumpliendo la desigualdad anterior, forzosamente $\lambda \geq 0$. De esta manera se justifica que todo eigenvalor de $\boldsymbol{S^c}$ es no negativo.

Propiedad 2: Sea $\boldsymbol{A}$ la matriz de dimensión $P \times P$ conformada por los eigenvectores de la matriz $\boldsymbol{S^c}$ por columnas, con la i-ésima columna dada por el eigenvector asociado al i-ésimo mayor eigenvalor. Entonces, la matriz $\boldsymbol{A}$ es una matriz ortogonal.

Demostración: Una matriz $\boldsymbol{A}$ es ortogonal si la multiplicación de $\boldsymbol{A'A = I}$, con $\boldsymbol{I}$ la matriz identidad. Entonces:

$$\begin{aligned} \boldsymbol{A'A} &= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1P} \\ a_{21} & a_{22} & \cdots & a_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ a_{P1} & a_{P2} & \cdots & a_{PP} \end{bmatrix} \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{p1} \\ a_{12} & a_{22} & \cdots & a_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1P} & a_{2P} & \cdots & a_{PP} \end{bmatrix} \\ &= \begin{bmatrix} a'_{1\cdot}a_{1\cdot} & a'_{1\cdot}a_{2\cdot} & \cdots & a'_{1\cdot}a_{P\cdot} \\ a'_{2\cdot}a_{1\cdot} & a'_{2\cdot}a_{2\cdot} & \cdots & a'_{2\cdot}a_{P\cdot} \\ \vdots & \vdots & \ddots & \vdots \\ a'_{P\cdot}a_{1\cdot} & a'_{P\cdot}a_{2\cdot} & \cdots & a'_{P\cdot}a_{P\cdot} \end{bmatrix} \end{aligned}$$

Pero como $a'_{i\cdot}a_{i\cdot} = 1$ para todo $i = 1, 2, \ldots, P$ y $a'_{i\cdot}a_{j\cdot} = 0$ para $i \neq j$, entonces:

$$\boldsymbol{A'A = I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \boldsymbol{I}$$

Por lo tanto, $\boldsymbol{A}$ es una matriz ortogonal.

Propiedad 3: Sean $\boldsymbol{X^c}$ y $\boldsymbol{A}$ las mismas matrices descritas anteriormente. Entonces, al ser $\boldsymbol{A}$ una matriz ortogonal, implica que $\boldsymbol{A}$ es también una matriz de rotación, por lo que la transformación $\boldsymbol{X^cA}$ conserva la misma distancia entre individuos que $\boldsymbol{X^c}$.

Propiedad 4: El vector de covarianzas de la matriz $\boldsymbol{X^c}$ y la i-ésima componente principal $z_{\cdot i}$ es denotado por:

$$\text{Cov}(\boldsymbol{X^c}, z_{\cdot i}) = \text{Cov}(\boldsymbol{X^c}, \boldsymbol{X^c}a_{i\cdot}) = \text{Cov}(\boldsymbol{X^c}, \boldsymbol{X^c})a_{i\cdot} = \boldsymbol{S^c}a_{i\cdot} = \lambda_i a_{i\cdot} \tag{5.9}$$

Por lo que la j-ésima entrada de este vector es equivalente a la covarianza entre la variable j de la matriz $\boldsymbol{S^c}$ y la i-ésima componente principal. De forma algebraica, esto queda expresado por:

$$\text{Cov}(X^c_{\cdot j}, z_{\cdot i}) = \lambda_i a_{ij} \tag{5.10}$$

Propiedad 5: De la propiedad anterior, se deriva que la correlación entre la variable j de $\boldsymbol{X^c}$ y la i-ésima componente principal se calcula como:

$$\begin{aligned}\text{Cor}(X^c_{\cdot j}, z_{\cdot i}) =& \frac{\text{Cov}(X^c_{\cdot j}, z_{\cdot i})}{\sqrt{\text{Var}(X^c_{\cdot j})\text{Var}(z_{\cdot i})}} \\ =& \frac{\lambda_i a_{ij}}{\sqrt{S^c_{jj}\lambda_i}} \\ =& \frac{\sqrt{\lambda_i}a_{ij}}{\sqrt{S^c_{jj}}}\end{aligned}$$

Propiedad 6: De resultados de álgebra lineal se obtiene que la traza de una matriz es igual a la suma de los eigenvalores de la misma. Esto interpretado para componentes principales, significa que la suma de las varianzas de cada variable de $\boldsymbol{X^c}$ es igual a la suma de las varianzas de las componentes principales, es decir:

$$\sum_{i=1}^{P} S^c_{ii} = \sum_{i=1}^{P} \lambda_i$$

Esto quiere decir que las componentes principales conservan la variabilidad de los datos.

Demostración: Dado que $\boldsymbol{S^c}$ se puede escribir como $\boldsymbol{ADA^{-1}}$ con $\boldsymbol{A}$ el conjunto de eigenvectores asociados a $\boldsymbol{S^c}$ acomodados por columna y $\boldsymbol{D}$ la matriz diagonal de eigenvalores de $\boldsymbol{S^c}$, entonces la varianza de la variable j de $\boldsymbol{X^c}$ se puede escribir como:

$$S^c_{jj} = (\boldsymbol{ADA^{-1}})_{jj}$$

Primero se calculará el valor de $\boldsymbol{ADA^{-1}}$ para poder utilizar sus entradas.

Dado que:

$$\boldsymbol{A}^{-1} = \frac{(\text{Adj}(\boldsymbol{A}))'}{\text{Det}(\boldsymbol{A})} = \frac{(\text{Adj}(\boldsymbol{A}))'}{|\boldsymbol{A}|}$$

con $\text{Det}(\boldsymbol{A})=|\boldsymbol{A}|$ el determinante de la matriz $\boldsymbol{A}$ y $(\text{Adj}(\boldsymbol{A}))'$ es la transpuesta de la matriz adjunta de $\boldsymbol{A}$. Entonces se sigue que:

$$\begin{aligned}\boldsymbol{S^c} = \boldsymbol{ADA^{-1}} =& \boldsymbol{AD}\frac{(\text{Adj}(\boldsymbol{A}))'}{\text{Det}(\boldsymbol{A})} \\ =& \frac{1}{\text{Det}(\boldsymbol{A})}\boldsymbol{AD}(\text{Adj}(\boldsymbol{A}))' \\ =& \frac{1}{|\boldsymbol{A}|}\boldsymbol{AD}(\text{Adj}(\boldsymbol{A}))'\end{aligned}$$

Y expresándolo en matrices definiendo a $\text{Adj}(\boldsymbol{A})_{ij}$ como la entrada del i-ésimo renglón y la j-ésima columna de la matriz adjunta de $\boldsymbol{A}$ se obtiene:

$$\boldsymbol{ADA^{-1}} = \frac{1}{|\boldsymbol{A}|}\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{P1} \\ a_{12} & a_{22} & \cdots & a_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1P} & a_{2P} & \cdots & a_{PP} \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_P \end{bmatrix}\begin{bmatrix} \text{Adj}(\boldsymbol{A})_{11} & \text{Adj}(\boldsymbol{A})_{12} & \cdots & \text{Adj}(\boldsymbol{A})_{1P} \\ \text{Adj}(\boldsymbol{A})_{21} & \text{Adj}(\boldsymbol{A})_{22} & \cdots & \text{Adj}(\boldsymbol{A})_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Adj}(\boldsymbol{A})_{P1} & \text{Adj}(\boldsymbol{A})_{P2} & \cdots & \text{Adj}(\boldsymbol{A})_{PP} \end{bmatrix}'$$

$$= \frac{1}{|\boldsymbol{A}|}\begin{bmatrix} a_{11} & a_{21} & \cdots & a_{P1} \\ a_{12} & a_{22} & \cdots & a_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1P} & a_{2P} & \cdots & a_{PP} \end{bmatrix}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_P \end{bmatrix}\begin{bmatrix} \text{Adj}(\boldsymbol{A})_{11} & \text{Adj}(\boldsymbol{A})_{21} & \cdots & \text{Adj}(\boldsymbol{A})_{P1} \\ \text{Adj}(\boldsymbol{A})_{12} & \text{Adj}(\boldsymbol{A})_{22} & \cdots & \text{Adj}(\boldsymbol{A})_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Adj}(\boldsymbol{A})_{1P} & \text{Adj}(\boldsymbol{A})_{2P} & \cdots & \text{Adj}(\boldsymbol{A})_{PP} \end{bmatrix}$$

$$= \frac{1}{|\boldsymbol{A}|}\begin{bmatrix} \lambda_1 a_{11} & \lambda_2 a_{21} & \cdots & \lambda_P a_{P1} \\ \lambda_1 a_{12} & \lambda_2 a_{22} & \cdots & \lambda_P a_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1 a_{1P} & \lambda_2 a_{2P} & \cdots & \lambda_P a_{PP} \end{bmatrix}\begin{bmatrix} \text{Adj}(\boldsymbol{A})_{11} & \text{Adj}(\boldsymbol{A})_{21} & \cdots & \text{Adj}(\boldsymbol{A})_{P1} \\ \text{Adj}(\boldsymbol{A})_{12} & \text{Adj}(\boldsymbol{A})_{22} & \cdots & \text{Adj}(\boldsymbol{A})_{P2} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Adj}(\boldsymbol{A})_{1P} & \text{Adj}(\boldsymbol{A})_{2P} & \cdots & \text{Adj}(\boldsymbol{A})_{PP} \end{bmatrix}$$

Dando como resultado:

$$\boldsymbol{S^c} = \boldsymbol{ADA^{-1}} = \frac{1}{|\boldsymbol{A}|}\begin{bmatrix} \sum_{i=1}^{P} \lambda_i a_{i1}\text{Adj}(\boldsymbol{A})_{1i} & \sum_{i=1}^{P} \lambda_i a_{i1}\text{Adj}(\boldsymbol{A})_{2i} & \cdots & \sum_{i=1}^{P} \lambda_i a_{i1}\text{Adj}(\boldsymbol{A})_{Pi} \\ \sum_{i=1}^{P} \lambda_i a_{i2}\text{Adj}(\boldsymbol{A})_{1i} & \sum_{i=1}^{P} \lambda_i a_{i2}\text{Adj}(\boldsymbol{A})_{2i} & \cdots & \sum_{i=1}^{P} \lambda_i a_{i2}\text{Adj}(\boldsymbol{A})_{Pi} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{P} \lambda_i a_{iP}\text{Adj}(\boldsymbol{A})_{1i} & \sum_{i=1}^{P} \lambda_i a_{iP}\text{Adj}(\boldsymbol{A})_{2i} & \cdots & \sum_{i=1}^{P} \lambda_i a_{iP}\text{Adj}(\boldsymbol{A})_{Pi} \end{bmatrix}$$

Recuérdese que el valor a_{ij} representa la j-ésimo loading utilizado para calcular a la i-ésima componente principal, λ_i representa la varianza de la i-ésima componente principal y $\text{Adj}(\boldsymbol{A})_{ij}$ representa el valor del i-ésimo renglón y j-ésima

columna de la matriz adjunta de $\boldsymbol{A}$.

Por lo tanto, la varianza de la j-ésima variable se puede calcular como:

$$S_{jj} = \boldsymbol{ADA}^{-\mathbf{1}}_{jj} = \frac{1}{|\boldsymbol{A}|}\sum_{i=1}^{P} \lambda_i a_{ij} \text{Adj}(\boldsymbol{A})_{ji}$$

Y el valor de la suma de las varianzas de las P variables de $\boldsymbol{X^c}$ es:

$$\begin{aligned}\sum_{j=1}^{P} S_{jj} =& \sum_{j=1}^{P} \boldsymbol{ADA}^{-\mathbf{1}}_{jj} \\ =& \frac{1}{|\boldsymbol{A}|}\sum_{j=1}^{P}\sum_{i=1}^{P} \lambda_i a_{ij} \text{Adj}(\boldsymbol{A})_{ji}\end{aligned}$$

Y al ordenar los términos con respecto a λ_i, se obtiene que $\sum_{j=1}^{P} S_{jj}$ es igual a:

$$\frac{1}{|\boldsymbol{A}|}\left[\lambda_1\left(\sum_{i=1}^{P} a_{1i}\text{Adj}(\boldsymbol{A})_{i1}\right) + \cdots + \lambda_P\left(\sum_{i=1}^{P} a_{Pi}\text{Adj}(\boldsymbol{A})_{iP}\right)\right]$$

Pero:

$$\sum_{i=1}^{P} a_{1i}\text{Adj}(\boldsymbol{A})_{i1} = \sum_{i=1}^{P} a_{2i}\text{Adj}(\boldsymbol{A})_{i2} = \cdots = \sum_{i=1}^{P} a_{Pi}\text{Adj}(\boldsymbol{A})_{iP} = \text{Det}(\boldsymbol{A}) = |\boldsymbol{A}|$$

Por lo tanto:

$$\sum_{j=1}^{P} S_{jj} = \sum_{i=1}^{P} \lambda_i \tag{5.11}$$

Y por lo tanto, las componentes principales conservan la variabilidad original de los datos $\boldsymbol{X^c}$.

5.3. Componentes principales bajo matriz de correlaciones

Hasta ahora, se ha trabajado con la matriz de varianzas y covarianzas $\boldsymbol{S^c}$ para obtener sus respectivos eigenvalores y eigenvectores y así poder calcular las componentes principales a utilizar en $\boldsymbol{X^c}$. Sin embargo, $\boldsymbol{S^c}$ no es la única opción para la obtención de las componentes principales de $\boldsymbol{X^c}$, pues además de

dicha matriz, se puede utilizar la matriz de correlaciones (de ahora en adelante denotada por $\boldsymbol{R}$) para obtener dichas componentes.

Las componentes principales calculadas utilizando la matriz de correlaciones se obtienen mediante:

$$\boldsymbol{z}_{\cdot\cdot} = \boldsymbol{X^s A^s} \tag{5.12}$$

donde $\boldsymbol{X^s}$ representa a la matriz de datos $\boldsymbol{X}$ estandarizada, $\boldsymbol{A^s}$ la matriz con los eigenvectores asociados a la matriz de correlaciones $\boldsymbol{R}$ ordenados por columna (del eigenvector asociado al mayor eigenvalor, al eigenvector asociado al menor eigenvalor) y $\boldsymbol{z}_{\cdot\cdot}$ representa la matriz de los *loadings* de cada componente principal para cada observación.

Claramente, todas las propiedades mencionadas anteriormente aplicarán para el caso en que se utiliza la matriz de correlaciones como base para calcular las componentes, sólo que ahora por matriz de datos se considerará a $\boldsymbol{X^s}$ y no a $\boldsymbol{X^c}$.

La elección entre utilizar la matriz de varianzas y covarianzas o la matriz de correlaciones dependerá básicamente de la naturaleza de los datos a utilizar, aunque normalmente se prefiere a la matriz de correlaciones. Los puntos a favor y en contra de utilizar cada una de las dos matrices propuestas para el cálculo de las componentes se mencionan a continuación.

1. Ventajas de utilizar la matriz de correlaciones $\boldsymbol{R}$ con respecto a la matriz de varianzas y covarianzas centrada $\boldsymbol{S^c}$:

   - Diferentes variables pueden tener una mejor interpretación al ser comparadas entre ellas.

   - Si algunas de las variables tuvieran una varianza mucho mayor en comparación a las demás, las primeras componentes principales podrían ser casi idénticas a las variables con mayor varianza y podrían no tomar en cuenta a las demás, por lo que no tendría mucho sentido realizar la técnica de componentes principales. Esto es un problema pues las componentes principales le darían mayor importancia a variables con una escala de medición mayor que a variables con escalas de medición menor; en otras palabras, las componentes principales son sensibles a la escala de medición de las variables.

     Para entender mejor el enunciado anterior, utilizando la matriz de varianzas y covarianzas se le dará mayor importancia a los datos de una variable medida en milímetros que a los mismos datos de la misma variable pero medida en metros o en kilómetros.

2. Ventajas de utilizar la matriz de varianzas y covarianzas centrada $\boldsymbol{S^c}$ con respecto a la matriz de correlaciones $\boldsymbol{R}$:

- En casos donde todas las variables sean de la misma naturaleza y se encuentren en la misma escala de medición, el uso de dicha matriz podría traer resultados confiables.

Dado que es de gran importancia el que se comprendan los efectos de utilizar una matriz de varianzas y covarianzas para datos con diferentes escalas de medición, se expondrá un ejemplo mostrando los posibles problemas de tomar esta decisión. El ejemplo expuesto a continuación es obtenido del libro *Principal Component Analysis* de *Jolliffe* del año *2002*, página *22*:

Supóngase un conjunto de observaciones con dos variables continuas cada una. Supóngase también que la primera variable puede ser medida en centímetros o en milímentros. En el caso en que la primera variable sea medida en centímetros, la matriz de varianzas y covarianzas de los datos es:

$$\boldsymbol{S^c} = \begin{bmatrix} 80 & 44 \\ 44 & 80 \end{bmatrix}$$

Y el valor de la primera componente principal para cualquier observación i viene denotado por:

$$z_{i1} = 0.707x_{i1} + 0.702x_{i2} \tag{5.13}$$

Es decir, la primera componente le da el mismo peso a las dos variables. Ahora, para el caso en que la primera variable es medida en milímetros, su respectiva matriz de varianzas y covarianzas queda calculada como:

$$\boldsymbol{S^c} = \begin{bmatrix} 8000 & 440 \\ 440 & 80 \end{bmatrix}$$

Mientras que el valor de la primera componente principal para cualquier observación i es calculado de la forma:

$$z_{i1} = 0.998x_{i1} + 0.055x_{i2}$$

Es decir, la primera componente toma prácticamente el valor de la variable con mayor varianza, lo que es un grave problema si se quisiera utilizar componentes principales de manera correcta, pues se busca que la escala de medida no influyera en el análisis y en ambos casos la primera componente principal tuviera el mismo valor. Es por esta causa, entre otras más, que la mayoría de las veces se utiliza la matriz de correlación y no la de varianzas y covarianzas para el cálculo de las componentes principales.

Como último punto, es importante mencionar que los eigenvectores y eigenvalores calculados mediante la matriz de correlaciones $\boldsymbol{R}$ y los calculados mediante la matriz de varianzas y covarianzas $\boldsymbol{S^c}$ no tienen relación alguna, por lo que no se podría obtener $\boldsymbol{A^s}$ a partir de $\boldsymbol{A^c}$ y viceversa (una matriz de correlaciones puede provenir de diferentes matrices de varianzas y covarianzas, por ejemplo). Son dos acercamientos totalmente a parte el uno del otro.

También es importante hacer énfasis en que la matriz de observaciones utilizada para calcular las componentes principales será la matriz de observaciones centrada ($\boldsymbol{X^c}$) para el caso de utilizar la matriz de varianzas y covarianzas, y la matriz de observaciones estandarizadas por variable ($\boldsymbol{X^s}$) para el caso de utilizar la matriz de correlaciones. En ningún momento se utilizará la matriz de observaciones original ($\boldsymbol{X}$) para el cálculo de componentes principales.

5.4. Interpretación de los resultados

En esta sección se presentarán las interpretaciones tanto de las componentes principales como de sus respectivas varianzas. Para el caso de las segundas, se recurrirá a métodos gráficos. Cabe mencionar también, que no siempre las componentes principales van a tener una interpretación sencilla o lógica, y pueden haber casos en que se puedan tener más de una interpretación. Esto se debe, como se verá a continuación, a que gran parte de su interpretación depende de la creatividad y enfoque del investigador.

Primero se hablará sobre la varianza de las componentes y cómo éstas pueden ayudar a elegir al mejor conjunto de componentes que cumplan con la característica de ser un conjunto pequeño y al mismo tiempo que posea un nivel de variabilidad grande.

Gracias a la propiedad vista anteriormente que demuestra que la suma de las varianzas de las variables originales es igual a la suma de las varianzas de las componentes principales:

$$\sum_{j=1}^{P} S_{jj}^{c} = \sum_{i=1}^{P} \lambda_i$$

Se verifica que se pueden utilizar a las varianzas de las componentes principales para explicar las varianzas de las variables originales.

Se define entonces al "porcentaje de variabilidad total explicada", como el porcentaje de la suma total de las varianzas de los datos que cubre un conjunto de componentes. Sea P el número total de variables y por ende de componentes principales que tiene un conjunto de datos. Entonces el porcentaje de variabili-

dad explicada de las primeras r componentes principales se calcula como:

$$\frac{\sum_{i=1}^{r} \lambda_i}{\sum_{i=1}^{P} \lambda_i} \tag{5.14}$$

El valor obtenido por la fórmula anterior describirá el porcentaje de la variabilidad total que describe el conjunto de las primeras r variables.

Este nuevo concepto es muy útil para el análisis de componentes principales, dado que un alto porcentaje de variabilidad explicada quiere decir que casi toda la información relevante de las observaciones se almacena dentro de esas r componentes principales, y ayuda al mismo tiempo a definir cuántas componentes y cuáles son necesarias tomar para tener algún nivel de varianza explicada específico.

Ahora que ya se definió al porcentaje de variabilidad total explicada, una pregunta importante es, ¿qué porcentaje de variabilidad es adecuado para considerar que el conjunto de componentes tomado describe en mayor parte la variabilidad completa de los datos? Esa pregunta no tiene una respuesta específica, dado que dependerá tanto de la experiencia del investigador, como del tipo de análisis y naturaleza de los datos. En algunos casos ese porcentaje puede ser 60 %, mientras que en otros casos 80 % o 85 %.

Existe una gráfica auxiliar en la elección de componentes principales. Consiste en comparar directamente la varianza de cada componente para así tener una idea del comportamiento de ellas, y por ejemplo, poder encontrar el número de componentes que tengan un impacto significativo en la varianza total de las variables. Dicha gráfica recibe el nombre de *scree plot* o "diagrama de codo" en español, y un ejemplo de ella es la gráfica de la figura 5.1 que se muestra a continuación.

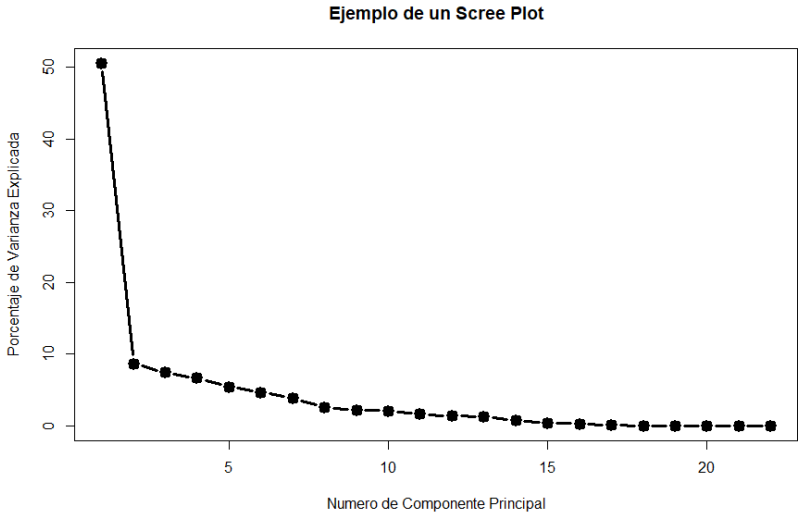

Figura 5.1: *Scree plot* de las componentes principales calculadas mediante la base de datos de entrenamiento mostrada en el apéndice.

Como se puede ver, la pendiente del cambio entre un eigenvalor y otro es significativa hasta el primer eigenvalor, donde después de éste, la pendiente del cambio de eigenvalores es muy pequeña. Por lo tanto se podría tomar como conjunto de componentes que explicaran a los datos únicamente a la primer componente principal, o bien se podría tomar el número de componentes necesarias para conseguir un porcentaje de variabilidad explicada específico. En esta figura, realizada a partir de la información de la base de datos de entrenamiento mostrada en el apéndice, el conjunto de componentes elegido de acuerdo a pendientes de cambio de eigenvalores que correspondía únicamente a la primer componente, tiene un porcentaje de variabilidad total explicada del 50 %, el cual se podría considerar un pequeño porcentaje. Seguramente, si el objetivo fuera explicar lo mayor posible con la menor cantidad de variables, se tomarían las primeras cuatro o cinco variables, que explicarían cerca del 80 % y aún así se reduciría el monto de variables de 15 a solo cuatro o cinco.

Normalmente se espera que la mayor variabilidad de las variables se encuentre en las primeras r componentes principales, con $r < P$, aunque esto claramente depende de la naturaleza de los datos y de qué tan correlacionadas se encuentren las variables.

Ahora se presentará un ejemplo gráfico que ayudará a entender mejor la transformación de variables ordinarias a componentes principales:

La figura 5.2 muestra la gráfica de dos variables para 100 observaciones. Se

aprecia que existe una alta correlación entre ellas, por lo que será de utilidad aplicar la técnica de componentes principales a los datos.

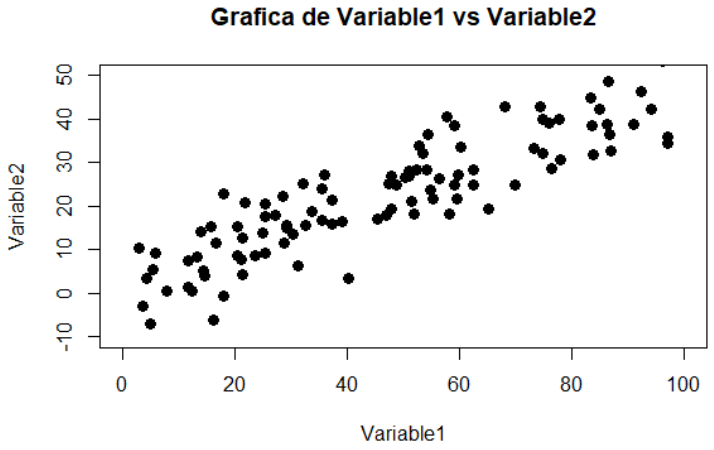

Figura 5.2: Gráfica de Variable1 *vs* Variable2

Una vez aplicada dicha técnica, la gráfica de ambas componentes se presenta en la figura 5.3.

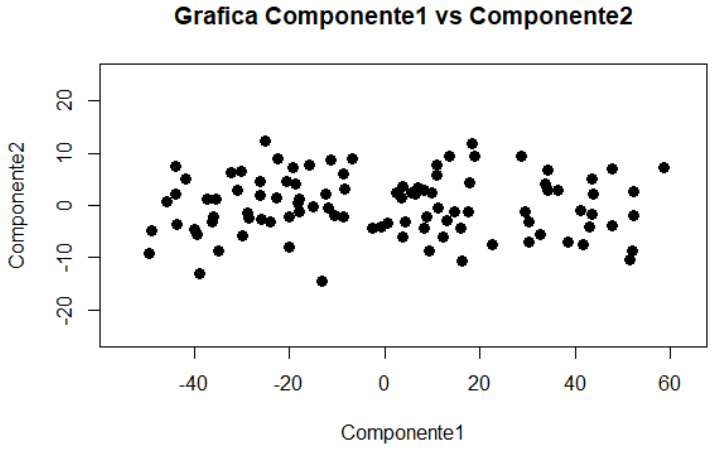

Figura 5.3: Gráfica de las componentes principales asociadas a Variable1 y Variable2.

Se puede apreciar, además de que tanto la distancia como la variabilidad

entre sujetos no se ha modificado, que la primer componente principal describe la mayor variabilidad entre los sujetos, dejando a la segunda componente una variabilidad mucho menor.

Normalmente en problemas reales, el número de variables a utilizar será mucho mayor a dos, por lo que esta gráfica únicamente dará una idea de la posición de los datos respecto a dos componentes principales. Aunque se pudiera graficar cualquier combinación de componentes principales, en la mayoría de las ocasiones se grafica a la componente uno con respecto a la componente dos, dado que esta combinación es la que presentará la mayor variabilidad al ser dichas componentes las que tienen la mayor varianza.

Otra utilidad que pudiera presentar la gráfica de las dos primeras componentes es utilizarla para detectar datos que pudieran ser *outliers*. Sin embargo, si se quisiera utilizar dicha gráfica para este motivo, se debe de tomar en cuenta el porcentaje de varianza explicada de las primeras dos componentes, dado que si este porcentaje fuera alto (dígase 80 % por ejemplo), será muy confiable el método para detectar *outliers*; pero si la varianza acumulada fuera pequeña (40 % por ejemplo), esta gráfica no será confiable para determinar si existen datos atípicos o no, dado que falta un gran porcentaje de variabilidad de los datos por describir.

Un último comentario se mencionará antes de terminar la sección, pues es importante saberlo y no alarmarse si se llegara a presentar la situación a mencionar. Supóngase que se tuvieran P variables que se quisieran analizar vía componentes principales, y al momento de aplicar componentes principales, q de estas P variables tuvieran signo positivo y las $P - q$ restantes signo negativo. El signo de las componentes principales es arbitrario, por lo que puede suceder que bajo un software distinto, las q variables positivas sean negativas mientras que las $P - q$ restantes, positivas. Esto no representa un problema, ya que los conjuntos de variables son los mismos. La interpretación es lo único que cambia bajo esta circunstancia, ya que se invertiría.

Al igual que los conjuntos de variables separados por signo serán los mismos, también lo serán las varianzas de cada componente.

5.5. Aspectos importantes a considerar

Los siguientes puntos son recomendaciones que el autor considera se deben de tomar en cuenta antes de realizar un análisis de componentes principales:

- La interpretación de las componentes será con respecto a los datos centra-

dos $\boldsymbol{X^c}$ si se utiliza la matriz de varianzas y covarianzas $\boldsymbol{S^c}$, y con respecto a los datos estandarizados $\boldsymbol{X^s}$ si se utiliza la matriz de correlaciones $\boldsymbol{R}$. En ningún caso se realizará con respecto a los datos originales dados por $\boldsymbol{X}$.

- Si se quisieran realizar predicciones con respecto a algún modelo de componentes principales, se deberán de utilizar los mismos valores utilizados en los datos originales para centrar o estandarizar los nuevos datos. Es decir, no se utilizará la media ni la desviación estándar de los nuevos datos, sino las obtenidas por los datos originales $\boldsymbol{X}$, y se calcularán las componentes con base en los *loadings* del modelo original.

- Si por objetivos del estudio se quisiera una vez calculadas las componentes principales, regresar los datos obtenidos a los datos originales $\boldsymbol{X}$, esto se puede realizar sin ningún problema, únicamente se debe de seguir considerando una vez obtenidos los datos en su escala original al multiplicarles su desviación estándar (para el caso estandarizado) y sumar su media a los datos, que los resultados del estudio en cuestión fueron obtenidos utilizando ya sea la matriz de correlaciones o la de varianzas y covarianzas. Es importante mencionar cuál de las dos se ha utilizado, dado que el uso de diferentes matrices pueden llevar a resultados completamente diferentes.

- Componentes principales será auxiliar a reducir dimensión manteniendo variabilidad sólo cuando las variables originales se encuentren correlacionadas. Si éstas no se encontraran correlacionadas, dichas variables tendrían un comportamiento similar a la gráfica 5.3, por lo que las componentes principales de los datos serían casi idénticas a las variables originales de los datos.

- Es de preferencia que la muestra de datos a utilizar no sea muy pequeña, ya que de serlo, podrían no ser muy confiables los resultados y conclusiones. No existe una regla sobre cuantas observaciones utilizar como mínimo, por lo que el concepto de "muy pequeño" o no dependerá de la experiencia de cada investigador.

- Si bien componentes principales es una herramienta creada para variables continuas, también puede ser utilizada en variables discretas tomando en cuenta una precaución, y es que la distancia entre los números discretos a utilizar tenga una interpretación lógica y de acorde a su respectivo número. Por ejemplo, si se fueran a utilizar los números uno, dos y tres para la variable discreta, entonces se tendría que cumplir que el valor tres realmente se encontrara al doble de distancia del valor uno que del valor dos, y que la interpretación del valor dos realmente se encontrara entre las

interpretaciones de los valores uno y tres.

Cabe resaltar que si la variable discreta únicamente se conformara de dos valores (cero y uno, por ejemplo), entonces no se tendría este problema y se pudiera utilizar directamente en componentes principales, ya que para variables binarias, el coeficiente *phi* es idéntico al coeficiente de correlación de Pearson, mismo que se utiliza para el cálculo de las componentes.

- Mientras mayor correlación tengan las variables originales entre ellas, menor será el número de componentes principales a utilizar para cubrir algún porcentaje de varianza específico, por lo que si lo que se quisiera es utilizar la técnica para encontrar un número pequeño de componentes que cubriera la mayor variabilidad posible, se recomienda utilizar variables que se encuentren fuertemente correlacionadas entre ellas.

Capítulo 6

Estadísticas de la NBA

6.1. Introducción al juego

A partir de este capítulo, se dejará de lado la base teórica de los modelos y métodos vistos anteriormente, para dar paso al análisis práctico del objeto de estudio por el cual se ha decidido realizar este trabajo: detectar a los factores que influyen en el desempeño de un equipo de baloncesto militante en la liga profesional estadounidense NBA (*National Basketball Association*, por sus siglas en inglés) y predecir el desempeño de los equipos para futuros años.

Este capítulo tiene el objetivo de familiarizar al lector con el conocimiento necesario para entender qué son, para qué sirven y cómo se calculan las estadísticas de la NBA que serán utilizadas como base para formar las variables predictoras necesarias para desarrollar el modelo.

Cabe resaltar que en este capítulo no se justificará el por qué se ha decidido utilizar estas variables, ya que estos motivos se mostrarán en el siguiente capítulo. En este capítulo únicamente se explicará el significado de las variables, así como su notación a utilizar para el análisis.

Antes de mostrar las estadísticas, se esclarecerán algunos conceptos respecto al baloncesto en la NBA:

- La NBA se compone de 30 equipos divididos en dos conferencias de 15 equipos cada una.
- El tiempo de juego de un partido se compone de 48 minutos, divididos en cuatro cuartos.
- La liga cuenta con dos etapas. Durante la primer etapa, los equipos se enfrentan entre ellos en un esquema "todos contra todos". El número total de juegos por equipo en esta etapa es de 82.

La segunda etapa, llamada postemporada, consiste en un sistema de eliminación directa entre los ocho mejores equipos de cada conferencia de acuerdo a sus respectivos porcentajes de victorias en la temporada, donde jugará el mejor equipo contra el peor, el segundo mejor contra el segundo peor y así sucesivamente. Los cuatro ganadores de cada uno de los enfrentamientos pasados se volverán a enfrentar entre ellos en un esquema similar al mencionado anteriormente, hasta que haya un equipo ganador para cada conferencia. Entonces, los dos equipos ganadores de su conferencia se enfrentarán, obteniendo de este enfrentamiento al equipo que será denominado el campeón de la NBA.

- Durante cualquier partido, no se permiten agresiones físicas contra los jugadores del equipo contrincante. Si llegasen a ocurrir, el jugador agresor será sancionado con una "falta" (*foul* en inglés). Si un jugador llegara a tener seis faltas en un sólo partido, será expulsado de éste.

- Cada lado de la cancha posee una línea llamada "línea de 3 puntos". Si un jugador encestara a una distancia menor a la línea de tres puntos del equipo rival, su enceste valdrá dos puntos; si el enceste se encontrara a una distancia mayor a la línea de tres puntos, el enceste le dará tres puntos a su equipo.

- Si a un jugador se le llegara a cometer una falta en el momento en que se encuentre realizando un tiro, tendrá como compensación el poder realizar tiros adicionales desde la línea de tiros libres. Los tiros libres, como su nombre lo dice, se realizan sin intervención de los jugadores del equipo rival. El número de tiros libres a realizar dependerá de la distancia donde haya tirado el jugador con respecto a la canasta, así como del hecho de que haya anotado o no el tiro al momento de que le realizaran la falta.

  Si el jugador encestó el tiro en el momento de recibir la falta, el jugador tendrá derecho a un tiro libre. Si el jugador falló el tiro y al jugador le realizaron la falta mientras tiraba desde fuera de la línea de tres puntos, recibirá un total de tres tiros libres; si el jugador falló a una distancia menor de la línea de tres, tendrá derecho a un total de dos tiros libres.

- el area más cercana a la canasta (denotada por el color anaranjado en la figura 6.1), será denotada como "la pintura".

- Un rebote se define como el evento de atrapar el balón después de que algún jugador haya tirado hacia la canasta y haya fallado. Si el jugador que tiró y el que obtuvo el rebote fueran del mismo equipo, entonces el rebote será ofensivo; si fuesen de distintos equipos, el rebote será defensivo.

- Un jugador tendrá el status de "veterano" cuando lleve tres años o más de antigüedad jugando en la liga, independientemente del número de equipos en que haya jugado.

- Un tiro de campo es equivalente a cualquier tiro que no haya sido tiro libre, es decir, que sea de dos o tres puntos (los únicos tiros de un punto son los tiros libres).
- El porcentaje de tiros de campo se define como:

$$\frac{\text{Tiros de campo encestados}}{\text{Tiros de campo intentados}} \cdot 100$$

- El porcentaje de tiros libres se define como:

$$\frac{\text{Tiros libres encestados}}{\text{Tiros libres intentados}} \cdot 100$$

- El porcentaje de tiros de tres puntos se define como:

$$\frac{\text{Tiros de tres puntos encestados}}{\text{Tiros de tres puntos intentados}} \cdot 100$$

La figura 6.1 presenta una cancha de baloncesto describiendo la localización de las líneas de tres puntos y de tiros libres descritas anteriormente.

Figura 6.1: Ubicación de tiros de 3 puntos y libres en una cancha de baloncesto

Existen más reglas y definiciones importantes para entender a profundidad cómo se juega el baloncesto en la NBA; sin embargo, las antes mencionadas son suficientes para poder entender a la perfección las estadísticas que servirán de apoyo para la realización del modelo.

6.2. Estadísticas elegidas

Ahora ha llegado el momento de presentar las estadísticas de la NBA que se utilizarán dentro del modelo. Cabe reiterar, que todas estas estadísticas son realizadas por equipo, no por jugador, dado que el análisis será con base en equipos. Así también, la mayoría de ellas serán resultado de un promedio por partido a lo largo de la temporada regular. Las estadísticas que se utilizarán para el cálculo del modelo son:

1. Proporción histórica de victorias del *coach*

   Consiste en la proporción de victorias que ha tenido el entrenador actual del equipo en cuestión a lo largo de toda su trayectoria como entrenador en la NBA. Se obtiene mediante la fórmula:

   $$\frac{\text{Total de victorias como entrenador de la NBA}}{\text{Total de juegos como entrenador de la NBA}}$$

2. Número de jugadores con experiencia en semifinales

   Consiste en el número de jugadores en la plantilla del equipo en cuestión que alguna vez en su carrera hayan participado cuando menos en una semifinal de la NBA. La estadística es de tipo contador.

3. Número de jugadores con tres años o más de antigüedad

   Como su nombre lo dice, consiste en el número de jugadores dentro de la plantilla actual del equipo en cuestión, que han jugado como mínimo los últimos tres años de manera ininterrumpida en el mismo equipo.

4. Minutos de veteranos

   Consiste en el promedio de minutos destinados a jugadores veteranos por partido.

5. Porcentaje de tiros de campo efectivo

   Variante del porcentaje de tiros de campo, con la característica de que ésta le dará una mayor importancia a los tiros de tres puntos. La fórmula para su cálculo es:

   $$\frac{\text{Tiros de dos puntos encestados} + 1.5(\text{Tiros de tres puntos encestados})}{\text{Tiros de campo intentados}} \cdot 100$$

6. Puntos

   Consiste en el promedio de los puntos realizados por el equipo en cuestión por partido.

7. Proporción de puntos sobre la pintura

   Es la proporción de los puntos del equipo que fueron obtenidos mediante tiros de campo realizados dentro de la zona de la pintura. Estos puntos, al ser mediante tiros de campo, excluyen a los puntos obtenidos por tiros libres. Su fórmula es:

$$\frac{\text{Total de puntos realizados en la pintura}}{\text{Total de puntos por partido}}$$

8. Porcentaje de tiros de campo del oponente

   Su fórmula es:

$$\frac{\text{Total de tiros de campo encestados por el oponente}}{\text{Total de tiros de campo realizados por el oponente}} \cdot 100$$

9. Puntos del oponente

   Es el promedio de puntos por partido que los equipos contrarios encestan al jugar contra el equipo en cuestión.

10. Estadística *clutch* 1

    Consiste en la diferencia entre el promedio de puntos anotados por el equipo en cuestión y el promedio de los puntos del oponente bajo la siguiente situación: el equipo en cuestión se encuentra perdiendo por cinco puntos o menos y restan tres minutos para que termine el juego.

11. Estadística *clutch* 2

    Consiste en el porcentaje de tiros libres del equipo en cuestión bajo la siguiente situación: el equipo en cuestión se encuentra ganando por un punto, resta un minuto para que termine el juego, y el resultado del juego termina en victoria para el equipo en cuestión.

12. Estadistica *clutch* 3

    Número de rebotes defensivos realizados por el equipo en cuestión bajo la siguiente situación: el equipo en cuestión se encuentra ganando por un punto, resta un minuto para que termine el juego, y el resultado del juego termina en victoria para el equipo en cuestión.

13. Estadística *clutch* 4

    Consiste en el promedio de puntos anotados por el equipo en cuestión bajo la siguiente situación: restan dos minutos de juego y la diferencia de puntos entre el equipo en cuestión y el equipo oponente es menor o igual a cinco.

14. Estadística *clutch* 5

    Consiste en el porcentaje de tiros de tres puntos encestados por parte del equipo en cuestión bajo la siguiente situación: restan dos minutos de juego y la diferencia de puntos entre el equipo en cuestión y el equipo oponente es menor o igual a cinco.

15. Estadística *clutch* 6

    Consiste en la proporción de victorias del equipo en cuestión sujeto a la siguiente situación: restan dos minutos de juego y la diferencia de puntos entre el equipo en cuestión y el equipo oponente es menor o igual a tres.

Ahora que ya se han presentado las estadísticas con las que se trabajará en el estudio, se mencionarán algunos comentarios respecto a la presente sección:

- Todas las estadísticas utilizadas fueron obtenidas a partir del sitio web oficial de la NBA, salvo unas cuantas que se enlistarán a continuación:
  - Proporción histórica de victorias del *coach*
  - Número de jugadores con experiencia en semifinales
  - Número de jugadores con tres años o más de antigüedad

  Las estadísticas mencionadas anteriormente, fueron creadas por el autor del documento, aunque la información necesaria para obtener los valores de estas fue obtenida en sitios confiables de la web.
- La imagen de la cancha de baloncesto fue obtenida del sitio web oficial del equipo de la NBA llamado "Phoenix Suns".

6.3. Una nota de gran importancia a considerar

Respecto a la estadística dos (jugadores con experiencia en semifinales), es muy importante remarcar que únicamente se tomaron en cuenta a jugadores que hayan promediado como mínimo 20 minutos por partido tanto en la temporada a evaluar como en la temporada en que el jugador alcanzó las semifinales de la NBA, y además haya jugado como mínimo el 60 % de todos los partidos posibles en la temporada a evaluar.

El filtro anterior se ha realizado con la idea de que la estadística mencionada tome en cuenta únicamente a jugadores que realmente hayan participado de

manera activa en el equipo tanto en la temporada actual como en la temporada en que logró el pase a semifinales, y por lo tanto hayan tenido un impacto significativo en ambas temporadas (¿de qué serviría un jugador con experiencia en semifinales, si en dichas semifinales únicamente jugó cinco minutos por partido?).

Así también, respecto a la estadística tres (jugadores con 3 años o más de antigüedad) se tomaron los mismos filtros de minutos por partido y juegos por temporada, sólo que para este caso, dichos requerimientos mínimos se debieron de haber registrado únicamente para la temporada actual.

Cabe mencionar que los valores utilizados para filtrar minutos y partidos jugados, fueron elegidos con base en la experiencia del autor del presente documento respecto a la NBA.

Capítulo 7

Caso Práctico: Aplicación a la NBA

7.1. Introducción al análisis

Ha llegado el momento de realizar el análisis de lo que fue el motivo de la realización de este documento, y en el cual se hará uso de las dos técnicas estadísticas mencionadas en los primeros cinco capítulos: el modelo de regresión logística multinomial y análisis de componentes principales.

Los objetivos principales de este estudio consistirán en determinar los factores con la mayor influencia en el desempeño de un equipo de baloncesto militante en la liga profesional NBA y predecir el desempeño de los equipos para futuros años.

Se utilizará el modelo de regresión logística multinomial para clasificar a los equipos de la liga de acuerdo a su desempeño en postemporada (llamada *playoffs* en inglés) de la siguiente manera:

- Si el equipo no clasificó a postemporada (*playoffs*) y además distó mucho de poder lograrlo, se clasificará en la categoría de equipos con desempeño pobre, grupo identificado por el número tres.
- Si el equipo no clasificó a *playoffs* pero no distó mucho de poder lograrlo, o si el equipo clasificó a *playoffs* pero fue eliminado en la primera ronda de estos, entonces el equipo se clasificaría en la categoría de equipos con desempeño medio, mismo que será identificado por el número dos.
- Si el equipo clasificó a *playoffs* y además resultó victorioso en la primera ronda de éstos, entonces el equipo se clasificará en la categoría de equipos con un desempeño alto, mismo que será identificado por el número uno.

Cabe destacar que se tomó la decisión de utilizar un modelo de regresión logística multinomial y no ordinal debido a que aunque la categoría tres se encuentra

acomodada de manera ordinal con respecto a las otras dos categorías, este acomodo ordinal no existe entre las categorías uno y dos. Un ejemplo del por qué no se encuentran ordenadas dichas categorías es el siguiente: pudiera ser que el octavo mejor equipo venciera al mejor equipo de una conferencia en la primera ronda de *playoffs*. Entonces, el octavo mejor equipo sería clasificado en la categoría número uno, mientras el mejor equipo de la temporada regular en la categoría dos.

Ahora, con respecto a cuánto es distar mucho o poco para que un equipo sea clasificado en la categoría dos o tres, se tomaron en cuenta las siguientes bases:

- Los tres equipos que no clasificaron a *playoffs* con los mejores porcentajes de victoria se clasificaron en la categoría dos.
- Los demás equipos sin avanzar a *playoffs* fueron clasificados en la categoría tres.

La decisión de elegir a dicha cantidad para separar las categorías mencionadas fue realizada con el objetivo de balancear el número de equipos en cada categoría, para así poder utilizar el punto de corte de máxima probabilidad al momento de clasificar observaciones (recuérdese que si un modelo de regresión logística multinomial estuviera desbalanceado, el modelo le dará mayor peso y por ende clasificará con mayor probabilidad a las categorías con mayor cantidad de observaciones que a las categorías con pocas observaciones, independientemente del verdadero valor de la variable de respuesta de cada observación).

La justificación del uso de componentes principales para el objetivo del estudio se revisará en la sección 2.4 del presente capítulo.

7.2. Preparación de la información para el caso

Para la elaboración de cualquier modelo estadístico, uno de los factores a los que se les debe de prestar mayor importancia para la búsqueda del mejor modelo posible, es encontrar al conjunto de variables que mayor explicación pueda dar respecto a la variable de respuesta, así como al conjunto de observaciones que puedan representar de buena manera a la población total. Así pues, se le dedicará una gran cantidad de tiempo y se le dará mucha importancia dentro de este estudio a la búsqueda de observaciones y variables que cumplan con estos requisitos.

7.2.1. Elección de observaciones

Primero se buscará el grupo de observaciones a utilizar. En este punto se tienen que tomar muchos factores en cuenta. Uno es, que se prefiere tomar un número grande de observaciones para realizar el modelo, ya que esta acción brindaría las siguientes ventajas:

- Reducir el efecto de posibles datos influyentes dentro del modelo.
- Permitir utilizar un mayor número de variables predictoras sin generar problemas de sobreajuste.
- Evitar sesgos por parte del modelo.

Por los motivos expresados anteriormente, se buscó utilizar la mayor cantidad posible de observaciones para el estudio en cuestión. Pero este no es el único factor que se debe de tomar en cuenta.

Dado que se busca utilizar una gran cantidad de observaciones, pero únicamente existen 30 equipos en la liga, una solución a este problema podría ser utilizar como observaciones a las estadísticas de los equipos de distintos años para la realización del modelo, aunque esta decisión podría resultar en un error, pues como se vio en el capítulo 1 del documento, la ecuación a maximizar y de la cual se obtienen los estimadores para el modelo deriva de la función de verosimilitud, que a su vez supone que todas las observaciones son independientes entre sí. Este es un problema para el estudio en cuestión, dado que al introducir mismos equipos de diferentes años, claramente se viola el supuesto de independencia entre observaciones. No existe una manera de aumentar el número de observaciones sin suponer independencia entre éstas (incluso, las 30 observaciones de cada año no son dependientes entre ellas, dado que el triunfo de alguna representa el fracaso de otras), lo que sí se podría realizar, sería minimizar el nivel de dependencia entre las observaciones lo mayor posible.

Lo que se propuso para aumentar el número de observaciones fue, en vez de utilizar la información de, dígase los últimos n años, se utilizarán observaciones de los últimos n años siempre y cuando los años sean números pares (o impares); es decir, se tomarán años no consecutivos.

Esto se realizó con la idea de reducir lo más posible el nivel de dependencia entre los mismos equipos de diferentes años, dado que al analizar la información de los equipos de manera longitudinal, se descubrió que del total de jugadores de un equipo que cumplían con las características de promediar como mínimo 20 minutos por partido y jugar como mínimo el 60 % de todos los partidos de la temporada (el porcentaje mencionado representa aproximadamente 50 partidos, la temporada consta de 82 juegos por equipo), cerca de un 45 % de ellos no militarían en ese equipo dentro de dos años, por lo que es muy probable que las estadísticas de ambos equipos no fueran tan similares, a comparación de las estadísticas del año contiguo donde aproximadamente un 23 % de los jugadores

del presente año no formarán parte del equipo durante el año siguiente.

Cabe mencionar que el tomar años no consecutivos no soluciona por completo el problema de dependencia entre observaciones, sino que lo minimiza. Para solucionar realmente dicho problema, se podrían utilizar modelos alternos como los llamados modelos lineales generalizados mixtos, que permiten correlación entre observaciones y por lo tanto se podrían utilizar observaciones de años consecutivos, o bien utilizar modelos para datos longitudinales; sin embargo, dichos modelos se encuentran fuera del alcance del presente trabajo, por lo que el tomar las observaciones de cada dos años será como se hará frente al problema de dependencia entre observaciones.

Ahora, uno podría argumentar que si en vez de utilizar las observaciones de cada dos años, se tomaran de cada tres o más años, entonces se podría prácticamente solucionar el problema de dependencia entre mismos equipos de diferentes años. Esta aseveración es cierta, sin embargo existe otro problema que más que del modelo en sí, pertenece al área a analizar (en este caso la NBA), y se le debe de prestar gran atención.

A diferencia de otros deportes o incluso otras ligas, la NBA se encuentra en una constante pero paulatina evolución. Por evolución de la NBA se refiere, más que a cambios en las reglas del juego (que suceden muy ocasionalmente), a la forma en que éste se juega.

Esto es de gran importancia para el modelo a analizar, dado que éste se basará en estadísticas que decidan el nivel de desempeño de un equipo; pero si gracias a esta evolución, una estadística que hace tiempo fuera fundamental para decidir el desempeño de un equipo, ya no lo fuera en la NBA moderna, entonces los resultados del modelo podrían dictar que esa variable es importante para el desempeño del equipo, cuando en el presente actual, el impacto de la estadística mencionada para el desempeño de un equipo podría incluso llegar a ser nulo, lo que generaría problemas al momento de realizar predicciones basadas en el modelo.

Un ejemplo de esta evolución es, por ejemplo, los tiros de tres puntos. Antes, los equipos con el mejor desempeño en la liga podían tener diferentes fortalezas, como un buen juego debajo de la canasta, un equipo que penetrara hacia la canasta o un equipo con buenos tiradores de tres puntos; hoy en día, todos los equipos con un desempeño alto en la liga poseen buenos tiradores de tres puntos, lo que significa que la importancia de estos tiros es mucho mayor hoy en día con respecto a antes.

Otro ejemplo importante de cómo ha evolucionado la liga tiene que ver con cierto tipo de jugadores. Existen cinco posiciones diferentes para los jugadores en el baloncesto. Estos son base, escolta, alero , ala-pivot y pivot, también llamada centro. Los jugadores que juegan la posición de centro son caracterizados

por su gran estatura y fuerza, así como su habilidad para atrapar rebotes y defender la canasta.

Ha resultado pues, que mientras en el pasado todo equipo importante poseía entre sus jugadores a un muy buen centro, ahora los equipos importantes sin un buen centro son más que los equipos importantes con uno. Y en el caso de los equipos importantes que aún poseen a un centro de calidad, este jugador ha cambiado su estilo de juego a comparación de los centros de hace algunos años. Mientras que antes rara vez algún centro poseía un buen tiro de media o larga distancia, ahora para que un jugador pueda ser considerado como un buen centro es un requisito contar con estas características.

Por estos motivos y otros más, se ha decidido que para obtener efectividad en el poder predictivo del modelo, se debe de contar con una cota inferior, un año en el que a partir del cual se pueda tomar información para utilizarse en el modelo. Es difícil saber con exactitud a partir de qué año fue que se hizo más notorio el cambio en el estilo de juego en la NBA dado que los cambios han sido paulatinos, y realizar un estudio para intentar encontrar la mejor cota requeriría de un lapso de tiempo invertido muy grande. Por este motivo, el autor del documento ha decidido fijar dicha cota inferior con base en su conocimiento del área en cuestión (la NBA), al año 2013. Por lo tanto, a partir de dicho año se podrá tomar información para el modelo.

Dado que ya se ha puesto todo sobre la mesa respecto a la elección de observaciones en el modelo, se concluye entonces que se utilizará la información de todos los equipos de los años 2013 y 2015 para realizar el modelo, haciendo que la muestra de entrenamiento se componga de 60 observaciones. Además, dado que uno de los dos objetivos principales del estudio es predecir el nivel de desempeño de equipos futuros, se utilizará la información de los 30 equipos del año 2017 para obtener el poder predictivo del modelo.

7.2.2. Variables predictoras

Una de las medidas más importantes a realizar para que un modelo sea efectivo es encontrar a las mejores variables predictoras posibles para el estudio, pues de esta manera los resultados del modelo serán lo más similar posible a la realidad. Ahora que se han definido las observaciones a utilizar en el modelo, se procederá a elegir el conjunto de variables candidatas a ser utilizadas en el modelo.

De entre todo el universo de estadísticas que pudieran fungir como variables predictoras del modelo, se intentó restringirse a variables que fueran lo más

sencillas posibles en términos de interpretación. Al mismo tiempo, se buscaron variables que no fueran funciones de otras estadísticas, y en caso de serlo, que fueran funciones de estadísticas que no fueran candidatas a ser utilizadas en el modelo.

El primer paso de esta búsqueda consistió en evaluar rápidamente todas las estadísticas encontradas dentro de la página oficial de la NBA, así como algunas otras creadas por el autor del documento. Dado que el número de variables aspirantes a ser utilizadas en el modelo era demasiado grande (superaba las 982, de las cuales aproximadamente 30 fueron creadas por el autor y las restantes obtenidas del sitio web oficial de la NBA), se prefirió un análisis informal univariado de la relación de las variables aspirantes con respecto a la variable de respuesta, que en este caso será la categoría en la que se clasificó al equipo.

Dicho análisis informal consistió en ordenar de forma ascendente o descendente a las observaciones a utilizar en el modelo de acuerdo a cierta variable, y una vez realizada dicha acción, verificar si las observaciones ordenadas se podrían ver agrupadas por las categorías de la variable de respuesta o no.

Un ejemplo de que lo anterior sucediera, sería si por ejemplo, la mayoría de las observaciones con los valores más grandes de la variable a analizar tuvieran como variable de respuesta el valor tres (valor que representa a los equipos con desempeño pobre), mientras que la mayoría de las observaciones con los valores más pequeños de la variable a analizar tuvieran como variable de respuesta el valor uno (valor que representa a los equipos con un desempeño alto).

Se vuelve a reiterar que aunque este no es un buen método para encontrar a las variables que mejor influyan en la variable de respuesta, sí fue de ayuda para desechar las variables que parecieran no tener relación alguna con la variable de respuesta.

El paso anterior se realizó con la finalidad de no destinar una gran cantidad de tiempo a analizar variables en las que es claro no tienen relación alguna con la variable de respuesta.

Una vez terminado el primer paso para descartar variables, el grupo de posibles variables predictoras se redujo a 66. El segundo filtro consistió (nuevamente un método de selección de variables informal) en obtener los promedios y desviaciones estándar de las variables para cada categoría de la variable de respuesta. Cabe mencionar que ninguna de las variables candidatas a ser predictoras era categórica.

Una vez obtenidos dichos promedios y desviaciones estándar por categoría, se crearon intervalos de la siguiente manera: el intervalo asociado a la categoría uno para alguna variable, consistirá en los valores que se encuentren entre el promedio de la variable para la categoría uno menos su respectiva desviación

estándar, y el promedio de dicha variable más su respectiva desviación estándar. Posteriormente, se analizaron detenidamente los tres intervalos obtenidos por variable y se seleccionaron a las variables en las que no se traslaparan sus intervalos de ninguna de las tres categorías. Nuevamente se repite que éste es un método muy informal, ya que ni siquiera se utilizó un nivel de confianza para dichos intervalos.

Una vez realizado el paso anterior, se realizó un análisis de correlación entre las variables seleccionadas para identificar variables candidatas que tuvieran una alta correlación entre ellas. Si alguna variable candidata no tuviera una correlación tan alta (entre -0.70 y 0.70) con alguna otra de las demás variables candidatas a ser explicativas, entonces esa variable seguiría siendo candidata a ser utilizada en el modelo; en cambio, si alguna variable tuviera una alta correlación (entre -1 y -0.70 ó 0.70 y 1) con alguna otra variable o grupo de variables candidatas, entonces dentro de este grupo de variables, la variable candidata que tuviera la mayor correlación con la variable de respuesta seguiría siendo candidata a ser utilizada en el modelo, mientras que las demás variables de dicho grupo se descartarían. Nuevamente se recuerda que este método fue muy informal ya que la correlación utilizada como método de elección de variables dentro de un conjunto de ellas con alta correlación fue la correlación de Pearson, que únicamente es correcta entre variables continuas mientras que la variable de respuesta es categórica, y además, únicamente identifica correlaciones de tipo lineal.

Lo realizado en el párrafo anterior tiene el objetivo de que el modelo evite problemas de multicolinealidad lo más posible, ya que si por ejemplo, hubiesen dos variables con correlación muy alta, el modelo no podría identificar el efecto individual verdadero de cada una de estas variables, y tomará a las dos variables como si fueran la misma.

Una vez terminado el proceso anterior, se habría encontrado el grupo de variables con las cuales se trabajará de ahora en adelante y que fueron mencionadas en el capítulo anterior, grupo constituido por 15 variables.

Cabe mencionar que aunque el proceso que se utilizó para encontrar a las 15 variables candidatas a ser utilizadas en el modelo no tuvo mayores dificultades técnicas, fue muy lento y se requirió de un gran número de horas para realizarlo debido a la enorme cantidad de variables a analizar.

Otra nota que es de gran importancia hacer mención, y que aplica generalmente para cualquier estudio que involucre la búsqueda de variables candidatas a ser utilizadas en un modelo de regresión logística multinomial, es que se debe de tener mucho cuidado en evitar tomar en cuenta posibles variables predictoras cuyo valor fuese resultado del valor de la variable de respuesta en la que se clasificó a la observación, pues si ese fuera el caso, no tendría utilidad alguna el modelo (¿cómo se podría predecir una variable de respuesta mediante variables

predictoras que dependen de ella misma?).

El autor del documento fue extremadamente minucioso en revisar que ninguno de los valores de las variables candidatas a ser predictoras fueran consecuencia del valor de su variable de respuesta.

7.2.3. Variable respuesta

Ahora que se han elegido las observaciones a ser utilizadas en el modelo y sus respectivas variables, en la mayoría de los casos se procedería directamente a la elaboración del modelo. Sin embargo, se revisará otro detalle antes de proceder a dicha elaboración. Este detalle es respecto a la variable de respuesta de las observaciones.

Se ha decidido utilizar la información de los equipos de los años 2013 y 2015 para elaborar el modelo; sin embargo, las reglas impuestas para decidir el valor de la variable de respuesta obligan a tener 11 observaciones en las categorías dos y tres, mientras que para la categoría uno dicho número disminuye a ocho. Esto ocasiona que de las 60 observaciones totales a utilizar en la elaboración del modelo, el número de observaciones dentro de las categorías uno, dos y tres sea 16, 22 y 22, respectivamente.

La diferencia entre el número de observaciones de la categoría uno con respecto a las categorías dos y tres, como se dijo en los capítulos uno a cuatro referentes a regresión logística multinomial, podrían generar errores en la predicción de observaciones dependiendo de la manera en que se decida clasificarlas, y podría ocasionar que algunas de las observaciones que debieran ser clasificadas en la categoría en la que se utilizó un número de observaciones menor a las demás para la elaboración del modelo (en este caso la categoría de equipos con desempeño alto, denotada por el número uno), sean clasificadas en las categorías en las que se utilizó un número mayor de observaciones (en este caso, las categorías dos y tres).

Este es un grave problema para el estudio, dado que para el investigador la categoría de la variable de respuesta de mayor importancia es la categoría uno, y dado que uno de los objetivos principales del estudio era predecir el desempeño de nuevos equipos, el que equipos con posible categoría uno sean clasificados en categorías con menor desempeño representaría un fracaso para el estudio. Debido a estos motivos, se ha tomado la decisión de intentar balancear lo mayor posible el número de observaciones dentro de cada categoría para la implementación del modelo y así tratar de evitar la ocurrencia de clasificaciones erróneas debido a disparidad en estos números de observaciones.

Una posible solución al problema anterior podría ser fijar el número de equipos de todos los años a diez equipos por categoría, de tal manera se tendrían para el estudio 20 observaciones por categoría (diez por cada año). Sin embargo, se optó por seguir utilizando la división anterior (originalmente ocho observaciones para la categoría uno y 11 para las otras dos categorías), pero realizando algunas modificaciones que dependan tanto del porcentaje de victorias como del resultado del equipo en la primera ronda *playoffs*.

Las medidas que se tomaron, junto a sus respectivas justificaciones para realizarlas, se mencionan a continuación:

- Los equipos que hayan perdido en la primera ronda de *playoffs* cuatro juegos a tres (las rondas de *playoffs* se juegan a ganar cuatro partidos), y además el equipo contrincante haya tenido una proporción de victorias dentro de la liga mayor a 0.70, se clasificarán en vez de en la categoría dos, en la categoría uno.

  Existe algo de lógica para volver a clasificar a los equipos que cuenten con las características anteriores. En primer lugar, si una serie se alarga hasta el séptimo juego significa que ambos equipos se mantuvieron casi al mismo nivel de juego a lo largo de la serie, y si el equipo contrincante tuvo una proporción de victorias tan alta como 0.70 (esta proporción es muy alta; por año, sólo dos equipos de los 30 de la liga llegan a rebasarla en promedio) o más, significa que el equipo perdedor estuvo a la altura de uno de los mejores equipos de la liga, y tal vez si hubiera jugado contra cualquier otro equipo en la primera ronda, pudiera haber salido victorioso y avanzado a la segunda ronda de *playoffs*, lo que implicaría obtener una categoría de uno en la variable de respuesta.

El argumento para utilizar esta nueva implementación consiste también en que, para el caso de las observaciones que fueron cambiadas de la categoría dos a uno, realmente el equipo estuvo muy cerca de vencer a su adversario. Esto hace pensar que tal vez incluso hubo otros factores externos al baloncesto que pudieron haber sido lo que le impidió al equipo haber alcanzado la segunda ronda, por lo que ambos equipos deberían merecerse pertenecer a la categoría uno.

Otro argumento utilizado para preferir el método descrito para asignar el valor de la variable de respuesta de esa manera y no mediante la clasificación de diez observaciones por categoría por año, es que según la experiencia del investigador referente a la NBA, existen años donde el número de equipos con desempeño alto es muy grande, mientras que en otros años dicho número es pequeño.

Por lo que si se utilizara el esquema de clasificación de diez observaciones por categoría por año, en el momento de utilizar diferentes años podría suceder que equipos con estadísticas muy buenas hayan sido clasificados en la categoría dos (debido a que fue un año con una gran cantidad de equipos fuertes), mientras

que equipos con estadísticas promedio o no tan buenas fueron clasificados en la categoría uno (debido a que ese año hubo muy pocos equipos fuertes). Estos casos, además de presentar incongruencias al ser interpretados, reducirían la efectividad del modelo, pues éste utilizará información de diferentes años.

Bajo el esquema de clasificación utilizado, además de los ocho equipos que ingresan de manera natural a la categoría uno (al avanzar a la segunda ronda de *playoffs*), únicamente se clasificarán en la categoría uno a los equipos fuertes de los años fuertes y que además quedaron muy cerca de avanzar a la segunda ronda de *playoffs*.

Terminado de explicar la nueva implementación al esquema de clasificación de la variable respuesta, las siguientes observaciones fueron cambiadas de la categoría dos a la uno:

1. 2013 Dallas Mavericks
2. 2013 Memphis Grizzlies

Esto da como resultado un total de 18 observaciones en la categoría uno, 20 en la categoría dos y 22 en la categoría tres para la implementación del modelo.

Es importante dejar en claro para el lector que, si el modelo de regresión logística multinomial no le diera mayor peso a las probabilidades de las categorías que poseen un mayor número de observaciones, estas medidas alternas nunca se hubieran implementado. Dadas las circunstancias, era necesario aumentar el número de observaciones en la categoría uno. Este simplemente fue el método que el autor del documento consideró mejor para preservar la efectividad del modelo y al mismo tiempo seguir conservando las interpretaciones que se les quiere dar a las categorías de la variable de respuesta.

7.2.4. Componentes principales en regresión logística multinomial

Hasta ahora ya se cuenta con todo lo necesario con respecto a información (observaciones y variables) para poder realizar una regresión logística multinomial con la cual se pueda medir el efecto de las variables del modelo en la variable de respuesta, así como predecir la categoría de desempeño de futuras observaciones.

Sólo queda un problema por resolver: se tomaron 15 posibles variables a utilizar en el modelo. Si todas a primera instancia parecen ser significativas ¿cuáles se deberían de utilizar? Y otra pregunta aún más interesante. Se sabe que si se

introducen P variables a un modelo, entonces para el modelo y el objeto de estudio, lo único que podrá influir en el resultado de la variable de respuesta serán las variables ingresadas en el modelo, así como sus interacciones, términos cuadráticos, etcétera. Todo lo demás no introducido en el modelo, se supone que no afecta a la variable de respuesta.

Se sabe que, en la vida diaria existen un sin fin de factores que influyen en el desempeño de un equipo de baloncesto, desde factores internos del baloncesto hasta externos como son los familiares. Ahora, ¿qué sucedería si el objetivo del estudio fuera crear un modelo lo más cercano posible a la realidad, dentro de las limitaciones que el investigador tuviera? Entonces, deberían de ingresarse muchas más variables al modelo.

Esta tarea sería complicadísima de realizar, debido a que además de encontrar a las variables, se debería de tener un número de observaciones lo suficientemente grande como para que no se sobreajuste el modelo y no arroje resultados incoherentes (como errores estándar enormes para los coeficientes estimados). Además de esto, se requeriría que todas las variables fueran no correlacionadas entre ellas, lo que complicaría aún más el trabajo.

Sin embargo, aunque al autor del documento le fue imposible encontrar información exterior al tema de baloncesto para cada equipo que pudiera afectar su respectivo desempeño, se tiene interés en saber cómo sería el comportamiento de todas las variables que se encontraron (las 15 finales) si se utilizaran de manera conjunta en un modelo. La justificación por lo que se quisiera realizar un modelo con todas las variables posibles es la siguiente:

El modelo está pensado para ser de utilidad en el área gerencial de los equipos de baloncesto; es decir, la intención hipotética del modelo sería que los *general managers* de los equipos lo utilicen para reforzar su plantilla lo más inteligente posible y sin la necesidad de invertir mucho dinero. En otras palabras, se busca que en vez de que los *general managers* busquen jugadores por simple talento o popularidad o porque tienen potencial, busquen jugadores que les ayuden a completar las estadísticas necesarias para que el equipo pueda llegar a tener un desempeño alto según el modelo implementado.

Es por esto que se busca un modelo con el mayor número de variables posibles. Mientras más variables se tengan en las cuales concentrar la atención de los *general managers*, mayores opciones podrán tener para formar un equipo competitivo. Por ejemplo, si según el modelo se debiera de aumentar el porcentaje de tiros de tres puntos, pero los únicos agentes libres (es decir, jugadores aún sin contrato con algún equipo) con un buen porcentaje de tiros de tres fueran caros, el equipo tendría problemas financieros para contratarlos; si en cambio, según el modelo se debe de aumentar el porcentaje de tiros de tres, o bien el número de jugadores con experiencia en semifinales, o bien el porcentaje de tiros libres, entonces sería más probable contratar jugadores que brinden las

estadísticas necesarias para un desempeño alto sin sacrificar tanto las finanzas del equipo.

Es por este motivo que es atractivo para el autor del documento encontrar un buen modelo con la mayor cantidad de variables posibles. Pero si no es posible aumentar el número de observaciones (como se vio en este caso), entonces los resultados del modelo tendrían muy poca confiabilidad. Lo que se debe de realizar es ingresar un número pequeño de variables al modelo, pero que estas variables contengan la información de todas las variables que se quisieran ingresar (en este caso, las 15 explicadas en el capítulo anterior). Entra entonces el análisis de componentes principales.

La técnica de componentes principales aplica únicamente para variables continuas o discretas con ciertas características. En este análisis sólo se utilizarán variables de esta índole, por lo que es factible utilizar dicha técnica en el estudio.

Además, el utilizar componentes principales como variables predictoras para el modelo de regresión logística multinomial garantizaría el cumplimiento de uno de los supuestos más difíciles de lograr que es la no correlación entre variables, implicando así ausencia de multicolinealidad en el modelo.

Por lo tanto, si se eligiera un conjunto de Q componentes principales para ser utilizadas como variables predictoras del modelo, con Q mucho menor a P, entonces el modelo tendría un número pequeño de variables predictoras y los resultados del modelo podrían ser más confiables y, al mismo tiempo, las Q componentes principales elegidas conservarían parte de la información de las P variables originales.

Por supuesto, el realizar dicha técnica dentro del modelo de regresión logística multinomial implicaría cambios en las interpretaciones explicadas en el capítulo cuatro, es por esto que en todo momento se llevará de la mano al lector para intentar esclarecer lo más posible dichos cambios.

7.3. Elaboración del modelo

Ahora que se ha definido todo lo necesario y se han mostrado las intenciones del estudio, se procederá a realizar el análisis del desempeño de los equipos. Para realizar estos análisis, se presentarán dos modelos finales. El primero será el mejor modelo encontrado utilizando componentes principales como variables predictoras, y el segundo será el mejor modelo encontrado utilizando las varia-

bles originales como predictoras. El único motivo para presentar ambos modelos es comparar sus desempeños para concluir si el utilizar componentes principales disminuye la eficacia del modelo o no, así como mostrar sus respectivas ventajas y desventajas.

El primer paso para el desarrollo de los modelos será obtener los patrones de covariables a utilizar y la cantidad de observaciones dentro de cada uno de ellos; sin embargo, al utilizarse únicamente variables continuas en cualquiera de los dos modelos (el de componentes principales como predictores y el de las variables originales) salvo las estadísticas dos y tres que son contadores, el número de patrones a utilizar será igual al número de observaciones y se tratarán a las observaciones como si se distribuyeran de manera *N-asintótica*.

Una alternativa para poder haber tratado al conjunto de observaciones como si pertenecieran a una distribución *M-asintótica* hubiera sido el categorizar a las variables continuas, ya que de esa manera se le podría haber dado credibilidad a los *p-values* de las estadísticas de resumen Ji-cuadrada de Pearson y Devianza; sin embargo, dado que únicamente se contaba con 60 observaciones para la implementación del modelo, se hubieran tenido que realizar muy pocas particiones en las variables para evitar categorías con poca densidad de observaciones, y el hecho de contar con pocas particiones por variable pudiera haber causado una pérdida de información significativa en las variables. Por esta decisión, se optó por no categorizar las variables.

Una vez definidos los patrones de covariables, se procedería a obtener el mejor modelo encontrado utilizando algún método como el *stepwise*. A continuación se vuelven a revisar los dos objetivos de este estudio, que son:

1. Predecir el nivel de desempeño de futuros equipos.
2. Determinar los factores que influyen en el desempeño de los equipos de la NBA y su grado de influencia en ellos.

Con el segundo punto como el de mayor interés para el autor.

El primer punto se logra al encontrar un modelo que pueda pronosticar correctamente la categoría de desempeño tanto de las observaciones con las que se elaboró el modelo, como de observaciones futuras y ajenas a dicha elaboración. El segundo punto, en cambio, se logra al encontrar un modelo robusto, con buen ajuste y además en el que todas sus variables sean significativas, esto con el fin de brindar mayor veracidad a las inferencias realizadas utilizando los coeficientes de dichas variables, inferencias que serán la base para cumplir con el objetivo dos.

Por lo tanto, si un modelo tuviera las mejores estadísticas de ajuste, esto no sería un motivo suficiente para argumentar que dicho modelo es el mejor para el estudio en cuestión, pues el que un modelo se ajuste bien a los datos no implica

que clasifique correctamente observaciones ajenas a las utilizadas en su elaboración. Un ejemplo claro de esto es lo que sucede al sobreajustar un modelo. Aunque la mayoría de las veces el ajuste sea perfecto, las predicciones de nuevos datos fallan de manera garrafal.

Es por esta razón que el autor del documento ha decidido en vez de utilizar un método tradicional como el método *stepwise* para intentar encontrar al mejor modelo, optar por utilizar un método propio, con restricciones adaptadas a los objetivos particulares del estudio, para intentar encontrar al mejor modelo posible que cumpla ambos objetivos.

A continuación se explicará de manera muy coloquial el funcionamiento de dicho método, aunque si se quisiera profundizar en dicho asunto, se puede consultar el código utilizado para este método dentro de los apéndices del documento.

Dado que son dos objetivos muy diferentes los del estudio, y con el fin de hacer más probable que los modelos encontrados cumplan con éstos de manera satisfactoria, el método de búsqueda se basó en encontrar modelos que cumplan con ciertas características. Para los modelos que utilizan componentes principales como variables predictoras, las características que se buscaron en los modelos fueron:

- Poder predictivo para observaciones de entrenamiento mayor o igual a 85 %.
- Poder predictivo para observaciones de validación mayor o igual a 85 %.
- Significancia en todas las variables predictoras menor o igual a 0.05 bajo prueba de Wald.
- Significancia en todas las variables predictoras menor o igual a 0.05 bajo pruebas de cocientes de verosimilitud.

Mientras que para modelos que utilizan a variables originales como variables predictoras, las mismas características fueron buscadas y además se añadió una nueva, que es:

- Correlación entre variables no mayor a 0.6 o menor a -0.6.

Mientras que los primeros dos puntos intentarán cubrir el objetivo uno del estudio, el tercer y cuarto punto intentarán cubrir el segundo objetivo. El punto referente a correlación también funciona como auxiliar para lograr el objetivo dos, ya que al no utilizar correlaciones grandes se intenta evitar a toda costa el problema de multicolinealidad entre variables predictoras.

Recuérdese que la teoría de análisis de componentes principales respalda que dichas componentes son independientes entre sí, por lo que no hace falta incluir el requisito de correlación para los modelos que utilizan dicha técnica.

El motivo por el cual se decide utilizar modelos que cumplan con significancia de sus variables a un nivel 0.05 tanto para la prueba de Wald como para cocientes de verosimilitud es, que aunque se menciona en la sección 4 del capítulo 1 que diversos investigadores han recomendado ampliamente el utilizar la prueba del cociente de verosimilitud sobre la prueba de Wald, el encontrar un modelo en el que ambas pruebas coincidan bajo el mismo nivel de significancia proyectará una mayor robustez al modelo en cuestión. Esto a su vez implica que se puede tener mayor seguridad, al realizar inferencias respecto a las variables, de que los resultados de dichas inferencias son de fiar; es decir, son aproximadas a la realidad.

Así pues, el algoritmo utilizado para la búsqueda del mejor modelo, para el caso en que se utilizaron componentes principales como variables predictoras, fue el siguiente:

1. Antes de calcular las componentes principales de las 15 variables, se analizó la posibilidad de incluir interacciones entre las variables siempre y cuando la interpretación de dichas interacciones tuviera sentido. Las interacciones elegidas a ser incluidas en el modelo fueron:

   - Proporción histórica de victorias del coach, número de jugadores con experiencia en semifinales y número de jugadores con tres años o más de antigüedad (interacción de segundo orden).
   - Minutos de veteranos y proporción histórica de victorias del coach.
   - Minutos de veteranos y número de jugadores con experiencia en semifinales.
   - Minutos de veteranos y número de jugadores con tres años o más de antigüedad.

   Cabe recordar que la interacción entre variables fue de manera multiplicativa, y si se mencionaron anteriormente interacciones de segundo orden, entonces también se tomaron en cuenta todas las interacciones de orden inferior entre las mismas variables.

2. Una vez calculadas las interacciones mencionadas, se procedió a realizar el análisis de componentes principales entre las 22 variables (las 15 originales más las siete interacciones).

   Es importante mencionar que para la realización de este estudio, no se supusieron cambios en los promedios ni en las desviaciones estándar de las observaciones de diferentes años, por lo que la estandarización para llevar a cabo la técnica de componentes principales fue realizada con los mismos parámetros tanto para las observaciones del año 2013 como las del 2015.

3. El siguiente paso consistió en realizar un análisis exploratorio de las 22 componentes, mismo que consistió en graficar, por cada componente principal, la distribución de sus valores por nivel de referencia; en otras palabras, se obtuvieron tres curvas de densidad por cada componente principal.

   Una vez realizado esto para cada componente principal, se analizó con detalle el conjunto de gráficas obtenido, y de éste se eligió a un subconjunto de componentes las cuales presentaron la mayor diferencia posible entre sus respectivas curvas de densidad.

4. A continuación se realizó un algoritmo cuyo propósito fue calcular todos los posibles modelos de regresión logística multinomial utilizando como variables predictoras a todos los posibles subconjuntos del conjunto de componentes principales elegidos durante el análisis exploratorio de datos, y mostrando como *output* o salida, a los poderes predictivos tanto de las observaciones de entrenamiento como las de validación. La categoría de referencia utilizada en los modelos para la variable de respuesta fue la categoría asociada al nivel de desempeño alto; es decir, la denotada por el número uno.

   El motivo por el cual se prefirió un algoritmo que calcule todas las posibles combinaciones de componentes con alto potencial de ser significativas fue que, realmente no se cuentan con interpretaciones adecuadas para las componentes principales calculadas debido al gran número de variables que las conforman, lo único de lo que se tendría idea sería de las varianzas de cada componente. Sin embargo, el que la magnitud de una varianza para una componente sea pequeña o grande no tiene relación alguna con el hecho de que la componente se encuentre relacionada con la variable de respuesta o no. Podría ser incluso, que alguna de las componentes con mayor varianza no sea de ayuda para clasificar a las observaciones mientras la componente con menor varianza de todas sí lo fuera.

   Por estos motivos, se decidió utilizar un programa que calcule todas las posibles combinaciones de variables predictoras con sus respectivos poderes predictivos.

   El código realizado está programado para calcular un total de:

   $$\sum_{i=1}^{Q} \binom{Q}{i}$$

   Regresiones logísticas multinomiales, con Q el número de componentes principales consideradas de importancia en el análisis exploratorio de datos.

   Posteriormente, se tomó el conjunto de modelos que cumplieran con la res-

tricción de poder predictivo impuesta (85 % para bases de entrenamiento y validación).

5. Para cada modelo dentro del conjunto obtenido, se verificará si cumple o no con los requisitos de significancia en todas sus variables a un valor menor o igual a 0.05 utilizando tanto la prueba de Wald, como las pruebas de cocientes de verosimilitud. Si el modelo llegara a cumplir con dichos requisitos, entonces ingresará al grupo de modelos finales.

6. Una vez obtenidos todos los modelos que cumplieran con todos los requisitos tanto de significancia de variables como de poder predictivo, se tomará al modelo que posea el mejor desempeño en los cuatro requisitos y se utilizará como el modelo final para componentes principales.

El caso para el modelo que utilizará a variables originales como variables predictoras sigue el mismo procedimiento que el modelo que utiliza componentes principales, con las siguientes excepciones:

- Aunque en este caso las combinaciones se harán sobre las variables originales más las interacciones, no se podrá aceptar en el grupo final de modelos a alguno que tenga como variable predictora a una interacción y que a la vez no contenga por sí solas a todas las variables que la conforman.

- Las correlaciones de las variables a ser utilizadas en el modelo final no pueden ser ni mayores a 0.6, ni menores a -0.6. Si este llegara a ser el caso, se tomará al siguiente modelo con los mejores valores para los primeros cuatro requisitos que cumpla además con el requisito referente a correlación entre variables.

A continuación se procedió a realizar el algoritmo anterior. Como conclusión, los resultados de dicho algoritmo fueron muy diferentes a los esperados por el autor: ningún modelo, tanto para el caso de componentes principales como variables predictoras como el de variables originales como predictoras, cumplió con todos los requisitos señalados.

Dada esta situación, se optó por reducir el nivel de rigidez en algunos de los requisitos de la siguiente manera. El mínimo poder predictivo de los modelos candidatos a ser elegidos como el modelo oficial se redujo de 85 %, a 80 % tanto para las observaciones de entrenamiento como para las de validación.

El motivo por el cual se optó por reducir la rigidez en el poder predictivo del modelo y no en la significancia de sus variables es que, en cuanto a predicción, unos cuantos puntos más o menos no marcarían una diferencia importante dado que se contempla que incluso dicha mal clasificación pudiera ser producto de factores externos para algunas observaciones. Mientras tanto, unos puntos menos en significancia de variables sí podrían llegar a marcar una diferencia importante en cuanto a robustez del modelo, ya que uno de los principales objetivos es realizar inferencias respecto a los coeficientes de las variables, por lo

que es de gran importancia que todas las variables sean significativas.

El motivo por el cual se debe de garantizar significancia en todas las variables explicativas del modelo que utiliza componentes principales es debido a que el hecho de que una sola componente no sea significativa tendría un impacto global en todas las estadísticas originales, y no se podrían realizar inferencias respecto a ninguna variable original.

Al reducir pues, el poder predictivo para ambos tipos de observaciones y ambos tipos de modelos de 85 % a 80 %, se llega a obtener únicamente un solo modelo bajo componentes principales que cumple con los cuatro requisitos, mientras que para el modelo bajo variables originales, el número de modelos que cumplen con los cinco requisitos (los cuatro originales más el requisito de correlación) también fue uno. Por lo tanto, se tomarán ambos modelos como los "modelos finales", y se procederá a analizarlos con mayor detalle.

A continuación se mostrarán los resultados obtenidos mediante estos algoritmos. Cabe mencionar que para el caso de la prueba de Wald, se tomó el estadístico de Wald con menor significancia en cada variable predictiva (recuérdese que cada variable predictiva tiene asociados $J - 1 = 2$ coeficientes) para verificar su correspondiente significancia en el modelo.

- Características del modelo propuesto por el autor para el caso en que se utilizaron componentes principales como predictores:

  - Componentes principales a utilizar: Componentes 1, 2, 3 y 15.
    Como nota a mencionar, el modelo que incluía a las primeras cuatro componentes no logró cumplir con alguno de los requisitos de ajuste o predicción y por lo tanto fue descartado. Así también, la varianza que explican las componentes 1, 2, 3 y 15 (55.56 %, 8.65 %, 7.41 % y 0.38 %, respectivamente) representa el 67.01 % del total de varianza de las variables.

  - Curvas de densidad de las componentes elegidas: Figura 7.1

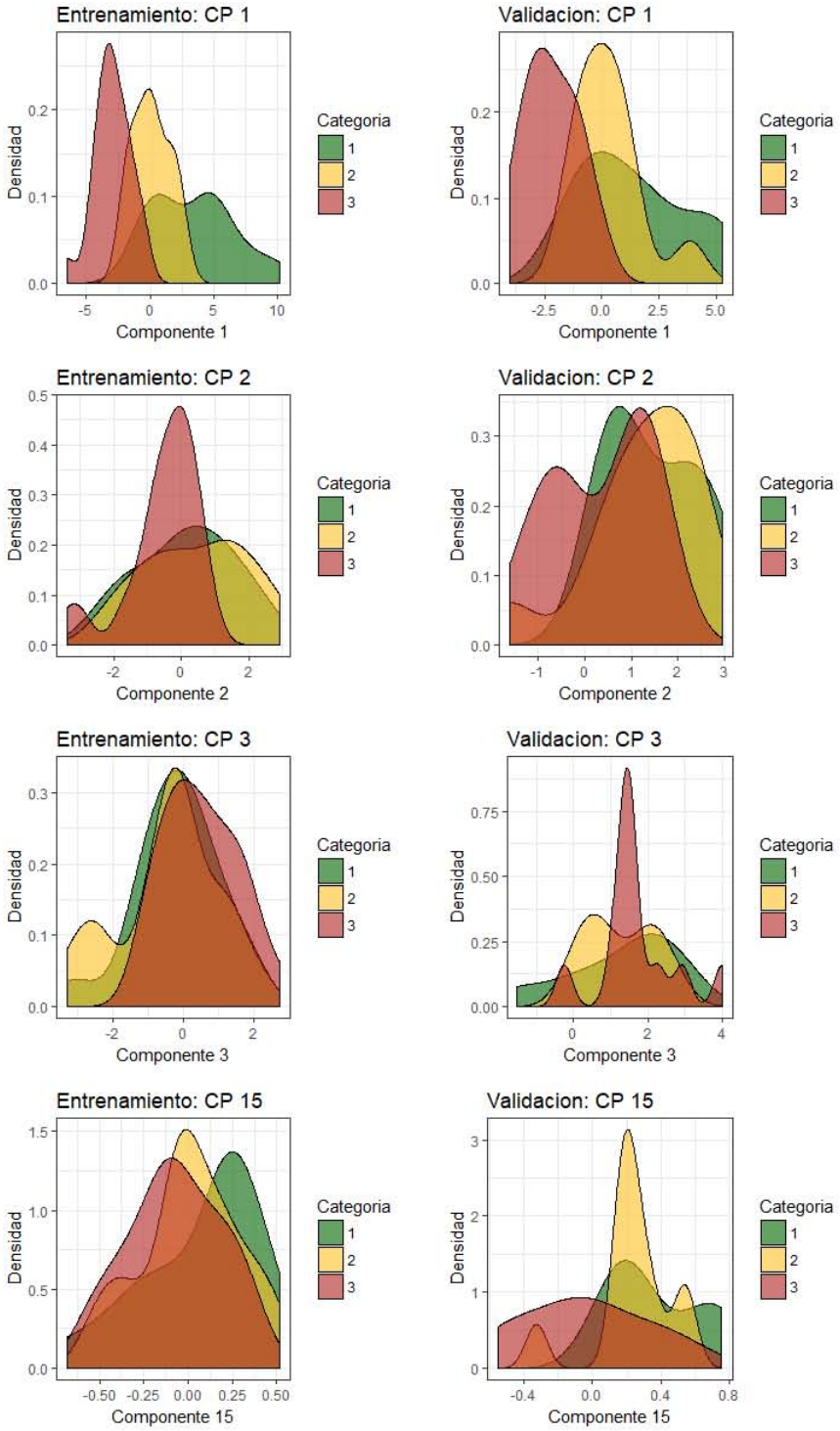

Figura 7.1: Curvas de densidad para el modelo con componentes principales para bases de datos de entrenamiento y validación.

- Coeficientes y errores estándar estimados: Tablas 7.1 y 7.2

Tabla 7.1: Coeficientes estimados del modelo que utiliza componentes principales (CP).

Nivel de Respuesta	Intercepto	CP 1	CP 2	CP 3	CP 15
Categoría 2	3.0096	-1.2855	-0.9054	0.5912	-5.7493
Categoría 3	-0.0889	-4.9321	-4.1828	2.7687	-8.0436

Tabla 7.2: Errores estándar estimados del modelo que utiliza componentes principales (CP).

Nivel de Respuesta	Intercepto	CP 1	CP 2	CP 3	CP 15
Categoría 2	1.0673	0.4003	0.4255	0.4563	2.6098
Categoría 3	1.8075	1.8477	1.8414	1.1476	4.2300

- Poder predictivo de las observaciones de entrenamiento: 51 de 60 observaciones clasificadas correctamente (85 %).
- Poder predictivo de las observaciones de validación: 24 de 30 observaciones clasificadas correctamente (80 %).
- Resultados pruebas de Wald: Tabla 7.3
  Mayor *p-value* para marcar la significancia de una variable: 0.0278
  Significancia en el que todas las variables son significativas: 0.03

Tabla 7.3: Valores de Wald para el modelo que utiliza componentes principales (CP).

Nivel de Respuesta	Intercepto	CP 1	CP 2	CP 3	CP 15
Categoría 2	2.8199	-3.2112	-2.1277	1.2958	-2.2029
Categoría 3	-0.0492	-2.6694	-2.2715	2.4126	-1.9015

- Resultados de las pruebas de cocientes de verosimilitud:
  De los 14 modelos posibles que únicamente contuvieron a alguna o algunas de las cuatro componentes elegidas (es decir, existen 14 modelos que utilizan únicamente como predictores a una, dos o tres de las cuatro componentes principales anteriores), en todos los modelos se rechazó por lo menos a un 0.03 de significancia la hipótesis nula referente a que el modelo completo (el modelo con las cuatro componentes principales) sea similar a estos modelos.

  El motivo por el que se decide calcular de esta manera a las pruebas de cocientes de verosimilitud es, que dicho proceso es la única manera que asegura que toda variable ingresada en el modelo sea

estrictamente significativa.

Por lo tanto, se considerarán a todas las componentes elegidas como significativas para motivos de este estudio.

Mayor *p-value* para marcar la significancia o no significancia de una variable: 0.0260
Significancia en el que todas las variables son significativas: 0.03

- Características del modelo propuesto por el autor para el caso en que se utilizan a las variables originales como predictoras:

  - Variables a utilizar: Variables número 2, 5 y 9, que corresponden a:

    - Número de jugadores con experiencia en semifinales
    - Porcentaje de tiros de campo efectivo
    - Puntos del oponente

    Respectivamente.

  - Curvas de densidad de las variables elegidas: Figura 7.2

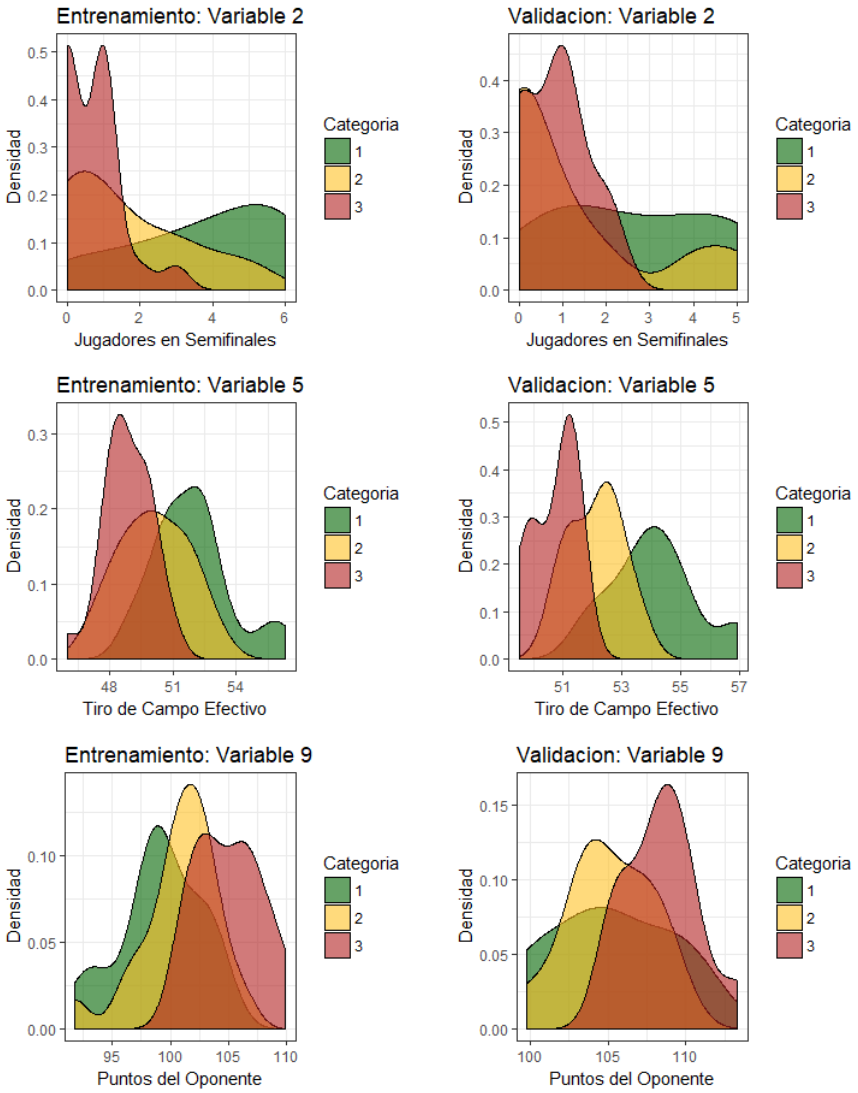

Figura 7.2: Curvas de densidad para el modelo con variables originales para bases de datos de entrenamiento y validación.

- Coeficientes y errores estándar estimados: Tablas 7.4 y 7.5

Tabla 7.4: Coeficientes estimados del modelo que utiliza variables originales como variables predictoras.

Nivel de Respuesta	Intercepto	Variable 2	Variable 5	Variable 9
Categoría 2	10.8549	-0.4503	-0.8254	0.3234
Categoría 3	19.8568	-1.5296	-2.7602	1.1758

Tabla 7.5: Errores estándar estimados del modelo que utiliza variables originales como variables predictoras.

Nivel de Respuesta	Intercepto	Variable 2	Variable 5	Variable 9
Categoría 2	13.4803	0.2403	0.3402	0.1610
Categoría 3	3.6609	0.6314	0.7082	0.3442

- Correlaciones entre las variables elegidas: Tabla 7.6. En ésta se observan correlaciones no mayores a 0.5 ni menores a -0.5.

Tabla 7.6: Matriz de correlaciones de las variables elegidas.

	Variable 2	Variable 5	Variable 9
Variable 2	1	0.4648	-0.4003
Variable 5	0.4648	1	-0.1662
Variable 9	-0.4003	-0.1662	1

- Poder predictivo de las observaciones de entrenamiento: 50 de 60 observaciones clasificadas correctamente (83 %).
- Poder predictivo de las observaciones de validación: 26 de 30 observaciones clasificadas correctamente (87 %).
- Resultados pruebas de Wald: Tabla 7.7
  Mayor *p-value* para marcar la significancia de una variable: 0.0156
  Significancia en el que todas las variables son significativas: 0.02

Tabla 7.7: Valores de Wald para el modelo que utiliza variables originales como variables predictoras.

Nivel de Respuesta	Intercepto	Variable 2	Variable 5	Variable 9
Categoría 2	0.8052	-1.8740	-2.4265	2.0084
Categoría 3	5.4240	-2.4226	-3.8973	3.4163

- Resultados de las pruebas de cocientes de verosimilitud:
  De los seis modelos posibles que únicamente contienen a alguna o algunas de las tres variables elegidas, en todos los modelos se rechazó por lo menos a un 0.01 de significancia la hipótesis nula referente a

que el modelo completo (con las tres variables) sea similar a estos modelos. Por lo tanto, se considerarán a todas las variables elegidas como significativas para motivos de este estudio.
Mayor *p-value* para marcar la significancia o no significancia de una variable: 0.0070
Significancia en el que todas las variables son significativas: Menor 0.01

A continuación se presentan algunas conclusiones de la búsqueda y elaboración de los modelos:

1. Por lo general, los modelos que utilizan variables originales como predictores tienden a ajustarse mejor a los datos que los modelos con componentes principales; sin embargo, el desempeño de los modelos con componentes principales no dista mucho del de los modelos en los que se utilizaron variables.

2. Ningún modelo con más de siete variables predictoras presentó significancia en todas sus variables a un 0.05 bajo la prueba de Wald o bajo pruebas de cocientes de verosimilitud.

3. De entre todos los modelos con seis o menos componentes principales como variables predictoras, únicamente 15 tuvieron un poder predictivo mayor a 80 % en ambas bases de datos (de entrenamiento y de validación). Dado que se calcularon una gran cantidad de modelos, se concluye que es realmente complicado encontrar modelos con ambos poderes predictivos altos.

4. De los 15 modelos mencionados anteriormente, sólo dos de ellos obtuvieron una significancia menor o igual al 0.05 para todas sus variables bajo la prueba de Wald; bajo pruebas de cocientes de verosimilitud, cinco de los 15 modelos obtuvieron dicha significancia. Únicamente un modelo tuvo a todas sus variables significativas bajo ambas pruebas, por lo que se concluye que es bastante complicado que un modelo garantice significancia de sus variables bajo ambas pruebas.

5. Respecto a los modelos que utilizan variables originales como variables predictoras, nuevamente se encontraron 15 modelos que cumplieran con los requisitos de poder predictivo y correlación, aunque de estos 15 modelos, únicamente tres obtuvieron la significancia deseada bajo pruebas de Wald, y solo uno bajo pruebas de cocientes de verosimilitud. Nuevamente solo un modelo tuvo a todas sus variables significativas bajo ambas pruebas, modelo que será el presentado.

   Con el propósito de evitar alargar el documento, se decidió por únicamente presentar a detalle los dos modelos que cumplieron con todos los requisitos de poder predictivo y significancia de variables, dejando a un lado a los demás.

6. Gracias al análisis exploratorio de datos realizado, se concluye que es muy probable que la distribución de las variables cambie conforme pasan los años, hecho que reduce la eficacia de los modelos presentados y principalmente sobre el que utiliza componentes principales, ya que para obtener las componentes principales del año 2017 se utilizaron los promedios y desviaciones estándar de los años 2013 y 2015.

7. Como conclusión final, encontrar un modelo con un buen desempeño tanto para validación como para ajuste tiende a ser un proceso muy complicado y con escasas opciones.

Es de vital importancia que, una vez elegidos los modelos, se prueben sus capacidades de ajuste mediante pruebas formales. En caso de que alguno de los dos modelos obtenga resultados positivos para dichas pruebas, se presentará al modelo como el modelo final. Si ambos modelos presentaran resultados positivos, se tomará como modelo final al que mejor se apegue a los objetivos del estudio. En cambio, si ambos modelos llegaran a fallar en algunas pruebas de ajuste, se optará por buscar algún otro modelo que sí las apruebe.

7.3.1. Comparación entre el método utilizado para encontrar modelos y métodos tradicionales

Con el motivo de brindar otra justificación del por qué la necesidad de utilizar un algoritmo como el que se implementó, se calcularon también modelos mediante métodos tradicionales con el fin de comparar resultados. El algoritmo tradicional consistió en primero utilizar la técnica *best subsets* para encontrar al mejor conjunto de variables, y posteriormente sobre ese conjunto se aplicó el método *stepwise* con criterio de Akaike para elección del modelo.

A continuación se compararán las propiedades principales de los dos modelos propuestos por el autor y los obtenidos por el método tradicional mencionado en el párrafo anterior. Cabe mencionar que las componentes, así como las variables a utilizar, serán mencionadas por orden de importancia para el caso de los modelos tradicionales según los resultados de *best subsets*.

Comparación entre métodos para modelos que utilizan componentes principales como variables predictoras:

Propiedades Principales	Método Propuesto por el Autor	Métodos Tradicionales
Componentes a Utilizar	1, 2, 3 y 15	1, 3, 2, 15 y 9
Significancia Máxima Verosimilitud	0.03	0.11
Significancia Wald	0.03	0.67
Poder Predictivo Entrenamiento	85 %	88 %
Poder Predictivo Validación	80 %	83 %

Comparación entre métodos para modelos que utilizan variables originales como variables predictoras:

Propiedades Principales	Método Propuesto por el Autor	Métodos Tradicionales
Variables a Utilizar	2, 5 y 9	2, 9, 6 y 13
Significancia Máxima Verosimilitud	0.01	0.05
Significancia Wald	0.02	0.01
Poder Predictivo Entrenamiento	83 %	92 %
Poder Predictivo Validación	87 %	80 %

Al comparar ambos métodos para las diferentes variables predictivas (componentes principales o variables originales) se ha encontrado que no existe un método que supere al otro, ya que cada uno presenta bondades así como imperfecciones.

El método tradicional, por ejemplo, llegó a mostrar en tres de cuatro ocasiones un mayor poder predictivo que el método sugerido por el autor. Sin embargo, también en tres de cuatro veces, la significancia de sus variables ya sea bajo la prueba de Wald o bajo la prueba de cocientes de verosimilitud fue mayor a las del método propuesto por el autor, y para el caso de los modelos con componentes principales, la diferencia entre las significancias de ambos métodos fue notoria.

Otra observación de interés, es que en general las variables utilizadas en el método propuesto por el autor fueron subconjunto de las utilizadas por el método tradicional, haciendo pensar que los modelos obtenidos mediante uno y otro

método no son tan diferentes entre ellos.

Sin embargo, se optó por utilizar los modelos obtenidos mediante el método propuesto por el autor dado que, al ser de vital importancia para el estudio las inferencias sobre los coeficientes, este método fue el que proporcionó las menores significancias para los mismos. No obstante, es importante enfatizar que la decisión de elegir dicho método fue personal, y en ningún momento se afirmó que el método propuesto fuese mejor que el tradicional.

7.4. Ajuste y predicción del modelo

Dado que se han elegido a priori los modelos que describirán el comportamiento de los equipos y con los cuales se realizarán predicciones, se procede ahora a evaluar el ajuste de ambos mediante pruebas formales.

7.4.1. Estadísticas resúmen

Ji-cuadrada de Pearson y Devianza

Como se había mencionado en el capítulo tres, cuando $M \approx N$ la distribución es de clase N-asintótica (es decir, al aumentar la muestra de observaciones, se incrementa el número de patrones de covariables y por ende los grados de libertad asociados a las distribuciones). Se había mencionado también que, cuando $M \approx N$, los *p-values* de χ^2 y D, así como su distribución, serían incorrectos.

Dado este problema y que además no se decidió categorizar a las variables predictoras continuas de los modelos, entonces tanto los *p-values* de χ^2 como los de D no serán de fiar para ningún modelo utilizado en este estudio al distribuirse de manera N-asintótica, y por lo tanto, no se presentarán.

Estadística de Hosmer-Lemeshow

Ahora se procede a calcular la estadística de Hosmer-Lemeshow para ambos modelos, estadística que es de fiar independientemente de la distribución de las observaciones (M-asintótica o N-asintótica).

Para el modelo con componentes principales como variables predictoras se tiene:

1. Estadística de Hosmer-Lemeshow: 6.0216
   Grupos utilizados: 10
   Grados de libertad: 16
   p-value: 0.9879

Mientras que para el modelo con variables originales como variables predictoras se tiene:

1. Estadística de Hosmer-Lemeshow: 18.119
   Grupos utilizados: 10
   Grados de libertad: 16
   p-value: 0.3170

Se optó por utilizar diez grupos dado que el número de observaciones no es tan pequeño. Si en vez de 60 observaciones se hubieran utilizado 30, el número de grupos a utilizarse hubiera sido menor.

El *p-value* resulta mayor a 0.05 para ambos modelos, por lo que no se rechaza la hipótesis nula que señala que el modelo ajusta bien a los datos.

Tablas de clasificación

Se había mencionado en la sección 2.4 del capítulo 3, que las tablas de clasificación no están relacionadas con el ajuste del modelo; sin embargo, dado que uno de los dos objetivos principales del análisis es predecir el nivel de desempeño de los equipos, se le dará importancia a esta herramienta.

A continuación se mostrarán las tablas de clasificación correspondientes tanto para el modelo con componentes principales como para el que utiliza variables originales como predictores. Cabe mencionar que el método del cálculo de las tablas consistió en clasificar a las observaciones en la categoría que tuviera la mayor probabilidad de pertenecer, sin tomar en cuenta ninguna otra restricción.

Además de la tabla de clasificación, se mostrará otra tabla donde se muestran los porcentajes de las observaciones clasificadas correctamente.

Para el modelo con componentes principales como variables predictoras se tiene:

1. Tabla de clasificación: Tabla 7.8
   Poder predictivo de las observaciones de entrenamiento: 85 %, correspondiente a 51 observaciones de 60 clasificadas correctamente.

Tabla 7.8: Tabla de clasificación: componentes principales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	14	4	0
	$Y = 2$	4	16	0
	$Y = 3$	0	1	21

2. Tabla de porcentajes de clasificación: Tabla 7.9
   Porcentaje de clasificaciones correctas para las categorías 1,2 y 3: 78 %, 80 % y 95 %, respectivamente.

Tabla 7.9: Tabla de porcentajes de clasificación: componentes principales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	78 %	22 %	0 %
	$Y = 2$	20 %	80 %	0 %
	$Y = 3$	0 %	5 %	95 %

Ahora, para el modelo que utiliza variables originales como variables predictoras se obtuvo lo siguiente:

1. Tabla de clasificación: Tabla 7.10
   Poder predictivo de entrenamiento: 83 %, correspondiente a 50 de 60 observaciones clasificadas correctamente.

Tabla 7.10: Tabla de clasificación: variables originales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	14	4	0
	$Y = 2$	3	15	2
	$Y = 3$	0	1	21

2. Tabla de porcentajes de clasificación: Tabla 7.11
   Porcentaje de clasificaciones correctas para las categorías 1,2 y 3: 78 %, 75 % y 95 %, respectivamente.

Tabla 7.11: Tabla de porcentajes de clasificación: variables originales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	78 %	22 %	0 %
	$Y = 2$	15 %	75 %	10 %
	$Y = 3$	0 %	5 %	95 %

Se puede analizar que bajo ambos modelos, las tablas de clasificación se comportan de una manera muy similar, lo que hace pensar que tal vez ambos modelos tuvieron problemas para clasificar en su mayoría a las mismas observaciones; sin embargo esto es sólo una teoría. En los análisis de diagnóstico se profundizará un poco más respecto a ello.

Como conclusión de esta sección, el desempeño de las tablas de clasificación es bueno para ambos modelos, aunque es ligeramente mejor para el modelo que utiliza como predictores a componentes principales. A continuación se procederá a la elaboración de curvas ROC.

Curvas ROC

Las curvas ROC (*Receiving Operator Characteristic*) son de gran utilidad en estudios de clasificación de dos niveles, para elegir un punto de corte el cual dicte si una observación debe ser clasificada en una categoría o en la otra.

Sin embargo, se hacen dos observaciones del párrafo anterior con respecto a los modelos en cuestión.

- La herramienta mencionada es de utilidad para poder elegir un punto de corte entre niveles o categorías, pero este punto de corte no es de mayor utilidad para los modelos presentes, dado que se optó por clasificar a las observaciones en la categoría que tuviera la mayor probabilidad de pertenencia.

- Se necesita que el número de categorías de la variable de respuesta fuera dos, mientras que los modelos a analizar constan de tres categorías para la variable de respuesta.

De cualquier forma, se procederá a utilizar dicha técnica con el fin de entender mejor el comportamiento de las probabilidades de las observaciones en cada una de las categorías de manera separada.

Se realizarán tres curvas ROC, en las que se evaluarán las siguientes situaciones:

- Categoría 1 vs categorías diferentes de 1 (2 y 3)
- Categoría 2 vs categorías diferentes de 2 (1 y 3)
- Categoría 3 vs categorías diferentes de 3 (1 y 2)

Una vez definidas las dos nuevas categorías para cada curva, se procederá a calcular para ciertos puntos de corte la especificidad y sensibilidad correspondiente.

Para los casos donde se aísla la categoría "j", se utilizarán las probabilidades $\hat{\pi}_{ij}$ y $1 - \hat{\pi}_{ij}$ para calcular las especificidades y sensibilidades de los puntos de corte. Así también, los puntos de corte a evaluar serán 0, 0.01, 0.02, 0.03, ..., 0.99 y 1.00.

El siguiente paso a realizar es encontrar el mejor punto de corte para los objetivos del estudio en cuestión. No existe una metodología para encontrar al "mejor" punto de corte, dado que este depende de los objetivos de cada estudio (puede ser que a algún estudio únicamente le sea de interés maximizar la sensibilidad, por ejemplo). Para efectos de este estudio, se considerará como mejor punto de corte al que maximice conjuntamente tanto sensibilidad como especificidad.

El motivo por el cual se decide definir de esta manera al mejor punto de corte se fundamenta en que, aunque en si el mayor interés de cada caso es potenciar la sensibilidad (el porcentaje de observaciones con categoría j clasificados correctamente), de nada serviría un punto de corte que ocasionara una especificidad pobre.

Así pues, la función a maximizar que se utilizó para encontrar el punto de corte óptimo para el estudio fue "sensibilidad + especificidad". Una vez que se encontró el punto de corte óptimo con su respectivo valor maximizado, se procedió a graficar la curva ROC del modelo y a calcular su AUC (*Area Under the Curve*).

Los resultados obtenidos, para cada uno de los dos modelos a evaluar, son los siguientes.

Para el modelo con componentes principales como variables predictoras se tiene:

Tabla 7.12: Curvas ROC: Modelo con componentes principales

Caso	"1 vs no 1"	"2 vs no 2"	"3 vs no 3"
Punto de corte óptimo	0.28-0.30	0.18-0.19	0.37-0.78
Sensibilidad	0.94	1.00	0.95
Especificidad	0.90	0.83	1.00
Suma de ambas	1.84	1.83	1.95
Figura Curva ROC	7.1 izquierda	7.1 derecha	7.2
AUC	0.9574	0.9288	0.9928

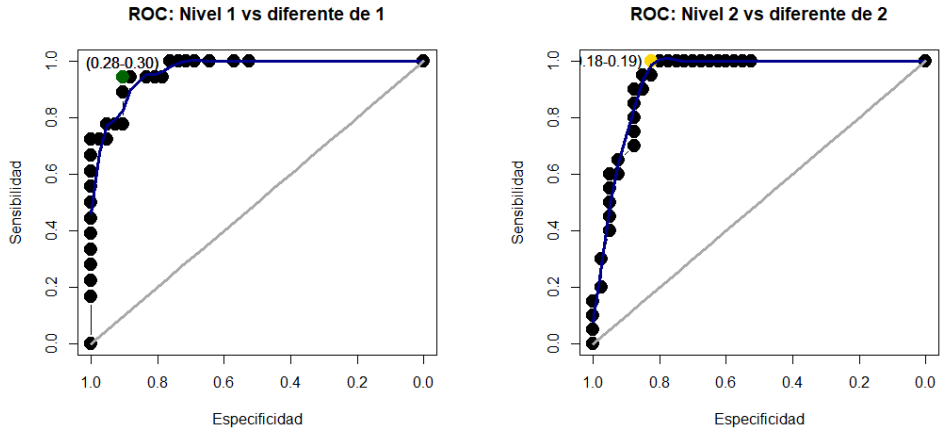

Figura 7.3: Izquierda: Curva ROC caso "Nivel 1 vs diferente de 1". Derecha: Curva ROC caso "Nivel 2 vs diferente de 2".

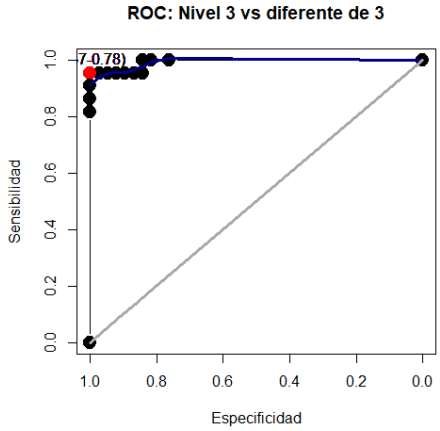

Figura 7.4: Curva ROC caso "Nivel 3 vs diferente de 3"

Mientras que para el modelo con variables originales como variables predictoras se tiene:

Tabla 7.13: Curvas ROC: Modelo con variables originales

Caso	"1 vs no 1"	"2 vs no 2"	"3 vs no 3"
Punto de corte óptimo	0.63-0.64	0.37	0.35-0.58
Sensibilidad	0.78	0.90	0.95
Especificidad	1.00	0.85	0.95
Suma de ambas	1.78	1.75	1.90
Figura Curva ROC	7.3 izquierda	7.3 derecha	7.4
AUC	0.9232	0.8729	0.9681

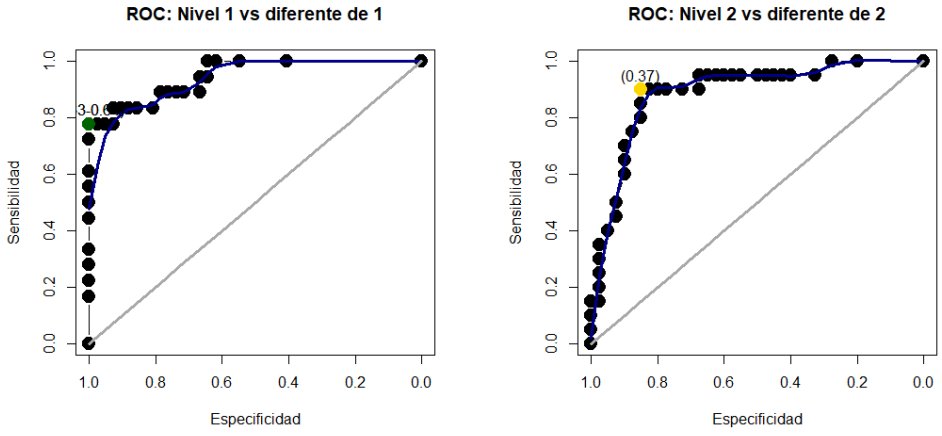

Figura 7.5: Izquierda: Curva ROC caso "Nivel 1 vs diferente de 1". Derecha: Curva ROC caso "Nivel 2 vs diferente de 2".

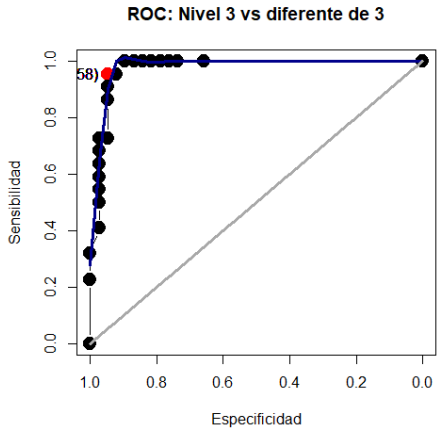

Figura 7.6: Curva ROC caso "Nivel 3 vs diferente de 3"

Las conclusiones obtenidas por las curvas ROC fueron las siguientes:

- Respecto a los intervalos de los puntos de corte:

  1. De los puntos de corte obtenidos, se podría interpretar que para el modelo con componentes principales se necesita una probabilidad $\hat{\pi}_{ij}$ de mínimo 0.31, 0.20 y 0.79 para que la observación i se encuentre, según el modelo, en el nivel j, con $j = 1, 2, 3$ respectivamente.

     En cambio, para el modelo con variables originales se necesitaría de una probabilidad $\hat{\pi}_{ij}$ de mínimo 0.65, 0.38 y 0.59 para que la observación i fuera catalogada, según el modelo, en la categoría j, con $j = 1, 2, 3$ respectivamente.

  2. Nótese que los intervalos de los mejores puntos de corte para el caso "Nivel 3 vs diferente de 3" en ambos modelos, tienden a ser muy grandes. Esto quiere decir que las probabilidades que arrojó el modelo asociadas a la categoría tres, π_{i3}, son en general o muy pequeñas, o muy grandes, y además el modelo no tiene dificultad para diferenciar ambas categorías (categoría tres contra categorías diferentes de tres).

- Respecto a sensibilidades y especificidades:

  1. Ambos modelos poseen buenos valores tanto de sensibilidades como de especificidades, aunque nuevamente el modelo que utiliza componentes principales presenta una ligera ventaja respecto al modelo que utiliza a las variables originales como predictoras.

2. Mientras que para el modelo con componentes principales el caso "Nivel 1 vs diferente de 1" no presenta mayor problema en cuanto a su sensibilidad y especificidad, el modelo que utiliza variables originales presenta dificultades en separar a un conjunto de observaciones con la categoría uno de tamaño considerable, del conjunto de observaciones catalogadas en dos, ya que bajo el punto óptimo, tan sólo el 78 % de las observaciones en la categoría uno fueron catalogadas correctamente.

   Para el caso "Nivel 3 vs diferente de 3", los resultados de las sensibilidades y especificidades (95 % cada uno para cada modelo) indican que en general, ambos modelos pueden separar claramente a las observaciones con categoría tres de las observaciones con otras categorías.

- Con respecto a los valores AUC:

  1. Los valores AUC de los casos del modelo que utiliza componentes principales como variables predictoras fueron 0.96, 0.93 y 0.99. Para el modelo que utiliza variables originales, 0.92, 0.87 y 0.97. Si bien se recuerda del capítulo 3, sección 2, valores de AUC entre 0.80 y 0.90 se consideran como excelentes, mientras que valores entre 0.90 y 1.00 se consideran excepcionales. Por lo tanto, según el análisis brindado por las curvas ROC y el AUC, ambos modelos se podrían considerar como excepcionales.

     Ambos modelos coinciden en, según sus respectivos valores "AUC", que las observaciones con mayor facilidad de clasificación corresponden a las observaciones con categoría tres, posteriormente las observaciones con categoría uno y finalmente las que poseen categoría dos.

Otras medidas de resúmen: *pseudo-*R^2

Para el modelo con componentes principales como variables predictoras se tiene:

1. *Pseudo-*R^2 del coeficiente de correlación de Pearson: 0.7456
2. *Pseudo-*R^2 como suma de cuadrados de una regresión lineal: 0.7456
3. *Pseudo-*R^2 con log-verosimilitudes: 0.7305

Mientras que para el modelo con variables originales como variables predictoras se obtuvo lo siguiente:

1. *Pseudo-*R^2 del coeficiente de correlación de Pearson: 0.6116

2. *Pseudo-*R^2 como suma de cuadrados de una regresión lineal: 0.6107

3. *Pseudo-*R^2 con log-verosimilitudes: 0.5864

Los resultados anteriores hacen pensar que aunque ambos modelos muestran valores de *pseudo-*R^2 buenos (recuérdese que para regresión logística multinomial una *pseudo-*R^2 con valor 0.6 representa un buen modelo, a diferencia de la R^2 para regresión lineal) y por lo tanto ambos modelos se pueden considerar con un buen desempeño, los resultados del modelo que utiliza componentes principales como variables predictoras son contundentemente mejores que los resultados obtenidos por el modelo que utiliza a las variables originales como variables predictoras.

7.4.2. Estadísticas de diagnóstico

Las estadísticas de diagnóstico serán de ayuda para localizar posibles *outliers* o datos influyentes. Antes de iniciar con el análisis de dichas estadísticas, se mencionarán algunos detalles correspondientes a las tablas de clasificación, esto con el propósito de buscar una relación entre los posibles *outliers* o datos influyentes que indiquen las estadísticas de diagnóstico, y las observaciones que fueron clasificadas incorrectamente por los modelos mostradas en las tablas de clasificación.

Las tablas de clasificación para ambos modelos son:

- Para el modelo que utiliza componentes principales como predictores: Tabla 7.14

Tabla 7.14: Tabla de clasificación: componentes principales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	14	4	0
	$Y = 2$	4	16	0
	$Y = 3$	0	1	21

- Para el modelo que utiliza variables originales como predictores: Tabla 7.15

Tabla 7.15: Tabla de clasificación: variables originales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	14	4	0
	$Y = 2$	3	15	2
	$Y = 3$	0	1	21

Ahora se mencionan las observaciones que fallaron en ser predecidas para cada modelo, dado que se tiene interés en conocer si éstas pueden ser consideradas como *outliers* o influyentes por las estadísticas de diagnóstico o no.

- Observaciones con predicciones erróneas por el modelo con componentes principales: Tabla 7.16

Tabla 7.16: Observaciones con predicciones erróneas: modelo con componentes principales.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Brooklyn Nets	1	2	0.3238	0.6734	0.0028
2013 Cleveland Cavaliers	3	2	0.0213	0.9122	0.0665
2013 Golden State Warriors	2	1	0.5477	0.4523	0.0000
2013 Houston Rockets	2	1	0.6836	0.3164	0.0000
2013 Portland Trail Blazers	1	2	0.3049	0.6951	0.0000
2013 Washington Wizards	1	2	0.3253	0.6742	0.0005
2015 Houston Rockets	2	1	0.5097	0.4830	0.0072
2015 LA Clippers	2	1	0.8040	0.1960	0.0000
2015 Portland Trail Blazers	1	2	0.1300	0.8699	0.0000

- Observaciones con predicciones erróneas por el modelo con variables originales: Tabla 7.17

Tabla 7.17: Observaciones con predicciones erróneas: modelo con variables originales.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Houston Rockets	2	1	0.6292	0.3707	0.0001
2013 Minnesota Timberwolves	2	3	0.0002	0.0204	0.9793
2013 New York Knicks	2	1	0.6178	0.3821	0.0000
2013 Portland Trail Blazers	1	2	0.1013	0.7889	0.1099
2013 Washington Wizards	1	2	0.4159	0.5830	0.0010
2015 Boston Celtics	2	3	0.0110	0.2923	0.6967
2015 LA Clippers	2	1	0.6079	0.3920	0.0001
2015 Milwaukee Bucks	3	2	0.0470	0.6289	0.3241
2015 Portland Trail Blazers	1	2	0.0611	0.6804	0.2585
2015 Toronto Raptors	1	2	0.2644	0.7297	0.0059

Las probabilidades $\hat{\pi}_{ij}$ de las tablas anteriores fueron redondeadas a cuatro decimales.

Ahora se procede al análisis de las estadísticas de diagnóstico.

Residuos

En esta sección se presentarán para los dos modelos a analizar tanto la gráfica de los residuos de Pearson, como la de los residuos de devianza, esto con el fin de localizar patrones de covariables que podrían ser *outliers*. Dado que para este estudio todos los patrones de covariables se componen de una sola observación, esto sería equivalente a localizar a los equipos que podrían ser *outliers*.

Otra consecuencia de que los patrones de covariables se compongan de una única observación es, que no se utilizarán supuestos de distribución en los residuos como percentiles para la búsqueda de datos atípicos.

Ahora, para el modelo que utiliza componentes principales como variables predictoras se tiene:

1. Gráfica de residuos de Pearson: Figura 7.5
2. Gráfica de residuos de Devianza: Figura 7.6

Cabe destacar que los puntos verdes de la gráfica de residuos de Pearson corresponden a residuos asociados a la categoría uno (desempeño alto), los puntos amarillos a la categoría dos (desempeño medio) y los puntos rojos a los de la categoría tres (desempeño pobre).

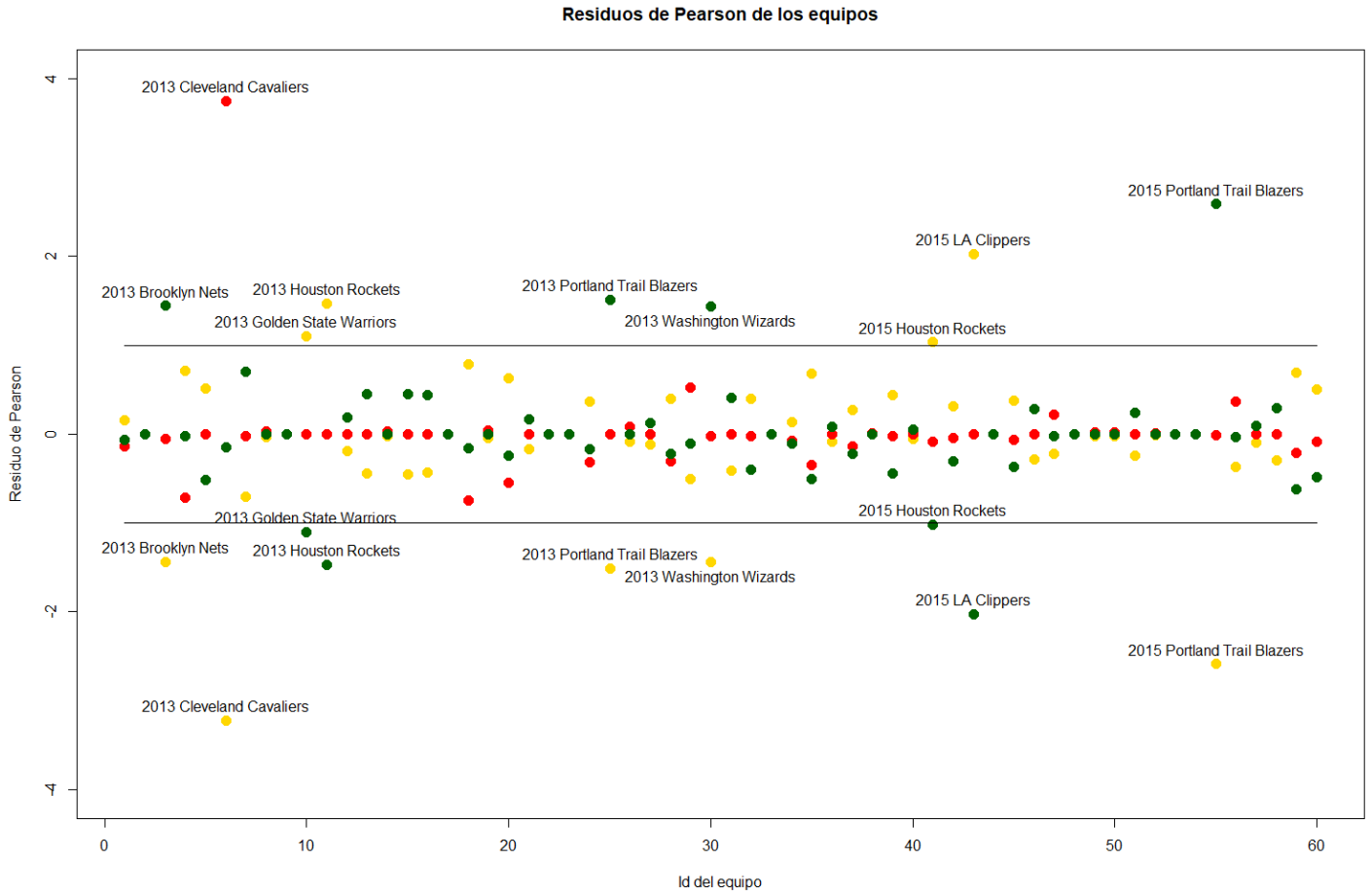

Figura 7.7: Residuos de Pearson de los equipos para el modelo con componentes principales. Se tienen tres residuos por cada observación o equipo, uno por cada categoría (verde para desempeño alto, amarillo para medio, rojo para bajo).

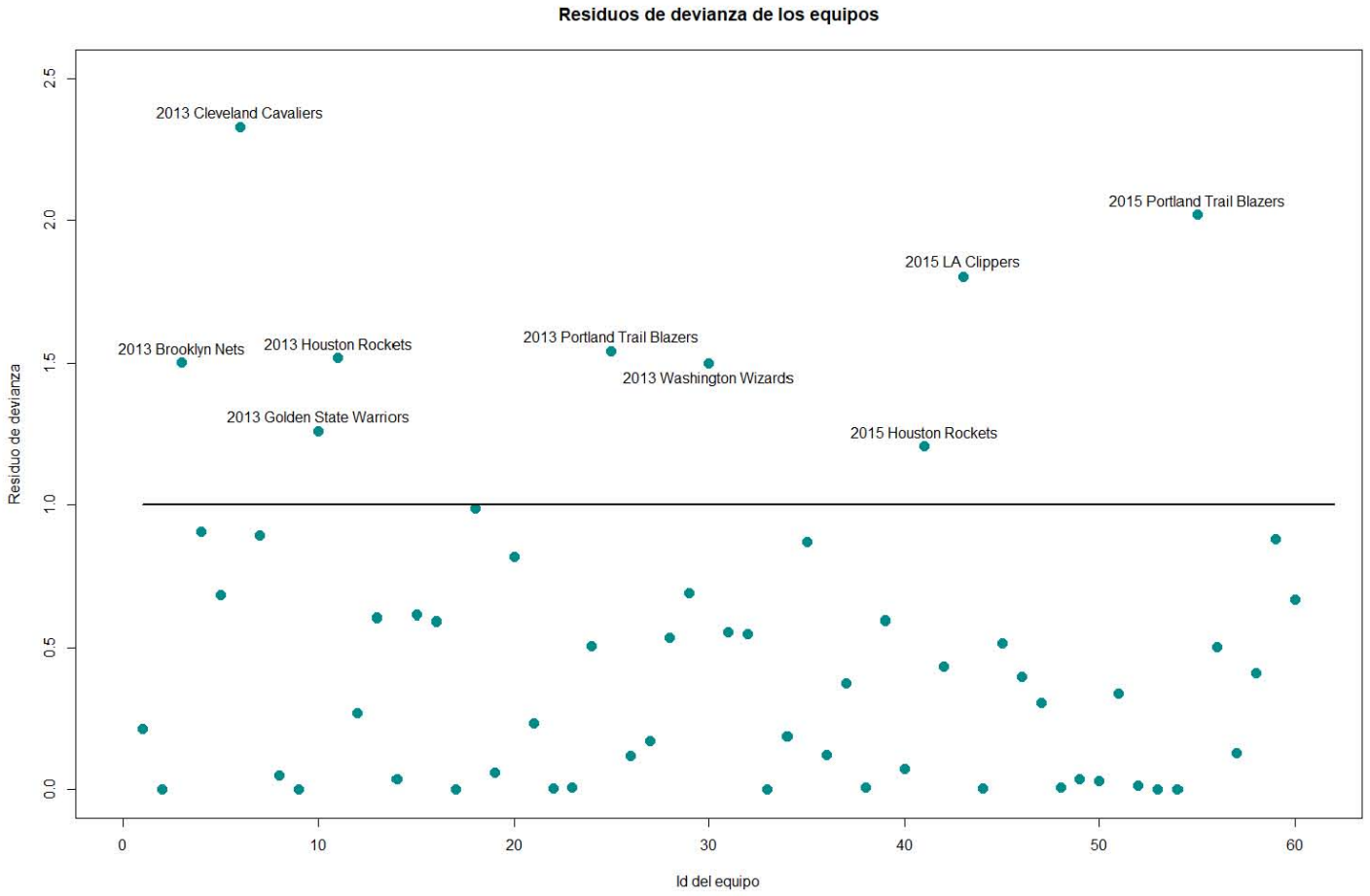

Figura 7.8: Residuos de devianza de los equipos para el modelo con componentes principales.

En cualquier modelo estadístico, contemplando ya que la base de datos de donde se obtuvo la información para dicho modelo es confiable, una de las propiedades que se buscaría en un análisis de diagnóstico, sería obtener el menor número de *outliers* posibles (¿de qué serviría un análisis que considerara a una tercera o cuarta parte de los datos como "atípicos"?). De esta manera, el análisis ganaría credibilidad al comportarse de una manera similar a la realidad (en la realidad no existe un porcentaje alto de datos atípicos; si existiera, entonces ya no serían "atípicos") y los resultados obtenidos por el modelo transmitirían mayor seguridad.

Las gráfica de los residuos de Pearson del modelo con componentes principales identifica claramente a las nueve observaciones que fueron clasificadas incorrectamente por el modelo y que se mencionan en la tabla 7.16.

Dentro de la gráfica se tienen tres residuos por cada observación o equipo, uno por cada categoría. Cada observación que tuvo un mal ajuste por el modelo posee dos residuos fuera de la banda que contiene a los residuos con comportamiento ordinario (banda entre los valores 1 y -1), un residuo positivo y otro negativo. Cabe mencionar que los límites de dicha banda fueron elegidos únicamente con el fin de hacer énfasis en la separación entre los residuos con escala grande y los residuos ordinarios, y en ningún momento se contempló alguna distribución para fijar los límites de la misma.

De entre los 18 residuos que se encuentran alejados de los demás, los colores de los residuos positivos corresponden a las categorías a las que pertenecen los equipos, mientras que los colores de los residuos negativos corresponden a las categorías en las que fueron clasificados dichos equipos por el modelo.

Se puede observar en la gráfica que, dentro de las nueve observaciones clasificadas incorrectamente, los residuos de seis de ellas se encuentran no muy alejadas de los residuos de las observaciones clasificadas correctamente. Esto se debe a que, a pesar de que el modelo las haya clasificado erróneamente, la probabilidad asociada a la verdadera categoría que poseen las observaciones (la observada, no la predecida) también es muy grande, aunque no tan grande como para ser pronosticada en dicha categoría. Estas son características de un buen modelo, ya que si bien tuvo errores, dichos errores no fueron tan grandes, y por lo tanto no son graves.

En cambio, los otros tres equipos restantes (2013 Cleveland Cavaliers, 2015 LA Clippers y 2015 Portland Trail Blazers) poseen residuos que se encuentran muy alejados de los demás. Esto se debe a que el modelo asignó a la categoría donde debería de haber sido clasificado el equipo una probabilidad muy pequeña. Esto implica una de dos cosas: que las tres observaciones podrían ser datos atípicos, o bien que el modelo simplemente tuvo problemas para ajustar equipos similares a éstos.

En el primer caso, se considera que el modelo no tiene problema alguno y el problema se debió a algún suceso "extraño" en los datos. En el segundo punto, el problema le es atribuido al modelo.

Por lo tanto, de ahora en adelante se le prestará una atención especial a estas tres observaciones, y al término del análisis de diagnóstico, se decidirá si dichas observaciones realmente son *outliers*, o simplemente son observaciones comunes a las que el modelo ha fallado en ajustar correctamente.

Ahora se proseguirá con el análisis de la gráfica de residuos de devianza:

Respecto a la gráfica de los residuos de devianza, también identifica a las mismas nueve observaciones clasificadas erróneamente. Al analizar esta gráfica con mayor detenimiento, se han llegado a las mismas conclusiones que la gráfica anterior: seis de las nueve observaciones clasificadas erróneamente parecen no tener características de observaciones atípicas, mientras las otras tres restantes podrían llegar a ser *outliers*, por lo que se les prestará detallada atención en las siguientes pruebas.

Ahora, el modelo que utiliza a las variables originales como variables predictoras obtuvo los siguientes resultados:

1. Gráfica de residuos de Pearson: Figura 7.7.
2. Gráfica de residuos de Devianza: Figura 7.8.

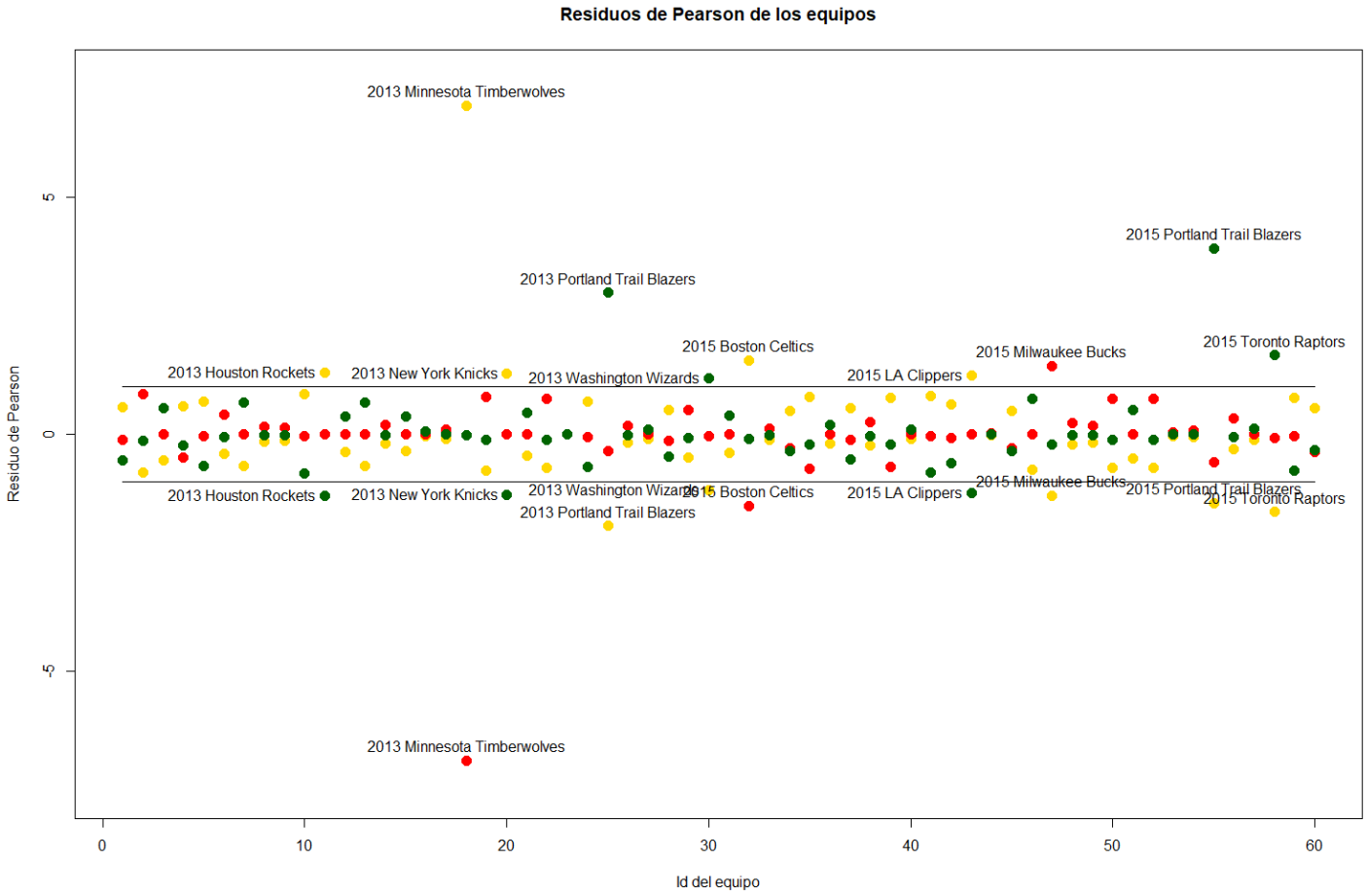

Figura 7.9: Residuos de Pearson de los equipos para el modelo con variables originales como variables predictoras.

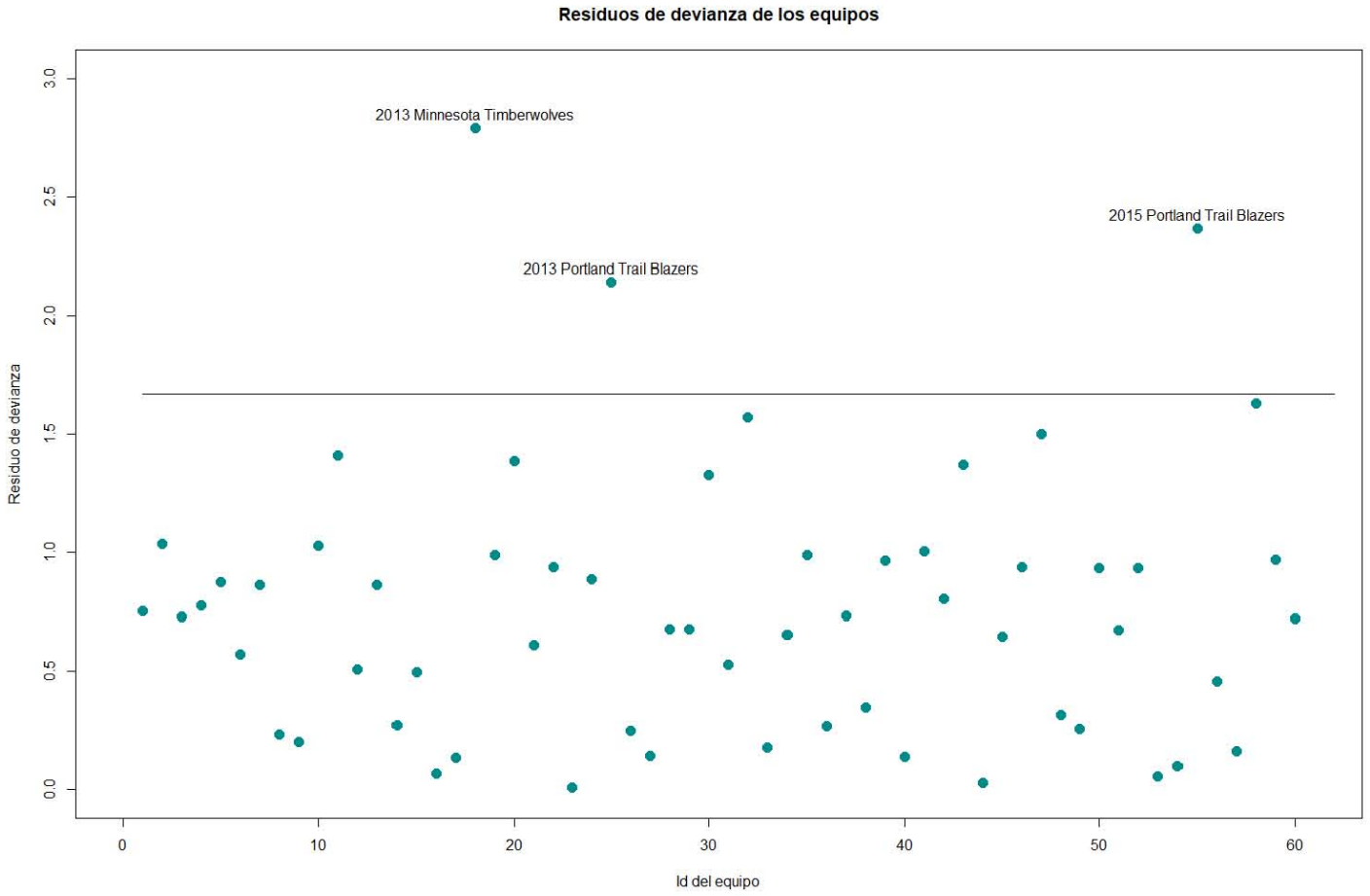

Figura 7.10: Residuos de devianza de los equipos para el modelo con variables originales como variables predictoras.

Al igual que en el modelo con componentes principales, la gráfica de residuos de Pearson para el modelo que utiliza variables originales como variables predictoras muestra a la mayoría de las observaciones mal clasificadas muy cerca del grupo de observaciones clasificadas correctamente, característica de un modelo con buen ajuste. La gráfica también muestra algunas observaciones que distan demasiado de las demás, indicando que las probabilidades ajustadas del modelo distaron mucho de lo observado en la realidad. Estas observaciones corresponden a los equipos:

- 2013 Minnesota Timberwolves
- 2013 Portland Trail Blazers
- 2015 Portland Trail Blazers

La observación 2013 Minnesota Timberwolves presenta un pésimo ajuste, por lo que se analizará detalladamente con futuras pruebas la posibilidad de que dicha observación sea un *outlier*. Con respecto a las otras dos observaciones (2013 Portland Trail Blazers y 2015 Portland Trail Blazers), se ha pensado por ahora en que mientras el equipo de 2013 parece no tener residuos graves, el equipo 2015 Portland Trail Blazers podría llegar a ser atípico, ya que ha sido clasificado incorrectamente en ambos modelos y con residuos grandes. Sin embargo la decisión de tomarlo o no como atípico no se concluirá sino hasta el final del análisis.

La gráfica de residuos de devianza ha vuelto a señalar a las mismas tres observaciones como atípicas, por lo que la posibilidad de tomar a cualquiera de las tres como *outlier* sigue abierta.

Como nota final al estudio, se ha concluido que ambos modelos se ajustan bien a los datos, aunque el modelo que utiliza componentes principales podría no tener errores tan graves como el de las variables originales.

Leverage values

En esta sección, se calcularán las estadísticas h_i y b_i, mismas que fueron mencionadas en el capítulo 3 sección 3.1 del presente documento y cuyo propósito es encontrar posibles observaciones influyentes.

Como se recordará, para el caso del modelo logístico con más de dos niveles en su variable de respuesta no se aseguró que dichas estadísticas fueran a dar realmente los resultados que se quisieran; únicamente fueron mencionadas por algunos autores como sugerencias de apoyo en la elaboración de estadísticas de diagnóstico.

Dado que los modelos a analizar se componen de $J = 3$ categorías para su

variable de respuesta, los resultados de dichas estadísticas podrían no ser confiables. Independientemente de esto, se calcularán sus valores y se intentarán realizar conclusiones respecto a ellas.

Para el modelo que utiliza componentes principales como variables predictoras se tiene:

- Gráfica de *leverage values*: Figura 7.9
- Gráfica de la estadística b_i: Figura 7.10

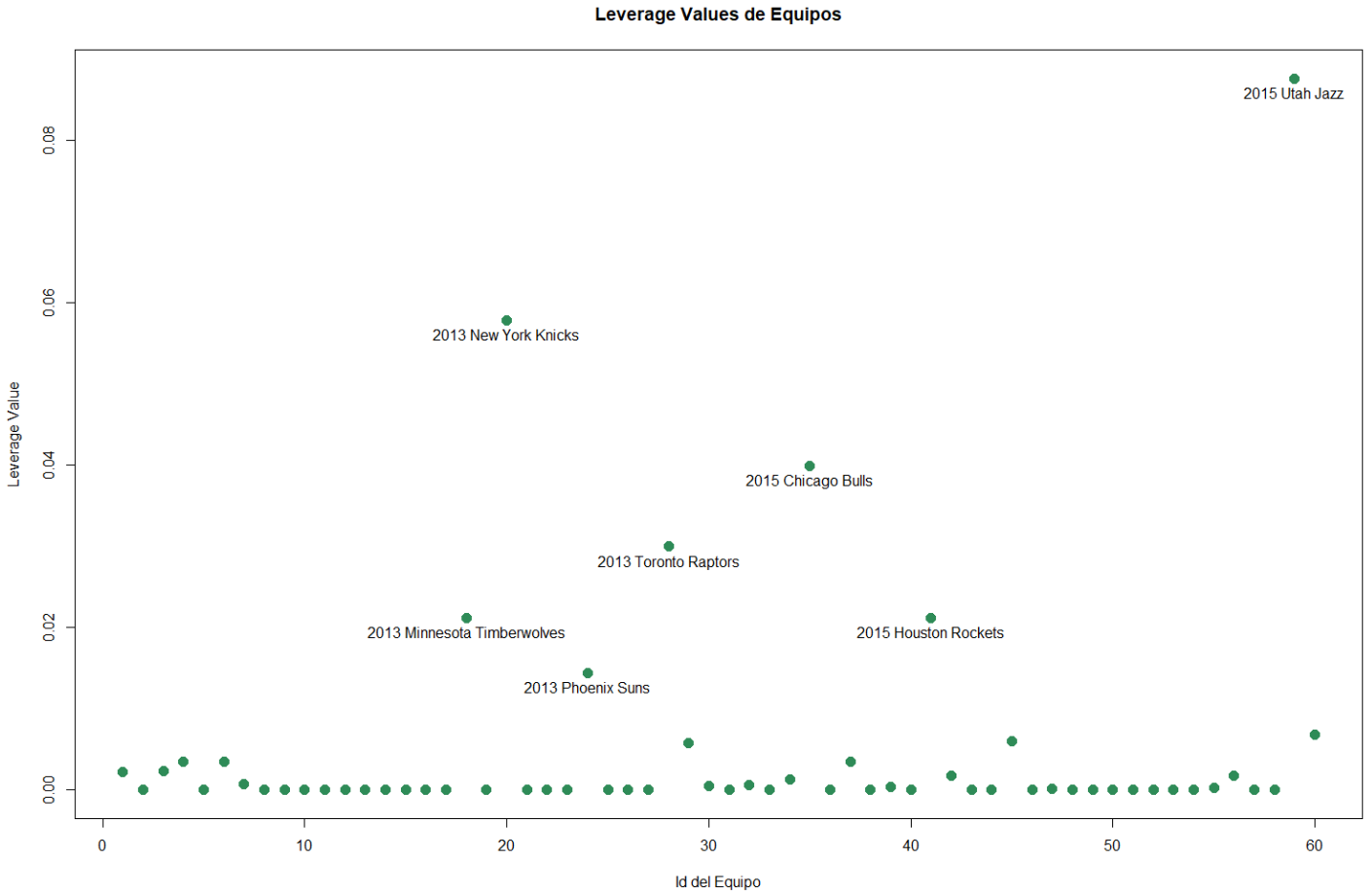

Figura 7.11: *Leverage values* de los equipos para el modelo con componentes principales.

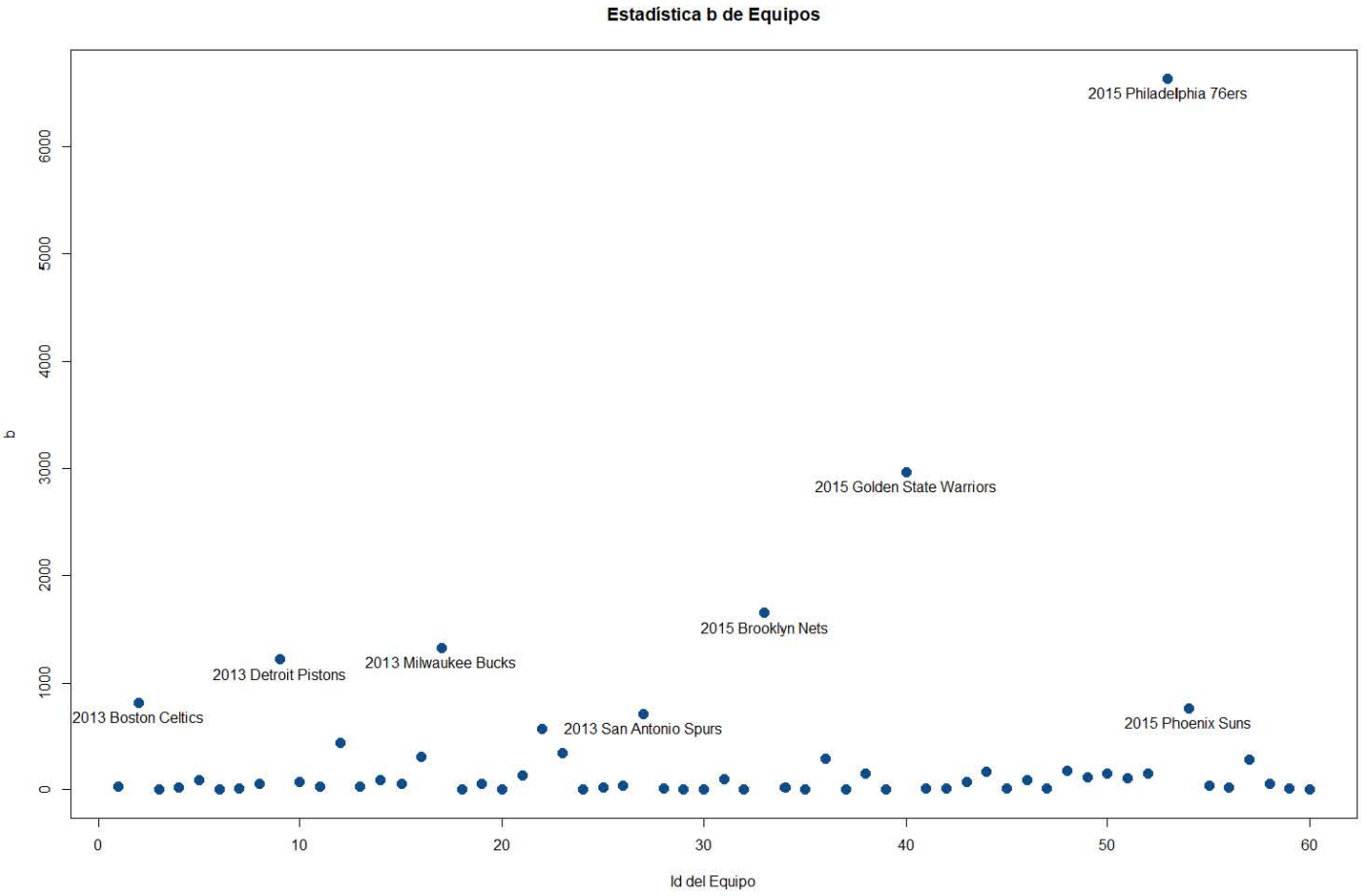

Figura 7.12: Estadística b de los equipos para el modelo con componentes principales.

La gráfica de *leverage values* sobresalta a las observaciones que se encuentran en la tabla 7.18.

Tabla 7.18: Equipos que la estadística h_i señala como lejanos a la media de equipos. Estudio realizado para el modelo que utiliza componentes principales como predictores.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Minnesota Timberwolves	2	2	0.0241	0.6152	0.3608
2013 New York Knicks	2	2	0.0559	0.7163	0.2278
2013 Phoenix Suns	2	2	0.0280	0.8795	0.0925
2013 Toronto Raptors	2	2	0.0477	0.8662	0.0861
2015 Chicago Bulls	2	2	0.2061	0.6858	0.1081
2015 Houston Rockets	2	1	0.5097	0.4830	0.0072
2015 Utah Jazz	2	2	0.2798	0.6796	0.0407

Para dar seguridad de que la estadística h_i ha funcionado de la manera deseada, se esperaría que la tabla mencionada anteriormente contuviera a los equipos que se encontraran muy lejos de la media de todos los equipos; sin embargo, al analizar con mayor detalle a los equipos de la tabla anterior, no se encontraron muestras suficientes que prueben que dichos equipos puedan distar de la media general.

Esta ausencia de evidencia, más el hecho de que todos los equipos de la tabla anterior hayan sido clasificados en la categoría dos, son motivos suficientes para argumentar que la estadística h_i no ha dado resultados congruentes, aunque se vuelve a recordar que esta conclusión no es de sorprender, dado que se sabía que el nivel de desempeño de dicha estadística era incierto para modelos que consideraran a un número de categorías mayor de dos en su variable de respuesta.

Por lo tanto, de ahora en adelante no se le dará importancia a los resultados de la estadística h_i.

Ahora, con respecto a la gráfica de b_i, los equipos que fueron remarcados por dicha gráfica se presentan en la tabla 7.19.

Tabla 7.19: Equipos que la estadística b_i señala como lejanos a la media de equipos. Estudio realizado para el modelo que utiliza componentes principales como predictores.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Boston Celtics	3	3	0	0	0.9999
2013 Detroit Pistons	3	3	0	0	0.9999
2013 Milwaukee Bucks	3	3	0	0	0.9999
2013 San Antonio Spurs	1	1	0.9857	0.0143	0.0000
2015 Brooklyn Nets	3	3	0	0	0.9999
2015 Golden State Warriors	1	1	0.9974	0.0026	0.0000
2015 Philadelphia 76ers	3	3	0	0	0.9999
2015 Phoenix Suns	3	3	0	0	0.9999

A diferencia de los resultados arrojados por h_i, se piensa que la estadística b_i ha mostrado buenos resultados, pues los equipos con los mayores valores de dicha estadística corresponden a equipos con un desempeño sobresaliente, o equipos con un desempeño pésimo.

De entre los ocho equipos sobresaltados por b_i, dos corresponden a equipos con desempeño sobresaliente y el resto a equipos con desempeño pésimo. Los dos equipos con desempeño sobresaliente:

- 2013 San Antonio Spurs
- 2015 Golden State Warriors

Aunque no son los equipos con la mayor probabilidad en la categoría uno, fueron los equipos que quedaron campeones en sus respectivos años. Esto implica que la estadística b_i logró identificar a los dos equipos ganadores de los años analizados.

Para el caso de los 6 equipos mencionados por b_i con un desempeño pobre, estos mismos coinciden con ser los equipos con la mayor probabilidad de pertenecer a la categoría tres (desempeño pobre), $\hat{\pi}_{i3}$. Entre ellos, se encuentran tres equipos del año 2013 y tres del 2015. Respecto a los equipos del año 2015, tres de los cuatro equipos con peor desempeño en ese año fueron identificados por b_i, mientras que de los tres equipos del año 2013 mencionados por b_i, uno de ellos fue el equipo con el peor desempeño en toda la liga (2013 Milwaukee Bucks).

Si estos 8 equipos llegaran a presentar valores altos en la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$, entonces serán considerados como datos influyentes. Si además de ser influyentes, llegaran a ser también atípicos (cualidad que hasta ahora no parecen poseer), se podrá plantear la eliminación de los mismos en el modelo.

En conclusión para el modelo que utiliza componentes principales, mientras que la estadística h_i no brindó resultados importantes, la estadística b_i tuvo un buen desempeño y ayudó a encontrar a los equipos que tanto para bien como

para mal, distaran considerablemente de la media de los equipos.

Una posible explicación del por qué de entre los ocho equipos mencionados por b_i únicamente dos son sobresalientes y seis pésimos, podría ser que, en general, el modelo considere a más equipos con un pésimo desempeño que equipos con un increíble desempeño. Por así decir, de entre los 18 equipos con categoría uno, todos o la gran mayoría corresponden a equipos con estadísticas buenas, pero de entre estos 18, dos de ellos tienen estadísticas muy superiores a los 16 restantes según el modelo en cuestión. El caso se puede replicar para los equipos pertenecientes a la categoría tres, pensando que de entre los 22 equipos con categoría tres, seis de ellos tuvieron un desempeño muy inferior a los restantes 16 según el mismo modelo.

A continuación, se analizarán ambas estadísticas para el modelo que utiliza a variables originales como variables predictoras:

- Gráfica de *leverage values*: Figura 7.11.
- Gráfica de la estadística b_i: Figura 7.12.

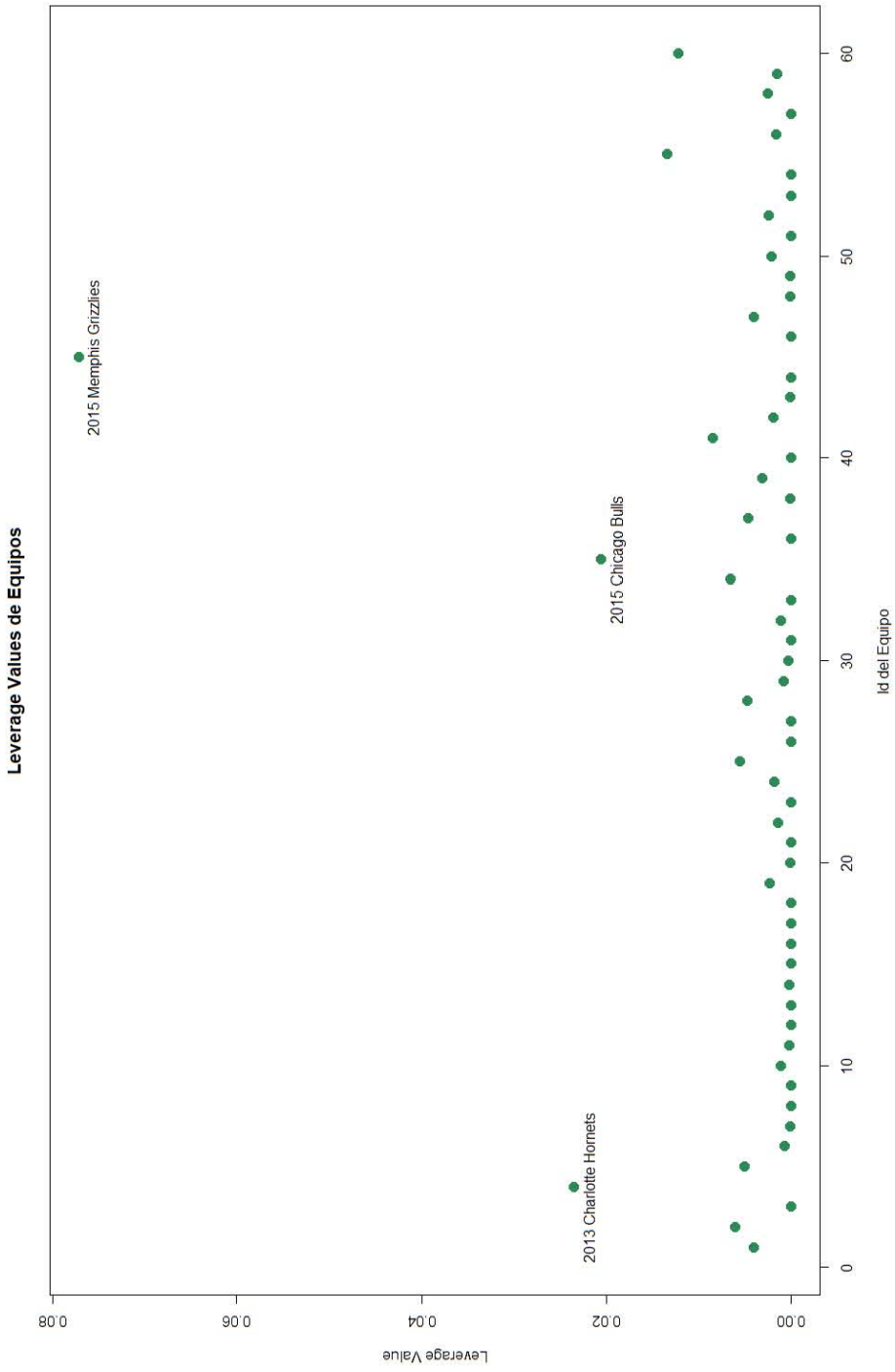

Figura 7.13: *Leverage values* de los equipos para el modelo con variables originales.

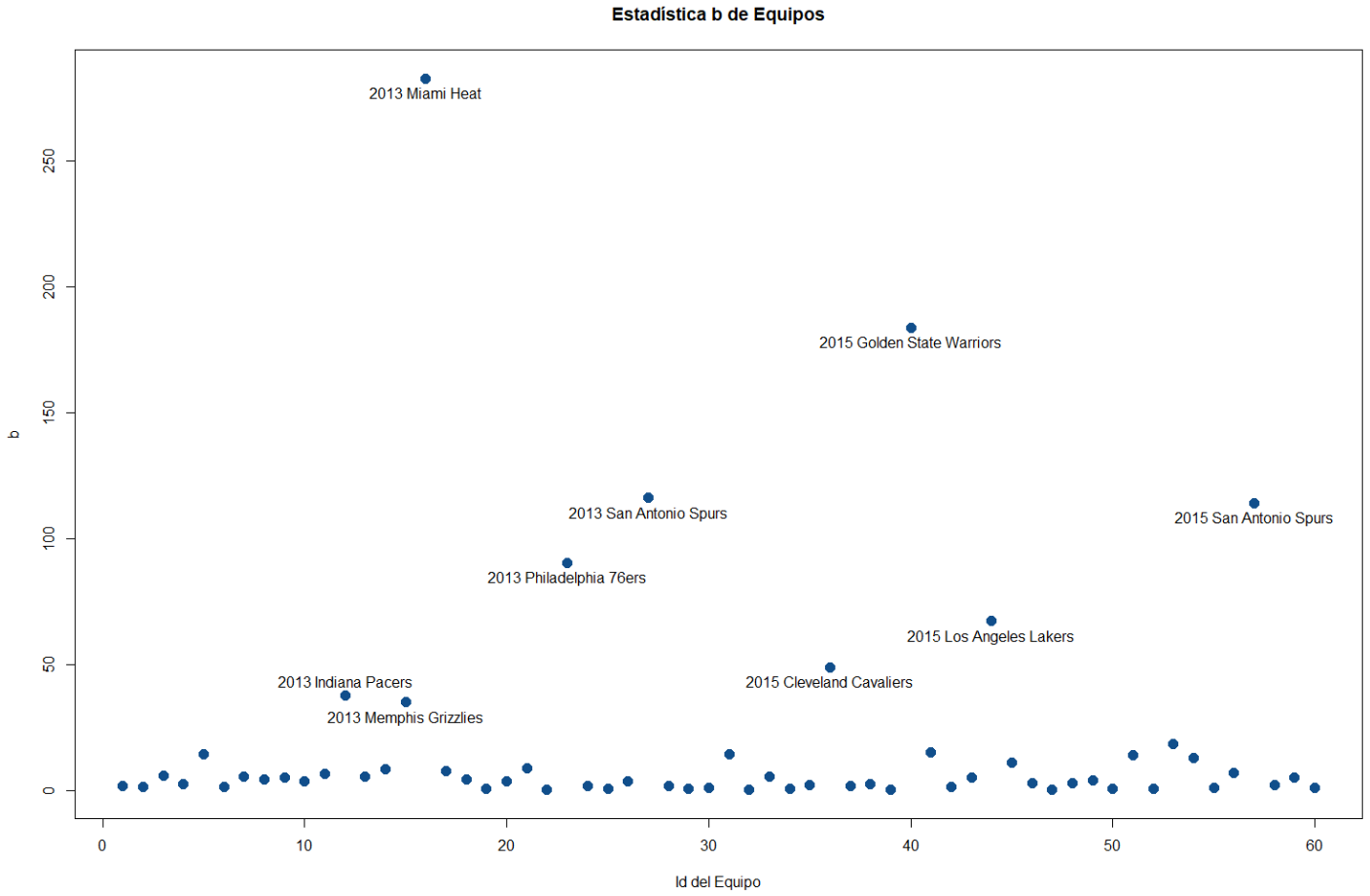

Figura 7.14: Estadística b de los equipos para el modelo con variables originales.

La gráfica de *leverage values* sobresalta a los equipos que se encuentran en la tabla 7.20.

Tabla 7.20: Equipos que la estadística h_i señala como lejanos a la media de equipos. Estudio realizado para el modelo que utiliza variables originales como predictores.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Charlotte Hornets	2	2	0.0574	0.7408	0.2018
2015 Chicago Bulls	2	2	0.0433	0.6144	0.3423
2015 Memphis Grizzlies	2	2	0.1104	0.8116	0.0780

De manera similar al otro modelo, no se alcanza a detectar alguna utilidad de los resultados de la estadística h_i, por lo que no se le dará importancia a ésta en las conclusiones finales del modelo salvo que la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$ marque como influyentes a dichas observaciones.

Ahora con respecto a b_i, los equipos que fueron resaltados por dicha gráfica se presentan a continuación en la tabla 7.21.

Tabla 7.21: Equipos que la estadística b_i señala como lejanos a la media de equipos. Estudio realizado para el modelo que utiliza variables originales como predictores.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Indiana Pacers	1	1	0.8796	0.1204	0.0000
2013 Memphis Grizzlies	1	1	0.8835	0.1165	0.0000
2013 Miami Heat	1	1	0.9977	0.0023	0.0000
2013 Philadelphia 76ers	3	3	0	0	0.9999
2013 San Antonio Spurs	1	1	0.9900	0.0099	0.0000
2015 Cleveland Cavaliers	1	1	0.9646	0.0354	0.0000
2015 Golden State Warriors	1	1	0.9905	0.0095	0.0000
2015 Los Angeles Lakers	3	3	0.0000	0.0004	0.9996
2015 San Antonio Spurs	1	1	0.9868	0.0132	0.0000

Al igual que con el modelo anterior, parece ser que b_i también funcionó de manera correcta para el modelo con variables originales, sólo que inversamente al modelo que utiliza componentes principales, el modelo en cuestión considera a una gran cantidad de equipos con desempeño sobresaliente y una cantidad muy pequeña de equipos con desempeño pobre .

De entre los nueve equipos que resalta b_i, los equipos que según el modelo tienen desempeño pobre fueron:

- 2013 Philadelphia 76ers
- 2015 Los Angeles Lakers

Ambos equipos fueron los segundos peores equipos de sus respectivos años, por lo que se puede pensar que en general, b_i no tuvo malos resultados en encontrar a los peores equipos.

Los equipos con desempeño sobresaliente señalizados por b_i para el año 2013, así como el papel que realizó cada uno de ellos en dicha temporada, se mencionan a continuación:

- 2013 Indiana Pacers: Perdedor en semifinales contra Miami Heat
- 2013 Memphis Grizzlies: Perdedor en octavos de final
- 2013 Miami Heat: Semifinalista
- 2013 San Antonio Spurs: Campeón de la NBA

Es decir, tres de los cuatro equipos obtuvieron tres de los cuatro mejores desempeños de toda la NBA.

Respecto al año 2015, los equipos con desempeño sobresaliente señalados por b_i y sus respectivos papeles realizados fueron:

- 2015 Cleveland Cavaliers: Semifinalista
- 2015 Golden State Warriors: Campeón de la NBA
- 2015 San Antonio Spurs: Perdedor en cuartos de final

Es decir, de entre los tres equipos mencionados se obtuvieron los dos equipos con los mejores desempeños en toda la NBA.

Respecto a esta sección, se tuvieron las siguientes conclusiones:

- La estadística b_i se comportó de manera superior a lo esperado para ambos modelos, mientras que los resultados de los *leverage values* h_i fueron incongruentes y carecieron de interpretabilidad para ambos modelos, por lo que no se le dará importancia a dicha estadística a menos que la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$ seleccione a las mismas observaciones como influyentes.
- Aunque pareciera que b_i cumplió con su función de encontrar las observaciones más lejanas de la media de manera satisfactoria, los equipos señalados por b_i no se considerarán como influyentes hasta corroborarlo con la estadística $\Delta\hat{\beta}_{pj}^{(-i)}$.

  Mientras tanto, los resultados de b_i se interpretarán como los mejores y peores equipos señalados por el modelo y los que sus variables explicativas se alejan más de la media, todo esto sin dar calidad de que sean

influyentes.

Deltas

En esta sección se tratan las estadísticas de diagnóstico presentadas en la sección 3.2 del capítulo 3, correspondientes a $\Delta\chi^{2(-i)}$, $\Delta D^{(-i)}$ y $\Delta\hat{\beta}_{pj}^{(-i)}$.

Se había mencionado anteriormente que tanto la estadística χ^2 como D no eran adecuadas para este estudio, dado que las conclusiones derivadas de éstas, basadas en *p-values*, son correctas únicamente cuando el número de observaciones es mucho mayor al número de patrones de covariables. Esto podría generar incertidumbre respecto a si sería correcto utilizar en este estudio a $\Delta\chi^{2(-i)}$ y a $\Delta D^{(-i)}$.

Sin embargo, en ningún momento se utilizan supuestos de distribución para utilizar este par de estadísticas, por lo que podrían ser de ayuda para dar una idea de cuales observaciones mejoran o empeoran el ajuste del modelo, todo esto sin utilizar *p-values* ni pruebas de hipótesis; es decir, dichas estadísticas se podrían utilizar como medidas descriptivas.

Por lo tanto, se presentarán las tres estadísticas para ambos modelos, no sin antes hacer una observación muy importante: la Ji-cuadrada de Pearson tiende a castigar mucho más a observaciones con un mal ajuste que la devianza. Esto podría llegar a ser un problema para estudios con un número de observaciones igual al de patrones de covariables, por lo que sería una buena idea fiarse más de la estadística $\Delta D^{(-i)}$ que de la estadística $\Delta\chi^{2(-i)}$ en este tipo de casos y en los que no se quisiera que una única observación influyera demasiado en el valor total de la estadística.

Respecto a $\Delta\hat{\beta}_{pj}^{(-i)}$, en apoyo a ella se creó una nueva estadística que ayudará a evaluar el efecto de la sustracción de algún patrón de covariables (una observación para este caso) mediante un único valor, dado que al sustraer una observación del estudio se tendrían que analizar $(J-1)(P+1)$ estadísticas $\Delta\hat{\beta}_{pj}^{(-i)}$.

Dicha nueva estadística, creada por el autor, será el promedio de todas las estadísticas $\Delta\hat{\beta}_{pj}^{(-i)}$ para algún i fijo, calculada mediante la ecuación:

$$\overline{\Delta\hat{\beta}}_{pj}^{(-i)} = \frac{1}{(J-1)(P+1)} \sum_{j=1}^{J-1} \sum_{P=0}^{P} \frac{\hat{\beta}_{pj} - \hat{\beta}_{pj}^{(-i)}}{z_{1-\alpha/2} \cdot \hat{\mathbf{SE}}(\hat{\beta}_{pj})}$$

Con (1-α) algún nivel de confianza elegido. En este caso, se trabajará con un nivel de confianza del 97 % para el modelo que utiliza componentes principales, y un 98 % para el modelo que utiliza a variables originales, pues las significancias

de las variables de dichos modelos fueron 0.03 y 0.02, respectivamente.

Cabe mencionar que la nueva estadística $\overline{\Delta\hat{\beta}_{pj}}^{(-i)}$ sigue los mismos principios que $\Delta\hat{\beta}_{pj}^{(-i)}$; es decir, se busca que su valor se encuentre entre menos uno y uno.

El análisis a realizar en ambos modelos consistirá en tres gráficas diferentes por modelo, una por cada estadística mencionada.

Además de las gráficas mencionadas, se podrían realizar algunas otras como combinación de las mismas; sin embargo, en el presente estudio las tres estadísticas se analizarán de manera individual.

Las gráficas del modelo que utiliza componentes principales como variables predictoras son:

- $\Delta\chi^{2(-i)}$: Figura 7.13
- $\Delta D^{(-i)}$: Figura 7.14
- $\overline{\Delta\hat{\beta}_{pj}}^{(-i)}$: Figura 7.15

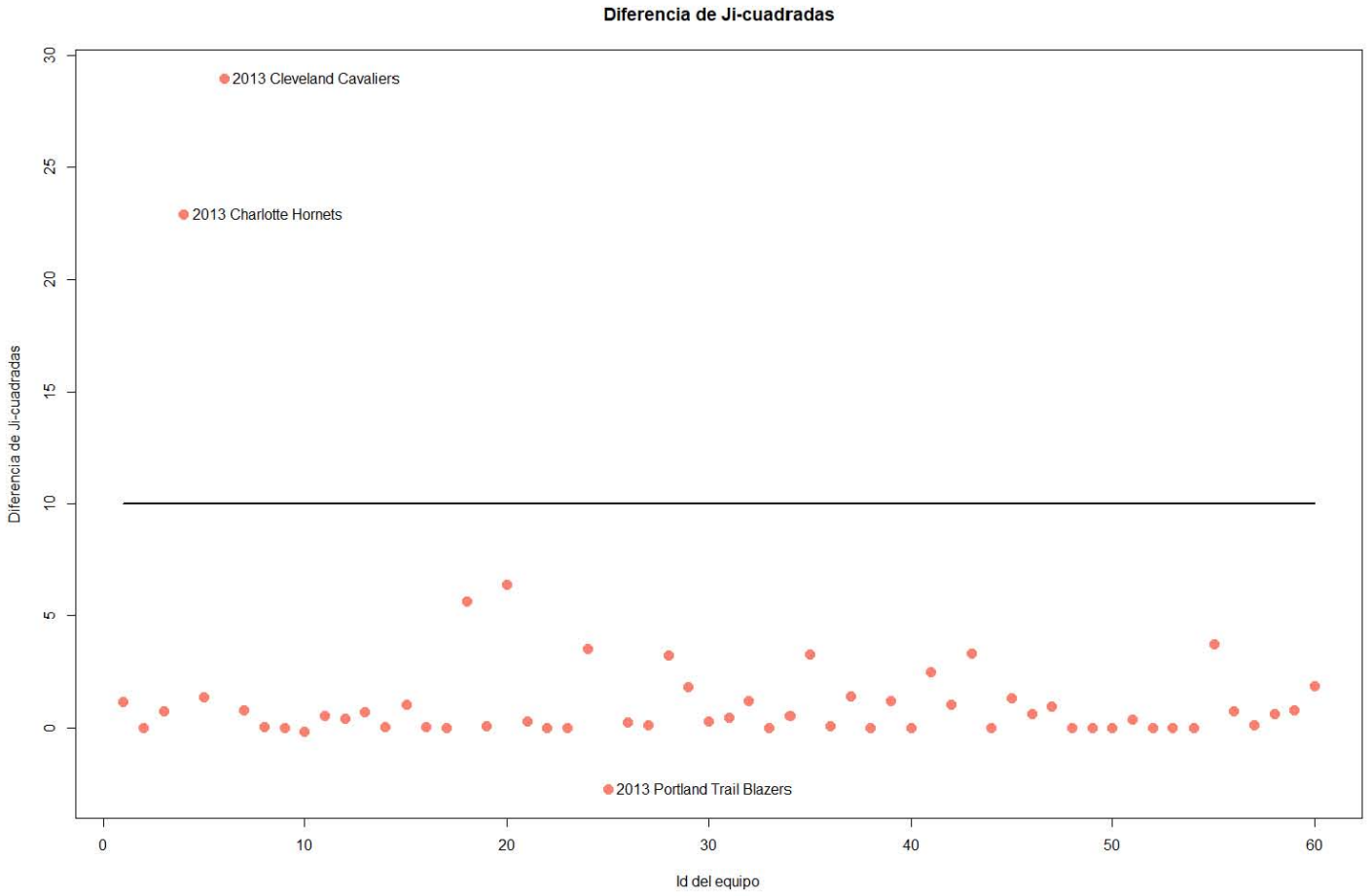

Figura 7.15: Gráfica de la estadística $\Delta\chi^{2(-i)}$ para el modelo que utiliza componentes principales.

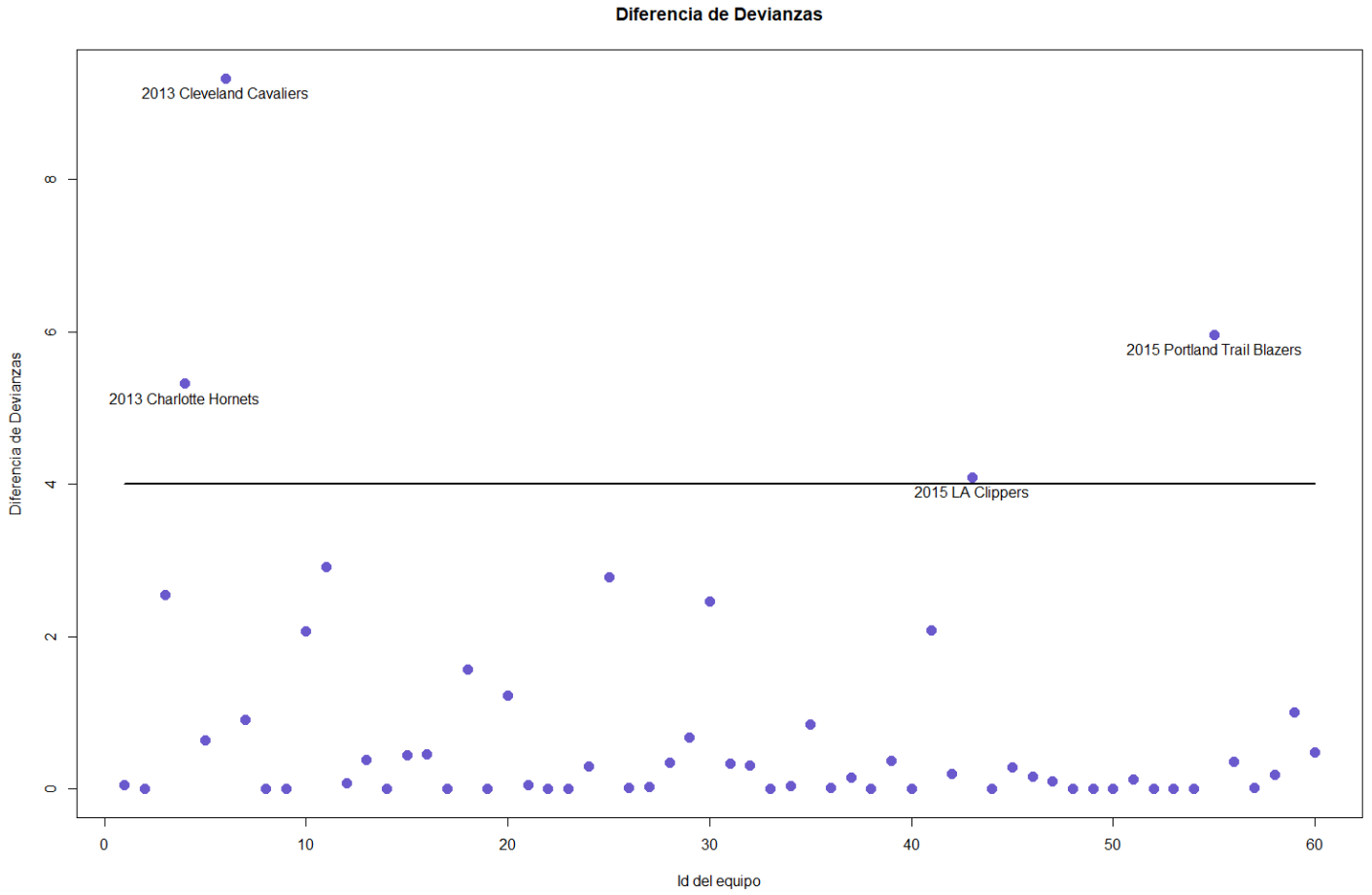

Figura 7.16: Gráfica de la estadística $\Delta D^{(-i)}$ para el modelo que utiliza componentes principales.

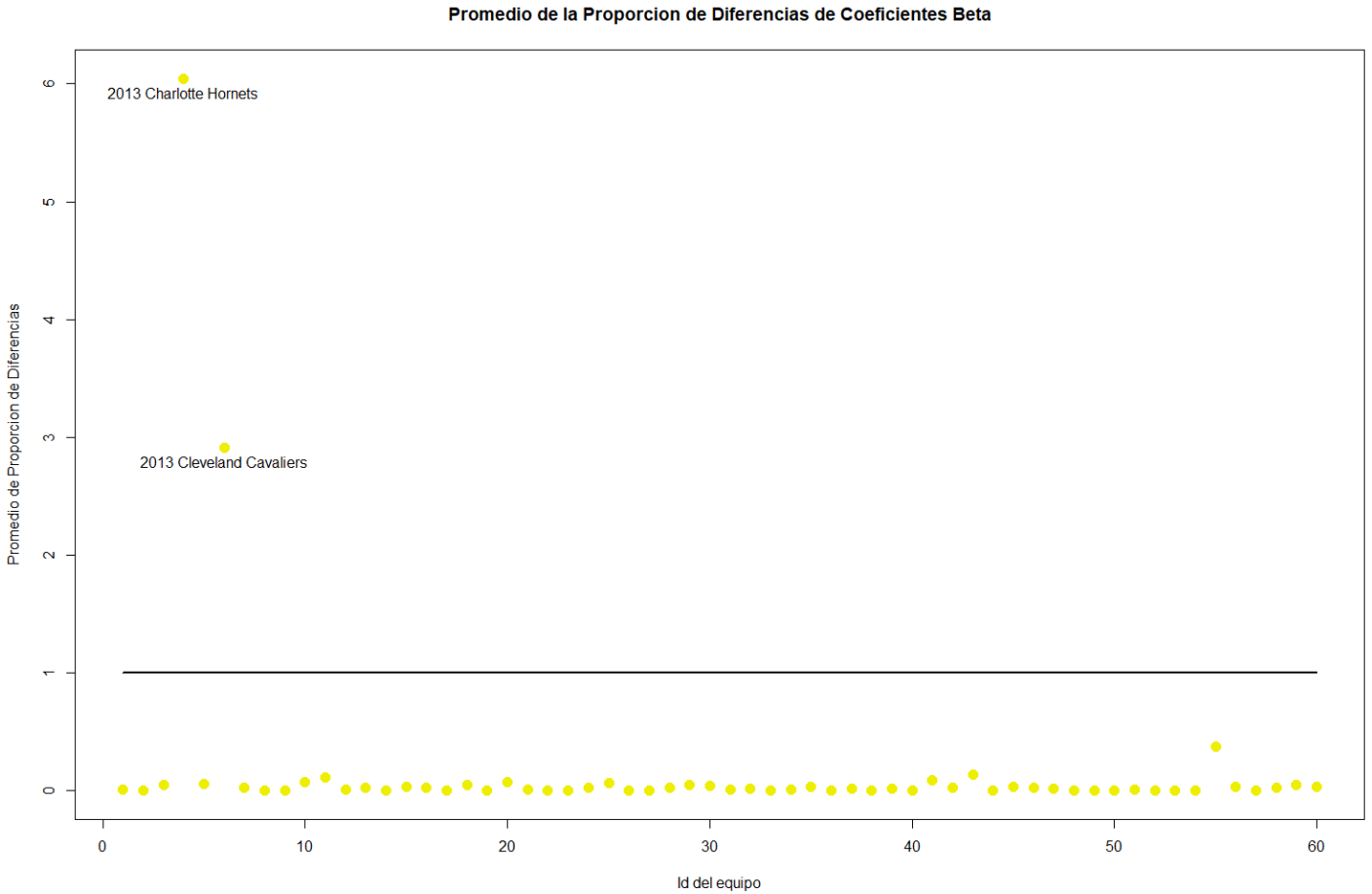

Figura 7.17: Gráfica de la estadística $\overline{\Delta\hat{\beta}_{pj}}^{(-i)}$ para el modelo que utiliza componentes principales.

Con respecto a las gráficas anteriores referentes al modelo que utiliza componentes principales como predictores, se concluyó lo siguiente:

- Respecto a la gráfica correspondiente a $\Delta\chi^{2(-i)}$, existen dos observaciones de especial interés dados los grandes valores que obtuvieron en esta estadística. Estas observaciones son 2013 Cleveland Cavaliers y 2013 Charlotte Hornets. Respecto al equipo residente en Cleveland no es de sorprender que tenga un gran valor en esta estadística, pues era uno de los equipos en los que el modelo se ajustó peor. Sin embargo, el equipo de Charlotte no tuvo residuos grandes, lo que implica que el remover esta observación causó una mejora en el ajuste de las demás observaciones. Por lo tanto, este equipo es un dato influyente. De ser este el caso, se esperaría también que al remover a este equipo del modelo los coeficientes se modificaran ampliamente. La tabla 7.22 señala con mayor detalle la información de ambas observaciones.

Tabla 7.22: Observaciones con grandes valores de $\Delta\chi^{2(-i)}$.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Charlotte Hornets	2	2	0.0007	0.6630	0.3363
2013 Cleveland Cavaliers	3	2	0.0213	0.9122	0.0665

- Respecto a la gráfica correspondiente a $\Delta D^{(-i)}$, se volvieron a presentar las mismas dos observaciones anteriores como *outliers*, más otras dos. Estas observaciones se presentan en la tabla 7.23.

Tabla 7.23: Observaciones con grandes valores de $\Delta D^{(-i)}$.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Charlotte Hornets	2	2	0.0007	0.6630	0.3363
2013 Cleveland Cavaliers	3	2	0.0213	0.9122	0.0665
2015 LA Clippers	2	1	0.8040	0.1960	0.0000
2015 Portland Trail Blazers	1	2	0.1300	0.8699	0.0001

Nótese que estas observaciones son las mismas a las que presentaron peor ajuste en probabilidades, salvo el caso de 2013 Charlotte Hornets. Por lo tanto, sólo se puede concluir que el equipo de Charlotte del año 2013 es un dato influyente, pues de los demás equipos ya se esperaba tuvieran un valor alto en esta estadística.

Por otro lado, se seguirá manteniendo la postura para el equipo 2015 Portland Trail Blazers de que pudiera ser atípico. El por qué de esta postura se explica a continuación.

Durante el 2015 LA Clippers y Portland Trail Blazers se enfrentaron en la primer ronda de *playoffs*, misma ronda que define la categoría de los equipos en este estudio. Al tercer juego LA Clippers llevaba dos de los cuatro partidos necesarios para ganar la serie, mientras que los Trail Blazers sólo llevaban uno. Sin embargo, al inicio del cuarto juego los dos mejores jugadores de LA Clippers sufrieron una lesión que no les permitió volver a jugar durante lo que restaba de la temporada. Posterior a este evento, Trail Blazers ganó los siguientes tres partidos consecutivos, y avanzó así a la siguiente ronda de *playoffs*, hecho que le hizo merecedor a la categoría uno dentro del presente estudio.

El autor del documento concluye que aunque el hipótetico caso de que los dos mejores jugadores de Clippers no se hubieran lesionado no hubiera garantizado una victoria de Clippers en la serie, sí la hubiera facilitado. Esto justifica la alta probabilidad que dio el modelo al equipo de Clippers de ser categorizada en uno, y también, el posible estatus de *outlier* hacia el equipo de Portland para el año 2015.

- Respecto a la gráfica correspondiente a $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$, los coeficientes de los modelos sin alguna de las 60 observaciones conservan aproximadamente los mismos valores que los coeficientes del modelo que contempla todas las observaciones, salvo en dos casos: cuando se elimina del modelo al equipo 2013 Charlotte Hornets y cuando se elimina a 2013 Cleveland Cavaliers.

  Dados estos resultados, se reafirma que el equipo 2013 Charlotte Hornets es un dato influyente. Aunado a esto, 2013 Cleveland Cavaliers también tiene calidad de influyente, pero además de eso, la observación también podría ser *outlier* dados los residuos tan grandes que ha presentado. En caso de que el experto en el área dictamine que el equipo efectivamente es atípico, se podría considerar la eliminación del mismo en el modelo al ser tanto influyente como atípico.

  Dejando de lado a este par de observaciones, el hecho de que los coeficientes no cambien para la mayoría de las observaciones sustraídas es muy bueno, ya que a excepción de estas dos observaciones, el modelo presenta una estabilidad sólida, pues al remover cualquiera de las otras 58 observaciones, el cambio máximo que sufren los coeficientes equivale a un 37 % de la distancia que se permite alejar considerando un intervalo de confianza del 97 %.

  Por lo tanto, en ninguna de las otras 58 observaciones los coeficientes se salen del intervalo de confianza aceptado, mientras que al remover 2013 Cleveland Cavaliers los coeficientes se alejan tres veces la distancia per-

mitida por el intervalo, y al remover 2013 Charlotte Hornets se aleja seis veces esta distancia de lo permitido.

Dados los resultados de las tres gráficas mencionadas anteriormente, se revisarán detalladamente a las observaciones 2015 Portland Trail Blazers y 2013 Cleveland Cavaliers con el fin de comprobar si pueden ser consideradas como observaciones atípicas o no. A la par, se confirma que tanto 2013 Cleveland Cavaliers como 2013 Charlotte Hornets son datos influyentes que afectan de manera negativa al modelo al hacer crecer la Ji cuadrada y la devianza del mismo, y por ende, empeorar el ajuste del modelo.

Después de analizar las observaciones desde un perfil enfocado al baloncesto de la NBA y no desde un perfil estadístico, el autor concluyó que el hecho de ser la única observación con una mal clasifición de gravedad en ambos modelos, más el cómo sucedieron las cosas durante la serie con LA Clippers, son evidencia suficiente para justificar que 2015 Portland Trail Blazers sea un dato atípico; sin embargo, no se encontraron suficientes motivos para decir lo mismo de 2013 Cleveland Cavaliers, por lo que este último tendrá calidad de observación común, y su mal clasificación será debida a algún defecto en el ajuste del modelo.

Así también, ninguna observación será removida del modelo, al no encontrar pruebas que llegaran a justificar dicha acción y ninguna poseer tanto calidad de atípica como de influyente.

Ahora que se ha terminado con las conclusiones de las gráficas *Delta* para el modelo con componentes principales, las gráficas para el modelo que utiliza variables originales como variables predictoras se presentan a continuación:

- $\Delta\chi^{2(-i)}$: Figura 7.16
- $\Delta D^{(-i)}$: Figura 7.17
- $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$: Figura 7.18

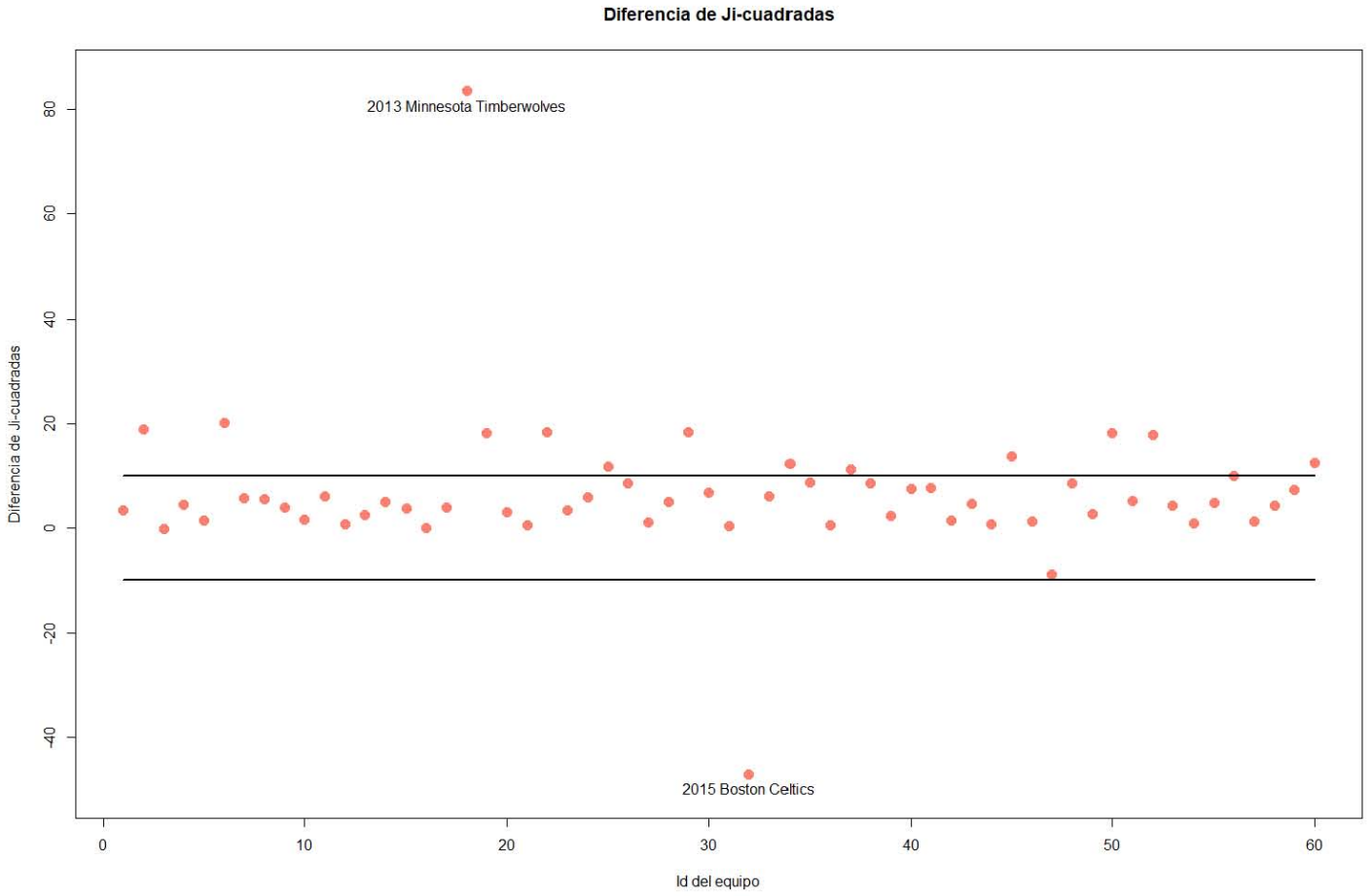

Figura 7.18: Gráfica de la estadística $\Delta\chi^{2(-i)}$ para el modelo que utiliza variables originales como predictores.

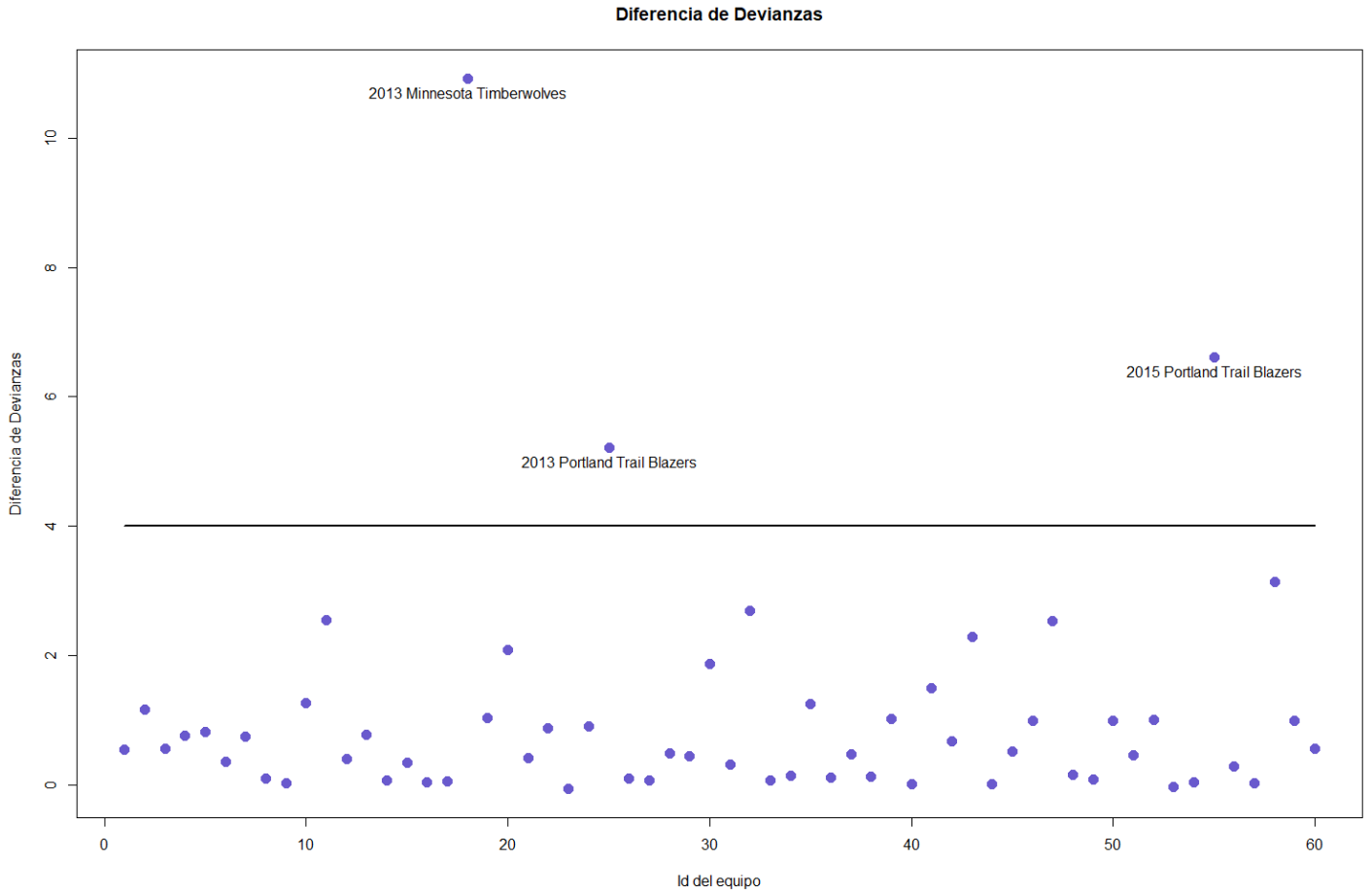

Figura 7.19: Gráfica de la estadística $\Delta D^{(-i)}$ para el modelo que utiliza variables originales como predictores.

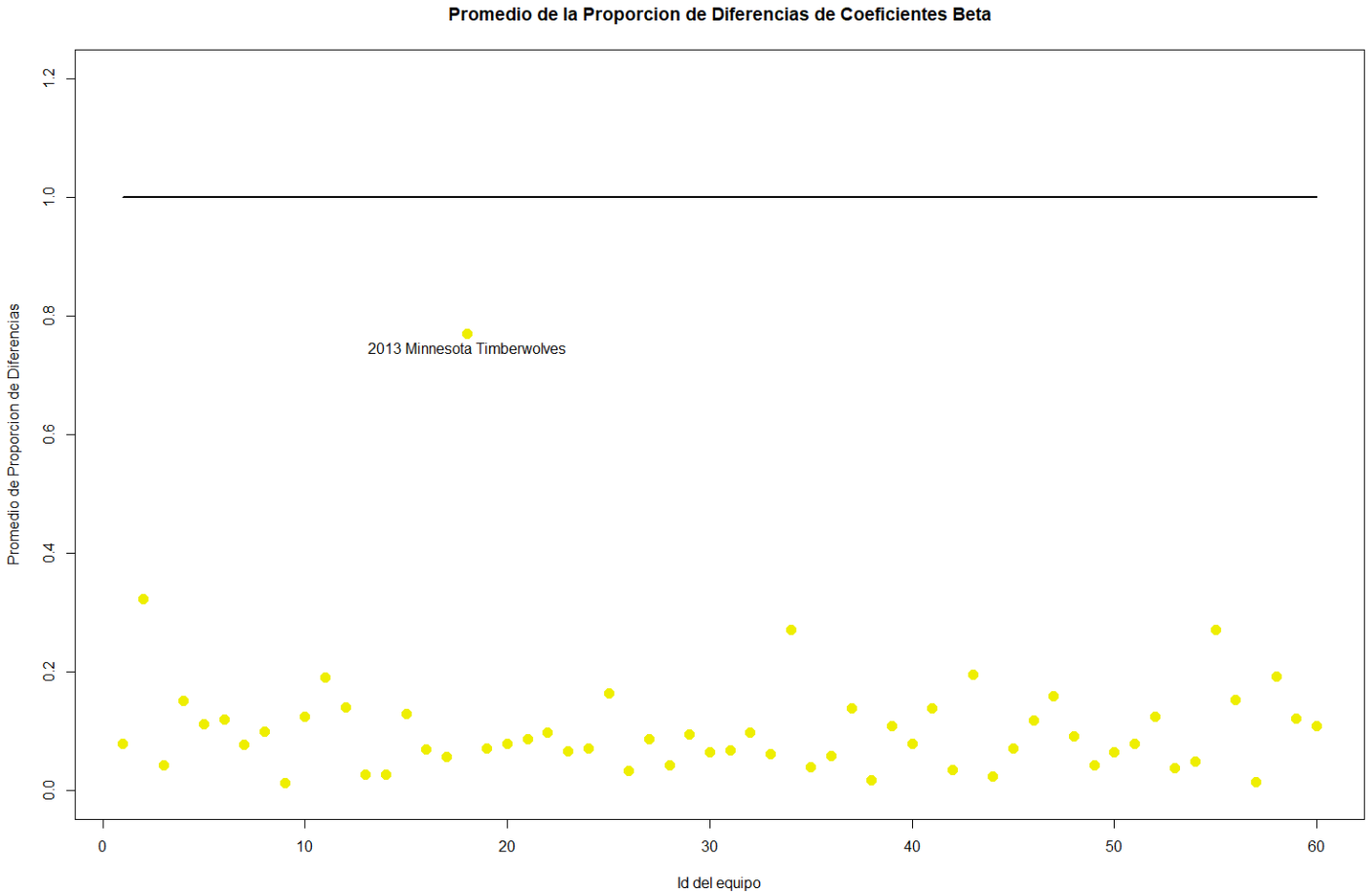

Figura 7.20: Gráfica de la estadística $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$ para el modelo que utiliza variables originales como predictores.

Respecto a las gráficas anteriores referentes al modelo que utiliza variables originales como predictores, se concluyó lo siguiente:

- Respecto a la gráfica correspondiente a $\Delta\chi^{2(-i)}$, se obtienen dos observaciones las cuales si se removieran al ajustar el modelo, éste tendría un cambio enorme en el ajuste referente a la estadística χ^2; dichas observaciones son 2013 Minnesota Timberwolves y 2015 Boston Celtics.

  Los detalles de dichas observaciones se encuentran en la tabla 7.24.

Tabla 7.24: Observaciones con grandes valores de $\Delta\chi^{2(-i)}$.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Minnesota Timberwolves	2	3	0.0002	0.0204	0.9793
2015 Boston Celtics	2	3	0.0110	0.2923	0.6967

  De estas dos observaciones, el equipo de Minnesota influye de manera positiva en $\Delta\chi^{2(-i)}$ mientras que el de Boston influye de manera negativa. Esto quiere decir que, si se llegara a remover la observación referente a Minnesota, el ajuste del nuevo modelo sin la observación i dado por $\chi^{2(-i)}$ sería mucho mejor al de la χ^2 perteneciente al modelo que contiene a todas las observaciones; mientras que si se llegara a remover 2015 Boston Celtics, el ajuste $\chi^{2(-i)}$ del nuevo modelo sin dicha observación sería peor que el ajuste de Ji-cuadrada del modelo que contempla a todas las observaciones, χ^2.

  Mientras que el caso que involucra a 2013 Minnesota Timberwolves es totalmente esperado dados los residuos tan grandes que poseía, el caso de tener observaciones que si se llegaran a eliminar afectaría en gran parte al ajuste del modelo no es lo buscado, pues esto hace de la efectividad del modelo parcialmente dependiente de una sola observación y es una clara señal de falta de robustez del modelo; es decir, el modelo aunque tuviera buenas características de ajuste, sería muy volátil de acuerdo a las observaciones específicas con las que se haya elaborado.

  También, nótese que en general, los valores de $\Delta\chi^{2(-i)}$ tienden a ser de mayor tamaño en este modelo que en el modelo que utiliza componentes principales. Esta como se había mencionado anteriormente no es una característica buena para un modelo.

- Respecto a la gráfica correspondiente a $\Delta D^{(-i)}$, las observaciones que presentaron valores altos son presentadas a continuación en la tabla 7.25.

Tabla 7.25: Observaciones con grandes valores de $\Delta D^{(-i)}$.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2013 Minnesota Timberwolves	2	3	0.0002	0.0204	0.9793
2013 Portland Trail Blazers	1	2	0.1013	0.7889	0.1099
2015 Portland Trail Blazers	1	2	0.0611	0.6804	0.2585

Aunque ya se había decidido que el equipo 2015 Portland Trail Blazers sería un *outlier*, se revisarán próximamente las otras dos observaciones con el fin de determinar si serán consideradas como *outliers* o no.

- Respecto a la gráfica correspondiente a $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$, los coeficientes de los modelos sin alguna de las 60 observaciones conservan aproximadamente los mismos valores que los coeficientes del modelo que contempla todas las observaciones.

  Estas son maravillosas noticias para el modelo, pues da completa certeza de estabilidad y robustez en el mismo, características que son necesarias para confiar en las inferencias realizadas a partir de los coeficientes y que normalmente son de gran importancia para el analista. Asismismo, estos resultados son una fuerte indicación de que el modelo no posee datos influyentes.

  Según la gráfica de $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$, el cambio máximo que podrían tener los coeficientes sería de aproximadamente un 70 % de lo permitido por los intervalos de confianza, y a partir de ese, el porcentaje bajaría a menos de un 40 % al remover cualquier otra observación.

Dados los resultados de las gráficas anteriores, se revisará el status de las observaciones 2013 Minnesota Timberwolves, 2013 Portland Trail Blazers y 2015 Boston Celtics. Sin embargo, al analizar con mayor detalle y desde un punto de vista enfocado al baloncesto de la NBA, no se encontraron evidencias para que alguna de las observaciones mencionadas sea considerada como *outlier*. Algunas de las razones para no categorizarlas en dicho status se mencionan a continuación:

- Respecto a 2013 Minnesota Timberwolves:
  Aunque es de las observaciones con menor nivel pertenecientes a la categoría dos, dicha observación tuvo resultados muy superiores a los de cualquier observación con categoría tres, por lo que claramente no pertenece a dicha categoría. Además, el equipo se desempeñó mejor que un par de observaciones con la misma categoría, garantizando así que no fuera un punto extremo de la categoría dos.

- Respecto a 2013 Portland Trail Blazers:
  El desempeño de este equipo se encontraba entre el límite de los equipos con categoría dos y el de los de categoría uno, por lo que no es de sorprender que terminara en la categoría uno. Por lo tanto, no hay motivos suficientes para considerarlo como *outlier*.

- Respecto a 2015 Boston Celtics:
  El desempeño del equipo fue el correspondiente a un equipo con categoría dos, encontrándose muy alejado de la categoría tres. Por lo tanto, los grandes residuos de la observación son consecuencia de un mal ajuste del modelo hacia la misma, y de ninguna manera se considerará como *outlier*.

- Respecto a 2015 Portland Trail Blazers:
  Los motivos para considerar a este equipo como un *outlier* ya fueron expuestos en las conclusiones de las gráficas para el modelo con componentes principales.

Por lo tanto, solo una de las cuatro observaciones anteriores será considerada como *outlier*; las demás serán consideradas como observaciones comunes y la causa de su mal clasificación se deberá a defectos del modelo en cuestión.

Ahora, respecto a si uno de los dos modelos presentó mejores resultados, se concluye que ambos presentaron sus ventajas y desventajas, pero ninguno se logró sobreponer al otro. En la gráfica de diagnóstico correspondiente a Ji-cuadradas, el desempeño fue mejor para el modelo que utiliza componentes principales, ya que el modelo con variables presentó a muchos puntos con mayor valor; en la gráfica referente a devianza, ambos modelos se comportaron de manera similar; y en cuanto a la gráfica correspondiente a los coeficientes beta, el modelo con variables originales presentó ventajas respecto al de componentes principales. Por lo tanto, se concluye que ambos modelos podrían considerarse con un buen desempeño en las estadísticas de diagnóstico analizadas.

7.4.3. Ajuste por validación externa

Dado que uno de los principales objetivos del estudio es pronosticar el nivel de desempeño de nuevas observaciones, esta sección toma gran importancia para el análisis de los modelos.

Para la realización de dicho ajuste, se tomarán a los equipos y estadísticas de la temporada de la NBA que inició en el año 2017. El motivo por el cuál se decidió tomar la información de esta temporada y no el de una temporada anterior es el mismo motivo por el cuál se decide tomar como observaciones para la elaboración de los modelos a los equipos y estadísticas de los años 2013

y 2015: para evitar lo mayor posible correlaciones altas entre observaciones. La tabla que contiene a los equipos del 2017 así como a sus respectivas estadísticas se puede consultar en el apéndice A.

Ahora pues se analizarán algunas herramientas mencionadas anteriormente, pero con la base de datos de validación.

Cabe mencionar que para esta sección se podría haber utilizado cualquiera de las estadísticas y pruebas aplicadas a los datos originales (a excepción de las estadísticas de diagnóstico de tipo "Delta"). Sin embargo, en este estudio únicamente se utilizaron las que se pensaron podrían ser de mayor utilidad de acorde a uno de los objetivos del análisis, que consiste en conocer el poder predictivo de los modelos.

Complementando el párrafo anterior, el motivo por el cual no se pueden obtener las estadísticas de tipo "Delta" para las observaciones de validación, es porque dichas observaciones no fueron utilizadas para elaborar al modelo, y el objetivo de las estadísticas de este tipo es encontrar, de entre las observaciones utilizadas para la elaboración del modelo, a aquellas que si se llegaran a remover causarían un gran cambio en ajuste o estabilidad.

Antes de iniciar con el análisis de validación externa, es importante recordar un punto importante respecto a las componentes principales utilizadas en el análisis, y es que para obtener las componentes principales de las nuevas observaciones, se debe de realizar el mismo proceso que se hizo con las observaciones originales; es decir, se deben de estandarizar los datos primero, pero utilizando los mismos valores de promedio y desviación estándar utilizados en las observaciones originales. Una vez que se encuentren estandarizados los nuevos datos utilizando los parámetros de las observaciones originales, se puede proceder a calcular las componentes principales de los nuevos datos con los *loadings* obtenidos con la base de datos original y con los datos estandarizados.

Ahora pues, se inicia con el análisis de validación externa. Dado que es de interés el poder predictivo de los modelos, únicamente se verificarán en esta sección las medidas correspondientes a las tablas de clasificación.

Las tablas relacionadas al modelo que utiliza componentes principales como variables predictoras, así como sus resultados, son los siguientes:

1. Tabla de clasificación:
   Tabla 7.26.
   Poder predictivo de validación:
   80 %, correspondiente a 24 observaciones de 30 clasificadas correctamente.
   Tabla de porcentajes de clasificación:
   Tabla 7.27.
   Poder predictivo para las categorías 1,2 y 3:

63 %, 73 % y 100 %, respectivamente.

Tabla 7.26: Tabla de clasificación para observaciones de validación: componentes principales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	5	3	0
	$Y = 2$	2	8	1
	$Y = 3$	0	0	11

Tabla 7.27: Tabla de porcentajes de clasificación para observaciones de validación: componentes principales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	63 %	37 %	0 %
	$Y = 2$	18 %	73 %	9 %
	$Y = 3$	0 %	0 %	100 %

Además, en la tabla 7.28 se pueden consultar las observaciones que fueron clasificadas erróneamente, así como sus respectivas probabilidades $\hat{\pi}_{ij}$.

Tabla 7.28: Observaciones con predicciones erróneas de los datos de validación: Modelo con componentes principales.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2017 Miami Heat	2	1	0.8296	0.1704	0.0000
2017 Milwaukee Bucks	2	3	0.0016	0.0870	0.9114
2017 New Orleans Pelicans	1	2	0.2113	0.7883	0.0003
2017 Oklahoma City Thunder	2	1	0.8096	0.1904	0.0000
2017 Philadelphia 76ers	1	2	0.0889	0.8756	0.0355
2017 Utah Jazz	1	2	0.2957	0.7033	0.0010

Respecto a las tablas de clasificación anteriores, se podría concluir que el modelo que utiliza componentes principales pareciera no haber realizado un trabajo tan sobresaliente en el momento de pronosticar a las observaciones de validación, pues únicamente obtuvo un poder predictivo de 80 %, a diferencia del 85 % que el modelo obtuvo para las observaciones de entrenamiento.

Sin embargo, al analizar con mayor detalle estas observaciones desde un enfoque orientado al baloncesto de la NBA, se ha concluido lo siguiente. Respecto a las observaciones:

- 2017 New Orleans Pelicans
- 2017 Oklahoma City Thunder
- 2017 Utah Jazz

los expertos de la NBA apostaban a que dichas observaciones obtuvieran la misma categoría que la que predijo el modelo; es decir, los expertos esperaban que 2017 New Orleans Pelicans y 2017 Utah Jazz fueran clasificados en la categoría dos, mientras que 2017 Oklahoma City Thunder fuera clasificado en la categoría uno. Por lo tanto, se podría pensar que la categoría de desempeño obtenida por estas observaciones fue "atípica", y su mal clasificación no repercutiría de gran manera en el poder predictivo del modelo.

Respecto a las otras tres observaciones:

- 2017 Miami Heat
- 2017 Milwaukee Bucks
- 2017 Philadelphia 76ers

No se puede decir lo mismo, ya que para el equipo de Miami, mientras que el modelo lo clasificó en la categoría uno dándole una alta probabilidad en esa categoría, se esperaba que las probabilidades de pertenecer a las categorías uno y dos fueran similares, con una ligera ventaja a la categoría dos.

Respecto al equipo de Milwaukee el error fue de mayor gravedad, ya que mientras que el modelo le asignó una probabilidad altísima de pertenecer a la categoría tres, el equipo en realidad fue clasificado en la categoría dos, y más aún, el equipo se quedó a casi nada de ser clasificado en la categoría uno, por lo que el ajuste en esta observación fue fatal.

Respecto al equipo de Philadelphia, el error también fue grave aunque no tanto como el de Milwaukee. Mientras se esperaba que Philadelphia tuviera probabilidades muy similares de pertenecer a la categoría uno y a la categoría dos, y su probabilidad de pertenecer a la categoría tres fuera prácticamente nula, el modelo ha clasificado al equipo enteramente en la categoría dos, dándole un peso insignificante tanto a las categorías uno como a la tres.

Terminado el análisis observación por observación, se ha concluido que aunque no se vea reflejado en el poder predictivo, el modelo ha tenido éxito al clasificar a las observaciones de validación, pues aunque el 20 % de sus observaciones fueron clasificadas incorrectamente, sólo un 10 % de los errores fueron graves, mientras el otro 10 % de los errores no fueron de gravedad.

Por lo tanto, se ha concluido que el modelo ha dado resultados agradables en cuanto a poder predictivo de validación.

A continuación se presentan las tablas relacionadas al modelo que utiliza a las variables originales como variables predictoras para las observaciones de validación:

1. Tabla de clasificación:
   Tabla 7.29.
   Poder predictivo de validación:
   87 %, correspondiente a 26 observaciones de 30 clasificadas correctamente.
   Tabla de porcentajes de clasificación:
   Tabla 7.30.
   Poder predictivo para las categorías 1,2 y 3:
   75 %, 91 % y 91 %, respectivamente.

Tabla 7.29: Tabla de clasificación para observaciones de validación: variables originales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	6	2	0
	$Y = 2$	1	10	0
	$Y = 3$	0	1	10

Tabla 7.30: Tabla de porcentajes de clasificación para observaciones de validación: variables originales

Categorías		Predecidos		
		$Y = 1$	$Y = 2$	$Y = 3$
Observados	$Y = 1$	75 %	25 %	0 %
	$Y = 2$	9 %	91 %	0 %
	$Y = 3$	0 %	9 %	91 %

La tabla 7.31 corresponde a las observaciones que fueron clasificadas erróneamente y sus respectivas probabilidades $\hat{\pi}_{ij}$.

Tabla 7.31: Observaciones con predicciones erróneas de los datos de validación: modelo con variables originales.

Equipo	Categoría	Predicción	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$	$\hat{\pi}_{i3}$
2017 Dallas Mavericks	3	2	0.1451	0.7945	0.0605
2017 New Orleans Pelicans	1	2	0.1787	0.7672	0.0541
2017 Philadelphia 76ers	1	2	0.4242	0.5741	0.0017
2017 San Antonio Spurs	2	1	0.7244	0.2756	0.0000

Al clasificar al 87 % del total de observaciones de validación correctamente se confirma que el modelo tiene una excelente capacidad de clasificación para observaciones nuevas. Más aún, de estas cuatro observaciones, la mal clasificación de dos de ellas, a nombrar:

- 2017 New Orleans Pelicans
- 2017 Philadelphia 76ers

no se consideran de gravedad. Los motivos que respaldan esta afirmación para el equipo de Nueva Orleans ya se han mencionado en el modelo anterior, dado que las probabilidades asociadas al equipo para ambos modelos fueron muy similares.

En cuanto al equipo de Philadelphia, también se había mencionado la predicción ideal para dicho equipo hubiera sido el dar un peso similar a las probabilidades asociadas a las categorías uno y dos, con una ligera ventaja para la categoría uno y dejando la probabilidad asociada a la categoría tres casi nula. En este caso pues, se ha cumplido que tanto las probabilidades asociadas a las categorías uno y dos fueran similares como que la probabilidad para la categoría tres fuera casi nula; sin embargo en vez de dar una ligera ventaja a la categoría uno, se le dio a la categoría dos. Es por eso que dicho equipo fue clasificado incorrectamente, sin embargo las probabilidades se acercaron mucho a la realidad, por lo que no se tomará a este error como uno de mucha gravedad.

Respecto a los equipos restantes:

- 2017 Dallas Mavericks
- 2017 San Antonio Spurs

Dichos equipos sí muestran mayor gravedad en su mal clasificación. Los motivos se presentan a continuación.

Para el caso del equipo de Dallas, el ajuste del modelo fue pésimo, ya que la probabilidad de que perteneciera a la categoría tres fue de únicamente 0.06, siendo que Dallas fue uno de los cuatro peores equipos en toda la liga, por lo cual se hubiera esperado que la probabilidad asociada a la categoría tres fuera mucho mayor a cualquier otra probabilidad.

Para el caso del equipo residente en San Antonio, el error no es tan grave como el del equipo anterior, aunque sí se hubiera esperado que la probabilidad mayor fuera la asociada a la categoría dos, seguida no muy de cerca de la categoría uno y con la probabilidad asociada a la categoría tres casi nula.

Terminado el análisis observación por observación, se ha concluido entonces que el poder predictivo del modelo ha sido excepcional, ya que presentó un 13 % de errores en las clasificaciones y sólo la mitad de ese porcentaje corresponde a errores importantes.

Como conclusión final de esta sección, ambos modelos muestran buenos resultados en su poder predictivo; sin embargo, el modelo que utiliza a las variables originales como predictores se ha desempeñado mejor que el modelo que utiliza componentes principales tanto en número de clasificaciones incorrectas, como en gravedad de dichas clasificaciones.

7.5. Conclusiones e interpretaciones

7.5.1. Conclusiones

A continuación se muestra una tabla comparativa de ambos modelos, esto con el fin de decidir cual de los dos será considerado como la mejor opción para el análisis de acuerdo a los dos objetivos del estudio.

Tabla 7.32: Tabla comparativa entre el modelo que utiliza componentes principales como predictores y el modelo que utiliza a variables originales como predictores.

Estadística	Modelo Comp. Principales	Modelo Variables
p-value C	0.9879	0.3170
Tablas clasificación	51 Aciertos (85 %)	50 Aciertos (83 %)
AUC	0.9574, 0.9288 y 0.9928	0.9232, 0.8729 y 0.9681
Pseudo-R^2 Pearson	0.7456	0.6116
Pseudo-R^2 Suma cuadrados	0.7456	0.6107
Pseudo-R^2 Log-verosimilitud	0.7305	0.5864
Errores graves $r(y_{ij}, \hat{\pi}_{ij})$	3 (5 %)	3 (5 %)
Errores graves d_i	3 (5 %)	3 (5 %)
Errores graves $\Delta\chi^{2(-i)}$	2 (3 %)	2 (3 %)
Errores graves $\Delta D^{(-i)}$	4 (7 %)	3 (5 %)
Errores graves $\overline{\Delta\hat{\beta}_{pj}}^{(-i)}$	2 (3 %)	0 (0 %)
Outliers	1 (2 %)	1 (2 %)
Datos influyentes	2 (3 %)	0 (0 %)
Errores v. externa	6 (20 %)	4 (13 %)
Errores graves v. externa	3 (10 %)	2 (7 %)

Como nota adicional a la tabla anterior, los valores han sido redondeados a cuatro decimales.

Ahora pues, ha llegado el momento de decidir cual será el modelo con el que se reportarán resultados y realizarán inferencias e interpretaciones. Dicho modelo será elegido con base en los objetivos del estudio que a continuación se vuelven a recordar:

1. Predecir el nivel de desempeño de futuros equipos.
2. Determinar los factores que influyen en el desempeño de un equipo de la NBA y su grado de influencia en el.

Con base en estos objetivos, y dados los resultados mostrados en la tabla 7.32, se ha optado como mejor modelo y con el cual se realizarán inferencias respecto a sus parámetros, al modelo que utiliza componentes principales como variables predictoras. Los motivos de esta elección se presentan a continuación:

Se ha comprobado mediante las técnicas utilizadas a lo largo de esta sección, que el comportamiento de ambos modelos ha sido sobresaliente tanto en el ámbito de ajuste como en el de poder predictivo, y que se podrían realizar inferencias respecto a sus coeficientes sin ningún problema dada la gran significancia que presentaron sus variables (0.03 y 0.02, respectivamente) más el hecho de que dichos coeficientes se modifiquen de manera casi despreciable, corroborado en las gráficas de la estadística $\overline{\hat{\beta}_{pj}^{(-i)}}$.

Por dichas razones, se podría utilizar cualquiera de los dos modelos para realizar inferencias y así cumplir con los dos objetivos del estudio satisfactoriamente; sin embargo, se ha optado por utilizar el modelo con componentes principales sobre el modelo que utiliza a las variables originales dado que el primero tiene la gran ventaja de poder realizar inferencias respecto a todas las variables, ya que cada componente está conformada por todas éstas, mientras que si se utilizara al modelo con variables originales como variables predictoras únicamente se podrían realizar inferencias sobre tres del total de variables, cuyo número se recuerda es 15.

Esta decisión ha sido personal y podría variar de acuerdo a los objetivos del análisis; sin embargo, viendo este estudio desde un punto de vista de una gerencia de algún equipo profesional, sería de gran interés conocer el nivel de influencia de la mayor cantidad de variables posibles, pues así podrían maximizar sus oportunidades, basadas en el modelo, de alcanzar alguna categoría deseada.

Cabe mencionar que bajo el mismo punto de vista gerencial, también influiría el factor económico del equipo, ya que si por ejemplo, el mejorar alguna estadística implicara destinar cantidades monetarias muy grandes y el equipo fuera un equipo pequeño, se tendría que recurrir a mejorar alguna otra estadística que tuviera un impacto similar en el desempeño, pero que fuera a requerir una cantidad monetaria mucho menor. Es por esto que sería muy provechoso para la

gerencia conocer el efecto en el desempeño del equipo de la mayor cantidad de estadísticas posibles, y es por ello que se ha decidido utilizar al modelo con componentes principales como el modelo final y en el cual se realizarán inferencias.

A continuación se menciona una recomendación que, aunque no es de importancia en el estudio actual dado que no se eliminó ninguna observación del modelo, sería de gran importancia tomar en cuenta en situaciones donde sí se realice:

Si se quisiera elaborar un nuevo modelo ignorando a algún dato atípico, dado que se han realizado dos técnicas estadísticas que involucran al conjunto entero de datos, podría haber dos maneras posibles de realizarlo. Estas son:

- Simplemente elaborar el modelo de regresión logística multinomial sin tomar en cuenta a la observación.
- Eliminar primero a la observación, posteriormente calcular las componentes principales y a continuación elaborar el modelo de regresión logística multinomial.

El primer punto sería bajo el pensamiento de que la observación únicamente es *outlier* bajo esa transformación de las componentes principales, mientras la segunda se fundamenta en que el dato también se comporta como atípico para el análisis de componentes principales, y por lo tanto tiene que ser removido antes de calcularlos.

Otro gran punto que separa a estos dos posibles caminos para elaborar el nuevo modelo sin la observación sería que, mientras el primer camino estandariza a las variables originales para obtener sus componentes principales respecto al total de observaciones, el segundo camino las estandariza sin tomar en cuenta a la observación eliminada.

Personalmente, el autor del documento sugiere al segundo camino como el más favorable en estudios que combinan a ambas técnicas estadísticas, sin embargo esta decisión seguramente dependerá de las características particulares de la observación y el tipo de estudio que se desee realizar.

7.5.2. Interpretaciones

Dado que los objetivos del análisis en cuestión consistían en lo siguiente:

1. Predecir el nivel de desempeño de futuros equipos.
2. Determinar los factores que influyen en el desempeño de un equipo de la NBA y su grado de influencia en el.

Se puede asegurar ahora que el primer objetivo se ha logrado exitosamente mediante el modelo oficial, al registrar un poder predictivo de 85 % para las observaciones de entrenamiento y un 80 % para las observaciones de validación, donde del 20 % de predicciones incorrectas, únicamente el 10 % fue de gravedad.

Teoría adicional necesaria para la interpretación de resultados

Ahora se prosigue a completar el segundo objetivo del estudio, que consiste en determinar los factores que mayor influencia tienen en el desempeño de un equipo de la NBA, así como el grado de influencia que poseen.

Para la realización de dicho objetivo, se utilizarán algunas de las herramientas mencionadas en el capítulo 4. Estas herramientas son los llamados "momios" y "cocientes de riesgos relativos".

En el capítulo mencionado, se había concluido que el momio entre el patrón de covariables i con la categoría j, y el patrón de covariables v con la categoría q, se calculaba como:

$$\frac{\hat{\pi}_{ij}}{\hat{\pi}_{vq}} = \frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}$$

con $j, q = 1, 2, \ldots, J$ y $i, v = 1, 2, \ldots, M$.

Y si el valor de dicho momio fuese algún número a, dicho momio se podría interpretar como "es a veces más probable que una observación del patrón de covariables i sea catalogado en la categoría j, a que una observación del patrón de covariables v sea catalogado en la categoría q".

Mientras que el cociente de riesgos relativos entre el patrón de covariables i con la categoría j y el patrón de covariables v con la categoría q, denotado como "$RRR_{(i,j),(v,q)}$", se expresaba de la forma:

$$\frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vq}}{\hat{\pi}_{vJ}}} = \frac{\frac{\frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pk}}}}}{\frac{\frac{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq}}}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}{\frac{1}{1+\sum_{k=1}^{J-1} e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pk}}}}} = \frac{e^{\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj}}}{e^{\sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq}}} = e^{\left(\sum_{p=0}^{P} x_{ip}\hat{\beta}_{pj} - \sum_{p=0}^{P} x_{vp}\hat{\beta}_{pq}\right)}$$

con $j, q = 1, 2, \ldots, J$ y $i, v = 1, 2, \ldots, M$.

Y si el valor de $RRR_{(i,j),(v,q)}$ fuese algún número a, se podría interpretar como "es a veces más probable que una observación dentro del patrón de covariables i sea catalogado en la categoría j y no en la J, a que una observación con el patrón de covariables v sea catalogado en q y no en J".

Esta interpretación será el pilar para el segundo objetivo del análisis. Ahora, dado que estas herramientas se obtienen a partir de los coeficientes de las variables del modelo en cuestión, las llamadas "$\hat{\beta}_{pj}$", es muy importante tener certeza de que dichos coeficientes sean lo suficientemente estables como para realizar inferencias respecto a ellos. En este caso, el hecho de que los coeficientes $\hat{\beta}_{pj}$ son lo suficientemente estables ya se comprobó mediante la gráfica de la estadística $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$ y pruebas de significancia de variables.

Ahora que se tiene plena seguridad en que las inferencias e interpretaciones realizadas mediante los momios y el cociente de riesgos relativos del presente modelo serán confiables, se procede a obtener la fórmula general para el cálculo del cociente de riesgos relativos de dicho modelo.

La fórmula general del cociente de riesgos relativos vista anteriormente, y aplicada para este caso en particular, queda denotada por:

$$RRR_{(i,j),(v,q)} = \frac{\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}}{\frac{\hat{\pi}_{vq}}{\hat{\pi}_{vJ}}} = \frac{e^{\sum_p CP_{ip}\hat{\beta}_{pj}}}{e^{\sum_p CP_{vp}\hat{\beta}_{pq}}} = e^{\left(\sum_p CP_{ip}\hat{\beta}_{pj} - \sum_p CP_{vp}\hat{\beta}_{pq}\right)}$$

con CP_{ip} haciendo referencia al valor de la p-ésima componente principal para la observación i y $\hat{\beta}_{pj}$ haciendo referencia al coeficiente asociado a la componente principal p para la categoría j.

Otras observaciones importantes para este caso en particular son:

- Los valores i y v pueden indicar los valores de cualquier observación (la observación no necesariamente debe de ser de algún patrón utilizado en la elaboración del modelo).
- $J = 1$. Es decir, la categoría uno es la categoría de referencia.
- $j, q = 2, 3$
- El indicador p puede tomar los valores $p = \{1, 2, 3, 15\}$, que son los números asociados a las componentes principales que fueron elegidas como variables predictoras del modelo oficial.

Sin embargo, si se utilizara dicha ecuación para hacer inferencias en el presente modelo, dichas inferencias serían realizadas respecto a las componentes

principales y, para cumplir con el segundo objetivo del estudio, se necesitan hacer inferencias respecto a las variables originales, no respecto a las componentes.

Lo que se hará entonces será ordenar a los valores de la ecuación del cociente de riesgos relativos anterior de tal manera que se puedan realizar inferencias respecto a las variables originales y no respecto a las componentes. Es importante dejar en claro que dicho proceso no modificará el valor de dicho cociente, únicamente se acomodarán de una manera diferente los valores pertenecientes a dicha ecuación.

Se procede pues a realizar dicho ordenamiento. Se sabe que las componentes principales son resultado de una combinación lineal de diferentes variables. Dichas variables pueden encontrarse estandarizadas o no; para este estudio, sí se estandarizaron las variables para obtener las componentes principales. Por lo tanto, la k-ésima componente principal asociada a la i-ésima observación (CP_{ik}) se puede representar de la manera:

$$CP_{ik} = l_{k1}x^s_{i1} + l_{k2}x^s_{i2} + \ldots + l_{kP}x^s_{iP} \tag{7.1}$$

donde l_{kp} hace referencia al *loading* que la k-ésima componente principal destina para la p-ésima variable, y x^s_{ip} hace referencia al valor estandarizado de la p-ésima variable para la observación i.

Es decir, para el estudio en cuestión, la k-ésima componente principal asociada a la i-ésima observación se obtiene mediante la ecuación:

$$\begin{aligned} CP_{ik} &= l_{k1}x^s_{i1} + l_{k2}x^s_{i2} + \ldots + l_{k22}x^s_{i22} \\ &= \sum_{p=1}^{22} l_{kp}x^s_{ip} \\ &= \sum_{p=1}^{22} l_{kp}\left(\frac{x_{ip} - \overline{x}_{\cdot p}}{\mathbf{SE}(x_{\cdot p})}\right) \end{aligned}$$

donde:

- x_{ip} representa a el valor de la p-ésima variable para la observación i.
- $\overline{x}_{\cdot p}$ representa el promedio de la p-ésima variable.
- $\mathbf{SE}(x_{\cdot p})$ representa la desviación estándar de la p-ésima variable.
- El número 22 corresponde al número de variables que conforman a las componentes principales.

Se debe de aclarar el por qué si en el capítulo cinco se habían presentado 15 variables, la ecuación anterior menciona 22. Esto es debido a que de las 22 variables mencionadas en la ecuación, las primeras 15 corresponden a las variables

originales, mientras que las siete restantes corresponden a las interacciones elegidas.

Otro punto importante a aclarar es que, aunque la observación i puede tanto pertenecer al conjunto de observaciones de entrenamiento como no, los promedios y desviaciones estándar utilizados en las componentes sí deben de ser los correspondientes a las observaciones de la base de datos de entrenamiento.

Ahora que se han definido a las componentes principales en términos de las variables originales, se procede a explicar en términos de las mismas variables

originales a las probabilidades $\hat{\pi}_{ij}$ para el presente estudio.

$$
\begin{aligned}
\hat{\pi}_{ij} &= \frac{\exp\left(\hat{\beta}_{0j} + \sum_{p=1}^{P} CP_{ip}\hat{\beta}_{pj}\right)}{1 + \sum_{a=1}^{J-1} \exp\left(\hat{\beta}_{0j} + \sum_{p=1}^{P} CP_{ip}\hat{\beta}_{pa}\right)} \\
&= \frac{\exp\left(\hat{\beta}_{0j} + \sum_{p} CP_{ip}\hat{\beta}_{pj}\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\beta}_{0a} + \sum_{p} CP_{ip}\hat{\beta}_{pa}\right)} \\
&= \frac{\exp\left(\hat{\beta}_{0j} + \sum_{p}\left(\sum_{w=1}^{22} l_{pw}x_{iw}^{s}\right)\hat{\beta}_{pj}\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\beta}_{0a} + \sum_{p}\left(\sum_{w=1}^{22} l_{pw}x_{iw}^{s}\right)\hat{\beta}_{pa}\right)} \\
&= \frac{\exp\left(\hat{\beta}_{0j} + (l_{11}x_{i1}^{s}\hat{\beta}_{1j} + \ldots + l_{1,22}x_{i22}^{s}\hat{\beta}_{1j}) + \ldots + (l_{15,1}x_{i1}^{s}\hat{\beta}_{15j} + \ldots + (l_{15,22}x_{i22}^{s}\hat{\beta}_{15j})\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\beta}_{0a} + (l_{11}x_{i1}^{s}\hat{\beta}_{1a} + \ldots + l_{1,22}x_{i22}^{s}\hat{\beta}_{1a}) + \ldots + (l_{15,1}x_{i1}^{s}\hat{\beta}_{15a} + \ldots + (l_{15,22}x_{i22}^{s}\hat{\beta}_{15a})\right)} \\
&= \frac{\exp\left(\hat{\beta}_{0j} + \left(\sum_{p} l_{p1}\hat{\beta}_{pj}\right)x_{i1}^{s} + \left(\sum_{p} l_{p2}\hat{\beta}_{pj}\right)x_{i2}^{s} + \ldots + \left(\sum_{p} l_{p22}\hat{\beta}_{pj}\right)x_{i22}^{s}\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\beta}_{0a} + \left(\sum_{p} l_{p1}\hat{\beta}_{pa}\right)x_{i1}^{s} + \left(\sum_{p} l_{p2}\hat{\beta}_{pa}\right)x_{i2}^{s} + \ldots + \left(\sum_{p} l_{p22}\hat{\beta}_{pa}\right)x_{i22}^{s}\right)} \\
&= \frac{\exp\left(\hat{\beta}_{0j} + \sum_{w=1}^{22}\left(\left(\sum_{p} l_{pw}\hat{\beta}_{pj}\right)x_{iw}^{s}\right)\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\beta}_{0a} + \sum_{w=1}^{22}\left(\left(\sum_{p} l_{pw}\hat{\beta}_{pa}\right)x_{iw}^{s}\right)\right)} \\
&= \frac{\exp\left(\hat{\alpha}_{0j} + \sum_{w=1}^{22} \hat{\alpha}_{wj}x_{iw}^{s}\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\alpha}_{0a} + \sum_{w=1}^{22} \hat{\alpha}_{wa}x_{iw}^{s}\right)} \\
&= \frac{\exp\left(\hat{\alpha}_{0j} + \sum_{w=1}^{22} \hat{\alpha}_{wj}\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right)}{1 + \sum_{a=2}^{3} \exp\left(\hat{\alpha}_{0a} + \sum_{w=1}^{22} \hat{\alpha}_{wa}\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right)}
\end{aligned}
$$

con

$$\hat{\alpha}_{wj} = \begin{cases} \hat{\beta}_{0j}, & \text{si } w = 0 \\ \sum_p l_{pw}\hat{\beta}_{pj}, & \text{si } w = \{1, 2, \ldots, 22\} \end{cases}$$

Es importante recordar que $p = \{1, 2, 3, 15\}$, números que hacen referencia a las componentes principales utilizadas en el modelo.

Una vez definidas las probabilidades $\hat{\pi}_{ij}$ en términos de las variables originales, se puede proceder a realizar el mismo tratamiento para la herramienta de mayor relevancia en esta sección: el cociente de riesgos relativos.

Así pues, el cociente de riesgos relativos entre la observación i con la categoría j y la observación v con la categoría q, queda denotado como:

$$\begin{aligned} RRR_{(i,j),(v,q)} &= \frac{\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}\right)}{\left(\frac{\hat{\pi}_{vq}}{\hat{\pi}_{vJ}}\right)} \\ &= \frac{\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i1}}\right)}{\left(\frac{\hat{\pi}_{vq}}{\hat{\pi}_{v1}}\right)} \\ &= \frac{\exp\left(\hat{\alpha}_{0j} + \sum_{w=1}^{22} \hat{\alpha}_{wj}\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right)}{\exp\left(\hat{\alpha}_{0q} + \sum_{w=1}^{22} \hat{\alpha}_{wq}\left(\frac{x_{vw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right)} \\ &= \exp\left\{(\hat{\alpha}_{0j} - \hat{\alpha}_{0q}) + \sum_{w=1}^{22}\left[\hat{\alpha}_{wj}\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right) - \hat{\alpha}_{wq}\left(\frac{x_{vw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\} \\ &= \exp\left\{(\hat{\beta}_{0j} - \hat{\beta}_{0q}) + \sum_{w=1}^{22}\left[\left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right) - \left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{vw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\} \end{aligned}$$

Aunque la forma general de este cociente tiene una apariencia algo complicada, existen algunos casos particulares donde dicha ecuación se llega a simplificar. Los casos particulares que cumplan con esa característica, y que además sean relevantes para el estudio, se mencionan a continuación.

- Caso $j = q$

$$\begin{aligned} RRR_{(i,j),(v,j)} &= \exp\left\{\sum_{w=1}^{22} \hat{\alpha}_{wj}\left[\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)-\left(\frac{x_{vw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\} \\ &= \exp\left\{\sum_{w=1}^{22}\left[\hat{\alpha}_{wj}\frac{1}{\mathbf{SE}(x_{\cdot w})}\left(x_{iw}-x_{vw}\right)\right]\right\} \\ &= \exp\left\{\sum_{w=1}^{22}\left[\left(\sum_{p} l_{pw}\hat{\beta}_{pj}\right)\frac{1}{\mathbf{SE}(x_{\cdot w})}\left(x_{iw}-x_{vw}\right)\right]\right\} \end{aligned}$$

- Caso $i = v$

$$\begin{aligned} RRR_{(i,j),(i,q)} &= \exp\left\{(\hat{\alpha}_{0j}-\hat{\alpha}_{0q})+\sum_{w=1}^{22}\left[\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)(\hat{\alpha}_{wj}-\hat{\alpha}_{wq})\right]\right\} \\ &= \exp\left\{(\hat{\beta}_{0j}-\hat{\beta}_{0q})+\sum_{w=1}^{22}\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\left[\left(\sum_{p} l_{pw}\hat{\beta}_{pj}\right)-\left(\sum_{p} l_{pw}\hat{\beta}_{pq}\right)\right]\right\} \\ &= \exp\left\{(\hat{\beta}_{0j}-\hat{\beta}_{0q})+\sum_{w=1}^{22}\left(\frac{x_{iw}-\overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\left[\sum_{p} l_{pw}\left(\hat{\beta}_{pj}-\hat{\beta}_{pq}\right)\right]\right\} \end{aligned}$$

Nótese también que

$$\begin{aligned} RRR_{(i,j),(i,q)} &= \frac{\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}\right)}{\left(\frac{\hat{\pi}_{iq}}{\hat{\pi}_{iJ}}\right)} \\ &= \left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iq}}\right)\left(\frac{\frac{1}{\hat{\pi}_{iJ}}}{\frac{1}{\hat{\pi}_{iJ}}}\right) \\ &= \frac{\hat{\pi}_{ij}}{\hat{\pi}_{qj}} \end{aligned}$$

Por lo que el cociente de riesgos relativos entre mismas observaciones y diferentes categorías tiene la misma interpretación que el momio de las mismas. Esto es de utilidad en el sentido de que la interpretación de momios es más sencilla que la del cociente de riesgos relativos.

Ahora, nótese que:

$$\begin{aligned}
\frac{\hat{\pi}_{iJ}}{\hat{\pi}_{ij}} &= \left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}\right)^{-1} \\
&= \left(\exp\left\{\hat{\alpha}_{0j} + \sum_{w=1}^{22} \hat{\alpha}_{wj}\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right\}\right)^{-1} \\
&= \exp\left\{-\left[\hat{\alpha}_{0j} + \sum_{w=1}^{22} \hat{\alpha}_{wj}\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\} \\
&= \exp\left\{-\left[\hat{\beta}_{0j} + \sum_{w=1}^{22}\left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\}
\end{aligned}$$

Por lo que si en el presente estudio se deseara realizar inferencias del cociente de riesgos relativos respecto a la categoría de referencia J con alguna otra j y no al revés, dicha fórmula quedaría descrita por:

$$\begin{aligned}
RRR_{(i,J),(v,J)} &= \frac{\left(\frac{\hat{\pi}_{iJ}}{\hat{\pi}_{ij}}\right)}{\left(\frac{\hat{\pi}_{vJ}}{\hat{\pi}_{vq}}\right)} \\
&= \frac{\exp\left\{-\left[\hat{\beta}_{0j} + \sum\limits_{w=1}^{22}\left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\}}{\exp\left\{-\left[\hat{\beta}_{0q} + \sum\limits_{w=1}^{22}\left(\sum_p l_{pw}\hat{\beta}_{pq}\right)\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\}} \\
&= \exp\left\{-(\hat{\beta}_{0j} - \hat{\beta}_{0q}) - \sum_{w=1}^{22}\left[\left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{iw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right) - \left(\sum_p l_{pw}\hat{\beta}_{pj}\right)\left(\frac{x_{vw} - \overline{x}_{\cdot w}}{\mathbf{SE}(x_{\cdot w})}\right)\right]\right\} \\
&= \left(RRR_{(i,j),(v,q)}\right)^{-1}
\end{aligned}$$

Es decir, este cociente es equivalente a la inversa del cociente de riesgos relativos entre la observación i con categoría j y la observación v con categoría q.

Interpretación de resultados

Ha llegado el momento de aplicar las herramientas anteriores a un caso específico y obtener inferencias respecto a ellas.

Para realizar dicho trabajo, se tomarán como ejemplo las estadísticas del equipo "Phoenix Suns" del año 2017, para hacer inferencias respecto a la temporada próxima a iniciar, la temporada 2018-2019. Las estadísticas de dicho equipo para el año 2017, así como otros valores de importancia para el estudio como

los *loadings* para cada una de las cuatro componentes principales a utilizar, los valores de cada uno de los coeficientes $\hat{\beta}_{pj}$, los promedios y las desviaciones estándar de cada variable, se muestran a continuación:

- Valores de la observación 2017 Phoenix Suns, así como promedios y desviaciones estándar de cada variable: Tabla 7.33.
- *Loadings* de las componentes principales elegidas: Tabla 7.34.
- Coeficientes $\hat{\beta}$ del modelo para las componentes principales: Tabla 7.35.
- Nuevos coeficientes $\hat{\alpha}$ asociados a las variables originales: Tabla 7.36.

Cabe mencionar que aunque el efecto encontrado en los cocientes de riesgos relativos es lineal (es decir, el efecto es el mismo independientemente del valor que tenga la variable) y en realidad no se tendría necesidad de utilizar los valores de alguna observación como ejemplo, se utilizará al equipo mencionado anteriormente con el fin de brindar mayor claridad, así como con el fin de proyectar un mejor entendimiento del efecto de las interacciones en dicho modelo.

Tabla 7.33: Valores del equipo 2017 Phoenix Suns.

# Var.	Estadística	Valor	Promedio	Desviación Est.
-	Categoría	3	-	-
Var1	Prop. Hist. Coach	0.2870	0.4986	0.1340
Var2	Jug. con experiencia en semifinales	0.00	1.92	2.01
Var3	Jug. 3 o más años de antigüedad	3.00	3.03	1.52
Var4	Minutos veteranos	25.50	39.12	6.65
Var5	% Tiros de campo efectivos	49.50	50.19	1.98
Var6	Puntos	103.90	101.84	4.05
Var7	Prop. puntos en la pintura	0.4437	0.4206	0.0342
Var8	% Tiros de campo del oponente	47.10	45.33	1.30
Var9	Puntos del oponente	113.30	101.84	3.99
Var10	Clutch 1	-1.50	-0.0117	1.00
Var11	Clutch 2	88.90	80.90	21.73
Var12	Clutch 3	0.3000	0.4783	0.2457
Var13	Clutch 4	3.60	4.45	0.6482
Var14	Clutch 5	28.60	28.39	7.88
Var15	Clutch 6	0.4090	0.5011	0.1339
Var16	Interacción coach y semifinales	0.00	1.12	1.34
Var17	Interacción semifinales y antigüedad	0.00	7.47	10.52
Var18	Interacción antigüedad y coach	0.8610	1.63	1.10
Var19	Int. antigüedad, coach y semifinales	0.00	4.58	7.23
Var20	Int. minutos veteranos y antiguedad	76.50	124.58	74.52
Var21	Int. minutos veteranos y semifinales	0.00	81.16	91.74
Var22	Int. minutos veteranos y coach	7.3185	20.08	7.62

Nótese que las estadísticas de Phoenix Suns para el año 2017 han sido pésimas. Sin embargo, dado que el equipo fue catalogado como el peor de los 30 equipos en ese año, no es de sorprender el comportamiento de las mismas.

Tabla 7.34: *Loadings* de las componentes principales elegidas.

# Var.	Estadística	CP 1	CP 2	CP 3	CP15
Var1	Prop. Hist. Coach	0.8455	0.1889	-0.0444	0.0297
Var2	Jug. con experiencia en semifinales	0.8260	-0.3965	0.0280	0.1123
Var3	Jug. 3 o más años de antigüedad	0.7887	-0.1812	0.1266	0.0964
Var4	Minutos veteranos	0.7352	0.2229	0.0934	-0.0361
Var5	% Tiros de campo efectivos	0.6943	0.3285	0.3050	-0.0009
Var6	Puntos	0.4626	0.4967	0.5326	0.0881
Var7	Prop. puntos en la pintura	-0.2567	-0.0956	0.1270	0.0113
Var8	% Tiros de campo del oponente	-0.6434	-0.1737	0.4787	0.0406
Var9	Puntos del oponente	-0.5202	0.0769	0.7132	-0.0796
Var10	Clutch 1	0.7210	0.4242	-0.2748	-0.0401
Var11	Clutch 2	0.2208	0.3303	0.3625	-0.0186
Var12	Clutch 3	-0.3670	-0.1855	-0.2485	0.0120
Var13	Clutch 4	0.4606	0.4934	-0.3306	-0.0162
Var14	Clutch 5	0.5169	0.1987	-0.1147	0.0396
Var15	Clutch 6	0.6727	0.2131	-0.1844	0.0088
Var16	Interacción coach y semifinales	0.8858	-0.3245	0.0446	0.0181
Var17	Interacción semifinales y antigüedad	0.8867	-0.4019	0.0907	-0.0899
Var18	Interacción antigüedad y coach	0.9114	-0.1259	0.0876	-0.0334
Var19	Int. antigüedad, coach y semifinales	0.8986	-0.3336	0.1019	-0.1591
Var20	Int. minutos veteranos y antiguedad	0.8483	-0.1270	0.1375	0.0200
Var21	Int. minutos veteranos y semifinales	0.8625	-0.3679	0.0443	0.0557
Var22	Int. minutos veteranos y coach	0.8992	0.1734	0.0171	-0.0326

Tabla 7.35: Coeficientes $\hat{\beta}_{pj}$ asociados a las componentes principales elegidas. Los coeficientes con un asterisco (*) implican que fueron significativos a un nivel 0.05 bajo pruebas de Wald, con dos asteriscos que fueron significativos al 0.03, y con tres asteriscos al 0.01.

Coeficiente	$\hat{\beta}_{02}^{***}$	$\hat{\beta}_{12}^{***}$	$\hat{\beta}_{22}^{*}$	$\hat{\beta}_{32}$	$\hat{\beta}_{15,2}^{**}$
Valor	3.0096	-1.2855	-0.9054	0.5912	-5.7493
Error estándar	1.0673	0.4003	0.4255	0.4563	2.6098
Estadística de Wald	2.8199	3.2112	2.1277	1.2958	2.2029
Coeficiente	$\hat{\beta}_{03}$	$\hat{\beta}_{13}^{***}$	$\hat{\beta}_{23}^{**}$	$\hat{\beta}_{33}^{**}$	$\hat{\beta}_{15,3}$
Valor	-0.0889	-4.9321	-4.1828	2.7687	-8.0436
Error estándar	1.8075	1.8477	1.8414	1.1476	4.2300
Estadística de Wald	0.0492	2.6694	2.2715	2.4126	1.9015

Tabla 7.36: Coeficientes $\hat{\alpha}_{pj}$ asociados a las variables originales

# Variable	Coeficiente	Valor	Coeficiente	Valor
Constante	$\hat{\alpha}_{02}$	3.0096	$\hat{\alpha}_{03}$	-0.0889
Var1	$\hat{\alpha}_{12}$	-1.4548	$\hat{\alpha}_{13}$	-5.3217
Var2	$\hat{\alpha}_{22}$	-1.3322	$\hat{\alpha}_{23}$	-3.2415
Var3	$\hat{\alpha}_{32}$	-1.3290	$\hat{\alpha}_{33}$	-3.5568
Var4	$\hat{\alpha}_{42}$	-0.8844	$\hat{\alpha}_{43}$	-4.0098
Var5	$\hat{\alpha}_{52}$	-1.0043	$\hat{\alpha}_{53}$	-3.9465
Var6	$\hat{\alpha}_{62}$	-1.2358	$\hat{\alpha}_{63}$	-3.5927
Var7	$\hat{\alpha}_{72}$	0.4264	$\hat{\alpha}_{73}$	1.9265
Var8	$\hat{\alpha}_{82}$	1.0339	$\hat{\alpha}_{83}$	4.8984
Var9	$\hat{\alpha}_{92}$	1.4785	$\hat{\alpha}_{93}$	4.8590
Var10	$\hat{\alpha}_{10,2}$	-1.2428	$\hat{\alpha}_{10,3}$	-5.7683
Var11	$\hat{\alpha}_{11,2}$	-0.2613	$\hat{\alpha}_{11,3}$	-1.3168
Var12	$\hat{\alpha}_{12,2}$	0.4240	$\hat{\alpha}_{12,3}$	1.8020
Var13	$\hat{\alpha}_{13,2}$	-1.1414	$\hat{\alpha}_{13,3}$	-5.1211
Var14	$\hat{\alpha}_{14,2}$	-1.1400	$\hat{\alpha}_{14,3}$	-4.0170
Var15	$\hat{\alpha}_{15,2}$	-1.2173	$\hat{\alpha}_{15,3}$	-4.7906
Var16	$\hat{\alpha}_{16,2}$	-0.9226	$\hat{\alpha}_{16,3}$	-3.0337
Var17	$\hat{\alpha}_{17,2}$	-0.2055	$\hat{\alpha}_{17,3}$	-1.7181
Var18	$\hat{\alpha}_{18,2}$	-0.8136	$\hat{\alpha}_{18,3}$	-3.4568
Var19	$\hat{\alpha}_{19,2}$	0.1218	$\hat{\alpha}_{19,3}$	-1.4748
Var20	$\hat{\alpha}_{20,2}$	-1.0093	$\hat{\alpha}_{20,3}$	-3.4332
Var21	$\hat{\alpha}_{21,2}$	-1.0697	$\hat{\alpha}_{21,3}$	-3.0407
Var22	$\hat{\alpha}_{22,2}$	-1.1155	$\hat{\alpha}_{22,3}$	-4.8506

Ahora que ya se ha expresado todo lo necesario, se procederá a realizar inferencias.

La temporada 2017-2018 ha terminado. Aún a 50 días para el inicio de la temporada 2018-2019 los equipos se preparan para enfrentar la próxima temporada lo mejor posible. Para el caso de Phoenix, el verano ha traído algunos grandes cambios para enfrentar la siguiente temporada, entre ellos:

- Contratación de Igor Kokoskov como el nuevo entrenador del equipo.
- Contratación del jugador veterano Trevor Ariza.
- Regreso de Brandon Knight, jugador que no había jugado por más de un año debido a una lesión.
- Reclutamiento de Mikal Bridges y Deandre Ayton, jugadores novatos con potencial de lograr un impacto positivo tanto defensivamente como ofensivamente del equipo.

Se tiene interés en conocer, dados los cambios realizados mas otras suposiciones, si el equipo tiene mejores oportunidades de poder pertenecer al grupo de equipos con nivel de desempeño medio, pues estos cambios han dado a los aficionados razones para creer lograr una mejoría notoria en este año respecto a los anteriores, donde continuamente han pertenecido al grupo de equipos con desempeño pobre.

Se procede pues a analizar el efecto de estos cambios respecto al nivel de desempeño del equipo.

- Porcentaje histórico de victorias del *coach*:

  Aunque el nuevo entrenador nunca había sido entrenador en jefe (por lo que realmente no existiría un valor para esta estadística para el equipo 2018 Phoenix Suns), se tienen altas expectativas de que vaya a realizar un buen trabajo. Por ello, se quisiera conocer qué tan probable es que aumentando el porcentaje histórico del entrenador en 0.1000, el equipo vaya a ser clasificado en la categoría dos en vez de en la tres, dejando como constantes a las demás variables.

  Para lo siguiente, se considerarán dos patrones de covariables con valores idénticos salvo en esta estadística; es decir, se fijarán las demás estadísticas. El mencionar que se fijarán a los valores de las demás estadísticas es importante, ya que esta estadística posee algunas interacciones que influirán en el resultado del cociente de riesgos relativos deseado. Además, en este caso se busca encontrar el efecto producido únicamente por un aumento en esta estadística y no en un conjunto de ellas.

  Así pues, si el resultado de este cociente fuera algún valor "a", este se interpretaría como:

  "Es a veces más probable que Phoenix sea clasificado en la categoría

dos a que sea clasificado en la tres si aumenta su porcentaje de victorias históricas del *coach* en 0.1000, a que sea clasificado en la categoría dos en vez de en la tres si no aumenta dicha estadística."

Así pues, al realizar algebraicamente dicho cociente de riesgos relativos, se eliminarán todos los términos de las variables a excepción del porcentaje histórico de victorias del *coach* (variable 1) y de todas sus interacciones (variables 16,18,19 y 22).

Entonces, i corresponde al patrón de covariables de Phoenix donde se aumenta en 0.1000 su porcentaje histórico de victorias, y v corresponde al patrón de covariables de Phoenix sin realizar ningún cambio, el cociente de riesgos relativos a calcular sería:

$$\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}}$$

que si se realiza un poco de álgebra, corresponde a la función exponencial de:

$$\begin{aligned}
&\left[(\hat{\alpha}_{12}-\hat{\alpha}_{13})\cdot\frac{1}{\mathrm{SE}(x_{\cdot 1})}\cdot(0.1)\right]\\
&+\left[(\hat{\alpha}_{16,2}-\hat{\alpha}_{16,3})\cdot\frac{1}{\mathrm{SE}(x_{\cdot 16})}\cdot((0.1)(x_{i2}))\right]\\
&+\left[(\hat{\alpha}_{18,2}-\hat{\alpha}_{18,3})\cdot\frac{1}{\mathrm{SE}(x_{\cdot 18})}\cdot((0.1)(x_{i3}))\right]\\
&+\left[(\hat{\alpha}_{19,2}-\hat{\alpha}_{19,3})\cdot\frac{1}{\mathrm{SE}(x_{\cdot 19})}\cdot((0.1)(x_{i2})(x_{i3}))\right]\\
&+\left[(\hat{\alpha}_{22,2}-\hat{\alpha}_{22,3})\cdot\frac{1}{\mathrm{SE}(x_{\cdot 22})}\cdot((0.1)(x_{i4}))\right]
\end{aligned}$$

Y al sustituir las expresiones por valores se obtiene:

$$\begin{aligned}
&\left[(-1.4548+5.3217)\cdot\frac{1}{0.1340}\cdot(0.1)\right]\\
&+\left[(-0.9226+3.0337)\cdot\frac{1}{1.3400}\cdot((0.1)(0)\right]\\
&+\left[(-0.8136+3.4568)\cdot\frac{1}{1.10}\cdot((0.1)(3))\right]\\
&+\left[(0.1218+1.4748)\cdot\frac{1}{7.23}\cdot((0.1)(0)(3))\right]\\
&+\left[(-1.1155+4.8506)\cdot\frac{1}{7.62}\cdot((0.1)(25.5))\right]\\
=&4.8565
\end{aligned}$$

Como nota, cabe destacar que se utilizó el caso especial $j = q$ expuesto en la página 208 del presente documento.

Regresando al análisis, se llega entonces a que:

$$\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}} = \exp\{4.8565\} = 128.5734$$

A lo que se concluye que es 128 veces más probable que Phoenix sea clasificado en la categoría dos en vez de la tres si se aumentara en 0.1000 al porcentaje histórico de victorias del entrenador, a ser clasificado en la categoría dos en vez de la tres si se mantiene constante dicha estadística.

Por lo tanto, sería de gran importancia para el equipo que en los primeros meses de la liga que estará por comenzar, el entrenador mantuviera un promedio de victorias similar a 0.380 para dictar si el entrenador Kokoskov está ayudando al equipo a ser clasificado en la categoría dos o no de la manera descrita en el cociente de riesgos relativos anterior, ya que en la temporada pasada la proporción histórica de victorias del *coach* fue de 0.2870.

Se ha mencionado únicamente a las categorías dos y tres porque a corto plazo el objetivo del equipo en cuestión es ser clasificado en la categoría dos, ya que hasta la fecha ha llevado un récord de cuatro años consecutivos siendo clasificados en la categoría tres.

Cabe resaltar que este cociente de riesgos relativos no concuerda algebraicamente con el expuesto en el capítulo cuatro dado que en vez de utilizar como denominadores a la categoría de referencia, se utilizaron otras categorías; sin embargo, el cálculo algebraico, así como las interpretaciones de dicho cociente, siguen el mismo procedimiento que el mencionado en el capítulo cuatro.

- Número de jugadores con experiencia en semifinales y Minutos de veteranos:

  Esta sección ronda en torno al nuevo jugador del equipo, Trevor Ariza. Ariza ha sido un jugador importante en las dos ocasiones que ha llegado a jugar en semifinales de la NBA, es por ello que se quiere conocer el impacto que tendrá esta nueva adición respecto a las posibilidades de que Phoenix sea clasificado en la categoría de equipos con desempeño intermedio y no en el pobre. Sin embargo, otra de las variables que cambiará significativamente con la adición de este jugador es la referente a los minutos jugados por veteranos, ya que Ariza ocupará un lugar en la rotación que antes

estaba destinada a jugadores novatos.

Ahora con la nueva adición de Ariza, se proyecta que el número de minutos destinados a jugadores veteranos aumentará como mínimo en 10 minutos, es decir aumentará de 25.5 a 35.5. Por lo tanto, se realizará un cociente de riesgos relativos para calcular el número de veces que es más probable que el equipo alcance la categoría dos y no la tres con la incorporación de Ariza, que lograr la misma clasificación pero sin la incorporación de dicho jugador.

Por lo tanto, se realizará el ya conocido cociente:

$$\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}}$$

Donde i representa al equipo con la adición de Trevor Ariza (suponiendo cambios únicamente en las dos variables mencionadas) y v representa el equipo sin la adición de Trevor.

Así pues, la ecuación correspondiente a este cociente de riesgos relativos es equivalente a la función exponencial del valor:

$$\sum_{g} \left[(\hat{\alpha}_{g2} - \hat{\alpha}_{g3}) \cdot \frac{1}{\mathrm{SE}(x_{\cdot g})} \cdot (x_{ig} - x_{vg}) \right] = 9.8687$$

Con $g = \{2, 4, 16, 17, 19, 20, 21, 22\}$, haciendo referencia a las dos variables que fueron modificadas (variables 2 y 4) más sus respectivas interacciones.

Así pues, el valor del cociente es:

$$\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}} = \exp\{9.8687\} = 19,316.21$$

Por lo que se puede deducir que es $19,316$ veces más probable que Phoenix vaya a ser clasificado en la categoría dos y no en la tres con la adición de Ariza a que vaya a ser clasificado en la categoría dos y no en la tres sin la adición de Ariza.

Es importante recordar que los resultados de este cociente de riesgos relativos tienen como hipótesis que la adición del jugador mencionado únicamente modificó las variables "número de jugadores con experiencia en semifinales" y "minutos de veteranos"; las demás variables se mantuvieron fijas suponiendo efecto nulo del jugador en ellas.

Estos son grandes resultados, ya que según el presente modelo, es mucho más importante para el equipo la adición de Trevor Ariza que el desempeño del nuevo entrenador, y se esperaría fuertemente que el equipo vaya

a ser clasificado en la categoría dos debido al resultado de cinco cifras que arrojó el presente cociente de riesgos relativos.

- Puntos:

  Como se había mencionado, Trevor Ariza no fue el único jugador nuevo en el equipo. Teóricamente el mejor de todos los novatos de la liga, Deandre Ayton, también vestirá el *jersey* de los Suns. De entre todas sus cualidades, su habilidad para encestar puntos es una de las más reconocidas. Este hecho, más el regreso de Brandon Knight al equipo, que en su última temporada completa promedió 19 puntos por partido, y la mejora continua de Devin Booker, actual capitán del equipo con tan sólo 21 años, hacen creer que el equipo promediará muchos más puntos de los que hizo la temporada pasada.

  Por lo tanto, se quiere encontrar un valor con el cuál medir la influencia del posible aumento en puntos por partido con respecto a las posibles categorías que podría tener el equipo. Dado que el autor piensa que el equipo podría promediar tres puntos más que la temporada pasada, se medirá el efecto de promediar dicha cantidad de puntos mediante cocientes de riesgos relativos.

  Dado que esta variable no posee ninguna interacción, el cociente de riesgos relativos requerido se podría calcular algebraicamente por:

$$\begin{aligned}\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}} &= \exp\left[(\hat{\alpha}_{62} - \hat{\alpha}_{63}) \cdot \frac{1}{\mathrm{SE}(x_{\cdot 6})} \cdot (3)\right] \\ &= \exp[1.7459] \\ &= 5.7308\end{aligned}$$

  Lo que implicaría que los puntos no son un factor tan importante para el mejoramiento del equipo, al sólo ser 5 veces más probable que el equipo llegara a ser clasificado en la categoría dos y no en la tres si el equipo promediara tres puntos más de lo que hizo anteriormente, a que si el equipo promediara la misma cantidad de puntos que antes.

- Porcentaje de tiros de campo del oponente:

  Dado que se ha visto que un factor ofensivo como los puntos no son tan importantes para el buen desarrollo del equipo, se comprobará ahora si bajo este modelo, la disminución en el porcentaje de tiros de campo del oponente es de importancia para clasificar al equipo en una mejor categoría o no.

El motivo por el cual se tiene interés en calcular el efecto de esta estadística es que se proyecta que Phoenix tendrá una mejor defensa que la del año pasado, pues mientras que Trevor Ariza y Mikal Bridges son especialistas en defender, Deandre Ayton y otro jugador joven del equipo, Josh Jackson, tienen el potencial para convertirse en excelentes defensores.

Se calculará entonces el efecto que tendría en las categorías el disminuir en un 2 % el porcentaje de tiros de campo del oponente. El valor del cociente de riesgos relativos correspondiente es:

$$\begin{aligned}\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}} &= \exp\left[(\hat{\alpha}_{82} - \hat{\alpha}_{83}) \cdot \frac{1}{\mathrm{SE}(x_{\cdot 8})} \cdot (-2)\right] \\ &= \exp[5.9454] \\ &= 381.9921\end{aligned}$$

Por lo que según el modelo, sería 381 veces más probable que Phoenix fuera clasificado en la categoría dos y no en la uno si promediara 2 % menos de porcentaje de tiros de campo del oponente, a que fuera clasificado en la categoría dos y no en la tres sin que se modificara dicha estadística.

- Proporción de puntos sobre la pintura:

El motivo de la inclusión de esta estadística es plenamente estratégico. Se sabe con anticipación que el tener una mayor proporción de puntos sobre la pintura es característica de equipos con desempeño pobre y tener un valor pequeño de dicha estadística es característica de equipos con buen desempeño gracias a los signos de los coeficientes. Sin embargo, se tiene interés por conocer la magnitud del impacto de dicha estadística respecto a las categorías.

Se calcularán los efectos de un aumento de 0.02 en la proporción de puntos sobre la pintura. El cociente asociado toma el valor:

$$\begin{aligned}\frac{\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}}{\frac{\hat{\pi}_{v2}}{\hat{\pi}_{v3}}} &= \exp\left[(\hat{\alpha}_{72} - \hat{\alpha}_{73}) \cdot \frac{1}{\mathrm{SE}(x_{\cdot 7})} \cdot (0.02)\right] \\ &= \exp[-0.8773] \\ &= 0.4159\end{aligned}$$

Por lo que se concluye que es $(0.4159)^{-1} = 2.4044$ veces más probable que algún equipo con una adición de 0.02 en la proporción de sus puntos sobre la pintura sea clasificado en la categoría tres en vez de la dos, a que el mismo equipo sin esa adición en la estadística mencionada sea clasificado en la categoría tres en vez de en la dos.

Respecto a los resultados anteriores, las conclusiones obtenidas se presentarán en el siguiente capítulo.

Capítulo 8

Conclusiones Finales

8.1. Conclusiones respecto a Phoenix Suns

Respecto a los análisis de los cocientes de riesgos relativos anteriores para el equipo de Arizona, las conclusiones inferidas fueron las siguientes:

1. La adición de Trevor Ariza al equipo tendrá un efecto enormemente positivo en cuanto al desempeño del equipo en cuestión.
2. El mejorar la defensa (visto por la estadística "porcentaje de tiros de campo del oponente") debe de ser el enfoque principal del equipo; el mejorar la ofensiva (visto por la estadística "puntos") repercute muy poco en el desempeño del equipo en cuestión.
3. El desempeño del entrenador sí influirá en el desempeño del equipo, aunque no es una prioridad para mejorar dicho desempeño.
4. Las estadísticas a las que el equipo les debe de dar mayor relevancia son minutos de veteranos y jugadores con experiencia en semifinales; posteriormente porcentaje de tiros de campo del oponente y, finalmente, porcentaje histórico de victorias del entrenador. Se podría llegar a ignorar a la variable puntos para el desarrollo positivo de un equipo, dado que su efecto es muy pequeño.
5. No se le debe de dar mayor relevancia a la proporción de puntos sobre la pintura para decidir alguna estrategia de juego, ya que su efecto en el cociente de riesgos es muy pequeño.

Cabe mencionar aquí, que todas las adiciones propuestas en las estadísticas y en los cocientes de riesgos relativos del capítulo anterior son alcanzables y realistas desde un punto de vista gerencial de la NBA y sobre todo desde la posición actual del equipo en cuestión, Phoenix Suns.

8.2. Conclusiones finales del documento, complicaciones y recomendaciones

Una vez finalizado el análisis tanto del modelo como del equipo elegido, se procede ahora a cerrar el presente documento con algunas conclusiones finales del trabajo realizado, así como complicaciones presentadas y recomendaciones que pudieran o no haber mejorado el presente trabajo y valdrían la pena tomarse en cuenta para la realización de futuros proyectos.

1. Conclusiones finales del trabajo:

   - Gracias a la realización del presente trabajo, se ha constatado que, normalmente será muy complicado encontrar un modelo que tenga excelentes cualidades tanto predictivas como de ajuste, ya que recuérdese que en un principio, ningún modelo pudo adecuarse a los requerimientos estrictos impuestos, por lo que se tuvo que reducir la rigidez de dichos requerimientos.

     En cambio, cuando únicamente se tomó en cuenta uno de los dos aspectos (ya sea el aspecto predictivo o el aspecto de ajuste) y bajo los mismos niveles de rigidez estrictos, se logró encontrar un gran número de modelos que cumplieran con dichos requerimientos. Por lo que se concluye que a pesar de que las variables se encuentren fuertemente correlacionadas con la variable de respuesta, la realización de un modelo multifacético la mayoría de las veces será desafiante.

2. Complicaciones presentadas:

   - El autor de este documento piensa que el análisis podría llegar a resultados más relevantes, si en vez de utilizar las estadísticas de los equipos de la temporada entera para ajustar el modelo, se utilizaran las estadísticas de los primeros 41 partidos de la temporada (el total de partidos por equipo es 82), pues seguramente el desempeño de los equipos durante los últimos partidos de la temporada se encontrará altamente correlacionado con la categoría final del mismo. Además, de esta manera se podrían realizar predicciones de los equipos a mitad de temporada, mismas que podrían ayudar en la toma de decisiones de la gerencia del equipo respecto a si es necesario mejorar la plantilla actual para asegurar la clasificación en cierta categoría o no.

     Sin embargo, realizar dicha tarea sería muy complicada, dado que los datos encontrados fueron los promedios finales por partido, y el hecho de obtener las estadísticas bajo una fracción de la temporada implicaría calcular dichas estadísticas de forma manual y revisando la información de cada partido jugado por cada equipo durante la

primera mitad de la temporada, labor a la que se le tendría que destinar una gran cantidad de tiempo.

Debido a todas esas complicaciones que se presentarían, se optó por realizar el análisis utilizando las estadísticas por temporada completa que presenta la página oficial de la NBA.

- Dada la naturaleza de los datos utilizados en el estudio (equipos de la NBA) fue imposible incluir un mayor número de observaciones, ocasionando así que se utilizara una cantidad reducida de ellas tanto en el entrenamiento del modelo, como en la validación del mismo.

  Así también, si se hubiera logrado emplear una mayor cantidad de observaciones en el modelo, la distribución de los coeficientes de las variables asociadas a éste se aproximaría cada vez más a una distribución normal, ocasionando así que algunas de las herramientas presentadas como los intervalos de confianza, pruebas de Wald y la Ji-cuadrada de Pearson, fueran más confiables.

3. Recomendaciones que podrían mejorar el presente trabajo:

   - Una sugerencia a realizar en posteriores estudios relacionados al tema sería trabajar desde un principio con las variables estandarizadas por año y no con las reales, pues se piensa que la causa por la que tanto las distribuciones como el dominio de las componentes principales distaran mucho entre los datos de entrenamiento y los de validación, fue que los datos de validación (año 2017) fueron estandarizados con la información de los datos de entrenamiento (años 2013 y 2015) cuando la distribución de ambas bases de datos eran diferentes.

     Así también, se podría sugerir utilizar pruebas formales para verificar que tanto los promedios como desviaciones estándar de las variables sean similares año con año.

   - Otra sugerencia para posteriores estudios sería asegurar la presencia de interacciones mediante pruebas de cocientes de verosimilitud y métodos gráficos, ya que en el presente estudio únicamente se buscaron las interacciones que tuvieran sentido bajo un punto de vista enfocado a la NBA y de manera directa se introdujeron al grupo de variables candidatas.

     Aunque la manera en que se eligieron a las interacciones seguramente no afectaría los resultados del modelo que utiliza como variables predictoras a las variables originales, tal vez podría afectar a los resulta-

dos del modelo que utilizó a componentes principales como variables predictoras del modelo.

- Utilizar variables que se pudieran calcular únicamente por jugador y no por equipo, pues de esa manera se podrían realizar predicciones de futuras temporadas incluso antes de que inicien las mismas.

- Respecto a la estadística $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$, se recomienda en vez de utilizar un nivel de confianza asociado a la significancia de las variables del modelo, utilizar un nivel de confianza fijo para todos los modelos, variables y observaciones eliminadas, pues al ser el objetivo de dicha estadística encontrar observaciones influyentes y dado que éstas no están relacionadas con la significancia de las variables, los resultados podrían ser incorrectos al utilizar diferentes niveles de confianza.

  Al asignar diferentes niveles de confianza de acuerdo a la significancia de las variables, los modelos con significancias pequeñas en sus variables tenderán a no aparentar tener observaciones influyentes, mientras que modelos con significancias grandes en sus variables presentarán valores más altos en dicha estadística, y por lo tanto, aparentarán tener una mayor cantidad de datos influyentes. El autor del documento recomienda entonces fijar el nivel de confianza a utilizar en la estadística $\overline{\Delta\hat{\beta}}_{pj}^{(-i)}$ a un mismo valor para todas las variables. A la vez, recomienda que dicho valor se designe con base en simulaciones.

- Por último, se recomienda para futuros proyectos relacionados con el tema de deportes, utilizar modelos más robustos como los modelos lineales generalizados mixtos y otras herramientas que permitan la correlación entre observaciones, ya que como se había mencionado, existe correlación entre el mismo equipo de diferentes años. Más aún, en el caso en se pudiera utilizar información de únicamente un año, los equipos de dicho año seguirían correlacionados, ya que el éxito de uno de ellos implica el fracaso de otros.

Con estos comentarios ha terminado este capítulo, y con ello el trabajo escrito. El autor de este documento espera que le haya sido de agrado al lector este trabajo, y lo más importante, que le haya sido de ayuda ya sea para tener un mejor conocimiento de la teoría detrás de estos dos métodos tan útiles como son la regresión logística multinomial y el análisis de componentes principales, o simplemente para ver una aplicación a la vida real de dichos métodos.

Apéndice A

Bases de Datos

Nomenclatura utilizada

A continuación se muestra la nomenclatura utilizada en las bases de datos presentadas:

- Cat.: Categoría del equipo.
- Var1: Proporción histórica de victorias del *coach*.
- Var2: Número de jugadores con experiencia en semifinales.
- Var3: Número de jugadores con tres años o más de antigüedad.
- Var4: Minutos de veteranos.
- Var5: Porcentaje de tiros de campo efectivo.
- Var6: Puntos.
- Var7: Proporción de puntos sobre la pintura.
- Var8: Porcentaje de tiros de campo del oponente.
- Var9: Puntos del oponente.
- Var10: Estadística *clutch* 1.
- Var11: Estadística *clutch* 2.
- Var12: Estadística *clutch* 3.
- Var13: Estadística *clutch* 4.
- Var14: Estadística *clutch* 5.
- Var15: Estadística *clutch* 6.

Base de datos de entrenamiento: Tablas A.1, A.2, A.3, A.4, A.5 y A.6

Base de datos de validación: Tablas A.7, A.8, A9 y A10

Tabla A.1: Base de datos de entrenamiento: Parte 1

Equipo	Cat.	Var1	Var2	Var3	Var4	Var5	Var6	Var7
2013 Atlanta Hawks	2	0.463	0	1	39.8	51.5	101	0.4059
2013 Boston Celtics	3	0.305	2	3	33.5	47.7	96.2	0.3929
2013 Brooklyn Nets	1	0.537	4	0	39.9	51.4	98.5	0.3898
2013 Charlotte Hornets	2	0.524	0	2	39.9	48.1	96.9	0.4169
2013 Chicago Bulls	2	0.6571	2	4	44.8	47.1	93.7	0.397
2013 Cleveland Cavaliers	3	0.616	1	3	35.3	47.9	98.2	0.4002
2013 Dallas Mavericks	1	0.5878	3	3	43	52.6	104.8	0.3998
2013 Denver Nuggets	3	0.439	0	4	42.6	49.7	104.4	0.4435
2013 Detroit Pistons	3	0.25	1	2	31	48.2	101	0.5139
2013 Golden State Warriors	2	0.526	0	3	37.4	51.7	104.3	0.4171
2013 Houston Rockets	2	0.5309	2	1	35.8	53.1	107.7	0.4726
2013 Indiana Pacers	1	0.6255	5	5	46.2	49	96.7	0.3971
2013 LA Clippers	1	0.5639	2	3	47.2	52.6	107.9	0.3892
2013 Los Angeles Lakers	3	0.5165	1	2	36.8	50.5	103	0.3981
2013 Memphis Grizzlies	1	0.61	6	4	45	49.4	96.1	0.4953
2013 Miami Heat	1	0.6597	6	6	48.3	55.4	102.2	0.4423
2013 Milwaukee Bucks	3	0.458	0	1	28.6	47.9	95.5	0.4304
2013 Minnesota Timberwolves	2	0.582	0	3	42.5	48.6	106.9	0.4425

Tabla A.2: Base de datos de entrenamiento: Parte 2

Equipo	Cat.	Var1	Var2	Var3	Var4	Var5	Var6	Var7
2013 New Orleans Pelicans	3	0.4103	0	2	31.8	49.5	99.7	0.4564
2013 New York Knicks	2	0.463	4	5	39.8	50.5	98.6	0.3398
2013 Oklahoma City Thunder	1	0.6328	4	5	38.8	52	106.2	0.3964
2013 Orlando Magic	3	0.2622	1	1	29.8	48.7	96.5	0.3948
2013 Philadelphia 76ers	3	0.232	0	1	28.3	47.5	99.5	0.4975
2013 Phoenix Suns	2	0.585	1	2	40.6	51.9	105.2	0.4002
2013 Portland Trail Blazers	1	0.4499	1	2	33.7	50.4	106.7	0.3533
2013 Sacramento Kings	3	0.341	0	3	38.6	48.4	100.5	0.4378
2013 San Antonio Spurs	1	0.6858	6	7	46.3	53.7	105.4	0.4355
2013 Toronto Raptors	2	0.4489	0	2	36.9	49.8	101.3	0.384
2013 Utah Jazz	3	0.4341	1	4	38.7	48.4	95	0.4368
2013 Washington Wizards	1	0.3673	2	3	41.7	50.6	100.7	0.4101
2015 Atlanta Hawks	1	0.5935	5	5	47.2	51.6	102.8	0.395
2015 Boston Celtics	2	0.4594	1	3	42	48.8	105.7	0.375
2015 Brooklyn Nets	3	0.244	0	1	35.9	49.2	98.6	0.411
2015 Charlotte Hornets	2	0.5041	0	2	41.7	50.2	103.4	0.404
2015 Chicago Bulls	2	0.512	3	4	36.3	48.7	101.6	0.404
2015 Cleveland Cavaliers	1	0.659	6	3	47.9	52.4	104.3	0.401
2015 Dallas Mavericks	2	0.5839	3	2	43.8	50.2	102.3	0.378
2015 Denver Nuggets	3	0.3828	1	3	26.5	48.9	101.9	0.447

Tabla A.3: Base de datos de entrenamiento: Parte 3

Equipo	Cat.	Var1	Var2	Var3	Var4	Var5	Var6	Var7
2015 Detroit Pistons	2	0.6016	1	2	43	49.1	102	0.424
2015 Golden State Warriors	1	0.8537	6	6	47.8	56.3	114.9	0.387
2015 Houston Rockets	2	0.521	5	4	43.3	51.6	106.5	0.363
2015 Indiana Pacers	2	0.58	3	4	42.9	49.7	102.2	0.42
2015 LA Clippers	2	0.5766	1	4	47.9	52.4	104.5	0.469
2015 Los Angeles Lakers	3	0.412	3	1	22.9	46	97.3	0.482
2015 Memphis Grizzlies	2	0.5975	5	4	41.5	47.7	99.1	0.427
2015 Miami Heat	1	0.6234	3	2	38.1	50.8	100	0.427
2015 Milwaukee Bucks	3	0.4797	1	2	36	49.9	99	0.438
2015 Minnesota Timberwolves	3	0.433	0	3	23.7	49.8	102.4	0.446
2015 New Orleans Pelicans	3	0.4638	0	3	45.7	49.8	102.7	0.467
2015 New York Knicks	3	0.278	1	1	34.9	48.3	98.4	0.448
2015 Oklahoma City Thunder	1	0.671	5	5	46.4	52.4	110.2	0.412
2015 Orlando Magic	3	0.499	0	2	33.2	50	102.1	0.44
2015 Philadelphia 76ers	3	0.191	0	1	22.5	48.7	97.4	0.425
2015 Phoenix Suns	3	0.273	1	2	39.7	48.7	100.9	0.442
2015 Portland Trail Blazers	1	0.4861	0	4	44.2	51.1	105.1	0.401
2015 Sacramento Kings	3	0.588	1	3	44.7	51	106.6	0.439
2015 San Antonio Spurs	1	0.6919	4	5	41.8	52.6	103.5	0.428
2015 Toronto Raptors	1	0.5097	0	5	45.3	50.4	102.7	0.43
2015 Utah Jazz	2	0.4756	0	4	31.7	50.1	97.7	0.423
2015 Washington Wizards	2	0.406	1	5	46.2	51.1	104.1	0.398

Tabla A.4: Base de datos de entrenamiento: Parte 4

Equipo	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15
2013 Atlanta Hawks	46.2	101.5	0.5	100	0.6	5	34.6	0.486
2013 Boston Celtics	46.5	100.7	-1.5	80	1	3.3	12	0.306
2013 Brooklyn Nets	45.8	99.5	-0.3	100	0.3	5.1	26.1	0.485
2013 Charlotte Hornets	44.2	97.1	-0.2	70.8	0.9	4.5	14.6	0.543
2013 Chicago Bulls	43	91.8	1.2	80	0.6	4.1	26.5	0.5
2013 Cleveland Cavaliers	45.2	101.5	-0.5	71.4	0.2	4.5	29.4	0.531
2013 Dallas Mavericks	46.4	102.4	0.4	77.8	0.7	4	28.3	0.559
2013 Denver Nuggets	45.7	106.5	-0.8	62.5	0.6	3.7	14.6	0.464
2013 Detroit Pistons	47	104.7	-1.8	100	0.7	3.3	18.8	0.375
2013 Golden State Warriors	43.6	99.5	2.3	100	0.6	4.9	34.6	0.559
2013 Houston Rockets	44.3	103.1	1.1	100	0.3	4.8	35.1	0.621
2013 Indiana Pacers	42	92.3	0.5	50	0.2	5	35.9	0.76
2013 LA Clippers	44.1	101	0.6	71.4	0.7	4.6	26.7	0.5
2013 Los Angeles Lakers	46.8	109.2	-1.1	100	0.5	3.8	31.3	0.435
2013 Memphis Grizzlies	45	94.6	0.5	100	0.7	4.4	29	0.75
2013 Miami Heat	45.7	97.4	1.5	75	0.3	5.3	33.3	0.5
2013 Milwaukee Bucks	46.8	103.7	-1.9	100	0	3.7	23.1	0.333
2013 Minnesota Timberwolves	47.1	104.3	-0.3	83.3	0.5	5.1	22.9	0.31
2013 New Orleans Pelicans	46.5	102.4	-0.5	85.7	0.5	4.3	30	0.472
2013 New York Knicks	45.8	99.4	-0.4	100	0.3	3.9	25.5	0.379
2013 Oklahoma City Thunder	43.6	99.8	1.3	91.7	0.8	5	35.6	0.563
2013 Orlando Magic	45.6	102	-1.3	85.7	0.5	5	17.8	0.346

Tabla A.5: Base de datos de entrenamiento: Parte 5

Equipo	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15
2013 Philadelphia 76ers	47.1	109.9	-0.1	60	0.6	5.1	44	0.565
2013 Phoenix Suns	45.6	102.6	-0.2	100	0.2	3.7	23.1	0.469
2013 Portland Trail Blazers	45.1	102.8	1.4	86.7	0.7	5.5	30.4	0.605
2013 Sacramento Kings	46.1	103.4	-0.8	100	0.1	4.6	32.4	0.367
2013 San Antonio Spurs	44.4	97.6	1.2	66.7	0.3	5.1	44.4	0.818
2013 Toronto Raptors	45	98	0	0	1	5	24.4	0.471
2013 Utah Jazz	47.3	102.2	-1.6	100	0.5	4.9	40.7	0.609
2013 Washington Wizards	45.8	99.4	0.5	66.7	0.3	5.6	33.9	0.421
2015 Atlanta Hawks	43.2	99.2	0.7	100	0.2	4.9	38.7	0.444
2015 Boston Celtics	44.1	102.5	0	85.7	0.4	5	29.8	0.433
2015 Brooklyn Nets	47.9	106	-1.5	71.4	0.8	3.8	23.1	0.4
2015 Charlotte Hornets	44.4	100.7	0.4	70	0.5	4.3	20.5	0.615
2015 Chicago Bulls	44.1	103.1	-0.4	72.7	0.7	3.5	35.5	0.571
2015 Cleveland Cavaliers	44.8	98.3	1.7	100	0.3	5	31.3	0.577
2015 Dallas Mavericks	45.1	102.6	0.6	57.1	0.4	3.9	21.2	0.471
2015 Denver Nuggets	46.1	105	-0.8	80	0.5	3.8	14.3	0.613
2015 Detroit Pistons	46.1	101.4	0.4	66.7	0.5	4.7	31.3	0.667
2015 Golden State Warriors	43.5	104.1	2.3	100	0.5	5.5	46.5	0.84
2015 Houston Rockets	45.9	106.4	-0.5	80	0.3	3.9	29.7	0.556

Tabla A.6: Base de datos de entrenamiento: Parte 6

Equipo	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15
2015 Indiana Pacers	44	100.5	0.5	100	0.7	4	19.6	0.462
2015 LA Clippers	43.4	100.2	0.6	92.9	0.4	5.4	28.2	0.606
2015 Los Angeles Lakers	47.3	106.9	-0.1	50	0.6	5	36.1	0.435
2015 Memphis Grizzlies	45.6	101.3	0.7	87.5	0.3	3.9	27.3	0.643
2015 Miami Heat	44.2	98.4	1.2	100	0.3	5	44	0.633
2015 Milwaukee Bucks	45.4	103.2	-0.8	71.4	0.5	3.7	34.8	0.586
2015 Minnesota Timberwolves	47.1	106	-0.5	80	0.3	4.3	15.9	0.447
2015 New Orleans Pelicans	46.8	106.5	-0.4	88.9	0.7	3.7	19.1	0.448
2015 New York Knicks	44.3	101.1	-0.9	71.4	0.3	4.4	25.6	0.448
2015 Oklahoma City Thunder	43.8	102.9	-0.4	60	0.1	4.8	30	0.486
2015 Orlando Magic	46	103.7	0	83.3	0.9	3.5	18.8	0.344
2015 Philadelphia 76ers	46.4	107.6	-2	0	1	2.9	22.6	0.048
2015 Phoenix Suns	46.7	107.5	-1.5	100	0.5	3.6	28.6	0.292
2015 Portland Trail Blazers	45.3	104.3	1	100	0.2	4.9	28.6	0.5
2015 Sacramento Kings	46.2	109.1	-0.9	87.5	0.2	3.9	23.8	0.433
2015 San Antonio Spurs	43.6	92.9	0.5	100	0	4.6	30.8	0.588
2015 Toronto Raptors	44.4	98.2	-0.1	100	0.3	4.6	34.3	0.531
2015 Utah Jazz	44.6	95.9	-0.2	42.9	0.7	4.8	29.6	0.344
2015 Washington Wizards	46.2	104.6	0	88.9	0.4	4.8	25	0.5

Tabla A.7: Base de datos de validación: Parte 1

Equipo	Cat.	Var1	Var2	Var3	Var4	Var5	Var6	Var7
2017 Atlanta Hawks	3	0.5768	0	2	25.2	51.2	103.4	0.3994
2017 Boston Celtics	1	0.506	3	2	29.5	51.8	104	0.3827
2017 Brooklyn Nets	3	0.244	1	1	39.2	51.4	106.6	0.4109
2017 Charlotte Hornets	3	0.4878	1	6	41.1	50.8	108.2	0.4039
2017 Chicago Bulls	3	0.506	0	1	26.8	49.7	102.9	0.4062
2017 Cleveland Cavaliers	1	0.6333	5	4	45.4	54.7	110.9	0.3986
2017 Dallas Mavericks	3	0.5716	2	4	31.3	51.3	102.3	0.3793
2017 Denver Nuggets	2	0.415	0	4	38.5	53.6	110	0.4455
2017 Detroit Pistons	2	0.587	1	4	43.9	51.2	103.8	0.42
2017 Golden State Warriors	1	0.8415	5	4	41.8	56.9	113.5	0.3885
2017 Houston Rockets	1	0.5293	2	3	47.5	55.1	112.4	0.3621
2017 Indiana Pacers	2	0.5137	2	1	42.3	52.5	105.6	0.4233
2017 LA Clippers	2	0.5798	0	3	36	52.7	109	0.4789
2017 Los Angeles Lakers	3	0.317	0	1	26.4	51.7	108.1	0.4801
2017 Memphis Grizzlies	3	0.524	2	3	30.7	50	99.3	0.426
2017 Miami Heat	2	0.6094	0	5	43.4	52	103.4	0.4342
2017 Milwaukee Bucks	2	0.4883	0	3	39.3	53.1	106.5	0.4451
2017 Minnesota Timberwolves	2	0.6011	1	3	47.7	52.3	109.5	0.4493
2017 New Orleans Pelicans	1	0.4591	1	3	46.8	54.1	111.7	0.4691
2017 New York Knicks	3	0.4475	1	1	39.5	51	104.5	0.4469
2017 Oklahoma City Thunder	2	0.622	4	2	43.3	51.4	107.9	0.4069

Tabla A.8: Base de datos de validación: Parte 2

Equipo	Cat.	Var1	Var2	Var3	Var4	Var5	Var6	Var7
2017 Orlando Magic	3	0.5439	1	4	41.5	51.2	103.4	0.441
2017 Philadelphia 76ers	1	0.229	1	2	28.4	53.5	109.8	0.4299
2017 Phoenix Suns	3	0.287	0	3	25.5	49.5	103.9	0.4437
2017 Portland Trail Blazers	2	0.4873	0	4	44.6	51.1	105.6	0.4072
2017 Sacramento Kings	3	0.5453	1	1	22.4	50.2	98.8	0.4403
2017 San Antonio Spurs	2	0.6947	5	5	35.7	50.7	102.7	0.4265
2017 Toronto Raptors	1	0.5256	4	4	32.8	53.9	111.7	0.4351
2017 Utah Jazz	1	0.525	0	3	38.8	52.7	104.1	0.4265
2017 Washington Wizards	2	0.6167	0	5	44.4	52.5	106.6	0.3968

Tabla A.9: Base de datos de validación: Parte 3

Equipo	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15
2017 Atlanta Hawks	46.9	108.8	-0.2	100	0.3	4.8	31.7	0.455
2017 Boston Celtics	44	100.4	-0.1	60	0.6	5.3	37.7	0.6
2017 Brooklyn Nets	46.6	110.3	0	63.6	0.6	4.9	24.6	0.424
2017 Charlotte Hornets	46.8	108	-1.9	100	0	3.8	16.1	0.345
2017 Chicago Bulls	47.2	110	0.2	83.3	0.4	4.5	23.1	0.545
2017 Cleveland Cavaliers	47.4	109.9	1.4	81.8	0.5	5.3	33.8	0.667
2017 Dallas Mavericks	46.9	105.4	-1.6	66.7	0	3.7	22.2	0.2
2017 Denver Nuggets	47.6	108.5	-0.1	90	0.5	5.2	27.3	0.545
2017 Detroit Pistons	45.9	103.9	-0.4	75	0.7	4.3	29.8	0.385
2017 Golden State Warriors	44.7	107.5	0	60	0.4	4.7	26.1	0.545
2017 Houston Rockets	46.2	103.9	1.1	100	0.6	4.6	30.6	0.818
2017 Indiana Pacers	46.5	104.2	-0.6	75	0.6	4.5	25	0.6
2017 LA Clippers	45.8	109	-0.6	83.3	0.1	4.6	37	0.56
2017 Los Angeles Lakers	45.6	109.6	-0.7	55.6	0.2	4.4	29.3	0.5
2017 Memphis Grizzlies	46.2	105.5	0	100	0	3.6	24.1	0.313
2017 Miami Heat	45	102.9	0.7	85.7	0.5	5	33.9	0.545
2017 Milwaukee Bucks	46.8	106.8	-1	83.3	0.8	4.4	16.7	0.629
2017 Minnesota Timberwolves	47.5	107.3	0.8	71.4	0.3	5.4	28.9	0.548
2017 New Orleans Pelicans	45.4	110.4	0.7	80	1	5.2	27.5	0.571
2017 New York Knicks	45.7	108	-0.5	100	0.3	4.7	23.5	0.391
2017 Oklahoma City Thunder	45.8	104.4	0.2	60	0.5	5	23.1	0.486

Tabla A.10: Base de datos de validación: Parte 4

Equipo	Var8	Var9	Var10	Var11	Var12	Var13	Var14	Var15
2017 Orlando Magic	46.8	108.2	-1.1	50	1.2	3.8	25.6	0.44
2017 Philadelphia 76ers	43.4	105.3	0.8	85.7	0.3	4.5	30.6	0.556
2017 Phoenix Suns	47.1	113.3	-1.5	88.9	0.3	3.6	28.6	0.409
2017 Portland Trail Blazers	44.7	103	0.1	75	0.6	4.7	27.5	0.529
2017 Sacramento Kings	47	105.8	0.3	75	0.3	5	37.2	0.515
2017 San Antonio Spurs	45.3	99.8	1.7	85.7	0.5	4.7	22.2	0.467
2017 Toronto Raptors	44.9	103.9	-0.1	100	0.4	5	24.1	0.515
2017 Utah Jazz	44.9	99.8	-0.5	100	0.8	4.8	27.5	0.476
2017 Washington Wizards	46.2	106	-1	90	0.6	4.9	14	0.433

Apéndice B

Códigos Utilizados para la Selección de Modelos

Código utilizado para la búsqueda de modelos con base en poder predictivo.

```
##############Se inicia el algoritmo para encontrar el mejor modelo##############
#BDModeloPCA corresponde a las componentes de la base de datos de entrenamiento
BaseM<-BDModeloPCA
#DatosAPredecir corresponde a las componentes de la base de datos de validacion
BaseP<-DatosAPredecir
attach(BaseM)
for(i in 1:22){
  TablaComb<-combn(c(1:22),i)
  PCT.Clasif.Modelo<-NULL
  PCT.Clasif.Prediccion<-NULL
  for(j in 1:dim(TablaComb)[2]){
    VarsaUsar<-NULL
    for(z in 1:i){
      VarsaUsar<-paste(VarsaUsar,"+CP",TablaComb[z,j],sep="")
    }
    VarsaUsar<-substr(VarsaUsar,2,nchar(VarsaUsar))
    VarsaUsar<-paste("~",VarsaUsar, sep="")
    Modelo<-multinom(as.formula(paste("Categoria",VarsaUsar)),
                     data=BaseM,maxit=800)
    PrediccionesM<-predict(Modelo,newdata=BaseM)
    PrediccionesP<-predict(Modelo, newdata=BaseP)
    TablaM<-table(Categoria,PrediccionesM)
    TablaP<-table(BaseP$Categoria,PrediccionesP)
    AuxErrM<-round(sum(diag(TablaM))/60,digits=3)
    AuxErrP<-round(sum(diag(TablaP))/30,digits=3)
    PCT.Clasif.Modelo<-rbind(PCT.Clasif.Modelo,AuxErrM)
```

```
    PCT.Clasif.Prediccion<-rbind(PCT.Clasif.Prediccion,AuxErrP)
  }
  A<-cbind(PCT.Clasif.Modelo,PCT.Clasif.Prediccion)
  colnames(A)<-c("PCT.Clasif.Modelo","PCT.Clasif.Prediccion")
  if(i>1 && i<22){
    b<-combn(c(1:22),i)
    b<-t(b)
    auxcolnames<-NULL
    for(j in 1:i){
      c<-paste("Col",j,sep = "")
      auxcolnames<-c(auxcolnames,c)
    }
    colnames(b)<-auxcolnames
    bla<- apply( b[ , auxcolnames ] , 1 , paste , collapse = "-" )
    rownames(A)<-bla
  }
  assign(paste("Modelocon",i,sep=""),A)
  View(paste("Modelocon",i,sep=""))
}
```

Código utilizado para la búsqueda de modelos con todas sus variables significativas con base en pruebas de cocientes de verosimilitud.

```
#Creacion del Likelihood Ratio Test o Devianza
#Recuerdese que esta prueba solo funciona bajo modelos anidados
LikelihoodRatioTest<-function(ModeloChico,ModeloGrande){
  DevianzaChico<-deviance(ModeloChico)
  DevianzaGrande<-deviance(ModeloGrande)
  g.l.Chico<-ModeloChico$edf
  g.l.Grande<-ModeloGrande$edf
  Valor<-DevianzaChico-DevianzaGrande
  gl<-g.l.Grande-g.l.Chico
  if (Valor<0 || gl<0){
    return("Error")
  }
  Pvalue<-1-pchisq(Valor,gl)
  return(Pvalue)
}
#El modelo a ser evaluado recibe el nombre de "ModelovencedorPCA"
LRT<-NULL
for(i in 1:(ModeloVencedorPCA$edf/2-2)){
  #En "combn" se introducen las componentes a utilizar en el modelo presente
  TablaComb<-combn(c(1,15,3,2),i)
  for(j in 1:dim(TablaComb)[2]){
    VarsaUsar<-NULL
    for(z in 1:i){
      VarsaUsar<-paste(VarsaUsar,"+CP",TablaComb[z,j],sep="")
    }
    VarsaUsar<-substr(VarsaUsar,2,nchar(VarsaUsar))
    VarsaUsar<-paste("~",VarsaUsar, sep="")
    ModeloAux<-multinom(as.formula(paste("Categoria",VarsaUsar)),
                        data=BDModeloPCA,maxit=800)
    LRT<-c(LRT, LikelihoodRatioTest(ModeloAux, ModeloVencedorPCA))
    names(LRT)[length(LRT)]<-as.character(VarsaUsar)
  }
}
View(LRT)
```

Bibliografía

1. Afifi, Abdelmonem A.; Susanne May; Virginia A. Clark. (2004). *Practical multivariate analysis*. Chapman & Hall.

2. Agresti, Alan. (2013). *Categorical Data Analysis*. Estados Unidos de América, John Wiley & Sons.

3. Agresti, Alan. (2015). *Foundations of Linear and Generalized Linear Models*. Estados Unidos de América, John Wiley & Sons.

4. Bellocco, Rino; Sara Algeri. (2013). *Goodness-of-fit tests for categorical data*. Obtenida de https://journals.sagepub.com/doi/pdf/10.1177/1536867X1301300210

5. Czepiel, Scott A. *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*. Obtenida de https://czep.net/stat/mlelr.pdf

6. Giugliano, Maria Maddalena. (2014). *Diagnostic measures for multinomial distance model* (Tesis doctoral). Obtenida de http://www.fedoa.unina.it/9889/1/Thesis.pdf

7. Hosmer, David W.; Stanley Lemeshow. (2000). *Applied logistic regression*. Estados Unidos de América, John Wiley & Sons.

8. Jolliffe, I. T. (2002). *Principal Component Analysis*. Estados Unidos de América, Springer.

9. Estadísticas de la NBA no creadas por el autor, así como información necesaria para la creación de estadísticas por parte del autor, obtenidas de https://stats.nba.com/

10. Figura número 6.1 obtenida de https://www.nba.com/suns/?tmd=1