UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

ALGUNOS MÉTODOS DE CLASIFICACIÓN ESTADÍSTICA: ANÁLISIS DE REGRESIÓN LOGÍSTICA, DE DISCRIMINANTE Y DE REGRESIÓN DE COX

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
ACTUARIO

P R E S E N T A:

SERGIO SERRANO AYVAR

DIRECTOR DE TESIS:
DRA. GUILLERMINA ESLAVA GÓMEZ

2011

Hoja datos del jurado

1. Datos del alumno
Serrano
Ayvar
Sergio
52 43 03 64
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
406005562

2. Datos del tutor
Dra.
Guillermina
Eslava
Gómez

3. Datos del sinodal 1
Dr.
José Luis
Castrejón
Caballero

4. Datos del sinodal 2
M. en C.
María del Pilar
Alonso
Reyes

5. Datos del sinodal 3
M. en C.
Gonzalo
Pérez
de la Cruz

6. Datos del sinodal 4
M. en C.
Sofía
Villers
Gómez

7.Datos del trabajo escrito
Algunos métodos de clasificación estadística:
análisis de regresión logística, de discriminante y de regresión de Cox
144p.
2011

Índice general

Índice de figuras

Índice de cuadros

Agradecimientos

A mis padres, Ricardo y Emilia por su amor incondicional y el sacrificio que han hecho por verme alcanzar este sueño, que a pesar de la distancia hasta el día de hoy siguen dándome aliento para seguir adelante, les doy gracias por ser como son y por ser unos padres maravillosos.

Le doy gracias a mi hermano Ricardo por estar a mi lado desde siempre, te pido disculpas por no ser un hermano ejemplar.

A mi familia que se encuentra en la costa Guerrerense, a mis tios Wilfrido Hernández y Guadalupe Ayvar por su apoyo y en especial a Homero Hernández por soportar nuestros desvelos y desordenes. Sin su ayuda no hubiera sido posible haber terminado esta etapa de mi vida.

A mi directora de tesis la Dra. Guillermina Eslava Gómez por la paciencia y la dirección de este trabajo, por su integridad y sentido de la calidad. Y gracias por los jalones de orejas. ¡Dios la bendiga!.

A mis sinodales por su dedicación a este trabajo, Dr. José Luis Castrejón Caballero, Mtra. María del Pilar Alonso Reyes, Mtro. Gonzalo Pérez de la Cruz y Mtra. Sofia Villers Gómez, que con sus valiosos comentarios me ayudaron a mejorar mi tesis. Gracias, por sus interesantes perspectivas y por haber dedicado tiempo a la revisión del trabajo.

A mis amigos de la facultad y a toda la gente que cree en mi.

Mi alma máter, la Universidad Nacional Autónoma de México y en especial a la Facultad de Ciencias por haberme dado una cómoda estancia y por los amigos que me permitiste hacer.

Gracias a todos.

Introducción

La Clasificación

La clasificación se ha convertido en una aplicación muy importante en la estadística, usualmente es utilizada para ordenar información de manera explícita y sistemática. Se puede distinguir dos formas de clasificación. La primera forma consiste en encontrar una agrupación natural de los datos en donde el número de grupos no está previamente definido. En la segunda forma se utilizan dos palabras clave; discriminar y clasificar, en donde discriminar se refiere a separar distintos conjuntos de objetos u observaciones, encontrando variables que expliquen tal separación. Clasificar, se refiere a asignar nuevos elementos (observaciones) en los grupos previamente bien definidos. La primera forma está relacionada con el análisis de conglomerados (*cluster analysis*), pero el término de clasificación es más usado para la segunda forma (Venables and Ripley 2002, pág. 331).

Debido a que los problemas de clasificación aparecen en diversas áreas, desde una simple actividad cotidiana hasta adquirir una formalidad científica, ha tenido como consecuencia el desarrollo de diversos métodos, entre ellos destacan los estadísticos y los no estadísticos. Algunas técnicas estadísticas suelen llegar a un modelo predictivo, donde se puede encontrar una cierta explicación de una regla de clasificación (discriminar), mientras que otros no la ofrecen, en cambio usan una decisión sin una explicación (clasificar). Por ejemplo, no es necesario explicar cómo una máquina lee e identifica códigos postales, pero en áreas como medicina, es necesario obtener explicaciones (Venables and Ripley 2002, pág. 331). Usualmente discriminar y clasificar pueden usarse de manera conjunta.

Los métodos de clasificación son un tema muy importante ya que su metodología es ampliamente usada en diversas áreas desde las económicas y sociales hasta las ciencias biológico-médicas. En el área de riesgo de crédito ha tenido un gran auge la implantación de métodos de clasificación con el fin de acelerar el proceso de selección de clientes, este tema es conocido como *credit scoring*.

El *credit scoring* tiene 60 años de historia en los países de primer mundo, aquí en México se tiene menos tiempo de haberse implementado y se está conociendo por parte del buró de crédito. Como en México no existe suficiente bibliografía acerca del *credit scoring* e implementación de modelos de clasificación, existe una gran importancia de implementar estas técnicas

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

en el área de crédito, ya que el crédito es un medio de financiamiento puede traer beneficios y progreso para cualquier persona que lo utilice ya sea física o moral.

El presente trabajo está dirigido a personas que tienen conocimientos sobre los métodos estadísticos y están interesadas en aplicarlos a problemas prácticos. La finalidad del mismo es que pueda servir como una pequeña guía que ilustra el uso de los modelos de clasificación. La aplicación de éstos se enfocará en el contexto del *credit scoring*, porque utilizan numerosos métodos de clasificación y además resulta ser un ejemplo fácil de explicar. Sin embargo, se presentan otros ejemplos en otros contextos donde los métodos de clasificación pueden ser usados.

¿Qué és el Credit Scoring?

El *credit scoring* consiste en un conjunto de modelos de decisión que tienen como objetivo principal ayudar a un prestador de crédito a determinar si una persona es merecedora de un crédito, si será un cliente potencialmente bueno o por el contrario si será malo, esto significa que será capaz de pagar el crédito o generará problemas por incumplimiento de pago.

Y uno se preguntará *¿cómo se hace?*, normalmente los prestadores de crédito extraen y reúnen información de la solicitud, del informe y de la experiencia de crédito, ya sea proporcionadas por sociedades de información crediticia (buros de crédito) o recolectadas por la misma institución financiera. Entre esa información es posible que se encuentren datos socio-demográficos (e.g. edad, estado civil, ocupación, etc.) y características socio-económicas (e.g. ingreso, dependientes económicos, educación, etc.). Además utilizan datos acerca del rendimiento económico como el historial del pago de cuentas, el volumen y el tipo de cuentas que tenga la persona, los pagos atrasados, solicitudes recientes, deudas pendientes y datos de cuánto tiempo lleva con las deudas actuales. Por medio del software disponible en el mercado son sometidos a un conjunto de modelos estadísticos y no estadísticos.

Muchos de estos métodos conducen a un *scorecard* o tarjeta de puntos, donde las características evaluadas por el modelo les proporciona un puntaje, que ayudará a determinar si a alguien se le otorga el crédito o ver cúal es la probabilidad de pagar o no pagar un préstamo en la fecha de vencimiento. Otras técnicas no terminarán en un *scorecard*, en cambio, indicarán directamente la posibilidad de que un cliente será bueno y valdrá la pena sólo si éste aceptase una oferta de crédito[1].

El *credit scoring* tiene la particularidad de ser un sistema de evaluación automático, rápido y objetivo para determinar el otorgamiento de los créditos, con capacidad de predecir la probabilidad de pago o no pago asociada a una operación crediticia. Estas técnicas ayudan a ver qué estrategias y medidas operativas se utilizarán para poder mejorar la rentabilidad y la

[1]Este tipo de análisis es utilizado principalmente para incrementar el límite de crédito del usuario, estas técnicas llevan el nombre de *behavioral scoring*. Véase Thomas et al. (2002), sección 1.1.

solvencia del negocio. Cada prestador puede usar su propio modelo de puntuación, modelos diferentes para tipos de crédito o un modelo de puntuación genérico desarrollado por una empresa de *credit scoring*, por lo que algunos prestadores evaluarán a un individuo como un cliente bueno mientras que otros no.

Mucho antes de la aparición del *credit scoring*, el juicio humano desenvolvía un papel fundamental en la concesión de un crédito. Los prestamistas usaban sus experiencias al observar el comportamiento de crédito de un consumidor, haciéndolo como base para evaluar nuevos clientes. Este proceso no sólo era lento sino también inviable ya que se incurría en el error humano.

En los Estados Unidos con la introducción de las tarjetas de crédito a finales de la década de los años sesenta el *credit scoring* se tornó en un tema de suma importancia. Con los avances de computación, almacenamiento y extracción de datos se fueron introduciendo modelos mucho más sofisticados, y muchos de estos métodos se emplean en otras áreas de conocimiento.

En México el *credit scoring* lleva sólo unas décadas de haberse introducido, se está conociendo por el buró de crédito, donde introdujo en sus servicios en línea un producto que se llama "Mi Score" que por solo $50 pesos más IVA calcula un puntaje de crédito[2].

Métodos y Problemas de los Métodos de Clasificación

Por ejemplo, el *credit scoring* trata de desarrollar distintos modelos aproximados a partir de los métodos de clasificación, para distinguir entre buenos solicitantes y malos solicitantes de crédito. Se han elaborado en las últimas décadas muchas técnicas de clasificación que han sido adoptadas por el *credit scoring*, que provienen principalmente de la minería de datos estadística.

Hand (2009) indica que la minería de datos se basa en analizar enormes bases de datos que fueron construidas por medio de la observación con la intención de encontrar relaciones inusuales. En el *credit scoring* la identificación de estos patrones son deseados para el propósito de identificar qué miembros de la cartera tienen características inusuales y predecir qué tipo de cliente será un nuevo solicitante.

Los métodos de clasificación son numerosos, entre ellos se distinguen los métodos estadísticos. Entre los métodos estadísticos se encuentran el análisis de discriminante, la regresión logística, los k-vecinos más cercanos y los árboles de clasificación. Por otro lado, existen los métodos no estadísticos, que provienen de un origen matemático o de aprendizaje máquina, como la programación matemática, las redes neuronales, los algoritmos genéticos y los sistemas expertos[3]. Otros son el resultado de la combinación de métodos, por ejemplo, los árboles

[2] *http://www.burodecredito.com.mx/miscore.html#2009-1.*

[3] Véase capítulo 5 de Thomas et al. (2002).

de regresión logística.

Para poder comparar entre modelos similares y distintos modelos de clasificación se han desarrollado algunos métodos. Uno de los más directos son las tasas de clasificación, éstos pueden ser calculados a partir de los resultados arrojados por el modelo. Las tasas de clasificación se crean a partir del número de observaciones al hacer una tabla de contingencia de la clase original y la clase pronosticada, cuadro 1, las tasas se calculan de acuerdo al cuadro 2.

	Clase observada		Total pronosticados
Clase pronosticada	Malos	Buenos	
Malos	a	b	a+b
Buenos	c	d	c+d
Total de observados	a+c	b+d	TOTAL=a+b+c+d

Cuadro 1: Ejemplo de una matriz de confusión. Indican cuantos individuos se pronostican buenos como buenos, buenos como malos, malos como buenos y malos como malos.

	Clase observada	
Clase pronosticada	Malos	Buenos
Malos	a/(a+b)	b/(a+b)
Buenos	c/(c+d)	d/(c+d)
Tasa global de clasificación correcta		tgcb=(a+d)/(a+b+c+d)

Cuadro 2: Cálculo de tasas de clasificación. Donde *tgcb* es la tasa global de clasificación correcta.

Otro método utilizado muy a menudo son las curvas ROC. La curva ROC, del inglés *receiver operating characteristic curve*, es una representación gráfica de las tasas de clasificación. Para poder comparar las curvas de diferentes modelos de clasificación se compara el área bajo la curva (AUC), el área da un pequeño esquema del rendimiento del clasificador. En donde, un buen clasificador tendrá un AUC mayor a 0.5. También otros métodos para medir el rendimiento son: Validación Cruzada, *Jackkinfing*, la distancia de Malahanobis y la estadística de Kolmogorov-Smirnov, el coeficiente de Gini y la aproximación Delta (Thomas et al. 2002, capítulo 7).

A pesar de que los métodos tienen un gran beneficio en áreas como el *credit scoring*, deben de tomarse en cuenta algunas limitaciones que pueden acarrear al emplearlos (Hand, 2009):

1. *El problema de selectividad.* Uno de los mayores problemas que pueden surgir al construir un modelo de *credit scoring* es, que el modelo pudo haber sido elaborado por una base de datos tendenciosa de clientes a quienes solamente se les otorgó el crédito. Esto

puede ocurrir porque a los candidatos que fueron rechazados por un proceso de selección pudieron no haber sido considerados en la construcción del modelo, portanto se produce una muestra sesgada, repercutiendo en la estabilidad del modelo ya que en el peor de los casos, el modelo construido calificará mal a las personas, traduciéndose en un costo para la entidad financiera.

2. *El problema de los datos fuera de fecha.* Este problema se refiere al cambio de patrones sobre el tiempo, lo cual implica que el modelo construido caducará con el paso del tiempo a causa de una diversidad de circunstancias (escenarios políticos, económicos, avances tecnológicos, etc.). A partir de este problema se han explorado aproximaciones desde el análisis de supervivencia hasta modelos más sofisticados de regresión logística.

3. *El problema del tipo de modelo.* En estadística se distinguen dos tipos de modelo; los *icónicos* y los *empíricos*. Los modelos icónicos son representaciones matemáticas simplificadas de la realidad. Los modelos empíricos tienen por objeto encontrar de manera conveniente o útil un resumen de los datos. Los modelos icónicos deberían de arrojar mejores resultados que los empíricos ya que los modelos icónicos permiten una aproximación razonable al sistema con base en funciones fundamentadas sin buscar un modelo extensivo y sin el peligro de añadir funciones superfluas, que por simple casualidad se ajustaron bien a los datos (por ejemplo una regresión). Los modelos empíricos son el resultado de buscar ampliamente funciones fundamentadas o el resultado de una restricción apriori a un conjunto particular (combinación lineal de variables explicativas). Este problema está presente ya que casi todos los modelos de predicción del *credit scoring* son empíricos, lo cual implica que el deterioro del modelo es inminente por el cambio de la sociedad debido a diversas circunstancias mencionadas en el problema número dos.

4. *El problema de la medición del rendimiento.* Quizás un problema muy frecuente es qué modelo de clasificación se puede elegir de todos los modelos ajustados. Este problema apunta a las medidas utilizadas para medir el rendimiento, ya que los distintos criterios utilizados para la elección del modelo pueden llevar a distintas conclusiones. En problemas de clasificación, las tasas de clasificación errónea calculadas a partir de los modelos antes mencionados, es una medida de rendimiento, ya que se busca minimizar dichas tasas y suelen ser diferentes a aquellos basados en la verosimilitud. Se debe de sopesar las debilidades entre la verosimilitud y aquellas medidas de rendimiento utilizadas a partir de las tasas de clasificación.

Objetivo y Delimitación

En los problemas de clasificación existen diversos modelos. El presente trabajo se enfocará en la exposición de los modelos de regresión logística, de discriminante y de regresión de Cox. El objetivo principal es presentar estos tres modelos e ilustrar su aplicación a dos conjuntos de datos. Un primer conjunto de datos pertenece al área de puntaje de crédito, del inglés *credit scoring* y el segundo pertenece al área de bioestadística. Los modelos serán

comparados en la teoría y en la aplicación. Además de ajustar estos modelos se comentan sobre sus similitudes y diferencias. Se escogió la regresión logística ya que ha probado ser una de las herramientas tradicionales más robustas y el análisis de discriminante porque es el modelo más antiguo para problemas de clasificación. Para poder llegar a este objetivo también se pasará por un proceso de aprendizaje de las bases de datos que se trabajarán, con el objetivo de disminuir e identificar las variables más significativas dentro del conjunto de datos.

Lamentablemente, no se puede introducir la regresión de Cox en la aplicación al *credit scoring* ya que no contiene ninguna variable de supervivencia (tiempos de falla). En esta base de datos solamente se ajusta el análisis de discriminante y la regresión logística, por un lado. Por otro con la base de datos de bioestadística sí se puede ajustar el modelo de regresión de Cox ya que la base de datos contiene tiempos de falla, se ajusta la regresión logística y la regresión de Cox y se comparan ambos métodos.

El vincular el análisis de supervivencia con métodos de clasificación no es una tarea fácil. En el *credit scoring* no es reciente la vinculación, Stepanova and Thomas (2002) desarrollaron algunos modelos de regresión de Cox y regresión logística para predecir incumplimiento o pago anticipado.

Esquema del Capitulado de la Tesis

En el capítulo 1 se describe el modelo de la regresión logística. En el capítulo 2 se describe el análisis de discriminante, otro método estadístico de clasificación. En el capítulo 3 se verá una breve introducción al análisis de supervivencia y se describe el modelo de regresión de Cox. En el capítulo 4 se aplican los modelos a las bases de datos. En el capítulo 5 se comparan los modelos ajustados de clasificación. En el capítulo 5.5 se concluye la tesis y se realiza una pequeña discusión.

Capítulo 1

Modelo de Regresión Logística

1.1. Introducción

Los modelos de regresión son una herramienta importante en el análisis de datos en muchas áreas de investigación, estos son aplicados para medir cuál es la relación entre una o varias variables explicativas X_j, $j = 1, 2, ..., p$ contra una variable respuesta Y e identificar cuáles subconjuntos de variables explicativas contienen información reduntante acerca de Y, así conociendo algún subconjunto, lo restante ya se considera como informativo o significante. Es usual encontrar problemas en donde la variable respuesta sea discreta tomando dos o más categorías. El uso de la regresión logística es ideal cuando se presenta esta situación, en donde la variable respuesta representa éxito o fracaso, o generalmente la presencia o ausencia de un atributo de interés.

En este capítulo se abordan diversos aspectos del modelo de regresión logística, por lo que en la sección 1.2 se define el modelo, seguida de la explicación de como se obtienen estos coeficientes (sección 1.2.1) y cómo se interpretan los coeficientes del modelo ajustado (sección 1.2.2). En la sección 1.3 se ven la pruebas de razón de verosimilitudes (sección 1.3.1) y de Wald (sección 1.3.2) para determinar si algún coeficiente no es estadísticamente significante. En la sección 1.4 se describen los intervalos de confianza. En la sección 1.5 se verán otras pruebas estadísticas, la estadística Ji-cuadrada de Pearson (sección 1.5.1) y la de Hosmer-Lemeshow (sección 1.5.2) para poder evaluar si el modelo ajusta bien a los datos o no. En la sección 1.6 se darán unos comentarios para poder seleccionar un buen modelo de regresión logística. En la sección 1.7 se dan algunos ejemplos en los cuales la regresión logística es usualmente aplicada.

1.2. Ajuste de la Regresión Logística Binaria

La regresión logística con respuesta binaria es la forma más simple y probablemente la más usada. El objetivo sería evaluar si un conjunto de las variables explicativas están significativamente relacionadas con las dos categorías de la variable respuesta.

Suponga una muestra de n observaciones independientes del par $(\mathbf{x}_i, y_i)$, $i = 1, ..., n$, donde y_i denota el valor de la variable respuesta, y $\mathbf{x}_i$ es el valor de las variables explicativas para el i-iésimo sujeto. Considere que Y, la variable respuesta, se codifica como 0 ó 1, representando la ausencia o presencia de la característica, respectivamente. Además, $\mathbf{X} = \{X_1, ..., X_p\}$ es el conjunto de variables explicativas, entonces el modelo de regresión logística es como sigue.

La probabilidad de que Y tome el valor de la categoría 1 se denota de la siguiente manera:

$$\begin{aligned} P(Y = 1|\boldsymbol{X} = \boldsymbol{x}) &= \pi(\boldsymbol{x}) \\ &= \frac{exp(\boldsymbol{\beta}\boldsymbol{x}^T)}{1 + exp(\boldsymbol{\beta}\boldsymbol{x}^T)} \\ &= \frac{exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}. \end{aligned} \tag{1.1}$$

La probabilidad de que Y tome el valor de la categoría 0 es:

$$P(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = 1 - \pi(\boldsymbol{x}), \tag{1.2}$$

en donde $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ es el vector de parámetros desconocidos del modelo. Observe que la ecuación (1.1) no es lineal en los parámetros, una manera equivalente de escribir esta ecuación es utilizando la transformación logito (en inglés *logit*), obteniendo así una función lineal en los parámetros:

$$\begin{aligned} g(\boldsymbol{x}) &= logito(\pi(\boldsymbol{x})) \\ &= log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) \\ &= \boldsymbol{\beta}\boldsymbol{x}^T \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \end{aligned} \tag{1.3}$$

Nótese que el rango de $\pi(\boldsymbol{x})$ en la ecuación (1.1) se encuentra entre 0 y 1, mientras que el rango de valores que toma la transformación logito en la ecuación (1.3) se encuentra en el intervalo $(-\infty, \infty)$. Haciendo una comparación de estas ecuaciones equivalentes, $\pi(\boldsymbol{x})$ resulta ser mucho más fácil de interpretar que su equivalente, ya que la medida de la probabilidad es mucho más comprensible que la misma transformación.

Por otro lado, la razón

$$\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}, \tag{1.4}$$

o bien

$$\begin{aligned}\frac{P(Y=1|\boldsymbol{X}=\boldsymbol{x})}{P(Y=0|\boldsymbol{X}=\boldsymbol{x})} &= exp(\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_px_p)\\ &= exp(\boldsymbol{\beta x}^T),\end{aligned} \tag{1.5}$$

es llamado momio (en inglés *the odds*), es utilizado para la construcción de una medida llamada *the odds ratio*, que suele utilizarse como una medida que cuantifica el riesgo. Esta medida en la literatura tiene traducción al español como la razón de momios.

1.2.1. Estimación de los Parámetros

La estimación de los parámetros en la regresión logística es llevada a cabo por el método de máxima verosimilitud. Como es sabido la función de verosimilitud expresa la probabilidad de los datos observados como una función de los parámetros desconocidos.

Considere una muestra de n observaciones independientes, la pareja observada ($\boldsymbol{x}_i$,y_i), $i = 1, ..., n$. Donde y_i es el valor de la variable respuesta Y, codificada con los valores 0 y 1. Sea $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ el vector de parámetros y $\boldsymbol{x_i} = (1, x_{i1}, ..., x_{ip})$ el valor observado del sujeto i, entonces para la pareja observada ($\boldsymbol{x}_i$,y_i) la probabilidad de que y_i tome el valor 1 es $\pi(\boldsymbol{x}_i)$ y de que tome el valor 0 es $1-\pi(\boldsymbol{x}_i)$. Donde $\pi(\boldsymbol{x}_i)$ es la función $\pi(\boldsymbol{x})$ evaluada en $\boldsymbol{x}_i$. Entonces para la pareja observada ($\boldsymbol{x}_i$,y_i) la probabilidad se puede expresar de la siguiente manera de acuerdo al modelo de probabilidad Bernoulli:

$$P[Y=y_i] = \pi(\boldsymbol{x}_i)^{y_i}\left(1-\pi(\boldsymbol{x}_i)\right)^{1-y_i}. \tag{1.6}$$

Por el supuesto de independencia entre las observaciones, la expresión de la función de verosimilitud queda:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(\boldsymbol{x}_i)^{y_i}\left(1-\pi(\boldsymbol{x}_i)\right)^{1-y_i}. \tag{1.7}$$

Obteniendo el logaritmo de la ecuación (1.7) queda la función de log-verosimilitud definida como:

$$\begin{aligned}L(\boldsymbol{\beta}) = ln\left(\ell(\boldsymbol{\beta})\right) &= \sum_{i=1}^{n} y_i ln\left(\pi(\boldsymbol{x}_i)\right) + \sum_{i=1}^{n}\left(1-y_i\right) ln\left(1-\pi(\boldsymbol{x}_i)\right)\\ &= \sum_{i=1}^{n} y_i ln\left(\pi(\boldsymbol{x}_i)\right) + \sum_{i=1}^{n} ln\left(1-\pi(\boldsymbol{x}_i)\right) + \sum_{i=1}^{n} -y_i ln\left(1-\pi(\boldsymbol{x}_i)\right)\\ &= \sum_{i=1}^{n} y_i ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) + \sum_{i=1}^{n} ln\left(1-\pi(\boldsymbol{x}_i)\right).\end{aligned} \tag{1.8}$$

Se empezará por escribir la ecuación (1.8) en términos de los parámetros. Recordando la transformación logito dada por la ecuación (1.3) y reescribiendo el signo del segundo término como sigue,

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}^T + \sum_{i=1}^{n} -\left(-ln\left(1-\pi(\boldsymbol{x}_i)\right)\right) \\ &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}^T - \sum_{i=1}^{n} -ln\left(1-\pi(\boldsymbol{x}_i)\right) \\ &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}^T - \sum_{i=1}^{n} ln\left(\frac{1}{1-\pi(\boldsymbol{x}_i)}\right). \end{aligned} \tag{1.9}$$

Sumando $\pi(x_i) - \pi(x_i)$ en el numerador del segundo término se tiene:

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}_i^T - \sum_{i=1}^{n} ln\left(\frac{1+\pi(\boldsymbol{x}_i)-\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) \\ &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}_i^T - \sum_{i=1}^{n} ln\left(1+\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) \\ &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}_i^T - \sum_{i=1}^{n} ln\left(1+exp\left\{ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)\right\}\right) \\ &= \sum_{i=1}^{n} y_i \ \boldsymbol{\beta}\boldsymbol{x}_i^T - \sum_{i=1}^{n} ln\left(1+exp(\boldsymbol{\beta}\boldsymbol{x}_i^T)\right) \\ &= \sum_{i=1}^{n} y_i \ (\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip}) - \sum_{i=1}^{n} ln\left(1+exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})\right). \end{aligned} \tag{1.10}$$

Según el método de máxima verosimilitud para encontrar los valores $\hat{\boldsymbol{\beta}}$ que maximizan a $L(\boldsymbol{\beta})$ con respecto al vector de parámetros $\boldsymbol{\beta}$, la función $L(\boldsymbol{\beta})$ dada en (1.10) tiene que derivarse con respecto a cada parámetro e igualarse a cero. Esto forma un sistema de ecuaciones.

Derivando con respecto a β_0 queda la siguiente expresión de (1.10)

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})}{1+exp(\beta_0+\beta_1 x_{i1}+...+\beta_p x_{ip})} \\ &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\boldsymbol{x}_i), \end{aligned} \tag{1.11}$$

y con respecto a β_j, $j = 1, ..., p$ la derivada queda como sigue

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij} \frac{exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})} \\ &= \sum_{i=1}^{n} x_{ij} \left[y_i - \pi(\boldsymbol{x}_i) \right].\end{aligned} \tag{1.12}$$

Igualando a cero 1.11 y 1.12 se tiene respectivamente

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\boldsymbol{x}_i) = 0, \tag{1.13}$$

$$\sum_{i=1}^{n} x_{ij} \left[y_i - \pi(\boldsymbol{x}_i) \right] = 0, j = 1, ..., p. \tag{1.14}$$

El sistema de $p + 1$ ecuaciones no es un sistema lineal, y no es posible despejar explícitamente los coeficientes de acuerdo a como está definido $\pi(\boldsymbol{x}_i)$. Esto nos lleva a recurrir a métodos numéricos para su solución. Estos métodos numéricos han sido implementados en el software estadístico y no serán revisados en este trabajo.

En la regresión logística es de interés la desviación estándar del coeficiente β_i, su principal uso es medir la significancia de dicho parámetro. Para poder obtener los estimadores de las varianzas y covarianzas es necesario obtener las segundas derivadas de la función de log-verosimilitud. Estas derivadas están expresadas en las ecuaciones (1.15) y (1.16), donde π_i denota $\pi(\boldsymbol{x}_i)$, $i = 1, ..., n$ y $j, u = 0, ..., p$.

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i), \tag{1.15}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{n} x_{ij} x_{iu} \pi_i (1 - \pi_i). \tag{1.16}$$

Los elementos de la matriz de información denotada como $\boldsymbol{I(\beta)}$, se define como el inverso aditivo de las derivadas de segundo orden representadas en las ecuaciones (1.15) y (1.16). Es decir, el elemento $(\boldsymbol{I(\beta)}_{ju})$ de la matriz de información está dado por

$$(\boldsymbol{I(\beta)}_{ju}) = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u}. \tag{1.17}$$

Si $u = j$, entonces, es un elemento de la diagonal de dicha matriz. Para calcular las varianzas y covarianzas de los coeficientes estimados, se obtiene la inversa de la matriz de información.

$$\boldsymbol{V(\beta)} = \left[\boldsymbol{I(\beta)} \right]^{-1}. \tag{1.18}$$

Los estimadores de las varianzas y covarianzas serán denotadas como $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$, estos son obtenidos al evaluar $\hat{\boldsymbol{\beta}}$ en la ecuación (1.18).

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \Big[\boldsymbol{I}(\hat{\boldsymbol{\beta}})\Big]^{-1}. \tag{1.19}$$

Los términos de la diagonal de la matriz son las varianzas es $\hat{V}(\hat{\beta}_j)$. Las covarianzas son los elementos fuera de la diagonal: $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_u)$, $j, u = 0, ..., p$ $j \neq u$. La estimación del error estándar del parámetro β_j, $j = 1, ..., p$ se denotan como

$$\hat{SE}(\hat{\beta}_j) = \Big(\hat{V}(\hat{\beta}_j)\Big)^{1/2}. \tag{1.20}$$

Una manera útil de representar la matriz de información es de manera matricial (Hosmer and Lemeshow 2000, pág. 35)

$$\hat{\boldsymbol{I}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}, \tag{1.21}$$

donde $\boldsymbol{X}$ es una matriz de dimensión $n \times (p+1)$ y $\boldsymbol{X}'$ es la matriz transpuesta de $\boldsymbol{X}$, que contiene los datos muestrales de cada sujeto u observación y $\boldsymbol{V}$ es una matriz de $n \times n$ con los elementos generales de $\hat{\pi}_i(1-\hat{\pi}_i)$. Será útil al representar los intervalos de confianza en la sección 1.4.

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ \vdots & & \ddots & \vdots & \\ 1 & x_{n1} & x_{2p} & ... & x_{np} \end{pmatrix}, \tag{1.22}$$

$$\boldsymbol{V} = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & ... & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & ... & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & ... & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}. \tag{1.23}$$

1.2.2. Interpretación de los Parámetros

La tarea después de ajustar un modelo es poder obtener información relevante. La tarea primordial después del cómputo y el análisis de la significancia estadística de los coeficientes estimados, es la interpretación de los valores de dichos coeficientes. Sin embargo, es conveniente primero evaluar el ajuste del modelo antes de intentar interpretarlo. Por ello, para introducir algunas formas de análisis del modelo ajustado, se supondrá que se ha evaluado tanto el ajuste general, como la significancia estadística de los parámetros, ya sea en el sentido estadístico o en el natural admisible, (Hosmer and Lemeshow 2000, pág. 47).

Siendo β_0, la constante del modelo y $\beta_1, \beta_2, ..., \beta_p$ los coeficientes de la regresión para las variables explicativas $X_1, X_2, ..., X_p$, de la ecuación (1.1) se deriva la siguiente expresión

$$\pi(\boldsymbol{x}) = \frac{1}{1 + \frac{1}{exp(\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_p x_p)}}. \tag{1.24}$$

Considere el caso cuando existe una sola variable explicativa ($p = 1$). En la regresión lineal el coeficiente β_1 representa la pendiente de la función, siendo igual a la diferencia entre la variable dependiente en $x + 1$ y x (i.e, $\beta_1 = y(x + 1) - y(x)$). En la regresión logística, el coeficiente representa la diferencia entre la transformación logito por una unidad en la variable independiente (i.e, $\beta_1 = g(x + 1) - g(x)$).

Naturalmente al analizar cuando se tiene una sola variable explicativa de la ecuación (1.24), se destacan las siguientes características:

- Un coeficiente positivo quiere decir que un aumento en la variable explicativa aumenta la probabilidad de la variable respuesta.
- Un coeficiente negativo quiere decir que un aumento en la variable explicativa disminuye la probabilidad de la variable respuesta.
- Un coeficiente grande quiere decir que esa variable explicativa influye fuertemente en la probabilidad de la variable respuesta.
- Un coeficiente pequeño (cercano a cero) quiere decir que esa variable explicativa tiene poca influencia en la probabilidad de la variable respuesta.

Por otro lado, dependiendo del tipo de variable explicativa se puede hacer este tipo de interpretación:

- Si X es una variable binaria. Entonces, puede decirse que la presencia de esta característica, dependiendo del signo, disminuye o aumenta la probabilidad de la variable respuesta.
- Si X es una variable continua. Se dice que esta variable dependiendo del signo, disminuye o aumenta la probabilidad de la variable respuesta a medida de que X aumenta.

Por otro lado, la razón de momios también es usado para poder interpretar los coeficientes del modelo. Por ejempo, si X es una variable binaria y se considera el momio de la transformación logito. La razón de momios (OR) se define como la razón del valor del momio cuando $X = 1$ entre el valor del momio cuando $X = 0$. Es decir,

$$OR = \frac{\pi(1)/(1 - \pi(1))}{\pi(0)/(1 - \pi(0)}. \tag{1.25}$$

Substituyendo el valor de las probabilidades se tiene

$$\begin{aligned} OR &= \frac{\frac{exp(\beta_0+\beta_1)}{1+exp(\beta_0+\beta_1)} / \frac{1}{1+exp(\beta_0+\beta_1)}}{\frac{exp(\beta_0)}{1+exp(\beta_0)} / \frac{1}{1+exp(\beta_0)}} \\ &= \frac{exp(\beta_0 + \beta_1)}{exp(\beta_0)} \\ &= exp(\beta_1). \end{aligned} \tag{1.26}$$

La interpretación de la razón de momios (OR) está basado en otra medida que cuantifica el riesgo, esta medida es llamada el riesgo relativo (*relative risk* (RR)). El riesgo relativo se define como

$$RR = \frac{\pi(1)}{\pi(0)}. \tag{1.27}$$

Entonces, la razón de momios se parece al riesgo relativo cuando $\frac{1-\pi(0)}{1-\pi(1)} \approx 1$. Ocurre cuando $\pi(x)$ es pequeña para $x = 0$ y $x = 1$ o cuando $\pi(0) \approx \pi(1) \approx 0{,}5$. Cuando el riesgo relativo es igual a 1 indica que no hay diferencia de riesgo entre aquellos que tienen la característica, $x = 1$, y aquellos que no, $x = 0$. Cuando $RR > 1$ el evento tiende a ocurrir más en el grupo donde presentan la característica, $x = 1$, que aquellos que no lo tienen, $x = 0$. Si $RR < 1$, entonces el evento tiende menos a ocurrir en aquellos que no presentan la característica, $x = 0$, que aquellos que la tienen, $x = 1$.

La interpretación para la razón de momios cuando X es una variable dicotómica es, de acuerdo con lo mencionado en el párrafo anterior; en la presencia de x ($x = 1$) la ocurrencia del evento de interés es de $\widehat{OR}$. Por ejemplo, si $\widehat{OR} = 2$ el evento de interés ocurre el doble en aquellos que presentan la característica ($x = 1$). Si el evento de interés ocurre el triple en aquellos que presentan la característica ($x = 1$), entonces $\widehat{OR} = 3$. Si la ocurrencia del evento de interés es de la mitad cuando se presenta la característica $x = 1$, entonces $\widehat{OR} = 0{,}5$. Se mencionan ejemplos más detallados en de Hosmer and Lemeshow (2000), ver página 50.

El cálculo de la razón de momios también es posible directamente de la tabla de contingencia de 2×2 de la variable explicativa contra la variable respuesta, dicha tabla se presenta en el cuadro 1.1.

	X=1	X=0	Total
Y=1=Presencia	a	b	a+b
Y=0=Ausencia	c	d	c+d
Total	a+c	b+d	n

Cuadro 1.1: Tabla de contingecia de 2×2 de Y y X, cuando X es una variable binaria

Entonces el cálculo de la razón de momios es como sigue

$$\widehat{OR} = \frac{a/c}{b/d}.$$

Cuando se trata de una variable categórica el cálculo de la razón de momios es similar sólo que ahora se compara contra la categoría de referencia. Hosmer and Lemeshow (2000) (sección 3.3), demuestran que la estimación de los coeficientes usando la regresión logística son iguales a las razones de momios usando tablas de contingencia debido a la codificación

de las variables *dummies*.

Cuando la variable del coeficiente ajustado es continua, la interpretación del coeficiente ahora depende de los valores que toma la variable. Como $g(x)$ es lineal, β_1 es igual a la diferencia entre de $g(x)$ en el incremento de $g(x)$ por una unidad, $\beta_1 = g(x+1) - g(x)$, para cualquier valor de x. Pero normalmente utilizan en un cambio para cualquier c arbitraria. Entonces el cambio de la función logito por incremento de c en x es el siguiente

$$\begin{aligned} g(x+c) - g(x) &= \beta_0 + \beta_1(x+c) - (\beta_0 + \beta_1 x) \\ &= c\beta_1. \end{aligned} \tag{1.28}$$

Para la expresión de la razón de momios de la ecuación (1.25) tomando el incremento de x en c arbitraria es

$$\begin{aligned} OR(c) = OR(x+c, x) &= \frac{exp(\beta_0 + \beta_1 x + c\beta_1)}{exp(\beta_0 + \beta_1 x)} \\ &= exp(c\beta_1), \end{aligned} \tag{1.29}$$

$\widehat{OR}(c)$ está dada cuando se obtiene $\hat{\beta}_1$, la estimación de los coeficientes es dada en la siguiente subsección. La interpretación es similar: por cada incremento en c unidades de esta variable el riesgo de que ocurra el evento de interés es de $\widehat{OR}(c)$.

1.3. Pruebas de Hipótesis sobre los Parámetros

La regresión logística también puede ser usada como un modelo exploratorio en la construcción de otros modelos. Un número de modelos anidados que son subconjuntos del modelo más grande o saturado, es comparado para determinar cuál es el más simple (parsimonioso[1]) y para predecir satisfactoriamente la probabilidad de caer en una de las dos categorías de la variable respuesta casi tan bien como el modelo original. Después de haber ajustado el modelo de regresión logística, la primera cosa que se necesita observar es la significancia estadística de las variables en el modelo. Por lo que es necesario hacer pruebas de hipótesis y determinar que variables explicativas son significativas en la variable respuesta. Regularmente se utiliza la prueba de Wald y el cociente de verosimilitudes. Estas pruebas se centran en probar la hipótesis $H_0 : \beta = 0$ *vs* $H_a : \beta \neq 0$.

1.3.1. Prueba del Cociente de Verosimilitudes

En la prueba del coeciente de verosimilitudes la hipótesis que se desea probar es

[1]Principio de parsimonia: En igualdad de condiciones la solución más sencilla que explique completamente un problema es probablemente la correcta (Guillermo de Ockham). Según este principio, cuando más de un modelo se ajuste a nuestras observaciones, siempre deberíamos quedarnos con el modelo más simple que explique nuestras observaciones con un grado adecuado de precisión.

$$H_0 : \beta_1 = ... = \beta_m = 0 \ vs \ H_a : \beta_i \neq 0 \text{ para algún } i = 1, ..., m.$$

donde m son el número de las variables no incluidas en el modelo ajustado. Básicamente la estadística de prueba es menos dos veces el logaritmo natural de la verosimilitud del modelo ajustado entre el modelo saturado (Hosmer and Lemeshow 2000, pág. 13), como se da en la expresión siguiente

$$G = -2ln\left[\frac{verosimilitud\ del\ submodelo\ ajustado}{verosimilitud\ del\ modelo\ saturado}\right]. \tag{1.30}$$

La estadística G se distribuye aproximadamente como una Ji-cuadrada con $(p-k)$ grados de libertad, en donde p es el número de variables explicativas del modelo saturado y k el número de variables del submodelo ajustado (Hosmer and Lemeshow 2000, pág. 38).

Otra manera de verlo es por medio de la devianza. La devianza no es más que $D = -2L(\boldsymbol{\beta}) = -2ln\ell(\boldsymbol{\beta})$. Entonces la ecuación (1.30) se convierte en

$$G = -2L_o(\boldsymbol{\beta}) - (-2L_a(\boldsymbol{\beta})).$$

1.3.2. Prueba de Wald

En comparación con la prueba del cociente de verosimilitudes la hipótesis que se desea probar en la prueba de Wald es

$$H_0 : \beta_i = 0 \ vs \ H_a : \beta_i \neq 0 \text{ para } i = 1, ..., p.$$

Esta prueba es obtenida al hacer el cociente del parámetro estimado bajo máxima verosimilitud con su error estándar, la razón resultante sigue una distribución aproximada a una normal estándar. El estadístico de prueba es

$$W_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}, \tag{1.31}$$

en donde el valor de p para dos colas es $P(|z| > W_i)$ donde z sigue una distribución normal estándar. Esta prueba por su fácil aplicación ha sido examinada. Sin embargo, la prueba del cociente de verosimilitudes resulta ser favorita sobre la prueba de Wald, pues la prueba del cociente utiliza la verosimilitud y el parámetro estimado (Agresti 2002, pág. 172).

1.4. Intervalos de Confianza de los Parámetros

El uso de los intervalos de confianza es una herramienta importante para poder evaluar la significancia del modelo de los parámetros de interés. La base para poder construir los estimadores de los intervalos es la misma teoría aplicada para formular pruebas de hipótesis, particularmente los estimadores de los intervalos para los parámetros son obtenidos a partir

de la prueba de Wald.

El intervalo de confianza con un nivel de significancia del $100(1-\alpha)\,\%$ para los parámetros es el siguiente:

$$\hat{\beta}_i \pm z_{1-\alpha/2}\hat{SE}(\hat{\beta}_i),\ i = 0, ..., p. \tag{1.32}$$

La constante β_0 es importante cuando se consideran los puntos y los estimadores de los intervalos para la transformación logito (la parte lineal del modelo) de regresión logística. La estimación para el logito es:

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p. \tag{1.33}$$

Por otro lado, la estimación de la varianza para el estimador de la función logito es como sigue:

$$\hat{V}(\hat{g}(\boldsymbol{x})) = \sum_{j=0}^{p} x_j^2 \hat{V}(\hat{\beta}_j) + \sum_{j=0}^{p}\sum_{k=j+1}^{p} 2x_j x_k \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k). \tag{1.34}$$

Entonces el intervalo de confianza para $\hat{g}(\boldsymbol{x})$ con $100(1-\alpha)\,\%$ es:

$$\hat{g}(\boldsymbol{x}) \pm z_{1-\alpha/2}\hat{SE}(\hat{g}(\boldsymbol{x})), \tag{1.35}$$

donde $\hat{SE}(\hat{g}(x))$, es la desviación estándar de la ecuación (1.34) (Hosmer and Lemeshow 2000, págs. 19 y 42). Usando notación matricial acerca del estimador de las varianzas de los coeficientes por parte de la matriz de información.

$$\hat{V}(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}. \tag{1.36}$$

La varianza en la ecuación (1.35) queda como:

$$\hat{V}(\hat{g}(\boldsymbol{x})) = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{x}. \tag{1.37}$$

1.5. Bondad de Ajuste del Modelo

También es importante en el modelo de regresión logistica saber si el modelo se ajusta a los datos. A continuación se dan algunas medidas de bondad de ajuste que se utilizan frecuentemente en el modelo de regresión logística. Suponiendo que el modelo ajustado contiene p variables explicativas y sea J el número de todos los posibles valores de $\boldsymbol{x} = (x_1, ..., x_p)$, estos posibles valores llevarán el nombre de configuraciones específicas. Nótese que algunas observaciones tendrán los mismos valores entonces $J < n$. Considerando el número de observaciones que toma una configuración específica $\boldsymbol{x} = \boldsymbol{x}_j$ por m_j observaciones, $j = 1, ..., J$. Donde $\sum_{j=1} m_j = n$. Y además si son contados con respecto al valor que toma y_i, exclusivamente éxitos, n_1, entre los m_i observados o exclusivamente fracasos, n_2, y sucesivamente para todas las configuraciones específicas. Se puede describir con facilidad estas medidas de

bondad de ajuste, que normalmente indicarán si el modelo ajustado describe adecuadamete o no, la relación entre la variable respuesta y las variables explicativas.

1.5.1. Prueba de Ji-cuadrada de Pearson

En la regresión logística existen distintas formas para medir la diferencia de los valores observados y los valores ajustados $(y-\hat{y})$. Para enfatizar el hecho de que los valores ajustados en la regresión logística son calculados para cada configuración específica y que dependa en la probabilidad estimada para esa configuración, se denota el valor ajustado para la j-ésima configuración como $\hat{y}_j$, donde

$$\hat{y}_j = m_j\hat{\pi}_j = m_j\frac{e^{\boldsymbol{\beta}\boldsymbol{x}_j^T}}{1+e^{\boldsymbol{\beta}\boldsymbol{x}_j^T}}. \tag{1.38}$$

Los residuos de Pearson se definen como

$$r(y_j,\hat{\pi}_j) = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}},\ j=1,...,J. \tag{1.39}$$

La estadística de prueba basada de estos residuales es la Ji-cuadrada de Pearson y queda como

$$X^2 = \sum_{j=1}^{J} r(y_j,\hat{\pi}_j)^2, \tag{1.40}$$

donde J es el número total de configuraciones específicas. La distribución asintótica de la estadística χ^2 bajo el supuesto de que el modelo ajusta bien a los datos, es una distribución Ji-cuadrada con $J-(p+1)$ grados de libertad. Ocurren problemas cuando el número de configuraciones $J \approx n$, una de las razones es que hay tantas configuraciones como observaciones, usar la distribución Ji-cuadrada con $J-(p+1)$ grados de libertad sería incorrecto según Hosmer and Lemeshow (2000), para ver a mayor detalle se sugiere revisar la pág. 146, de esta referencia.

1.5.2. Prueba de Hosmer-Lemeshow

Hosmer and Lemeshow (2000), proponen agrupar observaciones por medio de las probabilidades estimadas y compararlas con las observadas en la variable respuesta.

Sea $J=n$ tantas configuraciones como observaciones, haciendo una tabla con n columnas ordenadas de menor a mayor. Se agruparán las observaciones de acuerdo a percentiles. El número que usualmente se usa para los grupos es $g=10$ o deciles. Los g grupos deberían contener aproximadamente el mismo número de observaciones, las cuales deben de tener la misma configuración específica aproximadamente igual al número total de observaciones.

La estadística de prueba de Hosmer-Lemeshow es como sigue

$$\hat{C} = \sum_{k=1}^{g} \frac{o_k - n'_k \bar{\pi}_k}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}, \tag{1.41}$$

donde n'_k denota el número de observaciones en el grupo k, $k = 1, ..., g$. Donde c_k es el número de configuraciones específicas en el grupo k.

$$o_k = \sum_{i=1}^{c_k} y_i,$$

es el número de respuestas entre las c_k configuraciones específicas y

$$\bar{\pi}_k = \sum_{i=1}^{c_k} \frac{m_i \hat{\pi}_k}{n'_k},$$

es la probabilidad promedio estimada en el grupo k.

Hosmer y Lemeshow usando simulaciones extensas, demostraron que cuando $J = n$ y el modelo ajustado es el correcto, la estadística de prueba de Hosmer-Lemeshow se distribuye asintóticamente a una Ji-cuadrada con $(g - 2)$ grados de libertad (Hosmer and Lemeshow 2000, pág. 149). A valores grandes de la estadística (valores de p pequeños) indican una falta de ajuste en el modelo[2].

1.6. Selección del Modelo

Para poder discernir qué variables deberán incluirse en el modelo de regresión logística, los criterios pueden ser variados. Desde el punto de vista estadístico la construcción de un modelo incluye buscar aquel modelo que sea el más parsimonioso que pueda explicar los datos, además de una cierta lógica para el experto en el área de aplicación.

Habiendo ajustado un modelo que contenga múltiples variables se puede buscar la reducción por medio de la devianza como una medida de ajuste. Las pruebas de hipótesis sobre los parámetros del modelo son más útiles para ayudar a simplificar el modelo. La prueba de Wald y la prueba del cociente de verosimilitudes son las más usadas. Sin embargo, no se recomienda usar la prueba de Wald y es preferible usar la de verosimilitud (Agresti 2002, pág. 172). Además, la prueba de cociente de verosimilitudes puede ayudar a comparar submodelos y detectar si un conjunto de variables son estadísticamente significativas. Las pruebas de bondad de ajuste también son importantes para ver la estabilidad del modelo.

Un análisis complementario es utilizando procedimientos automáticos en donde las variables son incluidas o excluidas secuencialmente. Principalmente existen dos métodos de selección de variables el *backward* y el *forward*. La mayoría de los paquetes estadísticos tienen

[2] *Online SAS/STAT(R) 9.22 User's Guide.*

estas opciones de selección. Los procedimientos automáticos no son un sustituto para orientar la formulación de modelos. Debieran usarse complementariamente sólo para explorar si hay mejores modelos.

Otro critero comúnmente utilizado es el llamado Criterio de Información de Akaike (AIC del inglés *Akaike Information Criterion*). Con el AIC se evalúa tanto el ajuste del modelo a los datos como la complejidad del modelo. El AIC de define como sigue

$$AIC = -2L + 2k \tag{1.42}$$

donde L es la log-verosimilitud del modelo, k es el número de parámetros. Cuanto más pequeño es el AIC mejor es el ajuste. El AIC es muy útil para comparar modelos anidados, pero puede ser comparado en modelos que tengan distintas variables. Algunos paquetes estadísticos tienen la ventaja de manejarlos como un proceso automatizado para la selección de modelo (Venables and Ripley 2002, pág. 174).

Una vez que se obtiene un modelo en donde se cree que contiene las variables esenciales se debe de considerarar la necesidad de incorporar interacciones entre ellas. Esto se debe de hacer con cuidado ya que puede haber pérdida de interpretabilidad. Buscar interacciones plausibles sería lo más recomendable. A veces se necesita sopesar entre la complejidad de modelo y la interpretabilidad. Generalmente a mayor complejidad existe menor interpretabilidad.

En algunos casos habrá más de un modelo que describa los datos igual de bien. En estos casos queda al criterio del investigador elegir uno u otro, aunque puede ser recomendable utilizarlos todos y discutir las limitaciones que esto presenta desde el punto de vista inferencial.

Por otro lado, la significancia de los parámetros no debiera de ser el único criterio para incluir variables. A veces por propósitos del estudio una variable puede considerarse significativa aunque en un modelo ajustado no lo es. Mantenerla puede reducir tendencia en otras variables y es posible compararlo con otros estudios en donde se encontró significativo, quizás por un tamaño de muestra más grande (Agresti 2002, pág. 214).

No obstante debe de incluirse otras medidas que reflejen la calidad del modelo, las tablas 2×2 de clasificación, este tipo de aproximación es posible debido a la estrecha relación que guarda la regresión logística con el análisis de discriminante cuando la distribución de las variables explicativas es una normal multivariada dentro de los dos grupos de la variable respuesta (Hosmer and Lemeshow 2000, pág. 156; Cox and Snell 1989, sección 4.4). Esta relación será discutida en la sección 2.4. También de manera práctica se observa que la bondad de ajuste no se ve reflejada en buenas tasas de clasificación. Las curvas ROC también son utilizadas para la comparación de modelos, éstas son descritas en el capítulo 5.

1.7. Diferentes Ejemplos para usar la Regresión Logística

Éstos son algunos ejemplos en donde se puede modelar la variable respuesta de manera discreta:

- En el *credit scoring* (Thomas et al. 2002, capítulo 1; Agresti 2002, inicio del capítulo 5), cuando las personas solicitan un crédito, la institución financiera está interesada en modelar la probabilidad de que un cliente sea un cliente riesgoso, i.e., que el cliente no sea capaz de pagar cuando éste contraiga una obligación. Por ejemplo, se podría modelar la probabilidad de que un sujeto pague una cuenta en el plazo establecido, con algunas variables como el monto prestado, ingreso, ocupación, otras obligaciones, porcentaje de deudas pagadas en tiempo, y otros aspectos. También, otras opciones son detectar grupos o sectores de mercado en donde la institución financiera (pudiéndose tratar de una empresa distinta del sector financiero) ofrecerá sus productos. Se desea segmentar la cartera de clientes en distintos grupos, si la institución depende de ventas por catálogo, ella podría determinar a qué cliente le podría enviar un catálogo (enviarlo a un cliente potencial), modelando la probabilidad de vender como una función para predecir a que grupo pertenece usando datos del buró de crédito e índices de comportamiento de compra.

- En el transplante de órganos (Hand, Krzanowski and Crowder 2007) el peligro de presentar alguna complicación como el rechazo del tejido transplantado, a infección, así como en la calidad de vida futura del paciente. Medidas post-operativas Y pueden ser usadas para dividir pacientes en varias categorías o clases para reflejar las consideraciones antes mencionadas (por ejemplo codificar Y como 1=tuvo complicación, 0=no tuvo complicaciones) mientras que las medidas pre-operativas $\boldsymbol{X} = (X_1, ..., X_p)$ pueden usarse para predecir la clase post-operativa del paciente.

- En cualquier problema de clasificación, otros ejemplos sencillos pueden encontrarse en Johnson and Wichern (2007) página 576.

Capítulo 2

Análisis de Discriminante

2.1. Introducción

El análisis de discriminante se puede explicar en dos palabras clave: discriminar y clasificar. Discriminar y clasificar son técnicas multivariadas que consisten en separar distintos conjuntos de objetos u observaciones y asignar nuevos elementos a grupos previamente bien definidos. En cada caso los objetivos son:

- Discriminar.- Describir la separación del grupo ya sea gráfica o algebráicamente usando las variables para explicar las diferencias entre dos o más grupos. Los objetivos incluyen encontrar las variables (discriminantes) que influyen en la identificación de estos grupos y encontrar si existe una configuración de funciones de dichas variables en donde se explique mejor la separación.

- Clasificar.- Predecir o asignar a las observaciones externas o futuras a los grupos previamente definidos de manera correcta, los valores de un individuo medidos en un vector de observación son usadas para encontrar el grupo más probable al que este pertenece.

Las funciones creadas, a veces pueden servir como un clasificador de observaciones y también como, una regla que puede sugerir un procedimiento discriminatorio. En la práctica estos dos objetivos, discriminar y clasificar, frecuentemente se traslapan y pueden ser utilizadas de manera conjunta.

La estructura del capítulo es el siguiente. Se verá la discriminación desde el punto de vista de la teoría de decisiones en la sección 2.2, se darán las reglas de clasificación para dos grupos suponiendo la distribución normal multivariada en las variables de interés (sección 2.2.1) y se definirá el discriminante lineal y el cuadrático (secciones 2.2.2 y 2.2.3). En la sección 2.3 se dará la clasificación para más de 2 poblaciones, el costo asociado de clasificar una observación erróneamente (sección 2.3.1) y la regla de clasificación cuando se presenta este caso (sección 2.3.2). En la sección 2.4 se verá la semejanza del análisis de discriminante con la regresión logística. En la sección 2.5 se describirán algunas pruebas necesarias para probar la igualdad de matrices de varianzas-covarianzas entre 2 poblaciones.

2.2. Análisis de Discriminante visto desde la Teoría de Decisiones

Desde el punto de vista de la teoría de decisiones el objetivo principal es optimizar la clasificación de un individuo. Donde clasificar erróneamente puede implicar un costo. Por ejemplo, el puntaje de crédito ayuda al prestador a tomar decisiones: otorgar o no otorgar al solicitante un crédito. Para eso es necesario la construcción de una regla para aplicarla a los solicitantes. La regla se construye a partir de una muestra de datos observados de tamaño n.

En general, suponiendo que existen dos poblaciones, una observación puede provenir de alguna de las dos poblaciones distintas, las cuales se denotarán π_1 o π_2. La clasificación de una observacion depende del vector de los valores que toman las variables explicativas $\boldsymbol{x} = (x_1, x_2, ..., x_p)$. Se construye una regla de decisión que nos ayudará a caracterizar ciertos valores de $\boldsymbol{x}$ para la población π_1, y ciertos para la población π_2.

En el *credit scoring*, para detectar un cliente bueno deben de presentarse ciertas características que lo hacen aceptable o tolerable para la entidad financiera, mientras que un cliente malo tendría todas aquellas características del individuo que la entidad financiera trata de evitar. Se piensa que cada observación (cliente) es un punto en el espacio p-dimensional. Y se buscar dividir este espacio en 2 regiones, suponga que R es el conjunto de todos los posibles valores que podrían tomar las variables de $\boldsymbol{X}$. El objetivo es encontrar una regla que divida a R en dos subconjuntos R_1 y R_2. En donde si la observación cae en la región R_1 se asigna a la población π_1 y si cae en la región R_2 se asigna a la población π_2.

El *credit scoring* básicamente se trata de aceptar a los buenos y rechazar a los malos, minimizando el costo esperado del prestador. Habiendo dos poblaciones existen dos tipos de costos, que corresponen a los dos errores que se pueden cometer en la decisión:

1. Uno puede clasificar a alguien que es bueno como malo y no otorgarle el crédito.
2. Análogamente se puede clasificar a alguien que es malo como bueno y otorgarle el crédito.

En el primer caso la consecuencia es que se pierde la ganancia o el interés que pudo haber generado al otorgar el crédito al solicitante. En el segundo caso se pierde total o parcialmente el monto e interés del crédito. Sin pérdida de generalidad, suponga que π_1 es la población de créditos buenos y π_2 la población de créditos malos. Los costos se denotarán $C(2/1)$ y $C(1/2)$. Suponga que cada costo será igual para cada solicitante de crédito, es decir, $C(2/1)$ y $C(1/2)$ es igual para cualquier solicitante.

Sea $P[\boldsymbol{X} = \boldsymbol{x}|X \in \pi_1]$ la probabilidad condicional de que $\boldsymbol{X}$ tome un cierto valor dado que $\boldsymbol{X}$ pertenece a la población 1. Esta probabilidad se puede escribir como sigue:

$$P(\boldsymbol{x}|1) = P[\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{X} \in \pi_1] = \frac{P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X} \in \pi_1)}{P(\boldsymbol{X} \in \pi_1)}. \tag{2.1}$$

Análogamente esta probabilidad puede ser escrita para la población 2.

Si $P[\boldsymbol{X} \in \pi_1 | \boldsymbol{X} = \boldsymbol{x}]$ es la probabilidad condicional que $\boldsymbol{X}$ es bueno dado que tomó los valores $\boldsymbol{X}$. Se puede escribir esta probabilidad como:

$$P(1|\boldsymbol{x}) = P[\boldsymbol{X} \in \pi_1 | \boldsymbol{X} = \boldsymbol{x}] = \frac{P(\boldsymbol{X} \in \pi_1, \boldsymbol{X} = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}. \tag{2.2}$$

Si $P(\boldsymbol{X} \in \pi_1) = P_1$ y $P(\boldsymbol{X} = \boldsymbol{x}) = P(\boldsymbol{x})$, de la ecuación (2.1) y (2.2) se llega a la siguiente expresión:

$$P(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X} \in \pi_1) = P(\boldsymbol{x}|1)P_1 = P(1|\boldsymbol{x})P(\boldsymbol{x}). \tag{2.3}$$

La ecuación (2.3) puede escribirse como sigue

$$P(1|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|1)P_1}{P(\boldsymbol{x})}. \tag{2.4}$$

De igual manera haciendo el mismo procedimiento para la población 2 se obtiene

$$P(2|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|2)P_2}{P(\boldsymbol{x})}. \tag{2.5}$$

De (2.4) y (2.5) se llega a la siguiente relación:

$$\frac{P(1|\boldsymbol{x})}{P(2|\boldsymbol{x})} = \frac{P(\boldsymbol{x}|1)P_1}{P(\boldsymbol{x}|2)P_2}. \tag{2.6}$$

El costo esperado por tomar la decisión errónea de clasificación se escribe como:

$$\begin{aligned} C &= C(2/1) \sum_{\boldsymbol{x} \in R_2} P(\boldsymbol{x}|1)P_1 + C(1/2) \sum_{\boldsymbol{x} \in R_1} P(\boldsymbol{x}|2)P_2 \\ &= C(2/1) \sum_{\boldsymbol{x} \in R_2} P(1|\boldsymbol{x})P(\boldsymbol{x}) + C(1/2) \sum_{\boldsymbol{x} \in R_1} P(2|\boldsymbol{x})P(\boldsymbol{x}). \end{aligned} \tag{2.7}$$

La regla de clasificación óptima que minimizará el costo esperado de la ecuación (2.7), está dada por la partición de R en los subconjuntos de R_1 y R_2. Para asignar una observación $\boldsymbol{x}$ a R_1, el costo asociado es $C(1/2)P(x|2)P_2$. Si se asigna $\boldsymbol{x}$ a R_2 el costo asociado es $C(2/1)P(x|1)P_1$. Uno asigna $\boldsymbol{x}$ a R_1 donde

$$R_1 = \left\{ \boldsymbol{x} : \frac{C(1/2)}{C(2/1)} \leq \frac{P(\boldsymbol{x}|1)P_1}{P(\boldsymbol{x}|2)P_2} \right\}. \tag{2.8}$$

El tomar este criterio de clasificación conlleva al problema de que depende de los costos de clasificación errónea y éstos por lo regular no se conocen. Esto hace que se busque minimizar las probabilidades a un cierto nivel de riesgo. En el otorgamiento de crédito lo que se busca es minimizar el nivel de fallo conservando el porcentaje de aceptados a un nivel de riesgo.

Es equivalente a mantener la probabilidad de rechazar a los buenos aplicantes a un nivel fijo (Thomas et al. 2002, pág. 44).

Todo el procedimiento anterior puede suponer que las características son variables aleatorias continuas. La única diferencia sería que las probabilidades condicionales $P(\boldsymbol{x}|1)$ y $P(\boldsymbol{x}|2)$ son reemplazadas por las densidades condicionales $f(\boldsymbol{x}|1)$ y $f(\boldsymbol{x}|2)$ y las sumas por integrales. Así el costo esperado, cuando se divide la región R en los conjuntos π_1 y π_2 y se aceptan sólo aquellos en π_1 se vuelve

$$C = C(2/1) \int_{\boldsymbol{x} \in R_2} f(\boldsymbol{x}|1) P_1 + C(1/2) \int_{\boldsymbol{x} \in R_1}, f(\boldsymbol{x}|2) P_2, \tag{2.9}$$

y se asigna $\boldsymbol{x}$ a R_1 donde

$$R_1 = \left\{ \boldsymbol{x} : \frac{C(1/2)P_2}{C(2/1)P_1} \leq \frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|2)} \right\}. \tag{2.10}$$

2.2.1. Clasificación: Caso de la Normal univariada con la misma varianza

Éste es el caso más simple ya que sólo se cuenta con una variable explicativa X, la distribución condicional de la población de los buenos $f(x|1)$ es Normal con media μ_1 y varianza σ^2, y la distribución condicional de la población de los malos es Normal con media μ_2 y varianza σ^2. Entonces

$$f(x|1) = (2\pi\sigma^2)^{-1/2} exp\left(\frac{-(x-\mu_1)^2}{2\sigma^2} \right) \tag{2.11}$$

y la regla de decisión (2.10) es,

$$\begin{aligned} \frac{f(x|1)}{f(x|2)} &= \frac{exp\left(\frac{-(x-\mu_1)^2}{2\sigma^2} \right)}{exp\left(\frac{-(x-\mu_2)^2}{2\sigma^2} \right)} \\ &= exp\left(\frac{-(x-\mu_1)^2 + -(x-\mu_2)^2}{2\sigma^2} \right) \leq \frac{C(1/2)P_2}{C(2/1)P_1}. \end{aligned} \tag{2.12}$$

Despejando a x, la regla se convierte en asignar x a π_1 si

$$\Rightarrow x(\mu_1 - \mu_2) \leq \frac{\mu_1^2 - \mu_2^2}{2} + \sigma^2 ln\left(\frac{C(1/2)P_2}{C(2/1)P_1} \right). \tag{2.13}$$

2.2.2. Clasificación: Caso de la Normal multivariada con la misma matriz de covarianzas

Un caso más realístico es cuando se tienen dos poblaciones normales multivariadas con matriz de covarianza iguales, $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{ip})$ el vector de medias de la población $i = 1, 2$ y con matriz de varianzas y covarianza común $\boldsymbol{\Sigma}$, i.e.

$$f(\boldsymbol{x}|i) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}}}} exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)^T\right). \tag{2.14}$$

El procedimiento general para obtener la regla de clasificación es como sigue:

$$\begin{aligned}\frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|2)} &= \frac{exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right)}{exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right)} \\ &= exp\left\{-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T - (\boldsymbol{x} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right]\right\} \geq \frac{C(1/2)P_2}{C(2/1)P_1}.\end{aligned} \tag{2.15}$$

La región de clasificación para π_1, R_1, es el conjunto de las $\mathbf{x}$ que cumplen la condición en la expresión (2.15). Aplicando logaritmo a la ecuación (2.15) queda como sigue:

$$-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T - (\boldsymbol{x} - \boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right] \geq ln(k), \tag{2.16}$$

donde

$$k = \frac{C(1/2)P_2}{C(2/1)P_1}.$$

Expandiendo los términos de la parte derecha de la expresión (2.16).

$$-\frac{1}{2}\left[\boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T + \boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T + \boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2^T + \boldsymbol{\mu}_2\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T - \boldsymbol{\mu}_2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2^T\right]. \tag{2.17}$$

Reagrupando los términos de (2.17), la expresión en (2.16) queda como sigue

$$\boldsymbol{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T. \tag{2.18}$$

El primer término de la expresión (2.18) es una suma de valores ponderados $\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ y el segundo término es una constante. La expresión (2.18) es una función de puntaje, conocida como discriminante lineal.

Entonces la regla de clasificación[1] se convierte en

[1] Véase Anderson (2003), ecuación 6 pág.216

$$R_1 : \boldsymbol{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \geq ln(k). \tag{2.19}$$

En la práctica, las medias $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ y $\boldsymbol{\Sigma}$ son desconocidos, es común reemplazarlos usando la aproximación *plug-in*, en donde se sustituyen por los estimadores de máxima verosimilitud para las medias muestrales $\boldsymbol{\hat{\mu}_1}$, $\boldsymbol{\hat{\mu}_2}$ y la matriz muestral de varianzas-covarianzas $\boldsymbol{S}$. Entonces la regla queda como:

$$R_1 : \boldsymbol{x}\boldsymbol{S}_{pooled}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T - \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)\boldsymbol{S}_{pooled}^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \geq ln(k) \tag{2.20}$$

donde

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \frac{1}{n_i}\sum_{x_i \in i} x_i,\ i = 1, 2,$$

y

$$\boldsymbol{S}_{pooled} = \frac{1}{n_1 + n_2 - 2}\left\{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2\right\},$$

con $\boldsymbol{S}_i$ es la matriz de varianzas-covarianzas muestral de la población i

$$\boldsymbol{S}_i := \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \hat{\mu}_i)(x_{ij} - \hat{\mu}_i)^T,\ i = 1, 2.$$

2.2.3. Clasificación: Caso de la Normal multivariada con diferentes matrices de covarianzas

La restricción en el caso anterior es que las matrices de varianzas-covarianzas son las mismas para las dos poblaciones. Suponiendo que la matriz de covarianzas para π_1 es $\boldsymbol{\Sigma}_1$ y para π_2 es $\boldsymbol{\Sigma}_2$ se obtiene un modelo más aproximado a la realidad.

$$\begin{aligned}\frac{f(\boldsymbol{x}|1)}{f(\boldsymbol{x}|2)} &= \frac{|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}\, exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right)}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}\, exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right)} \\ &= |\boldsymbol{\Sigma}_2|^{\frac{1}{2}}\,|\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}\, exp\left\{-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T - (\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T\right]\right\}.\end{aligned} \tag{2.21}$$

La región de clasificación para π_1, R_1, es el conjunto de las $\boldsymbol{x}$ mayores o iguales que $k = \frac{C(1/2)P_2}{C(2/1)P_1}$. Aplicando logaritmo a la ecuación (2.21) queda la siguiente expresión:

$$-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)\boldsymbol{\Sigma_1}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T - (\boldsymbol{x} - \boldsymbol{\mu}_2)\boldsymbol{\Sigma_2}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T\right] - \frac{1}{2}ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) \geq ln(k). \tag{2.22}$$

Expandiendo el lado izquierdo de la desigualdad (2.22)

$$-\frac{1}{2}\left[\boldsymbol{x}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x}^T - \boldsymbol{x}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x}^T + \boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{x}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x}^T + \boldsymbol{x}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^T + \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x}^T - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^T\right]. \tag{2.23}$$

Reagrupando los términos de la ecuación (2.23), se obtiene la expresión siguiente:

$$-\frac{1}{2}\boldsymbol{x}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}^T + (\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}^T - \frac{1}{2}(\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^T) \geq ln(k). \tag{2.24}$$

La ecuación (2.24) se obtiene por las propiedades de asociatividad y distributividad presentadas en el Anexo A. Al final los coeficientes del termino lineal se duplican debido a la simetría de la matriz de varianzas-covarianzas, por lo que en notación puede ser expresado de una sola manera.

El primer término de la expresión resultante de la ecuación (2.24) es cuadrática en las variables explicativas, conocida como la función de discriminante cuadrática.

Entonces la regla de decisión[2] queda expresada como sigue:

$$R_1 : -\frac{1}{2}\boldsymbol{x}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}^T + (\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x}^T - C_0 \geq ln(k), \tag{2.25}$$

donde

$$C_0 = \frac{1}{2}ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}(\boldsymbol{\mu}_1\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2^T). \tag{2.26}$$

La función de discriminante cuadrática (QDA) aparenta ser una regla de decisión más general y se esperaría que tenga un mejor desempeño que el discriminante lineal. Thomas et al. (2002) en la página 47 menciona que en la mayoría de los casos no vale la pena tratar de conseguir ligeramente una mayor aproximación que pueda venir de la regla cuadrática.

De la misma manera se utilizan los estimadores de máxima verosimilitud para sustituir $\boldsymbol{\mu}_i$ por $\hat{\boldsymbol{\mu}}_{\boldsymbol{i}}$ y $\boldsymbol{\Sigma}_i$ por la matriz de covarianzas $\boldsymbol{S}_i$ de la población i.

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{i}} = \frac{1}{n_i}\sum_{x_i \in i} x_i,\ i = 1, 2,$$

$$\boldsymbol{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \hat{\mu}_i)(x_{ij} - \hat{\mu}_i)^T,\ i = 1, 2.$$

2.3. Clasificación de g Poblaciones

En este apartado se presenta el desarrollo de las reglas de clasificar una observación cuando se tienen g grupos o poblaciones.

[2]Johnson and Wichern (2007), ecuación (11-27) pág.593

2.3.1. El Método de Costo Mínimo Esperado de Clasificación Errónea

Sea $f_i(\boldsymbol{x})$ la densidad asociada a la población $\pi_i, i = 1, ..., g$. P_i la probabilidad apriori de la población π_i, $C(k/i)$ el costo de asignar una observación a π_k cuando este pertenece a π_i, para $i = 1, .., g$. Para $k = i$, $C(i/i) = 0$ pues no hay un costo de asignar correctamente una observación. Y finalmente R_k es el conjunto de las $\boldsymbol{x}$ clasificadas en π_k en donde las probabilidades de clasificación son dadas de la siguiente manera:

$$P(k/i) = P(clasificar\ \boldsymbol{x}\ a\ \pi_k/\pi_i) = \int_{R_k} f_i(\boldsymbol{x})d\boldsymbol{x}, \tag{2.27}$$

para $k, i = 1, 2, ..., g$; y

$$P(i/i) = 1 - \sum_{k=1, k\neq i}^{g} P(k/i). \tag{2.28}$$

Entonces el costo condicional esperado de clasificación errónea al asignar $\boldsymbol{x}$ proveniente de la población π_1 a π_2,π_3,...,π_g es

$$\begin{aligned} C(1) &= P(2/1)C(2/1) + P(3/1)C(3/1) + ... + P(g/1)C(g/1) \\ &= \sum_{k=2}^{g} P(k/1)C(k/1). \end{aligned} \tag{2.29}$$

Este costo condicional ocurre con probabilidad apriori P_i, la probabilidad de $\boldsymbol{x}$ pertenezca a π_1. De manera análoga se obtienen los costos condicionales $C(2), C(3), ..., C(g)$, multiplicando cada una por su probabilidad apriori se obtiene el costo global de clasificación errónea.

$$\begin{aligned} CGC &= C(1)P_1 + C(2)P_2 + ... + C(g)P_g \\ &= \sum_{i=1}^{g} P_i \left\{ \sum_{i=1, i\neq j}^{g} P(k/i)C(k/i) \right\}. \end{aligned} \tag{2.30}$$

Para obtener una clasificación óptima escogiendo una y excluyendo a las otras se lleva a escoger el mínimo. Las regiones de clasificación que minimizan a CGC en la ecuación (2.30) son definidas al asignar a $\boldsymbol{x}$ a la población $\pi_k, k = 1, ..., g$. Se considera aquella k que minimice la siguiente expresión.[3]

$$\sum_{i=1}^{g} P_i f_i(\boldsymbol{x})C(k/i), \tag{2.31}$$

Si llegara a ocurrir un empate, $\boldsymbol{x}$ puede asignarse a cualquiera de las poblaciones empatadas. Suponga que los costos de clasificación son iguales y además, sin pérdida de generalidad los costos valen 1, entonces el mínimo costo global esperado es:

[3]Anderson (2003), pág. 233, 234; Johnson and Wichern (2007), pág. 607

$$\sum_{i=1, i\neq k}^{g} P_i f_i(\boldsymbol{x}). \tag{2.32}$$

Usando el argumento con la cual se llega a (2.31), se asignará $\boldsymbol{x}$ a $\pi_k, k = 1, ..., g$ si cumple que sea la más pequeña.

La ecuación (2.32) será muy pequeña cuando el término omitido $P_k f_k(\boldsymbol{x})$ es el más grande. Consecuentemente, cuando los costos de clasificación errónea son iguales, el costo mínimo esperado puede ser expresado con más simpleza. La regla cuando se tienen costos unitarios queda como sigue:

Aloja $\boldsymbol{x}$ a π_k si

$$P_k f_k(\boldsymbol{x}) > P_i f_i(\boldsymbol{x}) \ \ \forall\ i \neq k, \tag{2.33}$$

o su equivalente

$$ln(P_k f_k(\boldsymbol{x})) > ln(P_i f_i(\boldsymbol{x})) \ \ \forall\ i \neq k. \tag{2.34}$$

2.3.2. Clasificación de Poblaciones con Distribución Normal Multivariada

Cuando las poblaciones se distribuyen Normal Multivariada con vector de medias $\boldsymbol{\mu}_i$ y matriz de varianzas-covarianzas $\boldsymbol{\Sigma}_i$. Es decir con función de densidad.

$$f_i(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p \cdot |\boldsymbol{\Sigma}|^{\frac{1}{2}}}} exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)^T\right),\ i = 1, ..., g. \tag{2.35}$$

Además, si $C(i/i) = 0$, $c(k/i) = 1, \forall\ k \neq i$, es decir, los costos de clasificación errónea son iguales, la regla de decisión se convierte en:

Asigna $\boldsymbol{x}$ a π_k si

$$\begin{aligned} ln(P_k f_k(\boldsymbol{x})) &= lnP_k - \frac{p}{2}ln2\pi - \frac{1}{2}ln\,|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \\ &= máx\left\{ln\ \ (P_i f_i(\boldsymbol{x}))\right\},\ i = 1, ..., g. \end{aligned} \tag{2.36}$$

Observe que la constante $-\frac{p}{2}ln2\pi$ puede ser ignorada, ya que es el mismo para todas la poblaciones. El *score* discriminante cuadrático para la i-ésima población

$$\delta_i(\boldsymbol{x})^Q = -\frac{1}{2}ln\,|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T,\ i = 1, ..., g. \tag{2.37}$$

El *score* cuadrático $\delta_i(\boldsymbol{x})^Q$ está integrado por el determinante de la matriz de varianzas-covarianzas $\boldsymbol{\Sigma}_i$, la probabilidad apriori y el cuadrado de la distancia de $\boldsymbol{x}$ a $\boldsymbol{\mu}_i$ la media de la población i. Usando la función de discriminante la regla de clasificación se convierte en:

Asigna $\boldsymbol{x}$ a π_k si el puntaje del *score* discriminante cuadrático $\delta_k(\boldsymbol{x})^Q$ es igual al máximo entre $\delta_1(\boldsymbol{x})^Q, \delta_2(\boldsymbol{x})^Q, ..., \delta_g(\boldsymbol{x})^Q$. Las funciones de discriminantes cuadráticas están dadas en la ecuación (2.37).

En la práctica $\boldsymbol{\mu}_i$ y $\boldsymbol{\Sigma}_i$ son desconocidos, en su lugar son sustituidos por sus estimadores de máxima verosimilitud $\hat{\boldsymbol{\mu}}_i$, $\boldsymbol{S}_i$. Entonces, la ecuación del *score* cuadrático en la expresión (2.37) queda

$$\hat{\delta}_i(\boldsymbol{x})^Q = -\frac{1}{2} ln\,|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i)\boldsymbol{S}_i^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i)^T + lnP_i,\ i = 1, ..., g. \tag{2.38}$$

Y la regla de clasificación es: asigna $\boldsymbol{x}$ a π_k si el *score* cuadrático

$$\hat{\delta}_k(\boldsymbol{x})^Q = máx\left\{\hat{\delta}_1(\boldsymbol{x})^Q, \hat{\delta}_2(\boldsymbol{x})^Q, ..., \hat{\delta}_g(\boldsymbol{x})^Q\right\},\ i = 1, ..., g. \tag{2.39}$$

La simplificación del modelo es posible si se supone la homogeneidad de las matrices de varianzas-covarianzas para todas la poblaciones. Suponiendo que $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j,\ i, j = 1, ..., g,$ $i \neq j$. Entonces al sustituir en la ecuación (2.37) se tiene

$$\begin{aligned} \delta_i(\boldsymbol{x})^Q &= -\frac{1}{2} ln\,|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \\ &= -\frac{1}{2} ln\,|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{x}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T + \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T - \frac{1}{2}\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i^T + lnP_i,\ i = 1, ..., g. \end{aligned} \tag{2.40}$$

Los primeros dos términos de la ecuación pueden ser omitidos, pues para $\delta_i(\boldsymbol{x})^Q$, $i = 1, ..., g$, son los mismos. Por tanto no influyen para el propósito de clasificación, los términos restantes son términos lineales y una constante. La expresión (2.40) queda simplificada en una función de *score* lineal como sigue

$$\delta_i(\boldsymbol{x}) = \boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{x}^T - \frac{1}{2}\boldsymbol{\mu}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i^T + lnP_i,\ i = 1, ..., g. \tag{2.41}$$

Generalmente cuando no se conocen los parámetros, la estimación del parámetro $\boldsymbol{\Sigma}$ es hecha por la siguiente relación

$$\boldsymbol{S}_{pooled} = \frac{1}{n_1 + n_2 + ... + n_g - g}\left((n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2 + ... + (n_g - 1)\boldsymbol{S}_g\right). \tag{2.42}$$

Esta matriz se llama *pooled covariance matrix*, es muy frecuente encontrarla denotada así por casi todos los libros de análisis multivariado y ésta es llamada matriz de varianzas-covarianzas agrupada. Entonces, la función del *score* lineal con los parámetros estimados es:

$$\hat{\delta}_i(\boldsymbol{x}) = \hat{\boldsymbol{\mu}}_i \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}^T - \frac{1}{2} \hat{\boldsymbol{\mu}}_i \boldsymbol{S}_{pooled}^{-1} \hat{\boldsymbol{\mu}}_i^T + lnP_i,\ i = 1, ..., g. \tag{2.43}$$

La regla de clasificación se convierte en asignar $\boldsymbol{x}$ a π_k al máximo entre las funciones de *score* lineal

$$\hat{\delta}_k(\boldsymbol{x}) = máx \left\{ \hat{\delta}_i(\boldsymbol{x}) \right\},\ i = 1, ..., g. \tag{2.44}$$

Las funciones de *score* lineal en la expresión (2.41) pueden ser comparadas dos a dos. Si $\delta_k(\boldsymbol{x})$ es la función de *score* más grande entre $\delta_1(\boldsymbol{x})$,...,$\delta_g(\boldsymbol{x})$, entonces

$$0 \leq \delta_k(\boldsymbol{x}) - \delta_i(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^T - \frac{1}{2}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^T + ln\left(\frac{P_k}{P_i}\right). \tag{2.45}$$

para $i = 1, ..., g$, $i \neq k$. Si $-ln\left(\frac{P_k}{P_i}\right) = ln\left(\frac{P_i}{P_k}\right)$, entonces la región de clasificación R_k consiste en las $\boldsymbol{x}$ que satisfacen

$$R_k : \delta_{ki}(\boldsymbol{x}) \geq ln\left(\frac{P_i}{P_k}\right), \tag{2.46}$$

donde

$$\delta_{ki}(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^T - \frac{1}{2}(\boldsymbol{\mu}_k + \boldsymbol{\mu}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_i)^T. \tag{2.47}$$

Entonces para las regiones de clasifiación R_1,R_2,...,R_g, estan separados por $(g-1)$ hiperplanos[4].

2.4. Análisis de Discriminante y Regresión Logística

El análisis de discriminante y la regresión logística son dos técnicas conceptualmente diferentes. Sin embargo, guardan una estrecha relación cuando se trata de utilizar estos dos métodos para problemas de clasificación.

En los problemas de clasificación para poder asignar o predecir la clase de pertenencia o población de una observación nueva, se utiliza la información en el vector de variables explicativas $\boldsymbol{x}$.

Si se restringe al caso en donde se tiene dos grupos o poblaciones $g = 2$. Por un lado el objetivo de la regresión logística es evaluar si un conjunto de las variables explicativas está significativamente relacionado con las dos categorías de la variable respuesta. $Y = 0, 1$.

[4]Véase Anderson (2003) sección 6.8

La distribución de $\boldsymbol{X}$ no es directamente relevante en la definición.

Por otro lado, en el análisis de discriminante, en cada población se tiene una distribución que proviene del conjunto de variables explicativas $\boldsymbol{X}$ pero no una variable respuesta Y. En el análisis de discriminante, se construye una regla de decisión que ayudará a caracterizar ciertos valores de $\boldsymbol{x}$ para cada población, en donde existen dos funciones de densidad $f_y(\boldsymbol{x})$, $y = 0, 1$. El énfasis que se hace es que estas dos densidades son bien definidas para las dos poblaciones. Cuando estas dos densidades provienen de la misma familia exponencial y en el caso más importante, cuando las densidades están definidas bajo una normal multivariada con la misma matriz de varianzas-covarianzas $\boldsymbol{\Sigma}$ y vector de medias $\boldsymbol{\mu}_0$ y $\boldsymbol{\mu}_1$, entonces la función de discriminante lineal se define como

$$\boldsymbol{x}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T - \frac{1}{2}(\boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_0\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0^T). \tag{2.48}$$

Otra cuestión en el análisis de discriminante es que hay dos probabilidades π_y, $y = 0, 1$ con $\pi_0 + \pi_1 = 1$. Éstas definen la probabilidad de pertenecer a la población $i = 0, 1$.

En Cox and Snell (1989), sección 4.4 se establece una aproximación de ambos métodos, representando cada población por una variable aleatoria $Y = 0, 1$ y para cada individuo por un vector de variables explicativas $\boldsymbol{X}$, incluyen la distribución de cada poblacion con $f_0(\boldsymbol{x})$ y $f_1(\boldsymbol{x})$ que especifican las densidades condicionales de $\boldsymbol{X}$ dado $Y = 0, 1$. Entonces, para un nuevo elemento $\boldsymbol{x}^*$ con Y desconocido, por el teorema de Bayes se expresa lo siguiente

$$\begin{aligned} P(Y = 1|\boldsymbol{X} = \boldsymbol{x}^*) &= \frac{f_1(\boldsymbol{x}^*)\pi_1}{P(\boldsymbol{X} = \boldsymbol{x})} \\ &= \frac{f_1(\boldsymbol{x}^*)\pi_1}{f_0(\boldsymbol{x}^*)\pi_0 + f_1(\boldsymbol{x}^*)\pi_1}, \end{aligned} \tag{2.49}$$

entonces,

$$ln\left[\frac{P(Y = 1|\boldsymbol{X} = \boldsymbol{x}^*)}{P(Y = 0|\boldsymbol{X} = \boldsymbol{x}^*)}\right] = ln\left[\frac{f_1(\boldsymbol{x}^*)}{f_0(\boldsymbol{x}^*)}\right] + ln\left[\frac{\pi_1}{\pi_0}\right], \tag{2.50}$$

define una ecuación de regresión logística en el cual las probabilidades apriori son aisladas en un sólo término. Cuando se tiene una función lineal como en la ecuación (2.48), resulta en una regresión logística lineal, solamente pasa cuando, pero no significa que sólo cuando, las funciones de distribuciónes condicionales de $\boldsymbol{X}$ son normal multivariada con la misma matriz de varianzas-covarianzas.

La formulación de la regresión logística por medio de la distribución de $\boldsymbol{X}$, es una manera de describir un valor de y para un individuo nuevo con un valor conocido de $\boldsymbol{x}$, si se toma el lado derecho de la ecuación (2.50), ésta expresa la regla de asignación del nuevo elemento $\boldsymbol{x}^*$ describiendo a cual de las dos poblaciones pertenece.

Otro punto importante es que cuando las poblaciones se distribuyen normal multivariada con la matriz de varianzas-covarianzas, la estimación de los coeficientes determinados por la máxima verosimilitud del método de discriminante y regresión logística guardan una semejanza dada en la siguiente relación (Cox and Snell (1989), pág. 137).

$$\hat{\boldsymbol{\beta}}_{(d)} = \hat{\boldsymbol{\beta}}_{(l)} SS_{res}/n \approx \hat{\boldsymbol{\beta}}_{(log)} SS_{res}/n, \tag{2.51}$$

donde $\hat{\boldsymbol{\beta}}_{(d)}$ son los coeficientes estimados por medio de la aproximación de discriminante, $\hat{\boldsymbol{\beta}}_{(l)}$ la estimación de los coeficientes obtenidos por sustituir los estimadores de máxima verosimilitud de $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\mu}}_0$ y $\hat{\boldsymbol{\mu}}_1$ en la ecuación (2.48) y $\hat{\boldsymbol{\beta}}_{(log)}$ los coeficientes ajustados por máxima verosimilitud en la regresión logística. SS_{res} es la suma de cuadrados de los residuales obtenida cuando y es regresada en $\boldsymbol{x}$.

En Hosmer and Lemeshow (2000) sección 2.6, se menciona que el método de estimación de los coeficientes $\boldsymbol{\beta}$ para la regresión logística aproximándolo, por el análisis de discriminante puede ser llevado de la siguiente manera

$$\beta_0 = ln\frac{\theta_1}{\theta_0} - 0{,}5(\mu_1 - \mu_0)\Sigma^{-1}(\mu_1 + \mu_0)^T, \tag{2.52}$$

donde $\theta_1 = P(Y = 1)$ y $\theta_0 = 1 - \theta_1$ y

$$\boldsymbol{\beta}_i = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}^{-1},\ i = 1, ..., p, \tag{2.53}$$

pero la distribución condicional de $\boldsymbol{X}$ dado $Y = 0, 1$ debe de es una normal multivariada, es decir, $\boldsymbol{X}|Y = i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ $i = 0, 1$. La estimación de los valores muestrales es de la misma manera antes mencionadas en las secciones anteriores de este capítulo. Sin embargo, ellos no recomiendan usar esos coeficientes estimados, sino los de máxima verosimilitud.

2.5. Prueba de Hipótesis para la Igualdad de Matrices de Varianzas Covarianzas

En muchas instancias es importante probar la homogeneidad de medias, otra es probar el supuesto de homogeneidad de varianzas, la presencia o ausencia de ella puede ser de interés para el investigador. En análisis de discriminante es útil para poder ver que tipo de análisis debe de ser implementado si el análisis lineal o el análisis cuadrático. Una de las pruebas más recurridas para probar el supuesto de homogeneidad de varianzas es la prueba de Bartlett.

2.5.1. Prueba Univariada

Las hipótesis a probar es:

$$H_0 = \sigma_1^2 = ... = \sigma_g^2 \ vs\ H_a = \sigma_i^2 \neq \sigma_j^2 \ para\ algún\ i \neq j.$$

La estadística de prueba, es[5]

$$B = (nlnS^2 - \sum_{i=1}^{g} n_i ln(S_i^2))/C, \tag{2.54}$$

donde

$$C = 1 + \frac{(\sum_{i=1}^{g} n_i^{-1}) - n^{-1}}{3(g-1)}, \tag{2.55}$$

S^2 es la varianza agrupada para g poblaciones

$$S^2 = \sum_{i=1}^{g} n_i \frac{S_i^2}{n},$$

y S_i^2, $i = 1, ..., g$ las varianzas muestrales de la población i

$$n = \sum_{i=1}^{g} n_i.$$

B es aproximadamente distribuida como una Ji-cuadrada con $(k-1)$ grados de libertad. Se rechaza H_0 si

$$P[B > \chi^2_{(k-1)}] < \alpha.$$

2.5.2. Prueba Multivariada

Esta prueba es la que será usada posteriormente para analizar la base de crédito en la sección 4.3.1.

Sea $\mathbf{\Sigma}_i$ la matriz de varianzas-covarianzas de la población i. Se considera probar

$$H_0 : \mathbf{\Sigma}_1 = ... = \mathbf{\Sigma}_g \ vs \ H_a : \mathbf{\Sigma}_i \neq \mathbf{\Sigma}_j \ p.a \ i \neq j.$$

Bajo la hipótesis nula, al menos dos de las matrices de varianzas-covarianzas difieren al menos en alguno de sus elementos. Sea

$$\boldsymbol{S}_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_i)(x_{ij} - \hat{\mu}_i)^T, \tag{2.56}$$

[5]Véase Rencher (2002), sección 7.3.1.

la matriz de varianzas-covarianzas muestral para la población i. La matriz de varianzas-covarianzas agrupada (*pooled covariance matrix*) es

$$\boldsymbol{S}_{pooled} = \frac{\sum_{i=1}^{g}(n_i - 1)\boldsymbol{S}_i}{\sum_{i=1}^{g}(n_i - 1)}. \tag{2.57}$$

Asumiendo una distribución normal multivariada para cada población. Bartlett propone una modificación al caso univariado de la estadística del cociente de verosimilitudes [6], dicha estadística sirve para probar la hipótesis $H_0 : \boldsymbol{\Sigma}_1 = ... = \boldsymbol{\Sigma}_g$ *vs* $H_a : \boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$ p.a $i \neq j$, y se representa como sigue

$$\Lambda = \prod_{i=1}^{g} \left(\frac{|\boldsymbol{S}_i|}{|\boldsymbol{S}_{pooled}|} \right)^{(n_i-1)/2}. \tag{2.58}$$

La estadística M de Box se basa en la aproximación a la Ji-cuadrada al obtener $M = -2ln\Lambda$ de la expresión 2.58. Sin embargo, usualmente se considera la siguiente estadística usando la corrección $(1 - c)$ como sigue[7]

$$L' = (1 - c)M = (1 - c) \left\{ (N - g)ln\,|\boldsymbol{S}_p| - \sum_{i=1}^{g}(n_i - 1)ln\,|\boldsymbol{S}_i| \right\}, \tag{2.59}$$

donde el factor de corrección c es

$$c = \frac{2p^2 + 3p - 1}{6(p + 1)(g - 1)} \left\{ \left(\sum_{i=1}^{g} \frac{1}{n_i - 1} \right) - \frac{1}{N - g} \right\}, \tag{2.60}$$

donde p es el número de variables y g el número de grupos. Bajo la hipótesis nula, la homogeneidad de las matrices de varianzas-covarianzas, L' es aproximadamente una Ji-cuadrada con $\frac{1}{2}p(p + 1)(g - 1)$ grados de libertad. Se rechaza H_0 al nivel α si

$$P[L' > \chi^2_{(\frac{1}{2}p(p+1)(g-1))}] < \alpha.$$

[6] Véase Anderson(2003), pág. 413; Johnson and Wichern (2007), pág. 310.

[7] Véase Johnson and Wichern (2007) sección 6.6 ecuación 6-52.

Capítulo 3

El Análisis de Supervivencia y el Modelo de Regresión de Cox

3.1. Introducción

El análisis de supervivencia ha desempeñado un papel fundamental en estudios de diversas áreas de la ciencia, en donde la principal variable de interés es la longitud de tiempo que transcurre antes de que un evento ocurra, normalmente llamado falla.

Se presenta una breve introducción y definiciones del análisis de supervivencia en la sección 3.2. En la sección 3.3 un estimador de la función de supervivencia y otro para la función de riesgo acumulada. En la sección 3.4 se presenta el modelo de regresión de Cox, seguida del ajuste del modelo en la sección 3.5. La interpretación del modelo se verá en la sección 3.6 y unas técnicas de validación del ajuste del modelo en la sección 3.7.

3.2. Conceptos de Análisis de Supervivencia

El análisis de supervivencia es una rama de la estadística que estudia datos con tiempos de vida o de falla. Por ejemplo, en el área médica sirve para predecir la probabilidad de supervivencia o de muerte de una enfermedad. En ingeniería, el empleo del análisis de supervivencia es utilizado para estimar el tiempo de vida, falla o descompostura de algún artefacto, componente o sistema eléctrico. Por esa razón, comúnmente el evento de interés es denotado como falla o muerte. Sin embargo, sus métodos pueden ser aplicados en las ciencias sociales y en el área financiera. Por ejemplo, en psicología podría susitarse el caso de estudiar el tiempo transcurrido antes de que un exconvicto recién liberado vuelva a cometer un crimen; en el *credit scoring* se puede construír un modelo para predecir el tiempo en que un cliente dejará de pagar su deuda; en los seguros, tiempo de las reclamaciones de indemnización de una cartera de trabajadores; incluyendo sus factores de riesgo que llevan al reclamo.

En cualquier caso el tiempo transcurre del origen de estudio hasta que se presenta la

falla, pero muchas veces un estudio es suspendido antes de que todas las observaciones hayan experimentado el evento de interés. Este tipo de datos u observaciones se les llama datos censurados. Realizar la investigación en la presencia de estos datos pueden ser complejo cuando se modela. Muchas de las técnicas del análisis de supervivencia fueron creadas para sustentar este tipo de observaciones.

3.2.1. Datos Censurados

Un dato censurado corresponde al caso cuando el tiempo de falla de un sujeto puede no ser observado en el horizonte de estudio y también puede ocurrir cuando la observación sale del estudio por causas ajenas al evento de interés. Por ejemplo, en un estudio médico el evento de interés es enfermar o morir, pero algunos pacientes pueden seguir vivos o libres de enfermedad al final del estudio. El tiempo exacto de falla de estos sujetos es desconocido. Éstos son llamados datos censurados o tiempo censurado.

De acuerdo a lo anterior suponga que, un dato no censurado del i-ésimo individuo en la muestra de tamaño n tiene un tiempo de falla t_i. Suponga que también hay un periodo de observación q_i, en el cual la observación de ese individo cae en q_i si la falla no ocurre. Entonces las observaciones consisten en la terna $(t_i, \boldsymbol{x}_i, c_i)$, donde $i = 1, ..., n$. El vector $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ denota el valor de las variables explicativas del sujeto i, y la variable indicadora c denota la censura, el cual $c_i = 0$ si $t_i \leq q_i$ (no censurado) y $c_i = 1$ si $t_i > q_i$ (censurado). Por otro lado, habrá individuos que tendrán un tiempo de censura t_i donde $t_i \leq q_i$.

En la mayoría de los casos q_i es conocida y pueden darse los siguientes tipos de censura:

Datos Censura del Tipo I

Este tipo de censura sucede cuando la falla no ha ocurrido al término del estudio que tiene un periodo fijo. Es decir, todas las censuras $q_i = t$, $i = 1, ..., n$, donde q_i es una constante determinada por el investigador. Dicho de otra manera la censura de tipo I ocurre cuando la muestra de n observaciones es seguida hasta un tiempo t, los eventos que ocurren antes de t son observados. Aquí, el número de sujetos que experimentan la falla es aleatoria, pero la duración del estudio es predeterminada.

Datos Censura del Tipo II

La observación de los eventos es medida hasta que ocurra un número fijo de r fallas o muertes, así q se convierte en una variable aleatoria. Este tipo de censura puede ocurrir en la ingeniería cuando se someten un grupo de componentes y son provados hasta que un número determinado de ellos fallen. Realizar ciertos experimentos pueden resultar altamente costosos y no pueden darse el lujo de perder todos los componentes. Por ejemplo, estudiar el tiempo de vida de motores y turbinas de aeronaves.

3.2.2. Distribuciones del Tiempo de Falla

En datos de análisis de supervivencia se mide el tiempo que transcurre desde el inicio de estudio al evento de interés, siendo esta muerte, falla, respuesta, recaída, desarrollo de una enfermedad, divorcio, incumplimiento, etc. Estos tiempos están sujetos a la aleatoriedad, y como variable aleatoria tiene una distribución. La distribución de los tiempos de falla pueden ser descritas o caracterizadas por tres funciones: la función de supervivencia, la función de densidad de probabilidad y la función *hazard* o de riesgo. Estas tres funciones son matemáticamente equivalentes, si una de ellas es dada se pueden derivar las otras funciones. En la práctica, estas tres funciones pueden ser utilizadas para ejemplificar distintas caraterísticas de los datos. Un problema básico en análisis de supervivencia es estimar estas funciones siendo una o varias, para expresar o detectar patrones en la población de estudio.

Sea T una v.a no negativa que representa el tiempo. La distribución de supervivencia de T denotada como $S(t)$ está definida como la probabilidad de que el individuo sobreviva más de t:

$$S(t) = P(T > t). \tag{3.1}$$

Si T es una v.a continua, la función $S(t)$ es una función continua estrictamente decreciente. Al ser T una v.a continua, la función de supervivencia es el complemento de la función de distribución acumulativa de T. $S(t) = 1 - F(t)$, donde $F(t) = P(T \leq t)$. Además es la integral de la función de densidad $f(t)$

$$S(t) = P(T > t) = \int_t^\infty f(u)du, \tag{3.2}$$

o bien,

$$f(t) = -\frac{d}{dt}S(t). \tag{3.3}$$

De la ecuación (3.3) la función se define como el límite de la probabilidad que un individuo falla en un pequeño intervalo por unidad de tiempo t. $f(t)$ es una función de densidad lo cual debe de cumplir que es una función no negativa con area bajo la curva igual a uno. Por otra parte, la función de supervivencia $S(t)$ además de ser estrictamente decreciente cumple con las siguientes propiedades[1]:

- La probabilidad de sobrevivir al tiempo cero es uno, es decir, todas las observaciones al tiempo $t = 0$ se encuentran con vida.

$$S(0) = \int_0^\infty f(u)du = 1. \tag{3.4}$$

[1]Lee and Wang (2003), pág. 9

- Todas las observaciones morirán eventualmente, es decir, la probabilidad de sobrevivir a tiempo infinito es cero.

$$S(\infty) = lim_{t \rightarrow \infty} \int_t^{\infty} f(u)du = 0. \tag{3.5}$$

Cuando T es v.a. discreta, otras técnicas son utilizadas. Las variables aleatorias discretas en el análisis de supervivencia se presentan debido al redondeo de las mediciones, la agrupación de fallas en intervalos, o cuando los tiempos de vida se refieren a un número entero de unidades. Normalmente la función de supervivencia se calcula:

$$S(t) = P(T > t) = \sum_{t > t_j} p(t_j). \tag{3.6}$$

Por otro lado, una función fundamental en el análisis de supervivencia, es la función de riesgo. Esta función obtiene otros diversos nombres en otras áreas de la ciencias como la tasa de fallo condicional en la fiabilidad (*reliability*), la fuerza de la mortalidad en la demografía, la función de intensidad en los procesos estocásticos, la tasa de fracaso específica por edad en la epidemiología, etc.

La función de riesgo se define como la tasa instántanea de fallo en cualquier tiempo t dado que el individuo ha sobrevivido a ese tiempo.

$$h(t) = lim_{\Delta t \rightarrow 0} \frac{P\left[t \leq T < t + \Delta t | T \geq t\right]}{\Delta t}. \tag{3.7}$$

Si T es v.a. continua entonces,

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} ln\left[S(t)\right], \tag{3.8}$$

de la función $h(t)$ se obtiene su función acumulada $H(t)$ como

$$H(t) = \int_0^t h(u)du = -ln\left[S(t)\right]. \tag{3.9}$$

De la ecuación (3.9) se establece una relación de la función de supervivencia $S(t)$ de $H(t)$ y $h(t)$

$$S(t) = exp\left[-H(t)\right] = exp\left[-\int_0^t h(u)du\right]. \tag{3.10}$$

La función de riesgo es muy útil en determinar las funciones de distribución apropiadas utilizando la infomación cualitativa acerca del mecanismo de falla y para describir el chance de presenciar el evento de interés a lo largo del tiempo, una medida que ayuda a reflejar esto es la razón de riesgos que será descrita en la sección 3.6. Y forma la base de muchas técnicas incluyendo el modelo de regresión de riesgos proporcionales de Cox que en la sección 3.5 se verá con más detalle.

3.3. Estimación no Paramétrica de las Funciones de Supervivencia y de Riesgo Acumulada

Cuando se trabaja con tiempos de vida es deseable estimar la función de supervivencia. Un estimador estándar para dicho análisis fue propuesto por Kaplan & Meier en 1958, llamado el estimador Producto Límite.

Suponga una muestra de n observaciones independientes denotada (t_i, c_i), $i = 1, .., n$ en donde la variable T denota el tiempo de supervivencia y C la variable que indica censura. Suponga que entre esas n observaciones se tiene m fallas ($m \leq n$). Ordenando las observaciones por tiempo de falla con $t_1 < t_2 < ... < t_m$, el número de elementos en riesgo de fallar a cada tiempo t_i es denotado por Y_i. Mientras que el número de fallas observadas en t_i son d_i. El estimador de Kapplan-Meier al tiempo t se define como

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right), \tag{3.11}$$

con

$$\hat{S}(t) = 1 \text{ si } t < t_1. \tag{3.12}$$

La varianza estimada de este estimador es calculada por medio de la fórmula de Greenwood[2]

$$\hat{V}\left[\hat{S}(t)\right] = \hat{S}(t)^2 \sum_{t_i} \frac{d_i}{Y_i(Y_i - d_i)} \text{ si } t < t_1. \tag{3.13}$$

Este estimador es una función escalonada que hace saltos hacia los tiempos del evento observado. El tamaño de los saltos depende del número de observaciones registradas en t_i y también del patrón de las observaciones censuradas después de t_i.

La estimación de la función de riesgo acumulada por medio de Kapplan-Meier es como sigue

$$\hat{H}(t) = -ln[\hat{S}(t)]. \tag{3.14}$$

Otro estimador alternativo para la función de riesgo acumulada muy utilizada es la del estimador de Nelson-Aalen

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i} \; si \; t_1 \leq t. \tag{3.15}$$

[2]Véase Klein and Moeschberger 2003, pág. 92 ecuación 4.2.2.

3.4. Modelo de Regresión de Cox

En problemas de análisis de supervivencia es muy común comparar dos o más poblaciones, normalmente se construyen las funciones básicas de análisis de supervivencia. Sin embargo, para un hacer análisis más detallado, el evento de interés puede ser representado por modelos en donde las variables explicativas son típicamente asociadas a parámetros desconocidos.

En esta sección, de los posibles modelos que pueden usarse para representar el efecto de las variables explicativas en el evento de falla se describe el modelo de regresión de Cox.

Sea un modelo de regresión considerando un vector de p variables explicativas, $\boldsymbol{X} = (X_1, X_2, ..., X_p)$, que afectan la respuesta, Y. El modelo de regresión de Cox o modelo de riesgos proporcionales (PHM) es su forma más simple se define de la siguiente manera

$$h(t; \boldsymbol{x}; \boldsymbol{\beta}) = \Psi(\boldsymbol{x}; \boldsymbol{\beta}) h_0(t) \tag{3.16}$$

en el cual el vector $\boldsymbol{X}$ es invariante en el tiempo para cada individuo. El vector $\boldsymbol{\beta}$ denota los parámetros de la distribución, el cual es necesario estimar. Suponga que se tienen distribuciones continuas para los tiempos de vida y son registradas con exactitud. Las 3 parametrizaciones de Ψ pueden ser consideradas de las siguientes formas (Cox and Oakes 1984, pág. 91):

1. Log lineal.
$$\Psi(\boldsymbol{x}; \boldsymbol{\beta}) = exp(\boldsymbol{\beta}\boldsymbol{x}^T) = exp(\beta_1 x_1 + ... + \beta_p x_p).$$

2. Lineal.
$$\Psi(\boldsymbol{x}; \boldsymbol{\beta}) = 1 + \boldsymbol{\beta}\boldsymbol{x}^T = 1 + \beta_1 x_1 + ... + \beta_p x_p.$$

3. Logística.
$$\Psi(\boldsymbol{x}; \boldsymbol{\beta}) = log(1 + exp(\boldsymbol{\beta}\boldsymbol{x}^T)) = log(1 + exp(\beta_1 x_1 + ... + \beta_p x_p)).$$

La parametrización log lineal es la forma más utilizada. Si se supone un modelo de regresión de Cox, la función de supervivencia bajo el modelo de regresión de Cox es

$$S(t; \boldsymbol{x}; \boldsymbol{\beta}) = [S_0(t)]^{\Psi(\boldsymbol{x}; \boldsymbol{\beta})}. \tag{3.17}$$

El cálculo se obtiene sustituyendo en la ecuación (3.16) la ecuación (3.10) (Hosmer and Lemeshow 1999, págs. 92,93).

3.5. Ajuste del Modelo de Regresión de Cox

La estimación de los parámetros en la regresión de Cox con h_0 conocida puede llevarse a cabo de la siguiente manera.

Suponga que tiene n observaciones independientes cada una contiene información sobre la longitud de tiempo en el cual el sujeto fue observado, los valores de las variables explicativas cuyos valores fueron determinados en el tiempo de observación de estudio, además se sabe si la observación es o no un dato censurado. Las observaciones entonces están denotadas por la terna $(t_i, \boldsymbol{x}_i, c_i)$, para $i = 1, ..., n$. La función de verosimilitud se obtiene al multiplicar la distribución de las ternas observadas, el valor de $f(t; \boldsymbol{x}; \boldsymbol{\beta})$ para una observación no censurada y un valor de $S(t; \boldsymbol{x}; \boldsymbol{\beta})$ para un valor censurado. En general una manera de denotar la distribución de cada terna en la verosimilitud es la siguiente expresión

$$f(t; \boldsymbol{x}; \boldsymbol{\beta})^c \times S(t; \boldsymbol{x}; \boldsymbol{\beta})^{1-c}, \tag{3.18}$$

donde $c = 0$ ó 1(no censurados, censurados). Como las observaciones son independientes, la función de verosimilitud es el producto sobre toda la muestra de la ecuación (3.18)

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(t_i; \boldsymbol{x}_i; \boldsymbol{\beta})^{c_i} \times .S(t_i; \boldsymbol{x}_i; \boldsymbol{\beta})^{1-c_i} \tag{3.19}$$

Para poder obtener la función de verosimilitud dada la función de supervivencia, se toma la expresión en la ecuación (3.8), entonces la función de densidad $f(t; \boldsymbol{x}; \boldsymbol{\beta})$ es

$$f(t; \boldsymbol{x}; \boldsymbol{\beta}) = h(t; \boldsymbol{x}; \boldsymbol{\beta}) S(t; \boldsymbol{x}; \boldsymbol{\beta}). \tag{3.20}$$

Al substituir la expresión dada en (3.20) en la función de verosimilitud (ecuación (3.19)) queda como

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} h(t; \boldsymbol{x}; \boldsymbol{\beta})^{c_i} \times S(t_i; \boldsymbol{x}_i; \boldsymbol{\beta}). \tag{3.21}$$

La función de log-verosimilitud se obtiene como:

$$L(\boldsymbol{\beta}) = log\left[\ell(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} c_i log\left[h(t; \boldsymbol{x}; \boldsymbol{\beta})\right] + \sum_{i=1}^{n} log\left[S(t_i; \boldsymbol{x}_i; \boldsymbol{\beta})\right]. \tag{3.22}$$

substituyendo según la relación (3.10)

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} c_i log\left[h(t; \boldsymbol{x}; \boldsymbol{\beta})\right] + \sum_{i=1}^{n} -H(t_i; \boldsymbol{x}_i; \boldsymbol{\beta}). \tag{3.23}$$

Partiendo de la ecuación (3.22) usando las ecuaciones (3.8) y (3.16) la función de log-verosimilitud si $\Psi(\boldsymbol{x}; \boldsymbol{\beta}) = exp(\boldsymbol{\beta}\boldsymbol{x}^T)$ puede quedar expresada en términos de las funciones base como sigue

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} c_i log\left[h_0(t)\right] + c_i \boldsymbol{\beta}\boldsymbol{x_i}^T + \sum_{i=1}^{n} log\left[exp(-H(t_i;\boldsymbol{x}_i;\boldsymbol{\beta}))\right] \\
&= \sum_{i=1}^{n} c_i log\left[h_0(t)\right] + c_i \boldsymbol{\beta}\boldsymbol{x_i}^T + \sum_{i=1}^{n} log\left[exp(-\boldsymbol{\beta}\boldsymbol{x_i}^T H_0(t))\right] \\
&= \sum_{i=1}^{n} c_i log\left[h_0(t)\right] + c_i \boldsymbol{\beta}\boldsymbol{x_i}^T + \sum_{i=1}^{n} log\left[exp(-H_0(t))\right]^{\boldsymbol{\beta}\boldsymbol{x_i}^T} \\
&= \sum_{i=1}^{n} c_i log\left[h_0(t)\right] + c_i \boldsymbol{\beta}\boldsymbol{x_i}^T + \sum_{i=1}^{n} \boldsymbol{\beta}\boldsymbol{x_i}^T log\left[S_0(t)\right]
\end{aligned}
\tag{3.24}
$$

.

Según el método de máxima verosimilitud esta función tiene que maximizarse con respecto a los parámetros. Sin embargo, en la práctica h_0 es desconocida por lo que la estimación de los parámetros es llevada de distinta manera.

Cox propuso usar la función de verosimilitud parcial (del inglés *partial likelihood function*), en donde especuló que dicha función tiene las mismas propiedades que un estimador de máxima verosimilitud (Hosmer and Lemeshow 1999, págs. 95,101).

$$
l_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{exp(\boldsymbol{\beta}\boldsymbol{x}_i^T)}{\sum_{j \in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)} \right]^{c_i}, \tag{3.25}
$$

donde la suma en el denominador es sobre los sujetos que están expuestos al riesgo al tiempo t_i denotado como $R_{(t_i)}$. Esto es que el conjunto de riesgo para todos los sujetos con tiempo de falla o censura es mayor o igual que el tiempo específico. La expresión tiene el supuesto que no hay tiempos de falla con empates.

Normalmente no se toma en cuenta a los datos censurados, se denotará m al número de individuos con tiempo de falla no censurado[3].

$$
l_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{exp(\boldsymbol{\beta}\boldsymbol{x}_i^T)}{\sum_{j \in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)}. \tag{3.26}
$$

Al obtener el logaritmo natural y derivar la función (3.25) crea un sistema de ecuaciones como sigue[4]

$$
\frac{\partial L_p(\boldsymbol{\beta_k})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \left\{ \boldsymbol{x}_{ik} - \bar{\boldsymbol{x}}_{w_i k} \right\}, \tag{3.27}
$$

[3]Véase Hosmer and Lemeshow 1999, pág 95 ecuación 3.18

[4]Para ver con más detalles los cálculos se recomienda ver Hosmer and Lemeshow (1999), págs. 96-98,101-102.

donde

$$w_{ij}(\boldsymbol{\beta}) = \frac{exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)}{\sum_{l \in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)}, \tag{3.28}$$

y

$$\bar{\boldsymbol{x}}_{w_i k} = \sum_{l \in R_{(t_i)}} w_{ij}(\boldsymbol{\beta})\boldsymbol{x}_{jk}. \tag{3.29}$$

Los elementos de la matriz $p \times p$ de la matriz de información están dadas por

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\frac{\partial^2 L_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}. \tag{3.30}$$

Las segundas derivadas son

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_k^2} = -\sum_{i=1}^{m} \sum_{j \in R_{(t_i)}} w_{ij}(\boldsymbol{x}_{jk} - \bar{\boldsymbol{x}}_{w_i k})^2, \tag{3.31}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_u} = -\sum_{i=1}^{m} \sum_{j \in R_{(t_i)}} w_{ij}(\boldsymbol{x}_{jk} - \bar{\boldsymbol{x}}_{w_i k})(\boldsymbol{x}_{ju} - \bar{\boldsymbol{x}}_{w_i u}). \tag{3.32}$$

La varianza estimada de los coeficientes estimados se obtienen al obtener la inversa de la matriz de información evaluada en los coeficientes estimados.

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{I}(\hat{\boldsymbol{\beta}})^{-1}. \tag{3.33}$$

3.5.1. Estimación de la Verosimilitud cuando se presentan empates

Usualmente cuando se trabajan con datos de supervivencia es usual encontrar empates en los tiempos del evento de interés. Se han propuesto por varios autores distintas funciones de verosimilitud. Entre ellos se encuentran Efron, Breslow y Cox.

Ordenando las observaciones por tiempo de falla (sin censura) con $t_1 < t_2 < ... < t_m$. Sea el número de fallas observadas en t_i son d_i, $i = 1, ..., m$. Sea D_i el conjunto de todos los individuos con tiempo igual a t_i . Sea $\boldsymbol{s}_i = \sum_{j \in D_i} \boldsymbol{x}_j$, donde $\boldsymbol{x}_j$ es el vector de los individuos que murieron en el tiempo t_i. Y $R_{(t_i)}$ los sujetos que estan expuestos al riesgo al tiempo t_i.

La función de verosimilitud parcial propuesta por Breslow[5] es la siguiente

$$l_1(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{exp(\boldsymbol{\beta}\boldsymbol{s}_i^T)}{\left[\sum_{j \in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)\right]^{d_i}} \tag{3.34}$$

[5]Klein and Moeschberger (2003), sección 8.4 ecuación 8.4.1

La aproximación de Efron es un poco más complicada, pero lleva a una mejor aproximación que la de Breslow, el paquete R utliza esta aproximación por *default*. La expresión para la aproximación de Efron[6] es

$$l_2(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{exp(\boldsymbol{\beta}\boldsymbol{s}_i^T)}{\prod_{j=1}^{d_i}\left[\sum_{k\in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_k^T) - \frac{j-1}{d_i}\sum_{k\in D_i} exp(\boldsymbol{\beta}\boldsymbol{x}_k^T)\right]} \tag{3.35}$$

Cox propone un modelo logístico para la función de riesgo como sigue

$$\frac{h(t,\boldsymbol{x},\boldsymbol{\beta})}{1-h(t,\boldsymbol{x},\boldsymbol{\beta})} = exp(\boldsymbol{\beta}\boldsymbol{x}^T)\frac{h_0(t)}{1-h_0(t)} \tag{3.36}$$

Sea Q_i el conjunto de todos los subconjuntos de d_i que pueden ser seleccionados del conjunto de riesgo $R_{(t_i)}$. Cada elemento de Q_i es un d_i-tuplo de individuos que pudieron haber estado en cualquier d_i fallas en el tiempo t_i. Sea $\boldsymbol{q} = (q_1, ..., q_{d_i})$ uno de los elementos de Q_i y $\boldsymbol{s}_q^* = \sum_{j=1}^{d_j} \boldsymbol{x}_{qj}$. La función de log verosimilitud[7] discreta está dada por

$$l_3(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{exp(\boldsymbol{\beta}\boldsymbol{s}_i^T)}{\sum_{\boldsymbol{q}\in Q_i} exp(\boldsymbol{\beta}\boldsymbol{s}_i^{*T})} \tag{3.37}$$

Cuando no se presentan colas las funciones de verosimilitudes presentadas se reducen a la expresión en la ecuación (3.26) (Klein and Moeschberger 2003, pág. 260).

3.5.2. Estimación de la Función de Supervivencia de un Modelo de Cox

Existen diversas estimaciones de la función de supervivencia para un modelo de Cox, pero están basados en el estimador de Breslow para la función de riesgo acumulada. Dicho estimador se produce a partir de la función de verosimilitud parcial de Cox,

$$\hat{H}_0(t) = \sum_{t_i \le t} \frac{1}{\sum_{j\in R_{(t_i)}} exp(\boldsymbol{\beta}\boldsymbol{x}_j^T)} \tag{3.38}$$

A partir de este estimador se ajusta un modelo de riesgos proporcionales a los datos para obtener la estimación de los parámetros $\hat{\boldsymbol{\beta}}$ y la varianza estimada $\hat{V}(\hat{\boldsymbol{\beta}})$. Sean $t_1 < t_2 < ... < t_m$ los tiempos de falla (datos no censurados) y d_i es el número de fallas al tiempo t_i, $i = 1, ..., m$. Sea

$$W(t_i, \boldsymbol{\beta}) = \sum_{j\in R_{(t_i)}} exp\left(\sum_{k=1}^{p} \beta_k x_{jk}\right) \tag{3.39}$$

El estimador de la función de riesgo acumulada[8] $H_0 = \int_0^t h_0(u)du$ esta dada por

[6]Klein and Moeschberger (2003), sección 8.4 ecuación 8.4.2
[7]Klein and Moeschberger (2003), sección 8.4 ecuación 8.4.3
[8]Klein and Moeschberger (2003), sección 8.8 ecuación 8.8.2.

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{W(t_i, \boldsymbol{\beta})} \tag{3.40}$$

Esta función es una función escalonada como la de Kaplan-Meier. Además este estimador se reduce al estimador de Nelson-Aalen de la sección 3.15.

3.6. Interpretación del Modelo

Habiendo ajustado un modelo de regresión es importante ver que información es redundante, con la finalidad de obtener un modelo parsimonioso que pueda explicar los datos. De la misma manera se utilizan las pruebas de Wald y la prueba del cociente de verosimilitudes vistas en la sección 1.3. Se prosigue con la interpretación de un modelo ajustado.

El modelo de regresión de Cox es usado cuando el objetivo principal es analizar los efectos de las variables en el tiempo de supervivencia. Considere el caso cuando se tiene una sola variable explicativa, $p = 1$. En un modelo de regresión de Cox al obtener el logaritmo de la función de riesgo, y hacer la diferencia para un cambio de $x = a$ a $x = b$ se obtiene

$$\begin{aligned} g(t; x = a; \beta) - g(t; x = b; \beta) &= \{ln[h_0(t)] + a\beta\} - \{ln[h_0(t)] + b\beta\} \\ &= a\beta - b\beta \\ &= (a - b)\beta. \end{aligned} \tag{3.41}$$

Esta diferencia es la manera de medir el efecto en el cambio en una variable explicativa. Sin embargo, no es fácil interpretarlo. Si se exponencia el resultado de la ecuación (3.41) queda como sigue

$$\begin{aligned} HR(t, a, b, \beta) &= exp(g(t; x = a; \beta) - g(t; x = b; \beta)) \\ &= \frac{h(t, a, \beta)}{h(t, b, \beta)} \\ &= exp((a - b)\beta). \end{aligned} \tag{3.42}$$

La expresión (3.42) es llamada *hazard ratio* o razón de riesgos. Éste juega la misma interpretación y explica los resultados de un análisis de supervivenvia de la misma manera que la razón de momios lo hace en la regresión logística (Hosmer and Lemeshow (1999), pág. 114).

Si el modelo tiene una sola variable explicativa ($p = 1$). El caso que que la variable independiente sea una variable binaria codificada como $X = 0, 1$. Se obtiene la razón de riesgos

$$HR(t, 0, 1, \beta) = exp(\beta) \tag{3.43}$$

Quedó de manera idéntica a la forma de la razón de momios de la regresión logística (sección 1.2.2) para una variable binaria. La diferencia es que, en el contexto, ésta es una razón de tasas de riesgos mientras que la otra es una razón de momios. La ventaja es que la razón de riesgos es una medida de la supervivencia que puede verse a lo largo del horizonte de tiempo t. Mientras que la razón de momios es una medida de comparación del evento de ocurrencia sólo al final del tiempo de estudio.

En el caso hipotético suponga que se estudia una cohorte de personas que es seguida durante 5 años, en donde se estudia la muerte de los individuos[9]. Bajo este supuesto se tiene el interés de estudiar el evento de muerte al final del estudio, usando una regresión logística. Suponga que X es una variable binaria que denota el sexo de la persona (1=masculino). Suponga que la razón de momios de la variable X es de 2. Su interpretación es que si la razón se acerca al riesgo relativo, entonces la probabilidad de ocurrencia del evento de muerte se presenta dos veces más en los hombres que las mujeres. Una razón de riesgos de dos significa que a cualquier tiempo del estudio, la tasa de muerte por unidad de tiempo de los hombres es el doble que las mujeres.

Si además es posible observar el tiempo de falla de cada individuo se puede estratificar la muestra y obtener el número de personas en riesgo, el número de muertes o fallas, las estimaciones de la función de riesgo

$$h_k(t) = \frac{d_k(t)}{n_k(t)},\ k = 0, 1,$$

y la razón de riesgos

$$HR(t) = \frac{h_1(t)}{h_0(t)}.$$

Por otro lado, si X es una variable categórica con l categorías, la construcción de la razón de riesgos es de la misma manera que en la regresión logística, se escoge un nivel de la variable como una categoría de referencia y construyen $l-1$ variables *dummies*, donde se compara la tasa de riesgo de cada categoría contra el grupo de referencia.

Por ejemplo, sea una variable X con $k = 1, 2, ..., l$ categorías, donde X_1 es la variable *dummy* que representa la categoría de referencia y las restantes son codificadas como 0 y 1.

El logaritmo de la función de riesgo, retomando las expresiones (3.41) y (3.42) ignorando la función hazard base es

$$g(t, X, \hat{\boldsymbol{\beta}}) = \hat{\beta}_1 X_2 + \hat{\beta}_2 X_3 + ... + \hat{\beta}_{l-1} X_l$$

El estimador de la función hazard comparando la categoría 2 de X (X_2) contra la categoría de referencia (X_1) aplicando la expresión (3.41) se tiene

[9]Véase Hosmer and Lemeshow (1999) en la página 116.

$$g(t, X = 2, \hat{\boldsymbol{\beta}}) - g(t, X = 1, \hat{\boldsymbol{\beta}}) = (\hat{\beta}_1 1 + \hat{\beta}_2 0 + ... + \hat{\beta}_{l-1} 0) - (\hat{\beta}_1 0 + \hat{\beta}_2 0 + ... + \hat{\beta}_{l-1} 0) = \hat{\beta}_1$$

Exponenciando se obtiene

$$\widehat{HR}(2, 1) = exp(\hat{\beta}_1).$$

Sucesivamente para l categorías se tiene

$$\widehat{HR}(3, 1) = exp(\hat{\beta}_2)$$

$$\vdots$$

$$\widehat{HR}(l, 1) = exp(\hat{\beta}_{l-1}).$$

Para una variable continua considere $a = x$ y $b = x + c$. De (3.42) el estimador de la razón de riesgos queda como:

$$\widehat{HR}(t, x, x + c, \beta) = exp(c\beta). \tag{3.44}$$

La interpretación es similar, si se tiene un $\widehat{HR}(5) = 1{,}5$ es que la tasa de riesgo se incrementa o disminuye en un 50 % por cada aumento de 5 en x e independiente del valor de x al cual el incremento es calculado[10].

3.7. Técnicas de Validación del Modelo

El siguiente paso después de haber ajustado un modelo de regresión en general es saber que tan bien ajusta a los datos. Este proceso es generalmente llamado como medición de la adecuación del modelo. Existen diversas técnicas para poder medir un modelo de regresión de Cox. Estas técnicas permiten desde checar el ajuste del modelo hasta el supuesto de proporcionalidad. El presente trabajo solamente estará interesado en analizar los residuales como una medida de ajuste del modelo pues no se está interesado en revisar la proporcionalidad del modelo.

La definición de residual en el modelo proporcional de Cox no es una tarea sencilla, porque la variable respuesta esta sujeta al tiempo del evento y las observaciones pueden estar incompletas o censuradas, es por eso que el análisis de regresión para tiempos de vida suele tratarse aparte de otros modelos de regresión. Además, su ajuste es distinto por el uso de la función de verosimilitud parcial. La combinación conjunta de datos, modelo y definición de ajuste es lo que hace difícil la tarea de definir un residual que otras técnicas estadísticas.

Las técnicas más usadas para ver el ajuste general del modelo de Cox son los residuales de Cox-Snell y los residuales de martingala.

[10]Un ejemplo puede verse en Hosmer and Lemeshow (1999), pág. 129

3.7.1. Residuales de Cox-Snell

Los residuales de Cox-Snell son utilizados para medir el ajuste de la regresión de Cox. Suponga que se ha ajustado un modelo de regresión considerando la terna (t_i, x_i, c_i), para $i = 1, ..., n$. Suponga que las variables explicativas $\boldsymbol{X} = (X_1, ..., X_p)$ son variables fijas (no dependientes de tiempo). Con $\hat{\boldsymbol{\beta}}$ los coeficientes estimados de la regresión de Cox, entonces los residuales de Cox-Snell son como siguen

$$r_i = \hat{H}_0(t_i) exp\left(\sum_{k=1}^{p} \hat{\beta}_k x_{kj}\right) \tag{3.45}$$

donde $\hat{H}_0(t_i)$ es el estimador de la función de riesgo acumulada de Breslow definida en la ecuación (3.40). Si el modelo es correcto y $\hat{\boldsymbol{\beta}}$ se acercan a los verdaderos valores $\boldsymbol{\beta}$ del modelo, r_j se ve como una muestra censurada de una distribución exponencial. Es decir

$$\begin{aligned} r_i &= \hat{H}(t_i, \boldsymbol{x}_i) \\ &= -ln\hat{S}(t_i, \boldsymbol{x}_i) \sim exp(\lambda = 1) \end{aligned} \tag{3.46}$$

Para checar que r_i tiene una distribución exponencial, se obtiene el estimador Nelson-Aalen para la función acumulada de riesgo de r_i ($\hat{H}_r(r_i)$). Si el estimador de la función de riesgo acumulada de r_i se acerca a la distribución exponencial ($\lambda = 1$), el estimador debe de parecerse a la función de riesgo acumulada de la exponencial. Entonces la gráfica de $\hat{H}_r(r_i)$ contra r_i debe de ser una línea que sale del origen con pendiente 1 (diagonal de 45°).

3.7.2. Residuales de Martingala

Este residual es una modificación de los residuales de Cox-Snell. Para definir este residual se supone que el i-ésimo individuo de la muestra, $i = 1, ..., n$, tiene un vector $\boldsymbol{x}_i(t)$ de posibles variables dependientes de tiempo. Sea $N_i(t)$ con valor de 1 al tiempo t si el individuo ha experimentado la falla y 0 que aún no ha experimentado la falla. $Y_i(t)$ es el indicador de que el individuo i esta bajo estudio por un tiempo antes del tiempo t. Y $\hat{\boldsymbol{\beta}}$ los coeficientes de la regresión y $\hat{H}_0(t)$ el estimador de Breslow de la función acumulada de riesgo. Los residuales de Martingala estan expresados en la siguiente expresión.

$$\hat{M}_i = N_j(\infty) - \int_0^{\infty} Y_i(t) exp(\hat{\boldsymbol{\beta}} \boldsymbol{x}_i^T) d\hat{H}_0(t_i). \tag{3.47}$$

Cuando hay censura por derecha y hay variables independientes de tiempo la expresión (3.47) se reduce a

$$\hat{M}_i = c_i - \hat{H}_0(t_i) exp\left(\sum_{k=1}^{p} \hat{\beta}_k x_{kj}\right) = c_i - r_i, \tag{3.48}$$

donde c_i es la variable que indica censura del individuo i[11]. Los residuales de Cox-Snell estan muy relacionados a los residuales de martingala cuando se presenta la censura por derecha y no hay variables que cambian con el tiempo. En R para poder graficar los residuales de Cox-Snell es necesario obtener los residuales de martingala y luego hacer la gráfica (Anexo B.3). No obstante, su uso principal es para poder graficarlos contra las variables explicativas y examinar si la variable necesita discretizarse o aplicarle una transformación.

[11]Klein and Moeschberger (2003), pág. 360 ecuación 11.3.2

Capítulo 4

Aplicación de los Modelos

4.1. Introducción

El propósito de este capítulo es ilustrar los modelos descritos en los capítulos previos, tomando como ejemplos una base de datos de crédito alemán y una base de datos de relacionada con la duración de vida después de un transplante de médula ósea. La base de crédito alemán puede encontrarse con el nombre *Determining the solidness of borrowers via credits-coring*[1] la cual presenta 1000 observaciones con 20 variables explicativas, es una buena base de datos ya que contiene numerosas observaciones. La base de datos de transplantes contiene datos de 137 pacientes de leucemia a los cuales se les hizo un transplante de médula ósea y se trabajará sobre el evento de muerte. De las variables explicativas de cada base de datos se tomará un conjunto de variables para someterlas a los distintos modelos, se describe la estructura de la base de datos y se busca un modelo simplificado, que ayude a explicar los eventos de interés, los clientes riesgosos más propensos a delinquir y cuales son los factores que afectan en la supervivencia del paciente.

En la sección 4.2.1, se describe la base de datos para la aplicación en el área de crédito de un banco alemán junto con las recodificaciones iniciales, en la sección 4.2.2 se describe la base de datos de 137 transplantes de médula anexando las recodificaciones de algunas variables. En la sección 4.3 son aplicados los modelos de la regresión logística y análisis de discriminante a la base de crédito. En la sección 4.4 son aplicados los modelos de regresión logística y regresión de Cox a la base de los 137 transplantes. En la sección 4.5 se concluye el capítulo resaltando los puntos más notables en el análisis de los modelos presentados.

[1]La base puede encontrarse en la siguiente página web:
http://www.statistik.lmu.de/service/datenarchiv/kredit/kredit_e.html

4.2. Descripción de las Bases de Datos

4.2.1. Base de datos de crédito aleman

Normalmente las bases de datos de *credit scoring* son grandes pudiendo llegar a contener más de 100,000 observaciones con 100 variables explicativas, por eso es importante obtener el modelo más parsimonioso que pueda ser utilizado para detectar con facilidad el tipo de cliente. Las variables se utilizan en Inglaterra dentro del *credit scoring* están expresadas en el cuadro 4.1 (Hand and Henley 1997).

Cuadro 4.1: Variables comúnmente utilizadas para determinar modelos de puntaje de crédito en Inglaterra (Hand and Henley 1997).

Variables
Antigüedad en el Domicilio
Estatus de la Propiedad
Telefono
Ingreso Anual
Tarjeta de crédito
Tipo de cuenta dentro del banco
Edad
Ocupación
Propósito del Crédito
Estado civil
Antigüedad del cliente en el banco
Antigüedad en el trabajo

Otras características como sexo y etnicidad no se toman en cuenta por la ley en Inglaterra, ya que algunos consideran que su uso es discriminatorio.

La base de datos de crédito contiene 1000 clientes de un banco alemán. Para cada persona u observación se tiene una variable respuesta binaria llamada *creditability* en donde clasifica si el cliente fue bueno o malo. Además, cuenta con 20 variables explicativas que describen las características económicas y demográficas del cliente. Esta base de datos puede encontrarse con el nombre de *Determining the solidness of borrowers via credit scoring.*

En Agresti (2002), se toma el problema 6.21 de la página 263. Se toman las variables corriente, duración del crédito, pago de los créditos pasados, propósito del crédito, sexo y estado civil.

En el cuadro 4.2 se presenta la codificación de las variables explicativas del problema en el libro de Agresti, cada porcentaje es el número de observaciones que toma la categoría de la variable explicativa con respecto al total de la categoría de la variable respuesta. En el caso de las variables continuas se pone el promedio que toma la variable.

Cuadro 4.2: Codificación original de las variables explicativas de la base de crédito alemán. La categoría 2 de la variable sexo/estado civil representa el 31 % con respecto al total, en esta categoría no hay manera de saber exáctamente si es hombre o mujer.

Variable Explicativa	Descripción	Categorías	% Buenos	% Malos	%Total
Kredit	Estatus del cliente	1 : solvente (bueno)			70
		0 : no solvente (malo)			30
Laufkont	Balance de cuenta corriente	1: no posee cuenta	45.00	19.86	27.40
		2: sin balance o débito	4.67	7.00	26.90
		3: 0 <= ... < 200DM o cuenta de cheques con al menos 1 año	15.33	49.71	6.30
		4: >= 200 DM or	35.00	23.43	39.40
Laufzeit	Duración del crédito	Continua	19.20	24.86	20.903
Moral	Pago de créditos anteriores	0: pago vacilante de créditos anteriores	8.33	2.14	4.00
		1: línea de crédito problemática/ tiene créditos en otros bancos	9.33	3.00	4.90
		2: no hay créditos previos / pagó todos sus créditos anteriores	56.33	51.57	53.00
		3: sin problemas con créditos actuales del banco	9.33	8.57	8.80
		4: pagó créditos anteriores del banco	16.67	34.71	29.30
Verw	Propósito del crédito	1: auto nuevo	5.67	12.29	10.3
		2: auto usado	19.33	17.57	18.1
		3: muebles	20.67	31.14	28
		4: radio / televisión	1.33	1.14	1.2
		5: aparatos electrodomésticos	2.67	2.00	2.2
		6:reparaciones	7.33	4.00	5
		7: educación	0.00	0.00	0
		8: vacaciones	0.33	1.14	0.9
		9: capacitación	11.33	9.00	9.7
		10: negocio	1.67	1.00	1.2
		0: otro	29.67	20.71	23.4
Famges	Estado civil / sexo	1: hombre: divorciado / vive aparte	6.67	4.29	5
		2: mujer: divorciada / vive aparte / casada	11.33	10.29	31
		2: hombre: soltero	25.00	18.43	-
		3: hombre: casado / viudo	48.67	57.43	54.8
		4: mujer: soltera	8.33	9.57	9.2

Se observa que en la variable respuesta el porcentaje de la población cosiderada como buenos clientes es del 70 % mientras que el 30 % no.

Limpieza de la Base de Datos

La presencia de datos faltantes es un problema muy común en la estadística, ya que conduce a problemas en el proceso de ajuste de los modelos y el problema se incrementa cuando se trabajan con clases desbalanceadas, sucede cuando el tamaño de una población constituye un porcentaje muy pequeño en la base de datos, o cuando se tiene muy pocos o nulos datos observados en una categoría de una tabla de contingencia (Hosmer and Lemeshow 2000, sección 4.5; Hand et al. 1997). Normalmente cuando se presentan datos faltantes en alguna variable explicativa se procura trabajar con otras variables, o bien usar algún método de imputación de valores ya sea por moda, promedio, etc. Esta base de datos no tiene valores faltantes.

Las variables fueron analizadas para buscar valores fuera de rango, o que no correspondan al tipo de variable.

En la variable de *balance de cuenta corriente*, no se encontró algún valor fuera de rango, hay pocas personas que presentan la tercera categoría de esta variable, pero su número no puede implicar problemas para el ajuste de los modelos. En la variable *duración del crédito* se observa que el rango de valores se encuentra de 4 a 73 y es muy irregular presentando un esquema de crédito con muchas observaciones en los múltiplos de 3, sus valores son congruentes. Se hizo una gráfica de las frecuencias de esta variable la cual se da en la figura 4.1.

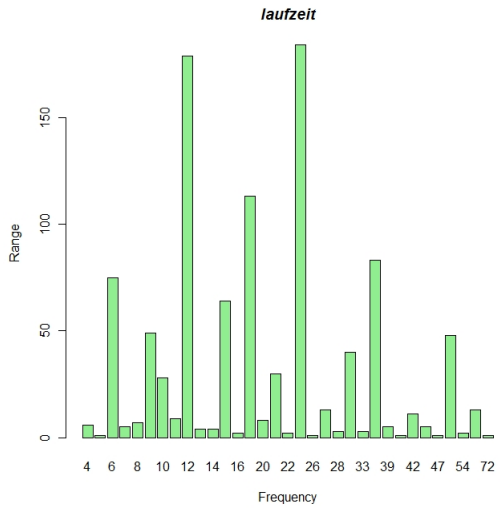

Figura 4.1: Frecuencias de la variable *laufzeit* (duración de crédito) en la base de crédito alemán. Se observan frecuencias altas en los múltiplos de 3.

En la variable del *pago de anteriores créditos* se observa que casi la mitad de la población eran clientes que no tienen historial crediticio o que sí pagaron sus créditos anteriores. Sin

embargo, la proporción de buenos y malos créditos de la variable respuesta con respecto a esta variable (cuadro 4.2) dice que el 56.3 % del total de los malos fueron a causa de esta característica, no se puede decir con certeza si ellos eran clientes sin historial o aquellos que tenían un buen historial crediticio. Y el 29.3 % son clientes del banco que pagaron sus deudas en la institución.

En la variable del *propósito de crédito* no se encontraron valores atípicos, pero la categoría 7: educación no presentó observaciones, esta variable deberá ser recodificada. Además, es una categórica sin orden. Por ejemplo, no se sabe si pedir un préstamo para tener un carro usado es mejor que para uno nuevo. Ya que al tomarla como ordinal se está asignando un puntaje. Se crearán variables dicotómicas tomando una clase considerando una categoría como referencia. Se aprecia que la proporción de las categorías 4, 5, 6, 8 y 10 de esta variable con respecto a la variable respuesta, presentan un porcentaje pequeño lo cual indica muy pocas observaciones, lo que se espera es que no contribuyan en el riesgo de fallo, por consiguiente se espera colapsar algunas categorías que esten relacionadas por esos conceptos.

En la variable del *estado civil/sexo*, sus valores se encuentran en rango ya que no toman valores atípicos. En esta variable no es posible distinguir hombres y mujeres ya que en la categoría dos están mezclados. Por ser una categórica sin orden se construirán variables dicotómicas tomando alguna como referencia.

Recodificación de las Variables

De acuerdo con las observaciones encontradas en la sección anterior se recodificaron algunas variables explicativas como sigue:

- *Laufkont* (Balance de Cuenta Corriente).- No se realizó ningún cambio.
- *Laufzeit* (Duración del crédito en meses).- No se realizó ningún cambio.
- *Moral* (Pago de antiguos créditos).- No se realizó ningún cambio
- *Verw* (Propósito del crédito).- Esta variable fue convertida en variables *dummies* tomando como la categoría de referencia a A0= Otro, A1= Nuevo carro, A2= Carro usado. Debido a que las categorias 4=Radio o televisión y 5= Aparatos electrodomésticos tenían muy pocas observaciones, fueron colapsadas con la categoría 3= Muebles haciéndola como la categoría A3, pues son bienes que se encuentran en la mayoría de las casas. Se considera A4= Reparaciones, A5= Vacaciones, A6= Capacitación y A7= Negocio.
- *Famges* (Edo. civil/sexo).- En esta variable explicativa se hicieron las variables *dummies*, donde se consideró la categoría 2 como categoría de referencia, ya que es donde se revuelven los hombres y mujeres.

4.2.2. Base Transplantes de Médula Ósea

Esta base de datos contiene 137 pacientes enfermos de distintos tipos de leucemia (AML, ALL), a los cuales se les hizo un transplante de médula ósea. Estos pacientes fueron tratados en 4 hospitales: 76 en el hospital universitario del estado de Ohio (OSU) en Columbus; 21 en la universidad de Hahnemann (HU) en Filadelfia; 23 en el hospital de San Vicente(SVH) en Sydney Australia; y 17 en el hospital de Alfred (AH) en Melbourne Australia.

El estudio consiste en transplantes que fueron realizados por estas instituciones desde marzo de 1984 a junio de 1989. El seguimiento o duración del estudio fueron 7 años. Posee 22 variables explicativas entre ellas estan los datos biológicos de la persona y el donante, así como variables médicas, las cuales presentan el desarrollo de síntomas o enfermedades.

El transplante de médula ósea es un tratamiento estándar para la leucemia aguda. El recuperamiento es un proceso complejo por lo que el pronóstico de recuperamiento puede depender de los factores de riesgo conocidos al momento del transplante, como la edad y sexo del paciente o el donante, el estado inicial de la enfermedad, el tiempo desde el diagnóstico al transplante, etc. El pronóstico final cambia debido al estado post-operativo del paciente como el desarrollo de otras enfermedades, o el recupermiento total. El transplante puede ser considerado como fallido cuando el paciente de leucemia recae en la enfermedad o muere durante su tratamiento.

En esta base de datos se tienen variables de tiempo a los cuales indica el tiempo que tomó en presentarse las características del paciente. Entre ellas se encuentre el estado de muerte del paciente. Esta base de datos puede ser encontrada en el libro de análisis de supervivencia de Klein & Moeschberger (2003) con el nombre de *Bone marrow transplantation for Leukemia* [2].

En el cuadro 4.3 se presenta la codificiación de las variables de tiempo asociadas al presentarse el desarrollo de síntomas post-operativos, en el cuadro 4.4 se encuentran las variables que describen al paciente. En esos cuadros se presenta cada porcentaje dentro de la categoría de la variable respuesta, i.e. es el número de observaciones que toma la categoría de la variable explicativa con respecto al total de la categoría de la variable respuesta. En el caso de las variables continuas se presenta el promedio que toma la variable en la categoría de la variable respuesta.

[2]No obstante los detalles del estudio pueden encontrarse en Copelan E. A., Biggs J. C., Thompson J. M., Crilley P., Szer J., Klein J. P., Kapoor N., Avalos B. R., Cunningham I., Atkinson K., Downs K., Harmon G. S., Daly M. B., Brodsky I., Bulova S. I., and Tutschka P. J. (1991). Treatment for Acute Myelocytic Leukemia with Allogeneic Bone Marrow Transplantation Following Preparation with Bu/Cy. *Blood* **78**, 838-843.

Cuadro 4.3: Codificación orignal de las variables de la base de 137 transplantes de médula ósea. Variables que describen el desarrollo de síntomas post-operativo del paciente.

Variable Explicativa	Descripción	Categorías	% Alive	% Death	%Total
g	Disease Group	1: ALL 2: AML Low Risk 3: AML High Risk	25.0 55.4 19.6	29.6 28.4 42.0	27.7 39.4 32.8
T1	Time To Death Or On Study Time	Continua	1538.42	355.72	839.16
T2	Disease Free Survival Time (Time To Relapse, Death Or End Of Study)	Continua	1502.69	283.79	782.03
C1	Death Indicator	1: Death 0: Alive			56 81
C2	Relapse Indicator	1: Relapsed 0: Disease Free	3.6 96.4	49.4 50.6	30.7 69.3
C3	Disease Free Survival Indicator	1: Dead Or Relapsed 0: Alive Disease Free	3.6 96.4	100.0 0.0	60.6 39.4
TA	Time To Acute Graft-Versus-Host Disease	Continua	1300.07	316.09	718.31
A	Acute GVHD Indicator	1: Developed Acute GVHD 0: Never Developed Acute GVHD	17.9 82.1	19.8 80.2	19.0 81.0
TC	Time To Chronic Graft-Versus-Host Disease	Continua	794.14	240.38	466.74
C	Chronic GVHD Indicator	1: Developed Chronic GVHD 0: Never Developed Chronic GVHD	57.1 42.9	35.8 64.2	44.5 55.5
TP	Time To Return of Platelets to Normal Levels	Continua	39.16	55.91	49.07
P	Platelet Recovery Indicator	1: Platelets Returned To Normal 0: Platelets Never Returned to Normal	98.2 1.8	80.2 19.8	87.6 12.4

Cuadro 4.4: Codificación de las variables explicativas de la base de 137 transplantes de médula ósea. Variables acerca del donante y del paciente.

Variable Explicativa	Descripción	Categorías	% Alive	% Death	%Total
Z1	Patient Age In Years	Continua	27.76	28.77	28.36
Z2	Donor Age In Years	Continua	27.48	28.91	28.33
Z3	Patient Sex	1 : Male	62.5	55.6	58.4
		0 : Female	37.5	44.4	41.6
Z4	Donor Sex	1 : Male	66.1	63.0	64.2
		0 : Female	33.9	37.0	35.8
Z5	Patient CMV Status	1 : CMV Positive	46.4	51.9	49.6
		0 : CMV Negative	53.6	48.1	50.4
Z6	Donor CMV Status	1 : CMV Positive	42.9	42.0	42.3
		0 : CMV Negative	57.1	58.0	57.7
Z7	Waiting Time to Transplant In Days	Continua	286.607	267.12	275.09
Z8	FAB	1 : FAB Grade 4 Or 5 and AML	19.6	42.0	32.8
		0 : Otherwise	80.4	58.0	67.2
Z9	Hospital	1: The Ohio State University	46.4	61.7	55.5
		2: Alferd	7.1	16.0	12.4
		3: St. Vincent	19.6	14.8	16.8
		4: Hahnemann	26.8	7.4	15.3
Z10	MTX Used as a Graft-Versus--Host-Prophylactic	1 : Yes	26.8	30.9	29.2
		0 : No	73.2	69.1	70.8

Un pequeño glosario del los términos que serán utilizados[3]:

- *Acute lymphoblastic leukemia* (ALL): es una forma de leucemia o cáncer de las células blancas caracterizada por el exceso de limfoblastos.
- *Acute myeloid leukemia* (AML): es un cáncer de la línea mieloide de las células de la sangre, caracterizada por un crecimiento rápido anormal de las células blancas que se acumulan en la médula ósea e interfieren con la producción normal de las células de la sangre.
- *Graft-versus-host disease* (GVHD): es una complicación del transplante de médula ósea el cual las células inmunes reconocen a la médula transplantada como foránea provocando un ataque inmunológico.
- *Cytomegalovirus* (CMV): es una forma de herpesvirus.

[3]Los términos se encuentran en Klein and Moeschberger (2003), sección 1.3.

- *French-American-British* (FAB) *classification*: es una clasificación basada en el criterio morfológico estándar. Aquellos que tengan clasificación M4 o M5 fueron considerados como posibles riesgos elevados de caer nuevamente en enfermos o de muerte a causa de tratamiento.

Limpieza de la base de datos

Esta base de datos no presentó valores faltantes, se revisa si esta base de datos contiene valores atípicos, y que sus valores corresponden al tipo de variable.

Del cuadro 4.3 se tienen diversas medidas de tiempo, la unidad de tiempo manejada es en término de días, por eso las variables de tiempo principalmente se encuentran entre el rango de 1 hasta 2640. En las variables indicadoras todos sus valores se encuentran en rango, tomando 1 presencia y 0 ausencia.

Del cuadro 4.4, en la variable de la edad del paciente sorprende que su rango de valores va desde los 7 a los 52 años. Mientras que la edad del donante esta entre los 2 y los 56 años. No es común que un menor de edad sea donante, lo que podría indicar es que algunos de estos donantes ya hayan fallecido. En el tiempo de espera en recibir el transplante sus valores van de 24 a 2616 días. Las otras variables dicotómicas se encuentran en rango. La variable Z9 es una variable categórica, los valores de esta variable se encuentran en rango. Como es una variable categórica no ordinal deberá ser recodificada.

Recodificación de variables

De acuerdo con las observaciones se recodificaron algunas variables explicativas como sigue:

- g:= Enfermedad del paciente. Se hicieron variables *dummies* quedando como gA1 la categoría de referencia representando el número 1 de la variable original, gA2 el número 2 y gA3 el número 3.
- Z_9:=Hospital. Esta solamente se hicieron las variables *dummies*, donde se consideró como categoría de referencia la número 1 representado como Z9A1, Z9A2 para la categoría 2 de la variable, Z9A3 para la categoría 3 y Z9A4 para la categoría 4.

4.3. Base de datos de Crédito Alemán

4.3.1. Aplicación de los Modelos de Regresión Logística y Análisis de Discriminante

A continuación se presenta la aplicación e interpretación de los modelos de regresión logística y análisis de discriminante. Se usa la recodificación del conjunto de variables explicativas descritas en la sección 4.2.1. Suponiendo que este conjunto de variables conforma el

modelo saturado.

En principio se ajustó el modelo de regresión logística usando el software R. Al obtener los coeficientes ajustados y el valor de p, se construyeron los intervalos de confianza con significancia del 5 % con el objetivo de hacer inferencia estadística sobre los parámetros ajustados.

En el cuadro 1.4 se muestran los datos obtenidos al ajustar la regresión logística, en donde se observa lo siguiente: hay menos variables que aumentan el riesgo de ser crédito malo estas variables son *famgesA1*, y *verwA4*. Las variables que tienen coeficientes grandes en valor relativo son *laufkont*, *laufzeit*, *moral*, *famgesA3*, *verwA1*, *verwA2*, *verwA3*. *Laufzeit* tiene un coeficiente estimado chico, quizás se debe por el amplio rango de valores que toma la variable. Las variables que son significativas son *laufkont*, *laufzeit*, *moral*, *famgesA3*, *verwA1*, *verwA2*, y *verwA3*.

Cuadro 4.5: Modelo 1: aplicación de la regresión logística al modelo saturado de la base de datos de crédito.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(intercept)	1.386673	0.319404	4.341	1.42E-05	0.7606	2.0127
Laufkont	-0.614226	0.068449	-8.973	2e-16	-0.7484	-0.4801
Laufzeit	0.043016	0.006737	6.386	1.71E-10	0.0298	0.0562
Moral	-0.391583	0.077099	-5.079	3.80E-07	-0.5427	-0.2405
FamgesA1	0.112572	0.349683	0.322	0.7475	-0.5728	0.7980
FamgesA3	-0.460420	0.17803	-2.586	0.0097	-0.8094	-0.1115
FamgesA4	-0.209135	0.293819	-0.712	0.4766	-0.7850	0.3668
verwA1	-1.406909	0.332756	-4.228	2.36E-05	-2.0591	-0.7547
verwA2	-0.576205	0.236669	-2.435	0.0149	-1.0401	-0.1123
verwA3	-0.744492	0.213903	-3.481	0.0005	-1.1637	-0.3252
verwA4	0.329088	0.366056	0.899	0.3686	-0.3884	1.0466
verwA5	-1.686947	1.117009	-1.51	0.131	-3.8763	0.5024
verwA6	-0.538666	0.300096	-1.795	0.0727	-1.1269	0.0495
verwA7	-0.928702	0.704782	-1.318	0.1876	-2.3101	0.4527

Para el modelo 1 se realiza la prueba de cociente de verosimilitudes por medio de la devianza nula y residual reportada en la salida de R y se obtuvo $P(\chi^2_{(13)} > G) = 6{,}14E-43$, se rechaza la hipótesis nula

$$H_0 : \beta_1 = ...\beta_p = 0 \ vs \ H_a : \beta_i \neq 0 \text{ para algún } i = 1, ..., p,$$

esto significa que hay coeficientes que no son cero. Y al realizar la prueba de Hosmer-Lemeshow (*HL = 5.005, d.f= 8, p=0.756*), el modelo no presenta falta de ajuste. Por otro lado, se realizó un procedimiento de eliminación hacia atrás removiendo una a una variables no significativas, dicho modelo llamado modelo 1.0 se muestra en el cuadro 4.6. Las variables fueron

eliminadas con base a la significancia estadística de la prueba de Wald y fueron removidas en el siguiente orden: *FamgesA1*, *FamgesA4*, *verwA4*, *verwA7*, *verwA5*, *verwA6*, *verwA2*.

Cuadro 4.6: Submodelo 1.0: afinación del modelo saturado de la base de crédito alemán se descartan las variables en el siguiente orden: *FamgesA1, FamgesA4 , verwA4, verwA7, verwA5, verwA6, verwA2*.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.008608	0.280095	3.601	0.000317	0.4596	1.5576
Laufkont	-0.606063	0.067239	-9.014	¡2e-16	-0.7379	-0.4743
Laufzeit	0.041561	0.006499	6.395	1.61E-10	0.0288	0.0543
Moral	-0.358813	0.07573	-4.738	2.16E-06	-0.5072	-0.2104
FamgesA3	-0.404843	0.15857	-2.553	0.010677	-0.7156	-0.0940
verwA1	-1.127278	0.307268	-3.669	0.000244	-1.7295	-0.5250
verwA3	-0.483871	0.176201	-2.746	0.00603	-0.8292	-0.1385

El submodelo anterior se compara con el modelo saturado, aplicando la prueba de cociente de verosimilitudes se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, a un $\alpha = 0{,}05$. Lo cual indica que alguna de las variables no incluidas es significativa (*Chi-squared 7, d.f. = 14.13832 , P value = 0.0488*). Se introdujo *verwA2* ya que fue la última variable que fue excluida, se aplica la prueba contra el modelo saturado y no rechaza la hipótesis nula (*Chi-squared 6 d.f. = 10.32812 , P value = 0.1115*). Lo cual *verwA2* es una variable significativa en el modelo.

Por otro lado, se construyó un modelo diferente, en este modelo se colapsaron las categorías *verwA5*, *verwA6* y *verwA7* creando la variable *verwA8*. Pues estas variables no son significativas con la prueba de Wald, y tienen el mismo signo negativo. Se construyó una variable *dummy* haciendo referencia a estas categorías y se ajustó el modelo de regresión logística. Se observó que los coeficientes obtenidos son casi de la misma magnitud en valor absoluto. Además, en donde se acepta la hipótesis nula (prueba de Wald) $H_0 : \beta_i = 0$ *vs* $H_a : \beta_i \neq 0$, son las mismas del modelo anterior. El modelo ajustado se llamará modelo 1.2, los resultados obtenidos al ajustar este modelo y sus intervalos de confianza al 5 % se muestran en el cuadro 4.7.

Cuadro 4.7: Modelo 1.2: ajuste de regresión logística en donde se colapsan categorías no significativas de la variable *verw* (propósito de crédito). *VerwA8=verwA5∪verwA6∪ verwA7.*

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.361089	0.318199	4.277	1.89E-05	0.7374	1.9848
Laufkont	-0.611373	0.068174	-8.968	¡2e-16	-0.7450	-0.4778
Laufzeit	0.043652	0.006688	6.527	6.73E-11	0.0305	0.0568
Moral	-0.38836	0.07681	-5.056	4.28E-07	-0.5389	-0.2378
FamgesA1	0.128725	0.35004	0.368	0.713064	-0.5574	0.8148
FamgesA3	-0.459447	0.177831	-2.584	0.009777	-0.8080	-0.1109
FamgesA4	-0.232182	0.292432	-0.794	0.427213	-0.8053	0.3410
verwA1	-1.413109	0.332661	-4.248	2.16E-05	-2.0651	-0.7611
verwA2	-0.577689	0.236627	-2.441	0.014633	-1.0415	-0.1139
verwA3	-0.744118	0.213886	-3.479	0.000503	-1.1633	-0.3249
verwA4	0.324876	0.365904	0.888	0.374609	-0.3923	1.0420
verwA8	-0.654887	0.281984	-2.322	0.02021	-1.2076	-0.1022

De la misma manera la prueba de Hosmer-Lemeshow no arrojó evidencia de falta de ajuste (*HL=2.843, d.f=8, p=0.943*).

Después se sometió el modelo saturado (modelo 1) y el submodelo 1.2 a la selección automática de variables por AIC en R con la función "step()". En los cuadros 4.8, 4.9, 4.10 y 4.11 se muestran los datos obtenidos y los intervalos de confianza de los submodelos respectivamente.

Cuadro 4.8: Submodelo 1.1: ajuste de la regresión logística del modelo 1, en donde se utiliza el método de selección de variables por AIC.

Regresión Logística	Coefficients:	exp(coef)	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.3634	-	0.3054	4.464	8.03E-06	0.7648	1.9620
Laufkont	-0.60683	0.5451	0.06807	-8.914	¡2e-16	-0.7402	-0.4734
Laufzeit	0.04257	1.0435	0.00667	6.382	1.75E-10	0.0295	0.0556
Moral	-0.38715	0.6790	0.07678	-5.042	4.60E-07	-0.5376	-0.2367
FamgesA3	-0.44159	0.6430	0.16021	-2.756	0.005847	-0.7556	-0.1276
verwA1	-1.41935	0.2419	0.32379	-4.383	1.17E-05	-2.0540	-0.7847
verwA2	-0.57226	0.5642	0.22458	-2.548	0.01083	-1.0124	-0.1321
verwA3	-0.77717	0.4597	0.2012	-3.863	0.000112	-1.1715	-0.3828
verwA5	-1.79285	0.1665	1.11523	-1.608	0.107924	-3.9787	0.3930
verwA6	-0.5481	0.5780	0.28846	-1.9	0.057425	-1.1135	0.0173

[Variables no seleccionadas: famgesA1, famgesA4, verwA4, verwA7.]

Cuadro 4.9: Submodelo 1.1: ajuste de la regresión logística del modelo 1, en donde se utiliza el método de selección de variables por AIC. Se aplica la exponencial a los coeficientes de los parámetros y a los límites del intervalo de confianza.

Regresión Logística	exp(coef)	exp(lim inf)	exp(lim sup)
(Intercept)	-	-	-
Laufkont	0.5451	0.4770	0.6229
Laufzeit	1.0435	1.0299	1.0572
Moral	0.679	0.5841	0.7892
FamgesA3	0.643	0.4697	0.8802
verwA1	0.2419	0.1282	0.4563
verwA2	0.5642	0.3633	0.8763
verwA3	0.4597	0.3099	0.6819
verwA5	0.1665	0.0187	1.4814
verwA6	0.578	0.3284	1.0175

En el submodelo 1.1 no permanecieron categorías que no eran significativas para las variables *famges* y *verw*, algunas que no eran significativas en la variable *verw* si permanecieron. La ausencia de estas características en este modelo quiere decir que fueron colapsadas con las categorías de referencia. Además en la segunda columna se contruye la razón de momios para las variables. Para *laufkont* vemos que el incremento en una unidad de esta variable el riesgo de que ocurra el evento de que sea un cliente malo es del 54.5 %. Para la variable *laufzeit* por cada incremento en una unidad de la variable el riesgo de sea un cliente malo es del 104.35 %. Para *moral* por cada incremento en una unidad de la variable el riesgo de que sea un cliente malo es del 67.9 %. Para *famgesA3* podemos observar que la ocurrencia de que el cliente sea malo es del 64.3 % en la presencia de la característica con respecto a la categoría de referencia. Para la variable *verw* se puede apreciar que las razones de momios de las *dummies* presentes se encuentran por debajo de 1, esto indica que en la presencia de la variable la ocurrencia de que sea un cliente malo sucede mas en la categoría de referencia.

Cuadro 4.10: Submodelo 1.2.1: ajuste de la regresión logística del modelo 1.2, en donde se utiliza la función de selección de variables por AIC. *VerwA8=verwA5* $\cup$ *verwA6* $\cup$ *verwA7*.

Regresión Logística	Coefficients:	exp(coef)	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.363936	-	0.304917	4.473	7.71E-06	0.7663	1.9616
Laufkont	-0.608922	0.5439	0.067976	-8.958	¡2e-16	-0.7422	-0.4757
Laufzeit	0.044395	1.0454	0.006673	6.653	2.87E-11	0.0313	0.0575
Moral	-0.386319	0.6796	0.076595	-5.044	4.57E-07	-0.5364	-0.2362
FamgesA3	-0.434221	0.6478	0.160083	-2.712	0.00668	-0.7480	-0.1205
verwA1	-1.474718	0.2288	0.326154	-4.522	6.14E-06	-2.1140	-0.8355
verwA2	-0.611415	0.5426	0.226661	-2.697	0.00699	-1.0557	-0.1672
verwA3	-0.818132	0.4413	0.203341	-4.023	5.73E-05	-1.2167	-0.4196
verwA8	-0.714575	0.4894	0.272843	-2.619	0.00882	-1.2493	-0.1798

[Variables no seleccionadas: famgesA1, famgesA4, verwA4.]

Cuadro 4.11: Submodelo 1.2.1: ajuste de la regresión logística del modelo 1.2, en donde se utiliza la función de selección de variables por AIC. *VerwA8=verwA5* $\cup$ *verwA6* $\cup$ *verwA7*. Se aplica la exponencial a los coeficientes de los parámetros y a los límites del intervalo de confianza.

Regresión Logística	exp(coef)	exp(lim inf)	exp(lim sup)
(Intercept)	-	-	-
Laufkont	0.5439	2.1518	7.1107
Laufzeit	1.0454	0.4761	0.6214
Moral	0.6796	1.0318	1.0592
FamgesA3	0.6478	0.5848	0.7896
verwA1	0.2288	0.4733	0.8865
verwA2	0.5426	0.1208	0.4337
verwA3	0.4413	0.3479	0.8460
verwA8	0.4894	0.2962	0.6573

En el submodelo 1.2.1 de la misma manera fueron removidas algunas categorías que no eran significativas obteniendo un modelo más simple. Los coeficientes de algunas variables son parecidos con respecto a otros modelos y se observa que al exponenciar los coeficientes, no hay un cambio notable.

En la prueba de Hosmer-Lemeshow para el submodelo 1.1 y el submodelo 1.2.1, no existe evidencia que indique falta de ajuste (***HL1.1****=7.173, d.f=8, p=0.518* **- *HL1.2.1****=5.638, d.f=8, p=0.687*). Realizando la prueba del cociente de verosimilitudes de ambos modelos contra los submodelos 1 y el submodelo 1.2 respectivamente, no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, esto significa que no son necesarias las variables que no aparecen en los dos nuevos modelos ajustados (submodelo 1.1 *P-value=0.4719* y submodelo 1.2.1 *P-value=0.6231*).

Las variables que siempre entran en los modelos son *laufkont*, *laufzeit* y *moral*, *famgesA3*, *verwA1*, *verwA2* y *verwA3*. De la variable *famges* la variable derivada que permanece en estos modelos es *famgesA3* que son los hombres casados o viudos. Para la de *verw* son auto nuevo (*verwA1*), auto usado (*verwA2*) y bienes para hogar (*verwA3*).

En el paquete estadístico SPSS se introdujo el modelo saturado a un método de selección automática llamada *forward stepwise selection* para ver si existe un modelo mejor, ya que el software R no tiene ese tipo de opción. Esta selección de variables se basa en probar la significancia usando el cociente de verosimilitudes o la estadística de Wald al añadir otra variable al modelo, este modelo es nombrado como modelo 1.3. Al introducir este modelo en SPSS se especificó que las variables *laufkont* y *moral* son factores, eso implica que el software construye las variables *dummies* internamente. Además, se habilitó la opción de tomar como clase de referencia el valor más pequeño. Mientras que las otras variables se usaron directamente pues ya se tenían construído las variables dicotomizadas.

En el cuadro 4.12 se encuentran los resultados obtenidos del último paso de esta selección

automática.

Cuadro 4.12: Submodelo 1.3: ajuste de la regresión logística por selección automática de variables (*stepwise forward selection*) introduciendo el modelo 1.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
Laufkont			80.0167	3.04405E-17		
Laufkont(1)	0.52496	0.18874	7.73616	0.00541	0.1550	0.8949
Laufkont(2)	1.08784	0.34034	10.21609	0.00139	0.4208	1.7549
Laufkont(3)	1.82565	0.20852	76.64693	2.04423E-18	1.4169	2.2343
Laufzeit	-0.04048	0.00662	37.30532	1.01008E-09	-0.0535	-0.0275
Moral			24.89344	5.28535E-05		
Moral(1)	0.05823	0.48632	0.01434	0.90467	-0.8950	1.0114
Moral(2)	0.88316	0.38107	5.37111	0.02047	0.1363	1.6301
Moral(3)	0.92948	0.44306	4.40087	0.03592	0.0611	1.7979
Moral(4)	1.48147	0.40221	13.56698	0.00023	0.6931	2.2698
verwA1	1.09974	0.30898	12.66775	0.00037	0.4941	1.7054
verwA3	0.4591	0.17784	6.66555	0.00982	0.1106	0.8077
FamgesA3	0.42867	0.16029	7.15150	0.00749	0.1145	0.7429
Constant	-0.47114	0.4214	1.24943	0.26366	-1.2973	0.3550

[Variables no seleccionadas: famgesA1, famgesA4, verwA2, verwA4, verwA5, verwA6, verwA7.]

Se observa que en el modelo resultante permanecen variables de manera muy similar a los modelos anteriores, los coeficientes estimados son distintos en cuanto a presentación y tamaño en comparación a los otros modelos que fueron ajustados en R. La prueba de Hosmer-Lemeshow no mostró evidencia de falta de ajuste (*HL=6.916, d.f=8, p=0.545*). Permanecieron algunas variables que no son significativas para *verw* lo que significa que fueron colapsadas con la categoría de referencia.

Una observación muy importante es que en todos los submodelos anteriores la prueba de Hosmer-Lemeshow no arroja evidencia de falta de ajuste, la prueba del cociente de verosimilitudes se observa que no tiene caso meter las variables que fueron removidas en los submodelos.

Por otro lado, no es recomendable basarse solamente en las pruebas de hipótesis, otra forma de analizar el ajuste de los modelos es por medio de las tasas de clasificación errónea. En el cuadro 4.13 en donde se obtienen las tasas de clasificación durante el proceso de selección automática de SPSS. Se observa que el último modelo presenta la mayor tasa de clasificación global correcta y tiene la tasa más alta de clasificación correcta (R-R), este modelo será comparado con otros modelos.

Cuadro 4.13: Tasas de clasificación de la selección automática (*stepwise forward selection*) en SPSS

Step	A-A	A-R	R-A	R-R	tgcb
1	100.00	0.00	100.00	0.00	70.0
2	89.57	10.42	64.33	35.66	73.4
3	90.14	9.85	61.00	39.00	74.8
4	90.42	9.57	60.66	39.33	75.1
5	90.14	9.85	60.00	40.00	75.1
6	88.71	11.28	55.00	45.00	75.6

A continuación se aplica el análisis de discriminantes para los modelos 1 y 1.2 en SAS. Se había iniciado el análisis en R, se quería encontrar los coeficientes del discriminante cuadrático, pero lamentablemente no fue posible obtenerlos. En SPSS ocurrió lo mismo, pero en el paquete de SAS sí se pueden obtener los coeficientes del *score* cuadrático, pero nos da una función para cada población. En SAS para encontrar la regla de clasificación ella asigna a la población que tenga el máximo entre los *scores* (sección 2.3). En R sólo nos da una función discriminante. En el discriminante lineal para que podamos obtener en SAS la misma función que en R, basta restar los coeficientes de los términos. Por conveniencia se trabajó en SAS ya que se pueden obtener las pruebas de hipótesis de las matrices de varianzas-covarianzas con facilidad.

Durante el análisis de discriminante para cada modelo se consideran los siguientes puntos:

1. Se especifican las probabilidades aposteriori como proporcionales al número de observaciones en cada población.

2. Se utiliza la prueba modificada de Bartlett ofrecida en SAS para poder determinar si las matrices de varianzas-covarianzas son homogéneas para las poblaciones. Esto es para encontrar si existe la evidencia de poder discernir entre un análisis lineal o cuadrático. No obstante al hacer esta prueba se ajustaran los dos modelos para ver que tan acertado es el modelo en cuanto a tasas de clasificación. Se realizará esta prueba a un nivel de significancia del 5 %

3. Se estimarán los parámetros de las funciones de densidad condicionales $f(\boldsymbol{X}|i)$, *i=0,1.* Asumiendo la distribución normal multivariada.

4. Se evalúan las observaciones en las funciones de discriminante ajustadas, se aplica la regla de clasificación en donde se asigna el grupo al que pertenece la observación y al final se obtienen las tasas de clasificación errónea para poder compararlos con el modelo de regresión logística.

Para el modelo 1 al realizar la prueba modificada de Bartlett se encontró evidencia suficiente para poder rechazar la hipótesis nula $H_0 : \boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$ *vs* $H_a : \boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$, *(L' = 361.275, d.f. = 91, p= < ,0001).* Dado este resultado se puede decir que las matrices de varianzas-covarianzas no son iguales para los de tipos de cliente. Se ha encontrado una diferencia en al menos un elemento de estas matrices. Si se rechaza la hipótesis nula esto sugiere que el

análisis de discriminante lineal no puede ser apropiado para estos datos. Entonces lo que se desea es conducirlo para un análisis de discriminante cuadrático. De igual manera en el modelo 1.2 la prueba modificada de Bartlett se rechaza $H_0 : \mathbf{\Sigma}_0 = \mathbf{\Sigma}_1$ *vs* $H_a : \mathbf{\Sigma}_0 \neq \mathbf{\Sigma}_1$, *(L'= 144.587; d.f.= 55; p=0.0001)*. Entonces, se toma el análisis de discriminante cuadrático.

La ventaja que tiene SAS es que al elegir la opción al realizar esta prueba de hipótesis, automáticamente SAS decide que tipo de análisis de discriminante va a llevar a cabo, se basa en el resultado de la prueba. Si no se rechaza la hipótesis nula este lleva a cabo el análisis de discriminante lineal. Si se rechaza la hipótesis nula este lleva a cabo el análisis de discriminante cuadrático. El paquete SAS no imprime la función de discriminante cuadrática pero aún asi se pueden obtener sus coeficientes.

Lo que ocurrió fue que para nuestra sorpresa, las tasas de clasificación del discriminante cuadrático mejoraron con respecto a la clasificación correcta de los malos de todos lo modelos anteriores. Sin embargo, la tasa de clasificación correcta de los buenos se ve afectado por dicho análisis. El análisis de discriminante lineal fue ajustado con el fin de ver como diferia con el análisis cuadrático. Las tasas de clasificación muestran un cierto parecido con la regresión logística como se observa en el cuadro 4.14.

Cuadro 4.14: Tasas de clasificación de los modelos anteriormente ajustados de la base de crédito alemán.

Modelo	Software	Método	A-A	A-R	R-A	R-R	tgcb
1	R	LR	89.57	10.43	53.33	46.67	76.70
1	SAS/R	LDA	88.14	11.86	53.33	46.67	75.70
1	SAS/R	QDA	78.71	21.29	42.67	57.33	72.30
1.0	R	LR	88.86	11.14	56.33	43.67	75.30
1.1	R	LR	90.00	10.00	54.33	45.67	76.70
1.2	R	LR	89.86	10.14	53.33	46.67	76.90
1.2	SAS/R	LDA	88.29	11.71	53.33	46.67	75.80
1.2	SAS/R	QDA	84.14	15.86	49.33	50.67	74.10
1.2.1	R	LR	89.43	10.57	53.33	46.67	76.60
1.3	SPSS		88.71	11.28	55.00	45.00	75.60

En el cuadro anterior se observa que todos los modelos ajustados son pésimos para poder identificar que personas resultarán ser malos créditos. Las tasas observadas para la clasificación de rechazados a aceptados (R-A) es del 50 %. Se observa que el discriminante cuadrático del modelo saturado presenta una tasa errónea ligeramente más baja que el resto de las demás reglas. Sin embargo, se degrada la tasa de clasificación correcta de buenos. El que presenta la mayor tasa de clasificación global correcta es el modelo 1.2.

A continuación se prueban otros modelos incorporando interacciones entre variables, y ver si existe una mejora en las tasas de clasificación. Al modelo saturado, se tratará de mejorarlo anexando interacciones de variables derivadas del análisis de discriminante cuadrático y se

ajusta una regresión logística, se procura mantener la tasa correcta para buenos créditos.

Como ejemplo, se considera la intersección del conjunto de variables que siempre entraron en los modelos anteriores y se prueban las interacciones sin importar el tamaño de los coe-ficientes obtenidos del modelo cuadrático. Los coeficientes cruzados de las combinaciones de estas variables en la función de *score* cuadrático, se presentan en el cuadro 4.15. Estas variables son *laufkont*, *laufzeit*, *moral*, *famgesA3*, *verwA1* y *verwA3*.

Cuadro 4.15: Coeficientes de las combinaciones de las variables en la función de *score* cuadrático que siempre entraron en los modelos anteriores: *laufkont*, *laufzeit*, *moral*, *famgesA3*, *verwA1* y *verwA3*

Combinación	Población	
	Malos	Buenos
Laufkont * Laufzeit	-0.009	0.002
Laufkont * Moral	0.188	0.094
Laufkont * FamgesA3	-0.090	0.013
Laufkont * verwA1	-0.057	0.170
Laufkont * verwA3	0.091	0.198
Laufzeit * Moral	0.007	-0.007
Laufzeit * FamgesA3	0.019	0.022
Laufzeit * verwA1	0.069	0.069
Laufzeit * verwA3	0.023	0.016
Moral * FamgesA3	0.152	0.160
Moral * verwA1	-0.052	-0.075
Moral * verwA3	-0.184	-0.151
FamgesA3 * verwA1	-0.633	0.317
FamgesA3 * verwA3	-0.014	-0.305
verwA1 * verwA3	-3.552	-4.975

En el cuadro 4.16 se presenta el modelo de regresión logística en R con las interacciones, este modelo se llamará modelo 1.4. Y se considera este modelo para poder derivar otros submodelos.

Cuadro 4.16: Modelo 1.4: ajuste de la regresión logística al modelo saturado aumentando las combinaciones de variables que han entrado en los modelos anteriores. Se presenta un problema de ajuste en la interacción *I(verwA1*verwA3)* se debe a que cuando ambas son uno no hay observaciones entre esas variables. Las variables resaltadas en negritas son las variables significativas del modelo.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.682233	0.71049	2.368	0.017899	0.2897	3.0748
Laufkont	**-0.450959**	**0.227328**	**-1.984**	**0.047286**	**-0.8965**	**-0.0054**
Laufzeit	0.03053	0.022212	1.375	0.169277	-0.0130	0.0741
Moral	**-0.825373**	**0.240107**	**-3.438**	**0.000587**	**-1.2960**	**-0.3548**
FamgesA1	0.049909	0.357027	0.14	0.888826	-0.6499	0.7497
FamgesA3	-0.053118	0.583083	-0.091	0.927415	-1.1960	1.0897
FamgesA4	-0.250338	0.302542	-0.827	0.407982	-0.8433	0.3426
verwA1	-0.702856	1.269093	-0.554	0.579698	-3.1903	1.7846
verwA2	**-0.602149**	**0.237382**	**-2.537**	**0.011193**	**-1.0674**	**-0.1369**
verwA3	-0.13353	0.659197	-0.203	0.839476	-1.4256	1.1585
verwA4	0.244846	0.364134	0.672	0.501324	-0.4689	0.9585
verwA5	-1.811125	1.121724	-1.615	0.106399	-4.0097	0.3875
verwA6	-0.479224	0.298742	-1.604	0.108683	-1.0648	0.1063
verwA7	-0.786298	0.702821	-1.119	0.263236	-2.1638	0.5912
I(Laufkont * Laufzeit)	-0.009265	0.00602	-1.539	0.123829	-0.0211	0.0025
I(Laufkont * Moral)	0.046686	0.067651	0.69	0.490132	-0.0859	0.1793
I(Laufkont * FamgesA3)	0.015927	0.140064	0.114	0.909468	-0.2586	0.2905
I(Laufkont * verwA1)	-0.321847	0.309271	-1.041	0.298032	-0.9280	0.2843
I(Laufkont * verwA3)	-0.317414	0.166339	-1.908	0.05636	-0.6434	0.0086
I(Laufzeit * Moral)	**0.017379**	**0.006282**	**2.766**	**0.005669**	**0.0051**	**0.0297**
I(Laufzeit * FamgesA3)	-0.013485	0.013898	-0.97	0.33191	-0.0407	0.0138
I(Laufzeit * verwA1)	0.036289	0.033604	1.08	0.280186	-0.0296	0.1022
I(Laufzeit * verwA3)	0.001498	0.015134	0.099	0.921136	-0.0282	0.0312
I(Moral * FamgesA3)	-0.072819	0.158598	-0.459	0.646131	-0.3837	0.2380
I(Moral * verwA1)	-0.223774	0.324053	-0.691	0.489849	-0.8589	0.4114
I(Moral * verwA3)	-0.047333	0.205971	-0.23	0.818245	-0.4510	0.3564
I(FamgesA3 * verwA1)	-1.104151	0.76578	-1.442	0.14934	-2.6051	0.3968
I(FamgesA3 * verwA3)	0.2915	0.375617	0.776	0.437716	-0.4447	1.0277
I(verwA1 * verwA3)	NA	NA	NA	NA	NA	NA

En el modelo 1.4 las variables significativas son las siguientes *laufkont*, *moral*, *verwA2* y las interacciones *I(Laufzeit*Moral)* y casi significativa *I(Laufkont*verwA3*), en la interacción *I(verwA1*verwA3)* se hizo un *crosstabs* de las dos variables y se observa que cuando ambas toman el valor 1, no presenta observaciones, esto ocasiona problemas de ajuste. Por otro lado, al realizar la prueba de razón de verosimilitudes con la devianza nula y la devianza residual se obtiene $P(\chi^2_{(27)} > 257{,}05) = 2{,}26E-39$. También ajusta un modelo derivado de éste por selección del modelo por AIC con la función "step()" en R, se obtiene el modelo del cuadro 4.17.

Cuadro 4.17: Submodelo 1.4.1: ajuste de la regresión logística con la función step del modelo 1.4 con los datos de crédito alemán.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.337332	0.511368	2.615	0.00892	0.3351	2.3396
Laufkont	**-0.309673**	**0.142122**	**-2.179**	**0.02934**	**-0.5882**	**-0.0311**
Laufzeit	0.033223	0.019959	1.665	0.09601	-0.0059	0.0723
Moral	**-0.776397**	**0.162129**	**-4.789**	**1.68E-06**	**-1.0942**	**-0.4586**
verwA2	**-0.55443**	**0.214849**	**-2.581**	**0.00986**	**-0.9755**	**-0.1333**
verwA5	-1.867473	1.11	-1.682	0.09249	-4.0431	0.3081
verwA6	-0.500254	0.280249	-1.785	0.07426	-1.0495	0.0490
I(Laufkont * Laufzeit)	-0.00898	0.005834	-1.539	0.12374	-0.0204	0.0025
I(Laufkont * verwA1)	**-0.413127**	**0.177714**	**-2.325**	**0.02009**	**-0.7614**	**-0.0648**
I(Laufkont * verwA3)	**-0.350255**	**0.081399**	**-4.303**	**1.69E-05**	**-0.5098**	**-0.1907**
I(Laufzeit * Moral)	**0.017189**	**0.006072**	**2.831**	**0.00464**	**0.0053**	**0.0291**
I(Laufzeit * FamgesA3)	**-0.015311**	**0.006808**	**-2.249**	**0.02453**	**-0.0287**	**-0.0020**
I(FamgesA3 * verwA1)	**-1.058384**	**0.537276**	**-1.97**	**0.04885**	**-2.1114**	**-0.0053**

Se realizó la prueba de cociente de verosimilitudes de este modelo contra el modelo 1.4 y no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, esto significa que el modelo reducido es tan bueno como el modelo saturado. (*Likelihood ratio test for MLE method, Chi-squared 15 d.f. = 6.542364 , P value = 0.9691*).

En el submodelo 1.4.1 las variables *laufzeit*, *verwA5* y la interacción *I(Laufkont*Laufzeit)* no son significativas. Se ajusta el modelo de regresión logística sin estas variables, este modelo se llamará submodelo 1.4.2, representado en el cuadro 4.18.

Cuadro 4.18: Submodelo 1.4.2: ajuste de la regresión logística removiendo las variables *laufzeit*, *verwA5* y la interacción *I(Laufkont * Laufzeit)* del submodelo 1.4.1.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	1.939557	0.257818	7.523	5.35E-14	1.4342	2.4449
Laufkont	-0.533347	0.071799	-7.428	1.10E-13	-0.6741	-0.3926
Moral	-0.844961	0.106237	-7.954	1.81E-15	-1.0532	-0.6367
verwA2	-0.411291	0.205839	-1.998	0.045704	-0.8147	-0.0078
I(Laufkont * verwA1)	-0.379383	0.174918	-2.169	0.030088	-0.7222	-0.0365
I(Laufkont * verwA3)	-0.298837	0.077019	-3.88	1.04E-04	-0.4498	-0.1479
I(Laufzeit * Moral)	0.020674	0.003094	6.683	2.34E-11	0.0146	0.0267
I(Laufzeit * FamgesA3)	-0.013368	0.006328	-2.113	0.034638	-0.0258	-0.0010
I(FamgesA3 * verwA1)	-0.918844	0.527653	-1.741	0.081617	-1.9530	0.1154

Se realiza la prueba del cociente de verosimilitudes contra el modelo 1.4, y no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, lo cual este modelo es tan bueno como el modelo 1.4 (*Likelihood ratio test for MLE method, Chi-squared 19 d.f. = 16.61787 , P value = 0.6157*). En el cuadro 4.19 se muestran las tasas de clasificación de

los modelos derivados del 1.4 en comparación con el modelo saturado.

Cuadro 4.19: Tasas de clasificación de los modelos ajustados introduciendo interacciones de variables *vs* el modelo saturado.

Modelo	Software	Metodo	A-A	A-R	R-A	R-R	tgcb	No.var. tec. distintas
1	R	LR	89.57	10.43	53.33	46.67	76.70	13
1	SAS/R	LDA	88.14	11.86	53.33	46.67	75.70	13
1	SAS/R	QDA	78.71	21.29	42.67	57.33	72.30	13
1.4	R	LR	89.14	10.86	49.67	50.33	77.50	13
1.4.1	R	LR	89.00	11.00	51.33	48.67	76.90	9
1.4.2	R	LR	88.43	11.57	51.33	48.67	76.50	7

El modelo 1 en regresión logística conserva la tasa más alta de clasificación correcta de los aceptados. Y el que tiene mejores tasas de clasificación es el modelo 1.4 ya que posee la mayor tasa de clasificación global. Sin embargo, este modelo contiene muchas variables e interacciones no significativas y el submodelo 1.4.1 tiene menos variables y las tasas son muy similares al modelo 4.

El submodelo 1.4.1 es mejor que el submodelo 1.4.2 ya que su tasa de clasificación correcta en la población de los buenos es ligeramente mayor y conserva los porcentajes de los malos. En comparación con respecto al modelo saturado es mejor pues se mejoró las tasas de clasificación y contiene menos variables técnicas distintas al modelo saturado, como puede verse en la última columna del cuadro 4.19. Siguiendo con la comparación se aprecia que los modelos tienen las siguientes variables en común: *laufkont*, *laufzeit*, *moral*, *famgesA3*, *verwA1*, *verwA2*, *verwA3*, *verwA5*, *verwA6*.

Lo que podría diferenciar el submodelo 1.4.1 con respecto al modelo saturado es una mejor interpretación del tipo de cliente deseado por la entidad financiera, las interacciones pueden ayudar a entender el tipo de cliente ideal. Se observa en este modelo que las variables que se repiten son *verwA1*, *famgesA3*, *laufkont*, *laufzeit*, y *moral*. Pero el cliente óptimo para otorgarle un crédito debería tener las características *famgesA3* y *verwA1* ya que aparecen repetidas en el modelo y su precencia ayuda a disminuir la probabilidad de ser un cliente riesgoso. Además, la interacción *famgesA3-verwA1* presenta en valor absoluto grande, lo cual en términos de la entidad financiera, estas características hacen del cliente tipo de cliente es más aceptable.

Analizando las interacciones del modelo, cuando son combinaciones de variables continua-binaria o binaria-binaria su interpretación es sencilla, como es el caso de la interacción *famgesA3-verwA1*, se puede decir que en la presencia de ambas disminuye fuertemente la probabilidad de caer en la categoría de cliente riesgoso. En el caso de continua-binaria, por ejemplo, *laufzeit-verwA3*, la interpretación es que disminuye la probabilidad con la presencia

de *verwA3* a medida que *laufzeit* aumenta. Ésta es una observación interesante, pues la variable *laufzeit* por sí sola aumenta la probabilidad de incumplimiento. Sin embargo, con la presencia de esta variable en conjunto actúa como un factor que suaviza el riesgo. En la práctica los préstamos a largo plazo no suelen ser menos riesgosos, y muchas entidades financieras hacen uso de medidas que pueden ayudar a disminuir o suavizar el riesgo, por ejemplo: el aumento de la tasa de interés y también el uso de pagos mínimos como en las tarjetas de crédito.

Cuando se tienen variables que son continuas con categóricas ordinales, su interpretación resulta más difícil, se debe a que sus categorías pueden estar tomando el mismo rango de valores, así contribuyendo la misma proporción para el riesgo de incumplimiento. En las figuras 4.2 y 4.3 se encuentran las gráficas de dispersión de las variables *laufzeit* con *laufkont* y con *moral* respectivamente. En ambas figuras se observa que el rango de valores de las categorías de las variables en la variable continua se encuentran muy similares entre ellos. Aunque en la categoría 3 de *laufkont* no abarca todo el rango de *laufzeit*, quizas habría un chance de colapsarla con alguna categoría. A partir del submodelo 1.4.1 se ajusta un nuevo modelo de regresión logística en donde se dicotomizan las variables *laufkont* y *moral* tomando la primera categoría de cada variable como la de referencia. En el cuadro 4.20 se muestran los promedios que toma *laufzeit* en las categorías que toman ambas variables.

La recodificación de las variables *laufkont* y *moral* es como sigue:

- *Laufzeit* (Duración del crédito en meses).- La categoría de referencia A1=no posee cuenta, A2=sin balance o débito, A3= $0 \leq ... < 200DM$, A4=$\leq 200DM$ o con cuenta de cheques por al menos 1 año.
- *Moral* (Pago de antiguos créditos).- La categoría de referencia A0=pago vacilante de créditos anteriores, A1=línea de crédito problemática/ tiene créditos en otros bancos, A2=no hay créditos previos / pagó todos sus créditos anteriores, A3=sin problemas con créditos actuales del banco, A4=pagó créditos anteriores del banco.

Cuadro 4.20: Promedio que toman la variable *laufzeit* en las categorías de *laufkont* y *moral*.

Categoría	Laufkont	Moral
A0	-	27.88
A1	21.34	22.69
A2	22.68	20.11
A3	17.35	26.22
A4	19.95	19.49

En los promedios de la duración de crédito en ambas variables son muy similares lo cual dificulta observar un patrón notable. En los modelos siguientes, se evalua el comportamiento de acuerdo a la significancia de las categorías de las variables *laufkont* y *moral* al dicotomizarlas y en su caso si existe la posibilidad de colapsar categorías no significantes. Por el momento

se suspende la eliminación de variables no significativas, será retomada cuando exista la posibilidad de afinar un modelo.

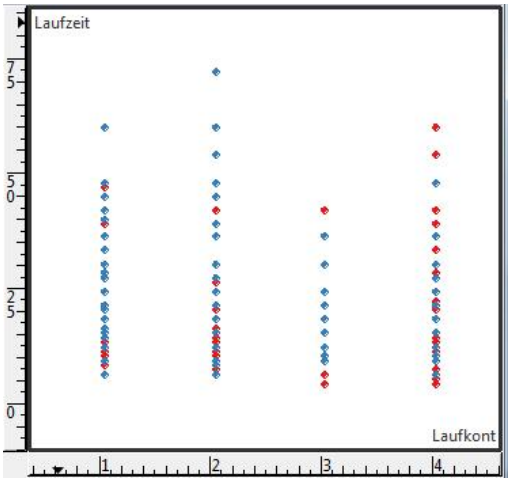

Figura 4.2: Gráfica de dispersión de las variables *laufkont* con *laufzeit*. Se observa que el rango de valores que toma cada categoría de la variable *laufkont* son similares. El color rojo de la gráfica representa los créditos malos y los de color azul los créditos buenos.

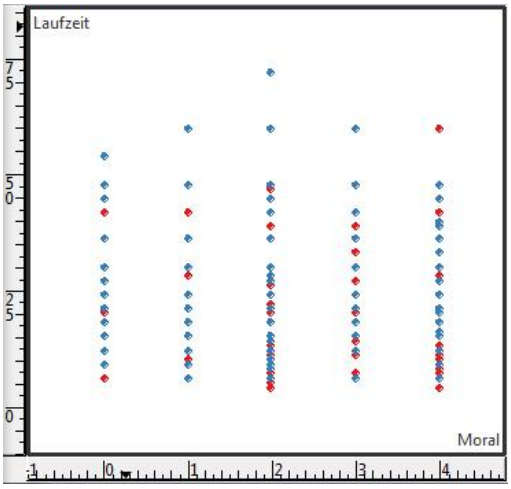

Figura 4.3: Gráfica de dispersión de las variables *laufkont* con *moral*. Se observa que el rango de valores que toma cada categoría de la variable *laufkont* son muy similares. El color rojo de la gráfica representa los créditos malos y los de color azul los créditos buenos.

Las variables *dummies* significativas en el modelo al dicotomizar ambas *laufkont* y *moral* son *laufkontA3*, *laufkontA1*, *moralA2* y *moralA4*. Al dicotomizar ambas, el rendimiento de las tasas de clasificación para los malos obtenida es degradada a 47 %. Se colapsaron las categorías que tuvieran una relación coherente en la variable *laufkont* colapsando la A1 con la de

referencia, luego colapsando la A2 y A3 y después ambas, pero se observó que colapsar A2 y A3 de *laufkont* tiene una mejora respecto al modelo que es sólo dicotomizar sin colapsar las variables pero se encuentra por debajo del submodelo 1.4.1.

Luego se deja *laufkont* como estaba originalmente y se dicotomiza *moral*, se observó un cambio notable en las tasas obteniendo una tasa de clasificación para los malos de un 49.3 %. Igualmente la significancia de la *moral* A2 y A4 se preservaron, se intentó colapsar categorías, y no se observó resultados favorables hasta que se colapsó la categoría A2 y A3 la tasa de los malos se degradó al 49 %, pero sigue siendo un mejor modelo que el 1.4.1. Se intentó colapsar la *moral* y *laufkont* al dicotomizar ambas variables pero las tasas de clasificación correctas bajan. En el cuadro 4.21 se muestra el resumen del procedimiento de los modelos ajustados y su respectiva descripción.

Cuadro 4.21: Tasas de clasificación de los modelos al dicotomizar y colapsar categorías a partir del submodelo 1.4.1

Modelo	Descripción	A-A	A-R	R-A	R-R	tgcb	No.var. tec. distintas
1.4.1	-	89.00	11.00	51.33	48.67	76.90	9
A	Dicotomizando ambas variables	89.86	10.14	53.00	47.00	77.00	14
B	Dicotomizando ambas, colapsando laufkont(A0-A1)	89.86	10.14	56.00	44.00	76.10	13
C	Dicotomizando ambas, colapsando laufkont(A2-A3)	89.71	10.29	52.67	47.33	77.00	13
D	Dicotomizando ambas, colapsando laufkont(A0-A1,A2-A3)	90.00	10.00	56.33	43.67	76.10	12
E	Dicotomizando sólo moral	89.43	10.57	50.67	49.33	77.40	12
F	Dicotomizando sólo moral, colapsando A2-A3	89.14	10.86	51.00	49.00	77.10	11
G	Dicotomizando sólo moral, colapsando A3-A4	89.86	10.14	53.33	46.67	76.90	11
H	Dicotomizando sólo moral, colapsando A0-A1 y A3-A4	89.86	10.14	53.33	46.67	76.90	10
I	Dicotomizando sólo moral, colapsando A0-A1 y A2-A3	88.71	11.29	51.67	48.33	76.60	10
J	Dicotomizando ambas, colapsando moral(A2-A3)	89.43	10.57	53.00	47.00	76.70	13
K	Dicotomizando ambas, colapsando laufkont(A2-A3) y moral(A2-A3)	89.71	10.29	53.67	46.33	76.70	12

Hasta ahora se llega a un punto en donde se puede mejorar el submodelo 1.4.1 (submodelo E), la prueba de Hosmer-Lemeshow no arroja evidencia de falta de ajuste ($HL = 2{,}404; d.f = 8; 0{,}966$). Sin embargo este modelo aún resulta difícil interpretar pues aún se presenta una interacción de variables continua-categórica. Se hizo otro análisis de este modelo. Se eliminó la interacción *I(Laufkont*Laufzeit)*, se eliminan variables y en el caso de categóricas se mandan a la clase de referencia para ver si se obtiene un mejor modelo. También se intentó colapsar

categorías no significativas de la variable *moral*. Finalmente el resultado de dicho análisis produjo los modelos que se muestran en el cuadro 4.22.

Cuadro 4.22: Tasas de clasificación de los modelos al dicotomizar sólo *moral* y eliminación de variables al colapsar categorías no significativas moralA1 y moralA3

Modelo	Descripción	A-A	A-R	R-A	R-R	tgcb	No.var. tec. distintas
1.4.1	Modelo sin dicotomizar	89.00	11.00	51.33	48.67	76.90	9
E	Dicotomizando sólo moral	89.43	10.57	50.67	49.33	77.40	12
E.1	Modelo E sin interacción	89.14	10.86	51.33	48.67	77.00	12
E.2	Modelo quitando laufzeit	89.00	11.00	51.33	48.67	76.90	12
E.3	Modelo E.2 sin I(FamgesA3*verwA1)	88.71	11.29	51.00	49.00	76.80	12
E.4	Modelo E.3 sin moralA1	89.14	10.86	51.67	48.33	76.90	12

El submodelo E.4 es un submodelo del E.1, por la prueba de coeficiente de verosimilitudes, no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, esto significa que no se rechaza que el el coeficiente del conjunto de variables es cero. (*Likelihood ratio test for MLE method Chi-squared 3 d.f. = 3.567857 , P value = 0.3121*). Con la prueba de Hosmer-Lemeshow y se obtiene un valor de p grande ($p - value = 0{,}768$), no hay evidencia suficiente que el modelo presenta carencia de ajuste. El submodelo E.4 se muestra en el cuadro 4.23.

Cuadro 4.23: Submodelo E.4: ajuste de la regresión logística, modelo afinado de E.2, se colapsaron categorías no significativas de la variable *moral* (moralA1 y moralA3).

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	2.42651	0.37429	6.48300	0.00000	1.69290	3.16013
Laufkont	**-0.48989**	**0.07266**	**-6.74300**	**0.00000**	**-0.63229**	**-0.34748**
moralA2	**-2.53443**	**0.40656**	**-6.23400**	**0.00000**	**-3.33128**	**-1.73757**
moralA3	**-1.99287**	**0.68313**	**-2.91700**	**0.00353**	**-3.33180**	**-0.65395**
moralA4	**-3.73628**	**0.50322**	**-7.42500**	**0.00000**	**-4.72260**	**-2.74996**
verwA2	**-0.49488**	**0.21580**	**-2.29300**	**0.02184**	**-0.91785**	**-0.07190**
verwA5	**-2.27974**	**1.16252**	**-1.96100**	**0.04988**	**-4.55828**	**-0.00120**
verwA6	-0.48765	0.27915	-1.74700	0.08065	-1.03478	0.05948
I(Laufkont * verwA1)	**-0.63849**	**0.16079**	**-3.97100**	**0.00007**	**-0.95365**	**-0.32334**
I(Laufkont * verwA3)	**-0.34200**	**0.08183**	**-4.18000**	**0.00003**	**-0.50239**	**-0.18162**
I(Laufzeit * moralA1)	-0.01624	0.01546	-1.05100	0.29327	-0.04654	0.01405
I(Laufzeit * moralA2)	**0.05645**	**0.01013**	**5.57100**	**0.00000**	**0.03659**	**0.07631**
I(Laufzeit * moralA3)	0.03536	0.02054	1.72200	0.08508	-0.00489	0.07562
I(Laufzeit * moralA4)	**0.08138**	**0.01536**	**5.29700**	**0.00000**	**0.05127**	**0.11149**
I(Laufzeit * FamgesA3)	**-0.01824**	**0.00640**	**-2.84900**	**0.00438**	**-0.03079**	**-0.00569**

Al quitar *verwA6* y mandarla a la categoría de referencia a la variable con interacciones, las tasas de clasificación correcta se degradan. El submodelo E.4 se puede obtener una mejor

información acerca del cliente. En el caso de la interacción *laufzeit-moral* si tengo *moral* baja "*moralA1*", al otorgar un plazo mayor para pagar dicho crédito las oportunidades para pagarlo aumentan. En la práctica esto implica que el monto de los pagos del crédito disminuyen, pero el pago de intereses también aumenta. En general debiera de suceder esto siendo un buen pagador, pero el modelo muestra que no siempre se da la relación, es decir, que aunque seas un buen pagador el plazo del crédito puede ser factor de no pagarlo, porque podrían cambiar las condiciones económicas a un corto plazo o tal vez del propio solicitante como quedarse sin trabajo, enfermarse, cambiar de residencia, etc.

Por otro lado, en la práctica debe de sopesarse entre la interpretabilidad y el ajuste de los modelos, a modelos mayormente sofisticados puede perderse interpretabilidad y modelos sencillos pueden perder ajuste. Acontece muy a menudo cuando se ajusta un modelo de regresión y a veces se introducen interacciones entre dos o más variables y cuando se incluyen modelos de optimización como las redes neuronales. También la construcción de la razón me momios en presencia de interacción no es de manera directa y no se profundizará en ese tema[4]

No obstante con el análisis anterior el submodelo E.1 explica información interesante. La decisión de otorgar un crédito no sería suficiente al considerar sólo el puntaje, pues se vio que estos modelos no pueden predecir de manera satisfactoria a los clientes malos, pues las tasas de clasificación se encuentran alrededor del 50 %. Para poder complementar el análisis del cliente debe de apegarse al resultado del estudio de crédito, mismo que considera la solvencia *moral*, su situación financiera, su capacidad de pago y aspectos cualitativos del solicitante, así como la entrega de la información que en cada caso aplique. Se puede establecer el cliente ideal pero resulta difícil expresar un cliente malo. El modelo 1 y el submodelo E.4 seguirán analizándose en el capítulo 5.

4.4. Base de transplante de Médula Ósea

Debido a que la base de datos presenta información pre-operativa y post-operativa del paciente. Es posible con la información post-operativa dividir a los pacientes en varias categorías o clases, las medidas pre-operativas pueden usarse para predecir la clase post-operativa del paciente, como lo mencionado en la sección 1.7 de esta tesis. Sin embargo, pueden modelarse distintos tipos de escenarios, esto va a depender del propósito del estudio. Por ejemplo, determinar la clase futura antes de que se realice el transplante, o bien, que factores antes y después del transplante pueden relacionarse con el evento de interés, por ejemplo muerte. Por otro lado, C1 y C2 definen dos medidas post-operativas las cuales explica la muerte del paciente y recaída en la leucemia. Estas dos medidas pueden ser usadas para definir el evento como en la variable C3 del cuadro 4.3, pues en términos de análisis de supervivencia puede decirse que el transplante falló si alguna de estas dos ocurre. Sin embargo la recaída del paciente en la leucemia puede ser un factor determinante en la calidad de vida del paciente.

[4]Para profundizar un poco sobre la razón de momios en presencia de interacción puede revisarse la sección 3.7 de Hosmer and Lemeshow (2000)

Lo cual hace pensar que C1 (muerte) depende de C2 (recaída). Por consiguiente, se toma en cuenta todas las variables antes del transplante y después del transplante, excluyendo las variables de tiempo de A, C y P.

Se desarrollan tres casos, el primero es que C1 depende de C2 y el segundo es que C1 no depende de C2 (se excluye). Se tratará de implementar interacciónes en los dos casos. Se considera como tercer caso un ejemplo implementado en Klein and Moeschberger (2003).

Para el tercer caso mencionado arriba, en el ejemplo 8.5 del capítulo 8, donde tomaron las variables siguientes: Z1, Z2, I(Z1*Z2), Z3, Z4, I(Z3*Z4), Z5, Z6, I(Z5*Z6), Z7, Z8 y Z10. Construyeron un modelo con el AIC y encontraron que las variables Z8(FAB) y la edad Z1, Z2 y I(Z1*Z2) debieran considerarse. En el capítulo 9 ejemplo 9.1, ajustan una regresión de Cox con estas variables explicativas incluyendo gA2 y gA3 con la variable de interés C3 (libre de enfermedad/ sobrevivencia). Este modelo en el ejemplo 11.1 de ese libro ven como es el ajuste del modelo con los residuales de Cox-Snell y encuentran que el modelo no ajusta mal. Para efectos prácticos, se incluirán las mismas variables: gA2, gA3, Z1, Z2, I(Z1*Z2), Z7, Z8 y Z10. Considerando el evento de muerte como el de interés (variable C1). Se compara el ajuste de la regresión de Cox y la regresión logística.

Por otro lado, la variable Z9 no debiera de ser considerada ya que la variable indica en que hospital fue realizado el estudio así, como el transplante. Los hospitales se encuentran distribuidos en varias regiones del globo, lo cual no tendría mucho sentido incluirla en el análisis. Sin embargo, si los hospitales estuvieran cerca entre ellos en diversos puntos de una ciudad puede ser crucial el estudio. Y si el tipo de enfermedad dependiera de la localización del individuo podría ser clave para un estudio epidemiológico. A veces el hospital estaría resaltando un rasgo en particular de la supervivencia del paciente, esto es que su supervivencia se deba a la calidad del hospital encuanto a especialistas en el área, avances tecnológicos, etc. Para efectos prácticos Z9 será incluida en los dos primeros casos para ilustrar los modelos.

4.4.1. Aplicación del Modelo de Regresión Logística

Primer Caso: C1 depende de C2

Se prosigue con el primer caso incluyendo la variable C2, en el cuadro 4.24 se obtiene el resumen del modelo saturado junto con los intervalos de confianza a un nivel de significancia del 5 %. Se toma la recodificación de las variables en la sección 4.2.2. Las variables que fueron significativas son C2, P, Z8, Z9A2 y Z9A3. Hay coeficientes grandes y aportan bastante para el riesgo de muerte. Hay otras variables que no son significativas pero tienen coeficientes grandes (gA2 y C).

Cuadro 4.24: Modelo saturado bajo la regresión logística de la base de datos de 137 transplantes de médula ósea. Se observa un problema de ajuste de las variables. Las variables resaltadas en negritas son las variables significativas.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.74780	1.64785	2.27400	0.02294	0.51800	6.97759
gA2	-1.29921	0.84355	-1.54000	0.12352	-2.95256	0.35415
gA3	-0.18496	0.81908	-0.22600	0.82135	-1.79036	1.42044
C2	**4.03709**	**0.95079**	**4.24600**	**0.00002**	**2.17354**	**5.90064**
A	0.62498	0.72933	0.85700	0.39149	-0.80451	2.05447
C	-0.75588	0.59567	-1.26900	0.20446	-1.92338	0.41163
P	**-3.73868**	**1.38291**	**-2.70300**	**0.00686**	**-6.44919**	**-1.02818**
Z1	0.05640	0.04889	1.15400	0.24867	-0.03942	0.15222
Z2	-0.00666	0.04098	-0.16300	0.87084	-0.08698	0.07365
Z3	-0.29089	0.56350	-0.51600	0.60571	-1.39536	0.81358
Z4	-0.23867	0.54880	-0.43500	0.66364	-1.31431	0.83697
Z5	-0.47834	0.61354	-0.78000	0.43560	-1.68088	0.72420
Z6	-0.57156	0.56951	-1.00400	0.31557	-1.68779	0.54468
Z7	-0.00014	0.00065	-0.21400	0.83071	-0.00141	0.00114
Z8	**1.39680**	**0.68178**	**2.04900**	**0.04049**	**0.06051**	**2.73309**
Z9A2	-0.66579	0.92950	-0.71600	0.47382	-2.48761	1.15604
Z9A3	**-1.98971**	**0.86179**	**-2.30900**	**0.02095**	**-3.67881**	**-0.30061**
Z9A4	**-3.17615**	**1.01818**	**-3.11900**	**0.00181**	**-5.17179**	**-1.18051**
Z10	NA	NA	NA	NA	NA	NA

El problema de ajuste del modelo saturado se debe al diseño de las variables dicotómicas, se encontró que $Z9A1 + Z9A2 = Z10$ para todas las observaciones. Se intentó cambiar la categoría de referencia en la variable Z9 y ocurrió lo mismo. Entonces fue considerado el modelo univariado incluyendo la variable Z10, para determinar su significancia estadística. El valor de p fue igual a 0.606, no se rechaza la hipótesis $H_0 : \beta = 0$ *vs* $H_a : \beta \neq 0$, por lo tanto esta variable no es significativa.

A partir de este modelo se realizó una afinación del mismo quitando una por una las variables que tenían un valor de p grande. El orden en el cual las variables fueron removidas fue el siguiente: Z10, Z2, Z7, Z4, gA3, Z3, Z9A2, Z5, A, Z6, Z1, C, gA2. En los cuadros 4.25 y 4.26 se presentan dos modelos, el primero es un submodelo intermedio del modelo saturado y el segundo es un submodelo afinado del submodelo intermedio. Las tasas de clasificación de estos modelos se encuentran en el cuadro 4.27.

Cuadro 4.25: Submodelo intermedio: eliminación manual de variables del modelo saturado usando la regresión logística, de los 137 transplantes de médula ósea.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.5722	1.1447	3.1200	0.0018	1.3286	5.8158
gA2	-1.0371	0.5406	-1.9190	0.0550	-2.0967	0.0225
C	-0.7572	0.5224	-1.4500	0.1472	-1.7811	0.2667
C2	3.6304	0.8652	4.1960	0.0000	1.9346	5.3262
P	-3.1916	1.1306	-2.8230	0.0048	-5.4076	-0.9756
Z8	1.2616	0.5819	2.1680	0.0302	0.1211	2.4021
Z9A3	-1.4243	0.7218	-1.9730	0.0485	-2.8390	-0.0096
Z9A4	-2.4471	0.9111	-2.6860	0.0072	-4.2329	-0.6613

Cuadro 4.26: Submodelo afinado: Se removieron las variables no significativas del modelo intermedio.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.0013	1.1430	2.6260	.0086	0.7610	5.2416
C2	3.8221	.8463	4.5160	.0000	2.1634	5.4808
P	-3.4398	1.1617	-2.9610	.0031	-5.7167	-1.1629
Z8	1.1123	.5546	2.0060	.0449	0.0253	2.1993
Z9A3	-1.1703	.6729	-1.7390	.0820	-2.4892	0.1486
Z9A4	-2.5248	.8395	-3.0080	.0026	-4.1702	-0.8794

El modelo intermedio presenta tasas muy similares de clasificación errónea y presenta mejores tasas de clasificación para la clase de muertos, pero la tasa de clasificación global es más pequeña en comparación con el submodelo afinado. Se realizó la prueba del cociente de verosimilitudes para comparar el submodelo con el modelo saturado y no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_p = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., p$, lo cual este modelo es tan bueno como el saturado (*Likelihood ratio test for MLE method Chi-squared 12 d.f. = 11.28997 , P value = 0.5042*). Se compararon también los modelos intermedio y saturado de la misma manera no se encontró evidencia suficiente para rechazar la hipótesis nula (*Likelihood ratio test for MLE method Chi-squared 1 d.f. = 2.656059 , P value = 0.1032*).

Cuadro 4.27: Tasas de clasificación de los modelos de la base de 137 transplantes

Modelo	Software	Metodo	M-M	M-V	V-M	V-V	tgcb
satudaro	R	LR	86.42	13.58	23.21	76.79	82.48
intermedio	R	LR	80.25	19.75	17.86	82.14	81.02
afinado	R	LR	76.54	23.46	10.71	89.29	81.75

Así como en la base de crédito se recurrió al método de selección automática de variables para ver si es posible encontrar un modelo mejor. Usando SPSS se ajustó la selección

automática de variables *stepwise forward selection*. Se crearon previamente las variables *dummies* y luego se introdujeron en el modelo, dicho modelo se encuentra en el cuadro 4.28, el modelo se llamará submodelo 1.

Cuadro 4.28: Submodelo 1: regresión logística del modelo saturado bajo la selección automática de la base de 137 transplantes.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
C2	3.633	.830	19.141	0.000	2.006	5.261
P	-3.443	1.143	9.067	0.003	-5.684	-1.202
Z8	1.093	.540	4.099	0.043	.035	2.150
Z9A4	-2.249	.802	7.861	0.005	-3.821	-.677
Constant	2.802	1.117	6.288	0.012	.612	4.992

Se observa que en el modelo resultante permanecen las mismas variables de manera muy similar a los modelos anteriores, los coeficientes estimados son similares en comparación al modelo afinado ajustado en R. En SPSS se ajustó la prueba de Hosmer-Lemeshow, y no muestra evidencia que el modelo no ajusta mal, (*HL=2.605, d.f=4, p=.626*). La prueba de Hosmer-Lemeshow da una pauta para decir si el modelo tiene falta de ajuste. La manera de comprobarlo es por medio de las tasas de clasificación errónea.

En el cuadro 4.29 se presentan las tasas de clasificación del proceso de selección automática de SPSS. Se concluye que hay dos modelos uno que tiene la tasa de clasificación global más alta último modelo, pero el modelo en el paso dos y paso tres tienen las mejores tasas de clasificación correcta para los vivos. Se toma el último modelo ya que tiene la mayor tasa de clasificación global.

Cuadro 4.29: Tasas de clasificación de la selección automática en SPSS de la base de los 137 transplantes. Se incluye la recaída (C2) en el modelo.

Step	M-M	M-V	V-M	V-V	tgcb
1	49.38	50.62	3.57	96.43	68.61
2	65.43	34.57	5.36	94.64	77.37
3	65.43	34.57	5.36	94.64	77.37
4	77.78	22.22	14.29	85.71	81.02

Por otro lado, muchas veces la agresividad del cáncer puede ser un factor importante para la muerte del paciente. Otro ejemplo, sería la edad del paciente influye en la rapidez de curación. Por eso, se efectúa un segundo análisis para ver si existen interacciones admisibles entre algunas variables. Se incluyen interacciones de las variables de edad del paciente, el recuperamiento, el tipo de enfermedad, si desarrolló un ataque inmunológico fuerte y su clasificación FAB. En el cuadro 4.30 se presentan las variables del modelo en R lo cual se

añaden las interacciones de variables al modelo intermedio, a este modelo se llamará modelo 2.

Cuadro 4.30: Modelo 2: regresión logística del submodelo intermedio con interacciones de la base de transplantes. Las variables resaltadas en negritas son las variables significativas del modelo

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.5930	1.3220	2.7170	0.0066	1.0019	6.1841
gA2	-0.7526	0.8214	-0.9160	0.3595	-2.3625	0.8573
C2	16.6400	6274.0000	0.0030	0.9979	-12280.4000	12313.6800
C	-0.6493	0.5602	-1.1590	0.2464	-1.7473	0.4487
P	**-4.7120**	**1.6710**	**-2.8190**	**0.0048**	**-7.9872**	**-1.4368**
Z8	**3.2570**	**1.4500**	**2.2460**	**0.0247**	**0.4150**	**6.0990**
Z9A3	-1.0760	0.8309	-1.2950	0.1952	-2.7046	0.5526
Z9A4	**-3.2390**	**1.0260**	**-3.1560**	**0.0016**	**-5.2500**	**-1.2280**
I(gA2 * Z8)	-1.7020	1.5020	-1.1330	0.2572	-4.6459	1.2419
I(gA3 * Z8)	NA	NA	NA	NA	NA	NA
I(P * Z1)	0.0410	0.0383	1.0700	0.2847	-0.0341	0.1161
I(A * C2)	12.6300	3783.0000	0.0030	0.9973	-7402.0500	7427.3100
I(A * P)	1.0760	0.7388	1.4570	0.1452	-0.3720	2.5240
I(C2 * gA2)	-16.2200	2873.0000	-0.0060	0.9955	-5647.3000	5614.8600
I(C2 * gA3)	-0.8186	3596.0000	-0.0002	0.9998	-7048.9786	7047.3414
I(C2 * P)	3.1650	5578.0000	0.0010	0.9996	-10929.7150	10936.0450
I(Z1 * gA3)	-0.0045	0.0273	-0.1650	0.8692	-0.0580	0.0490
I(Z8 * Z10)	**-2.9670**	**1.5520**	**-1.9120**	**0.0558**	**-6.0089**	**0.0749**

Se observa que el modelo tiene problemas de ajuste, una razón es por una gran variabilidad y coeficientes muy grandes en algunas variables, en la interacción I(gA3*Z8) no se ajusta el coeficiente. Éste es un problema muy común a la hora de ajustar un modelo de regresión. Muchas veces se debe a que se presentan muy pocas o nulas observaciones en alguna de las categorías en las dos variables. Por eso se hizo una tabla de contingencia de gA3 *vs* Z8 (cuadro 4.31). No hay indicios de un mal ajuste por la presencia de pocas observaciones en alguna de las celdas de la tabla de contingencia. Se sospecha que esta ocurriendo un problema de colinealidad entre variables pues en R muestra el siguiente mensaje "*Coefficients: (1 not defined because of singularities)*". Aumentan las posibilidades de obtener colinealidad o multicolinealidad al trabajar con muchas variables binarias. Se introdujeron menos variables, ajustando la regresión con las variables del modelo intermedio y la interacción I(gA3*Z8), el coeficiente esta vez sí es impreso, en dicho modelo no es un coeficiente significativo, por lo tanto no será incluida en el análisis.

Normalmente la recomendación cuando se presenta este tipo de problemas es incluir menos variables o buscar más observaciones. Pues el cálculo de los coeficientes es hecho por la inversión de una matriz [5]. Es un problema cuando existe singularidad, pues este tipo de matrices no tienen inversa. Cuando es multicolinealidad el cálculo sería inestable.

[5]Detalles de los métodos iterativos pueden encontrarse en Agresti (2002), pág. 194

Al modelo 2 se aplica la función "step()" en R, proceso automatizado de selección de modelo por AIC, se llamará submodelo 3 y se muestra en el cuadro 4.32.

Cuadro 4.31: Tabla de contingencia de la interacción I(gA3*Z8).

		FAB		Total
		Otherwise	FAB Grade 4 Or 5 and AML	
gA3 0	Count	74	18	92
	within gA3	80.4 %	19.6 %	100.0 %
	within FAB	80.4 %	40.0 %	67.2 %
	of Total	54.0 %	13.1 %	67.2 %
1	Count	18	27	45
	within gA3	40.0 %	60.0 %	100.0 %
	within FAB	19.6 %	60.0 %	32.8 %
	of Total	13.1 %	19.7 %	32.8 %
Total	Count	92	45	137
	within gA3	67.2 %	32.8 %	100.0 %
	within FAB	100.0 %	100.0 %	100.0 %
	of Total	67.2 %	32.8 %	100.0 %

Cuadro 4.32: Submodelo 3: regresión logística realizada por selección automática por AIC del Modelo 2 en la base de 137 transplantes de Leucemia.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.13480	1.25370	2.50000	0.01240	0.6775	5.5921
P	-3.99820	1.29540	-3.08600	0.00203	-6.5372	-1.4592
Z8	4.12380	1.45240	2.83900	0.00452	1.2771	6.9705
Z9A4	-3.28660	0.99550	-3.30100	0.00096	-5.2378	-1.3354
I(gA2 * Z8)	-2.87540	1.38180	-2.08100	0.03745	-5.5837	-0.1671
I(A * P)	1.34460	0.68420	1.96500	0.04939	0.0036	2.6856
I(C2 * P)	4.23970	1.02550	4.13400	0.00004	2.2297	6.2497
I(Z8 * Z10)	-3.93930	1.45740	-2.70300	0.00687	-6.7958	-1.0828

Este modelo es reducido y todas sus variables son significativas. Lo que se puede diferenciar de los otros submodelos es que en el submodelo 3 se podrá ver que tipo de variables influyen más en el evento de interés. En este modelo las variables Z8 y P interactúan varias veces, por lo tanto, si el paciente muestra esa información se ve afectado en el riesgo de muerte. Por otro lado, en este modelo se encuentran algunas variables que no son tan significativas, pero al interactuar con Z8 y P sí lo son. C2 es una variable significativa en el modelo intermedio, pero no fue significativa al reducir el modelo y entró como una interacción. Esto se debe por problemas de multicolinealidad o singularidad antes mencionados.

Segundo Caso: C1 no depende de C2

En el segundo caso no se incluyó C2 en el modelo saturado. De igual manera se realizó una eliminación manual de las variables. Se eliminaron las variables en el siguiente orden: Z5, A, Z2, Z4, Z9A2, Z6, Z3, gA3, Z1, Z7, Z9A3. Dicho modelo es presentado en el cuadro 4.33, se llamará modelo 4. La prueba de cociente de verosimilitudes de este modelo contra el modelo saturado sin incluir C2, no rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_p = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., p$, lo cual las variables no son significativas. (*Likelihood ratio test for MLE method Chi-squared 11 d.f.=6.87028 , P value=0.8095*).

Cuadro 4.33: Submodelo 4: eliminación manual de variables del modelo saturado excluyendo C2. Las variables fueron removidas en el orden siguiente: Z5, A, Z2, Z4, Z9A2, Z6, Z3, gA3, Z1, Z7, Z9A3.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.60970	1.11350	3.24200	0.00119	1.42724	5.79216
gA2	-1.33940	0.45210	-2.96300	0.00305	-2.22552	-0.45328
C	-0.92220	0.42700	-2.16000	0.03077	-1.75912	-0.08528
P	-2.71510	1.09900	-2.47100	0.01349	-4.86914	-0.56106
Z8	1.65010	0.50760	3.25100	0.00115	0.65520	2.64500
Z9A4	-1.64550	0.66680	-2.46800	0.01360	-2.95243	-0.33857

De la misma manera al submodelo afinado (submodelo 4) se incluyen interacciones, no se consideran las interacciones que tienen C2 y tampoco la interacción I(gA3*Z8) ya que había presentado problemas. A ese modelo se aplica la selección de variables por AIC y resultó ser casi el mismo modelo que el submodelo 4, dicho modelo se presenta en el cuadro 4.34 (submodelo 5).

Cuadro 4.34: Submodelo 5: Modelo 4 con interacciones que no tienen a C2, elimina la interacción I(gA3*Z8) y se aplica la selección de variables por AIC.

Regresión Logística	Coefficients:	Std. Error	Wald	p-value	inf	sup
(Intercept)	3.7561	1.1329	3.315	0.000915	1.5356	5.9766
gA2	-1.4287	0.4727	-3.022	0.002508	-2.3552	-0.5022
C	-1.0142	0.4438	-2.285	0.022284	-1.8840	-0.1444
P	-2.761	1.1133	-2.48	0.01314	-4.9431	-0.5789
Z8	2.4332	0.6743	3.609	0.000308	1.1116	3.7548
Z9A4	-2.2256	0.7842	-2.838	0.00454	-3.7626	-0.6886
I(Z8 * Z10)	-2.5245	0.9976	-2.531	0.011385	-4.4798	-0.5692

Tercer Caso: Modelo del libro

Se ajusta una regresión logística con las mismas variables (gA2, gA3, Z1, Z2, I(Z1*Z2), Z3, Z7, Z8 y Z10) al ejemplo 11.1 del capítulo 11 del libro de Klein and Moeschberger (2003), pero se toma el evento de muerte C1. En el cuadro 4.35 se presenta este modelo que se llamará modelo 6. Se centran las variables Z1, Z2 y Z7, la codificación es de la manera siguiente: Z1c=Z1-28, Z2c=Z2-28 y Z7c=(Z7/30)-9.

Cuadro 4.35: Modelo 6: Variables consideradas en el libro de Klein and Moeschberger (2003) de los ejemplo 11.1, considerando el evento de muerte C1.

Regresión Logística	Coefficients:	exp(coef)	Std. Error	Wald	p-value	inf	sup
(Intercept)	0.20857	-	0.46198	0.45100	0.65166	-0.69692	1.11405
gA2	-1.54560	0.2131	0.55774	-2.77100	0.00559	-2.63878	-0.45243
gA3	-0.42909	0.6511	0.61566	-0.69700	0.48583	-1.63577	0.77759
Z1c	-0.00251	0.9974	0.03369	-0.07400	0.94063	-0.06855	0.06353
Z2c	0.01535	1.0154	0.03122	0.49200	0.62288	-0.04584	0.07655
I(Z1c * Z2c)	0.00816	-	0.00291	2.79900	0.00512	0.00245	0.01386
Z7c	-0.01065	0.9894	0.01620	-0.65800	0.51085	-0.04241	0.02110
Z8	1.33348	3.7942	0.49043	2.71900	0.00655	0.37223	2.29472
Z10	0.23620	1.2664	0.45860	0.51500	0.60653	-0.66266	1.13506

Cuadro 4.36: Modelo 6: Variables consideradas en el libro de Klein and Moeschberger (2003) de los ejemplo 11.1, considerando el evento de muerte C1. Se aplica la exponencial a los coeficientes de los parámetros y a los límites del intervalo de confianza.

Regresión Logística	exp(coef)	exp(lim inf)	exp(lim sup)
(Intercept)	-	-	-
gA2	0.2131	0.0714	0.6361
gA3	0.6511	0.1948	2.1762
Z1c	0.9974	0.9337	1.0656
Z2c	1.0154	0.9552	1.0796
I(Z1c * Z2c)	-	-	-
Z7c	0.9894	0.9585	1.0213
Z8	3.7942	1.4510	9.9217
Z10	1.2664	0.5155	3.1114

Con este modelo se ejemplifica la razón de momios, sin embargo, la construcción de la razón de momios en presencia de interacción no es de manera directa y no se profundizará en ese tema[6]. La razón de momios para las variables *dummies* de g se puede observar que la ocurrencia de presentar el evento de muerte se presenta más en la categoría de referencia.

[6]Para profundizar sobre la razón de momios en presencia de interacción puede revisarse la sección 3.7 de Hosmer and Lemeshow (2000)

Para Z1c, Z2c y Z7c tienen una razón de momios cercano a 1, lo cual indica que el riesgo de la ocurrencia del evento no se incrementa. Nótese si hay un incremento en 56 unidades de estas variables la razón de momios de cada variable respectivamente es 0.8689, 2.3622 y 0.5508. La variable Z8 se observa que la razón de momios es grande, esto indica que en la presencia de la variable la ocurrecia del evento casi se cuadruplica. En la presencia de Z10 el riesgo de que ocurra el evento es mayor que cuando no se presenta.

A continuación se resume en el cuadro 4.37 las tasas de clasificación de los modelos presentados.

Cuadro 4.37: Tasas de clasificación de los modelos ajustados regresión logística para cada caso considerado. Caso1=C1 depende de C2; caso 2=C1 no depende de C2; caso3=ejemplo 11.1 de libro Klein and Moeschberger (2003) de la base de 137 transplantes de médula ósea.

Modelo	Software	Método	M-M	M-V	V-M	V-V	tgcb	No.var. tec. distintas
Caso 1								
satudaro	R	LR	86.42	13.58	23.21	76.79	82.48	17
intermedio	R	LR	80.25	19.75	17.86	82.14	81.02	7
afinado	R	LR	76.54	23.46	10.71	89.29	81.75	5
1	SPSS	LR	77.78	22.22	14.29	85.71	81.02	4
2	R	LR	81.48	18.52	12.50	87.50	83.94	12
3	R	LR	80.25	19.75	16.07	83.93	81.75	7
Caso 2								
4	R	LR	74.07	25.93	19.64	80.36	76.64	5
5	R	LR	72.84	27.16	16.07	83.93	77.37	6
Caso 3								
6	R	LR	81.48	18.52	44.64	55.36	70.80	7

Se observa que entre los modelos ajustados es el submodelo 2 que presentan mejores tasas de clasificación global. Pero ese modelo presentó problemas de ajuste en algunas interacciones y obtuvo variables con alta variablidad y coeficientes muy grandes, lo cual este modelo no sería el ideal. El submodelo 3 no presenta problemas de ajuste en cuanto a tasas de clasificación es muy similar al submodelo intermedio, pero en cuanto a tasas de clasificación global es igual al submodelo afinado. Resulta un poco difícil elegir un modelo entre estos tres, no obstante el submodelo intermedio y el submodelo 3 presentan mejores tasas de clasificación para el evento de muerte.

Para el caso 2 se observa que C2 es una variable que influye fuertemente pues las tasas de clasificación global se degradan. Se observa también que las variables gA2, C, P , A, Z8 siguen apareciendo en esos modelos. Lo cual los hace significantes. Para el ejemplo del libro se observa de manera muy similar a lo que pasa con la base de datos de crédito. Se dificulta distinguir a los que viven.

4.4.2. Aplicación del Modelo de Regresión de Cox

Se iniciará con un breve análisis descriptivo de la muestra utilizando la función de supervivencia empírica de Kapplan-Meier. En la figura 4.4 se muestra el ajuste general de toda la muestra y se observa que si $t > 781$ hay cada vez menos muertos. Al parecer éste es un indicio que existe un patrón al cual se debe la supervivencia, indicando la existencia de un mayor riesgo de morir durante los primeros 781 días posteriores al día del transplante.

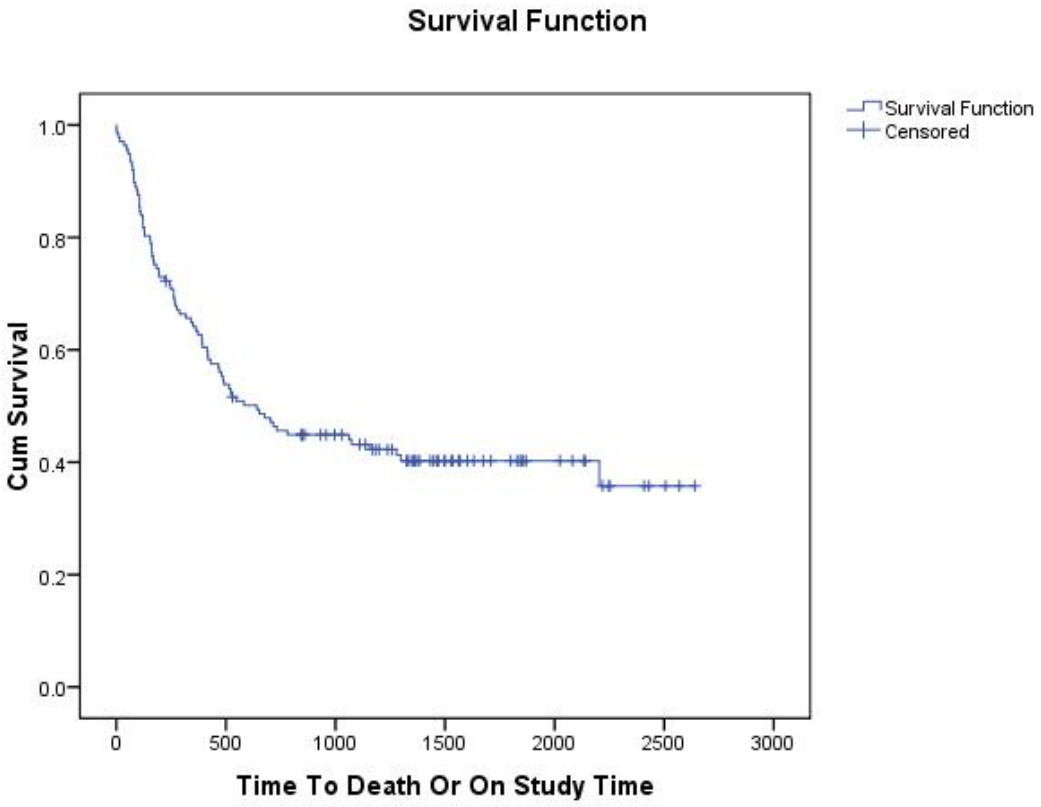

Figura 4.4: Función de Kapplan-Meier de toda la muestra de transplantes de médula ósea.

Caso 1: C1 depende de C2

A continuación se ajusta el modelo de regresión de Cox. En el cuadro 4.38 se muestra la regresión ajustada del modelo saturado (modelo S1). Las variables que son significativas son las variables C2, C, P, Z2, Z9A2, Z9A3. Se efectúa una afinación de las variables quitando una por una del modelo saturado, fueron retiradas las variables en el siguiente orden: Z10, Z3, Z4, Z7, gA3, Z6, Z1, Z5, Z9A3, gA2, A, Z8. La simplifiación se muestra en el cuadro 4.39, dicha afinación llevará el nombre de submodelo S2.

Cuadro 4.38: Modelo S1: modelo saturado bajo la regresión de Cox, de la misma manera usando como indicadora de dato censurado la variable C1, de la base de 137 transplantes de médula ósea. Las variables resaltadas en negritas son las variables significativas del modelo.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
gA2	-0.3818	0.4089	-0.9340	0.3500	-1.1833	0.4196
gA3	0.1721	0.4096	0.4200	0.6700	-0.6308	0.9749
C2	**0.9850**	**0.2820**	**3.4930**	**0.0005**	**0.4322**	**1.5378**
A	0.6529	0.3412	1.9140	0.0560	-0.0159	1.3216
C	**-1.0094**	**0.2652**	**-3.8060**	**0.0001**	**-1.5292**	**-0.4895**
P	**-1.2100**	**0.3482**	**-3.4750**	**0.0005**	**-1.8924**	**-0.5276**
Z1	0.0145	0.0246	0.5890	0.5600	-0.0337	0.0626
Z2	0.0377	0.0212	1.7750	0.0760	-0.0039	0.0793
Z3	-0.0125	0.2656	-0.0470	0.9600	-0.5331	0.5082
Z4	-0.0281	0.2647	-0.1060	0.9200	-0.5469	0.4907
Z5	-0.3684	0.2693	-1.3680	0.1700	-0.8962	0.1595
Z6	-0.1375	0.2670	-0.5150	0.6100	-0.6608	0.3857
Z7	0.0001	0.0004	0.3590	0.7200	-0.0006	0.0009
Z8	0.4531	0.3235	1.4010	0.1600	-0.1810	1.0873
Z9A2	**1.0310**	**0.4031**	**2.5580**	**0.0110**	**0.2410**	**1.8210**
Z9A3	-0.4297	0.3616	-1.1880	0.2300	-1.1384	0.2790
Z9A4	**-1.2953**	**0.4715**	**-2.7470**	**0.0060**	**-2.2194**	**-0.3712**
Z10	NA	0	NA	NA	NA	NA

Se efectúo la prueba del cociente de verosimilitudes entre el modelo saturado (modelo S1) y el submodelo S2, no se rechaza la hipótesis nula, $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$, esto quiere decir que este modelo es tan bueno como el modelo saturado. (*Likelihood ratio test for Cox regression & conditional logistic regression: Chi-squared 12 d.f. = 14.11269 , P value = 0.2936*).

Cuadro 4.39: Submodelo S2: afinación del modelo saturado de la base de 137 transplantes de médula ósea.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
C2	1.1276	0.2392	4.7100	0.0000	0.6588	1.5964
C	-0.7478	0.2456	-3.0500	0.0023	-1.2292	-0.2664
P	-1.4127	0.3009	-4.6900	0.0000	-2.0025	-0.8229
Z2	0.0402	0.0106	3.8000	0.0001	0.0194	0.0610
Z9A2	0.9790	0.3263	3.0000	0.0027	0.3395	1.6185
Z9A4	-0.9637	0.4335	-2.2200	0.0260	-1.8134	-0.1140

También se ajustó la regresión de Cox con las variables del submodelo intermedio de la regresión logística (cuadro 4.25) y al compararlo con el modelo saturado (modelo S1), dicho modelo tenían variables no significativas. Se aplicó la prueba de cociente de verosimilitudes entre estos dos modelos y se rechazó la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para

algún $i = 1, ..., m$, entonces, en el modelo se removieron variables significativas. (*Likelihood ratio test for Cox regression conditional logistic regression: Chi-squared 11 d.f. = 28.74803 , P value = 0.0025*). Se compararó el modelo intermedio con el modelo afinado, se rechaza la hipótesis nula, lo cual, alguna de las variables variables excluídas no es cero.(*Likelihood ratio test for Cox regression conditional logistic regression Chi-squared 1 d.f. = 4.26568 , P value = 0.0389*).

En SPSS se usa el método de selección automática de variables (*forward stepwise selection*), metiendo el modelo saturado. Antes fueron introducidas las variables dicotomizadas de las variables g y Z9. Se obtuvo el siguiente submodelo llamado S2.1, observado en el cuadro 4.40.

Cuadro 4.40: Submodelo S2.1: se usa la selección automática de variables (*forward stepwise selection*) en SPSS, metiendo el modelo saturado de la base de 137 pacientes de transplante.

Regresión de Cox	Coefficients:	Std. Error	Wald	p-value	inf	sup
C2	1.1268	.2391	22.2147	.0000	.6582	1.5954
C	-.7442	.2454	9.1944	.0024	-1.2252	-.2632
P	-1.4122	.3010	22.0179	.0000	-2.0021	-.8223
Z2	.0402	.0106	14.4254	.0001	.0194	.0609
Z9A2	.9708	.3260	8.8712	.0029	.3320	1.6097
Z9A4	-.9634	.4335	4.9388	.0263	-1.8131	-.1137

En el modelo resultante permanecen las variables exactamente iguales al modelo S2. Los coeficientes estimados son casi iguales en comparación al modelo ajustado en R.

En las figuras 4.5 y 4.6 se observan la funciones de supervivencia de los modelos S1 y S2. Además, que las funciones de supervivencia ajustadas son bastante similares entre ellas. En R la función de supervivencia estimada cuando se ajusta un modelo de Cox, por *default* es evaluada en la media de las variables explicativas usadas en el ajuste de la regresión[7].

[7]En véase en la ayuda de R la función "survfit.coxph" en el paquete "survival"

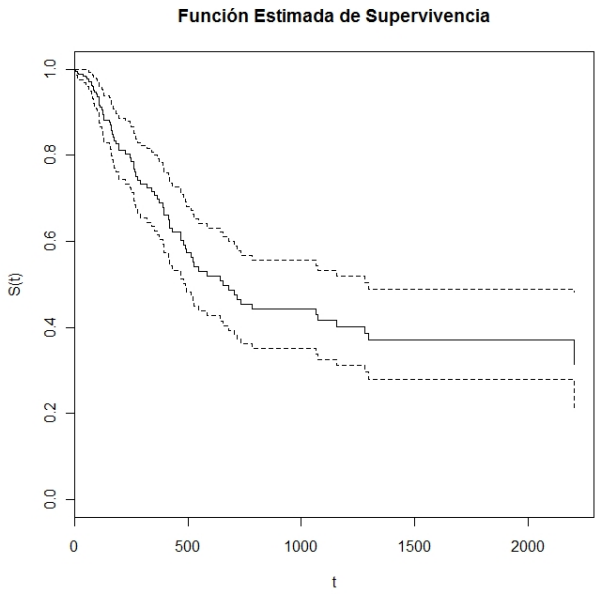

Figura 4.5: Función de supervivencia del modelo S1: modelo saturado considerando bajo la regresión de Cox, de la misma manera usando como indicadora de dato censurado la variable C1, de la base de 137 transplantes de médula ósea. Los valores de las variables son evaluados en la media.

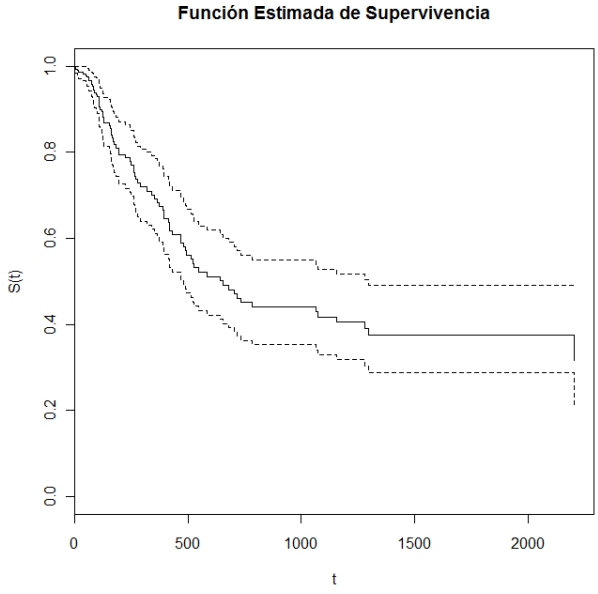

Figura 4.6: Función de supervivencia del submodelo S2: eliminación manual de variables del modelo saturado de la base de 137 transplantes de médula ósea. Los valores de las variables son evaluados en la media.

De igual manera que en la sección 4.4.1, se introducen las variables presentadas en el

submodelo 2 de regresión logística en el modelo de regresión de Cox. En el cuadro 4.41 es presentado el resumen de dicho modelo ajustado en R, este modelo se llamará S3. En la figura 4.7 se presenta la función de supervivencia del modelo S3. Se observa que es muy similar a los modelos anteriores. El modelo S3 también presenta problemas de ajuste en la misma interacción. Y otra es que la variable C2 antes positiva en el modelo S2 ahora es negativa.

Cuadro 4.41: Modelo S3: Se usan las variables del submodelo 2 de regresión logística de la base de 137 pacientes. Se observan que las mismas variables tienen problemas de ajuste que en modelo de regresión logística. La explicación antes mencionada es porque debe de existir una multicolinealidad entre variables. Las variables resaltadas en negritas son las variables significativas del modelo.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
gA2	0.09635	0.46910	0.20500	0.84000	-0.82303	1.01559
C2	**-1.90757**	**0.86580**	**-2.20300**	**0.02800**	**-3.60454**	**-0.21072**
C	**-0.89292**	**0.28050**	**-3.18300**	**0.00150**	**-1.44265**	**-0.34249**
P	**-3.15188**	**0.61980**	**-5.08600**	**0.00000**	**-4.36615**	**-1.93794**
Z8	0.19260	0.42780	0.45000	0.65000	-0.64588	1.03105
Z9A3	-0.27299	0.37890	-0.72100	0.47000	-1.01556	0.46938
Z9A4	**-1.59808**	**0.50180**	**-3.18500**	**0.00140**	**-2.58098**	**-0.61434**
I(gA2 * Z8)	0.16310	0.62570	0.26100	0.79000	-1.06305	1.38929
I(gA3 * Z8)	NA	NA	NA	NA	NA	NA
I(P * Z1)	0.00256	0.01960	0.13000	0.90000	-0.03594	0.04114
I(A * C2)	-0.93436	0.69170	-1.35100	0.18000	-2.28967	0.42134
I(A * P)	**1.53647**	**0.48770**	**3.15000**	**0.00160**	**0.58059**	**2.49230**
I(C2 * gA2)	-0.61452	0.69230	-0.88800	0.37000	-1.97113	0.74241
I(C2 * gA3)	-0.15004	0.62550	-0.24000	0.81000	-1.37595	1.07568
I(C2 * P)	**3.72261**	**0.77630**	**4.79500**	**0.00000**	**2.20104**	**5.24419**
I(Z1 * gA3)	0.02682	0.01550	1.73400	0.08300	-0.00351	0.05733
I(Z8 * Z10)	-0.24027	0.67620	-0.35500	0.72000	-1.56542	1.08519

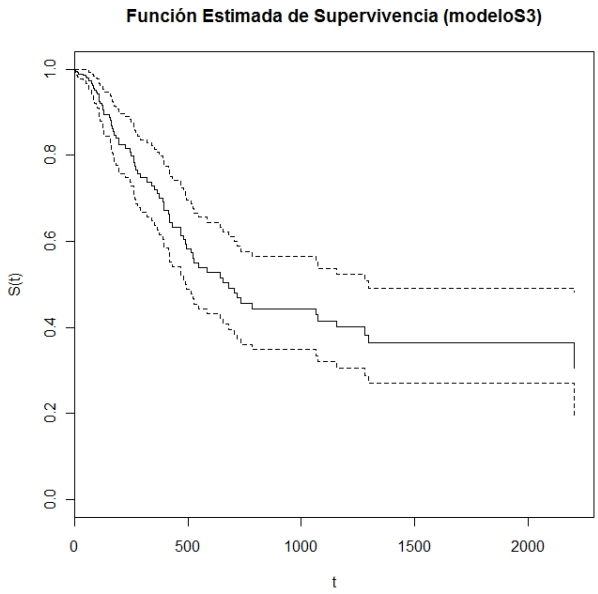

Figura 4.7: Función de supervivencia ajustada de modelo S3: Se usan las variables del modelo 2 de regresión logística de la base de 137 pacientes del modelo de Cox. Los valores de las variables son evaluados en la media.

Al modelo S3 de igual manera se le ajusta la función "step()" en R, a este modelo se le llamará submodelo S4 y se muestra en el cuadro 4.42. En la figura 4.8 se encuentra la función de supervivencia ajustada por dicho modelo.

Cuadro 4.42: Submodelo S4: modelo S3 por selección de AIC bajo la regresión de Cox de la base de 137 transplantes de médula ósea.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
C2	-1.97530	0.66504	-2.97000	0.00300	-3.27810	-0.67139
C	-0.92930	0.25837	-3.60000	0.00032	-1.43548	-0.42312
P	-3.02480	0.38397	-7.88000	0.00000	-3.77662	-2.27303
Z9A4	-1.31260	0.44227	-2.97000	0.00300	-2.17948	-0.44629
I(A * P)	1.13340	0.33849	3.35000	0.00081	0.46994	1.79675
I(C2 * P)	3.42130	0.72150	4.74000	0.00000	2.00719	4.83541
I(Z1 * gA3)	0.02670	0.00728	3.67000	0.00024	0.01252	0.04114

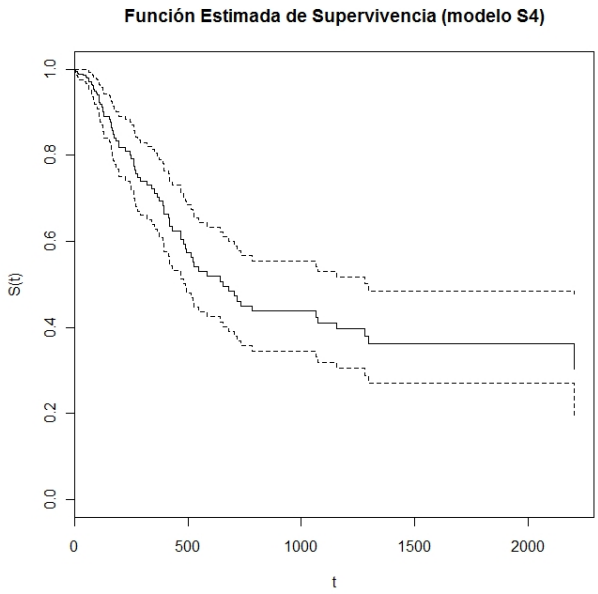

Figura 4.8: Función de supervivencia del submodelo S4: modelo S3 por selección de AIC bajo la regresión de Cox de la base de 137 transplantes de médula ósea. Los valores de las variables son evaluados en la media.

Se aplica la prueba del cociente de verosimilitudes para corroborar que los parámetros no incluidos no son significativos, el resultado fue que no se rechaza la hipótesis nula $H_0 : \beta_1 = ...\beta_m = 0$ *vs* $H_a : \beta_i \neq 0$ para algún $i = 1, ..., m$. Entonces, los parámetros eliminados no son significativos, (*Likelihood ratio test for Cox regression & conditional logistic regression Chi-squared 10 d.f. = 4.535651 , P value = 0.92*).

Luego en el modelo 3 de la regresión logística, se incluyen las mismas variables y se aplica la regresión de Cox, ese modelo se llamará submodelo S5. En el cuadro 4.43 se presenta dicho modelo y en la figura 4.9, se muestra la función de supervivencia ajustada del modelo. La gráfica expone la misma forma a todos los modelos anteriores. Se aplica la prueba del cociente de verosimilitudes, comparando el submodelo S5 con el submodelo S3 ya que el submodelo S5 es un modelo anidado del submodelo S3. Se encontró evidencia estadística que este modelo no es tan bueno como S3, esto implica que las variables removidas son significativas (*Likelihood ratio test for Cox regression & conditional logistic regression Chi-squared 10 d.f. = 29.73361 , P value = 9e-04*).

Cuadro 4.43: Submodelo S5: variables del modelo 3 de la regresión logística en el modelo de regresión de Cox, de la base de 137 transplantes de médula ósea.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
P	-2.022	0.366	-5.532	0.0000	-2.7394	-1.3046
Z8	0.41	0.282	1.453	0.1500	-.1427	.9627
Z9A4	-0.994	0.438	-2.271	0.0230	-1.8525	-.1355
I(gA2 * Z8)	-0.547	0.412	-1.326	0.1800	-1.3545	.2605
I(A * P)	0.648	0.322	2.015	0.0440	.0169	1.2791
I(C2 * P)	1.539	0.273	5.629	0.0000	1.0039	2.0741
I(Z8 * Z10)	-0.088	.573	-0.154	0.8800	-1.2111	1.0351

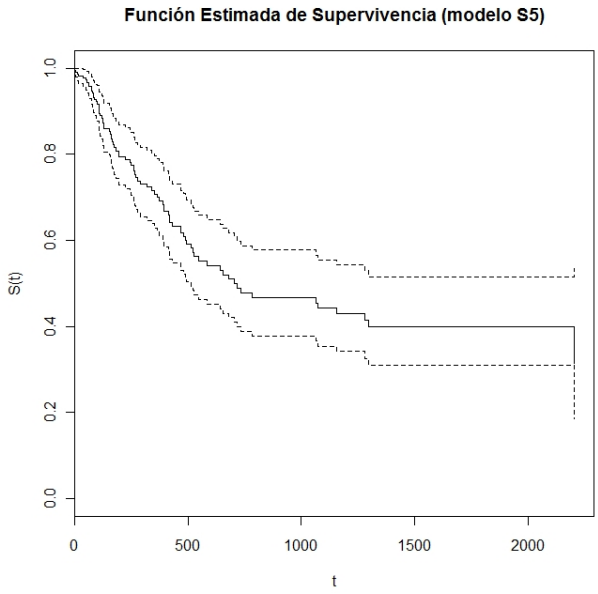

Figura 4.9: Función de supervivencia del submodelo S5: variables del modelo 3 de la regresión logística en el modelo de regresión de Cox, de la base de 137 transplantes de médula ósea. Los valores de las variables son evaluados en la media.

Dado el dato anterior el modelo S5 pierde poder predictivo sobre el evento de interés, pues se omite información. Por otro lado, una de las cosas más interesantes entre los modelo S2 y el modelo S4, destaca que, la variable C2 cambia de signo positivo en el submodelo S2 a signo negativo en el modelo S4. Otra información interesante son las interacciones que se encuentran en S4 pues las variables A, P no son directamente significativas y las variables que se repiten de dos a más veces son A, P, C2, Z8. Estas variables han sido significativas no solamente en la regresión de Cox sino también directa e inderectamente en la regresión logística. Por lo que estas variables serían factores determinantes para explicar la supervivencia del paciente.

Caso 2: C1 no depende de C2

Al no incluir en el modelo saturado a C2, y eliminar manualmente las variables se llegó a un modelo distinto (cuadro 4.44), a este ajuste se le llama submodelo S6. Como en la regresión logística para obtener el submodelo 5 del apartado anterior, se agregan las mismas interacciones que no continen la variable C2 y se aplica la selección de variables por AIC (submodelo 7 cuadro 4.45).

Cuadro 4.44: Submodelo S6: afinación del modelo saturado exluyendo la variable C2 de la base de transplantes de médula ósea.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
gA2	-0.9050	0.2624	-3.4500	0.0006	-1.4188	-0.3901
C	-1.0379	0.2493	-4.1600	0.0000	-1.5279	-0.5499
P	-1.3206	0.3099	-4.2600	0.0000	-1.9310	-0.7133
Z1	0.0474	0.0128	3.7100	0.0002	0.0227	0.0723
Z8	0.5348	0.2366	2.2600	0.0240	0.0714	0.9984
Z9A3	-0.7217	0.3496	-2.0600	0.0390	-1.4065	-0.0367
Z9A4	-1.2829	0.4555	-2.8200	0.0049	-2.1716	-0.3901

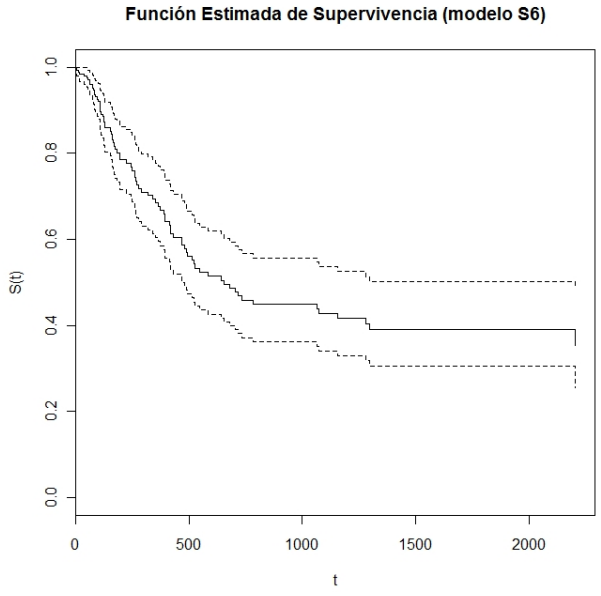

Figura 4.10: Función de supervivencia del submodelo S6: se eliminan variables una a una del modelo saturado exluyendo la variable C2 en el análisis de la base de transplantes de médula ósea. Los valores de las variables son evaluados en la media.

Cuadro 4.45: Submodelo S7: submodelo S6 con interacciones y reducción de modelo por medio de AIC.

Regresión de Cox	coef	se(coef)	z	p	lower 0.95	upper 0.95
gA2	-1.0089	0.2618	-3.8500	0.0001	-1.5219	-0.4959
C	-1.2189	0.2575	-4.7300	0.0000	-1.7237	-0.7133
Z1	0.0905	0.0152	5.9600	0.0000	0.0607	0.1204
Z8	0.7417	0.2352	3.1500	0.0016	0.2807	1.2027
Z9A3	-0.7090	0.3527	-2.0100	0.0440	-1.4004	-0.0182
Z9A4	-1.4038	0.4683	-3.0000	0.0027	-2.3218	-0.4861
I(P * Z1)	-0.0597	0.0101	-5.9000	0.0000	-0.0795	-0.0398
I(A * P)	0.8147	0.3331	2.4500	0.0140	0.1618	1.4676

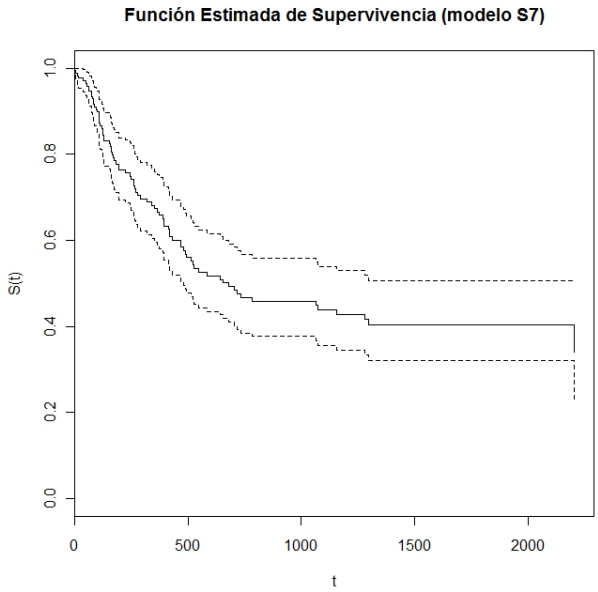

Figura 4.11: Función de supervivencia del submodelo S7: submodelo S6 con interacciones y reducción de modelo por medio de AIC, se añadieron las interacciones antes ya utilizadas y se redujo el modelo por selección de AIC. Los valores de las variables son evaluados en la media.

Caso 3: modelo del libro

Se considera el modelo mencionado en la sección 4.4.1. En el cuadro 4.44 se presenta el ajuste de la regresión de Cox. Con este modelo se ejemplifica la razón de riesgos, sin embargo, la construcción de la razón de riesgos en presencia de interacción no es de manera directa y no se profundizará en ese tema[8]. La razón de riesgos para las variables *dummies* de g se puede observar que la ocurrencia de presentar el evento de muerte se presenta más en la categoría

[8]Para profundizar sobre la razón de riesgos en presencia de interacción puede revisarse la sección 4.4 de Hosmer and Lemeshow (1999)

de referencia. Para Z1c, Z2c y Z7c tienen una razón de momios cercano a 1, lo cual indica que el riesgo de la ocurrencia del evento no se incrementa. Nótese si hay un incremento en 56 unidades de estas variables la razón de momios de cada variable respectivamente es 0.9041, 3.2053 y 0.6949. La variable Z8 se observa que la razón de momios es grande, esto indica que en la presencia de la variable la ocurrecia del evento se duplica. En la presencia de Z10 el riesgo de que ocurra el evento es mayor que cuando no se presenta.

Cuadro 4.46: Modelo S8: ajuste de la regresión de Cox considerando las variables del ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 transplantes de médula ósea.

Regresión de Cox	coef	exp(coef)	se(coef)	z	p	lower 0.95	upper 0.95
gA2	-1.062	0.3458	0.3794	-2.7995	0.0051	-1.8079	-0.3188
gA3	-0.3766	0.6862	0.3839	-0.9812	0.33	-1.1301	0.3757
Z1c	-0.0018	0.9982	0.0205	-0.0888	0.93	-0.0419	0.0383
Z2c	0.0208	1.0210	0.0183	1.1327	0.26	-0.0151	0.0564
I(Z1c * Z2c)	0.0024	-	0.0009	2.7049	0.0068	0.001	0.004
Z7c	-0.0065	0.9935	0.011	-0.5949	0.55	-0.0284	0.0149
Z8	0.7609	2.1402	0.2868	2.6533	0.008	0.1989	1.3231
Z10	0.2645	1.3028	0.2572	1.0285	0.3	-0.2395	0.7687

Cuadro 4.47: Modelo S8: ajuste de la regresión de Cox considerando las variables del ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 transplantes de médula ósea. Se aplica la exponencial a los coeficientes de los parámetros y a los límites del intervalo de confianza.

Regresión de Cox	exp(coef)	exp(lim inf)	exp(lim sup)
gA2	0.346	0.164	0.727
gA3	0.686	0.323	1.456
Z1c	0.998	0.959	1.039
Z2c	1.021	0.985	1.058
I(Z1c * Z2c)	-	-	-
Z7c	0.994	0.972	1.015
Z8	2.14	1.22	3.755
Z10	1.303	0.787	2.157

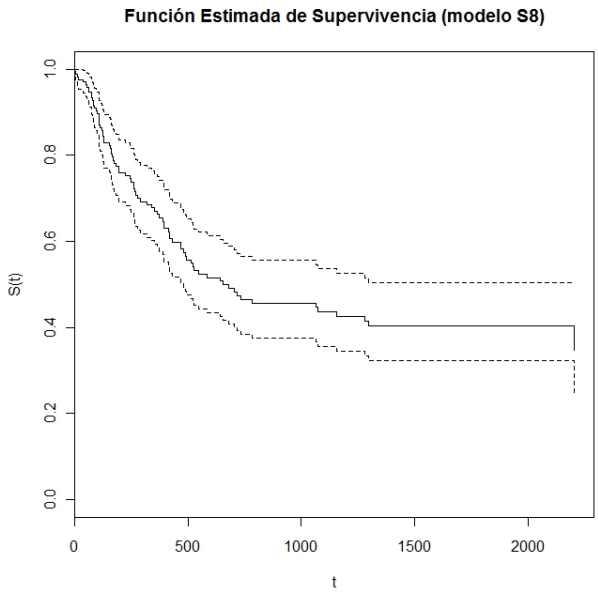

Figura 4.12: Función de supervivencia estimada del modelo S8: ajuste de la regresión de Cox considerando las variables del ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 transplantes de médula ósea. Los valores de las variables son evaluados en la media.

En R no se encontró una manera de agrupar las funciones de supervivencia de todos los modelos, para enfrentar esta dificultad fueron graficados en MS Excel representados en la figura 4.13. Como se puede observar las funciones ajustadas por los modelos son muy similares.

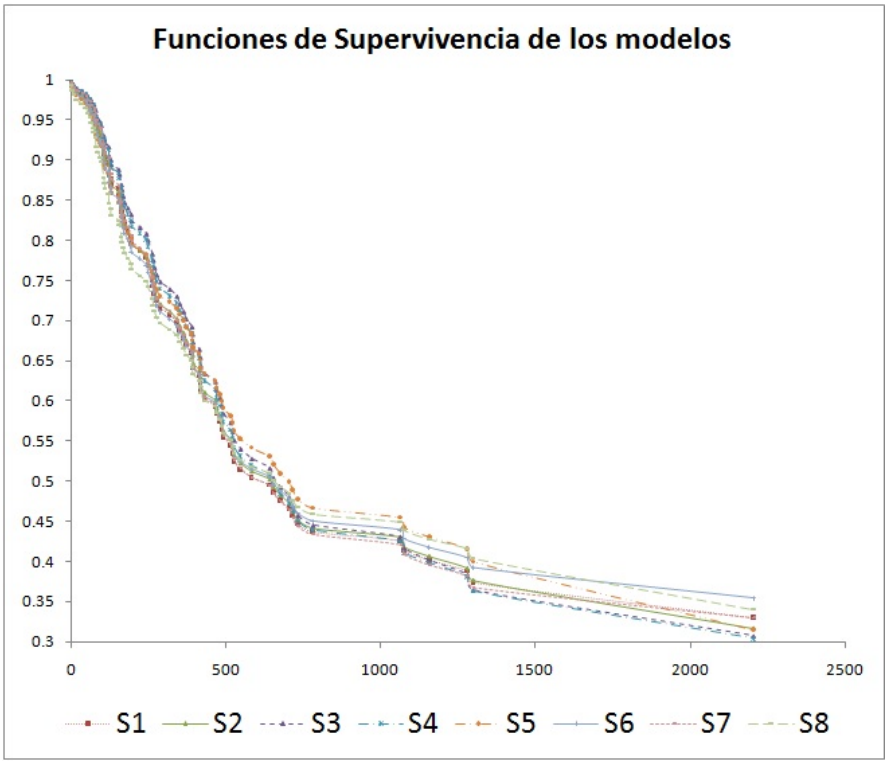

Figura 4.13: Funciones de supervivencia de los modelos de regresión de Cox ajustados $S_i(t)$, $i = 1, ..., 8$. Los valores de las variables son evaluados en la media.

Se analiza el ajuste general de los modelos presentados, los residuales de Cox-Snell se calculan y se grafican. En la figura 4.14 se presentan para los modelos S1,..., S4. Se observa que la función de riesgo acumulada de los residuales siguen la diagonal de 45 grados, lo cual no hay un indicio de falta de ajuste del modelo.

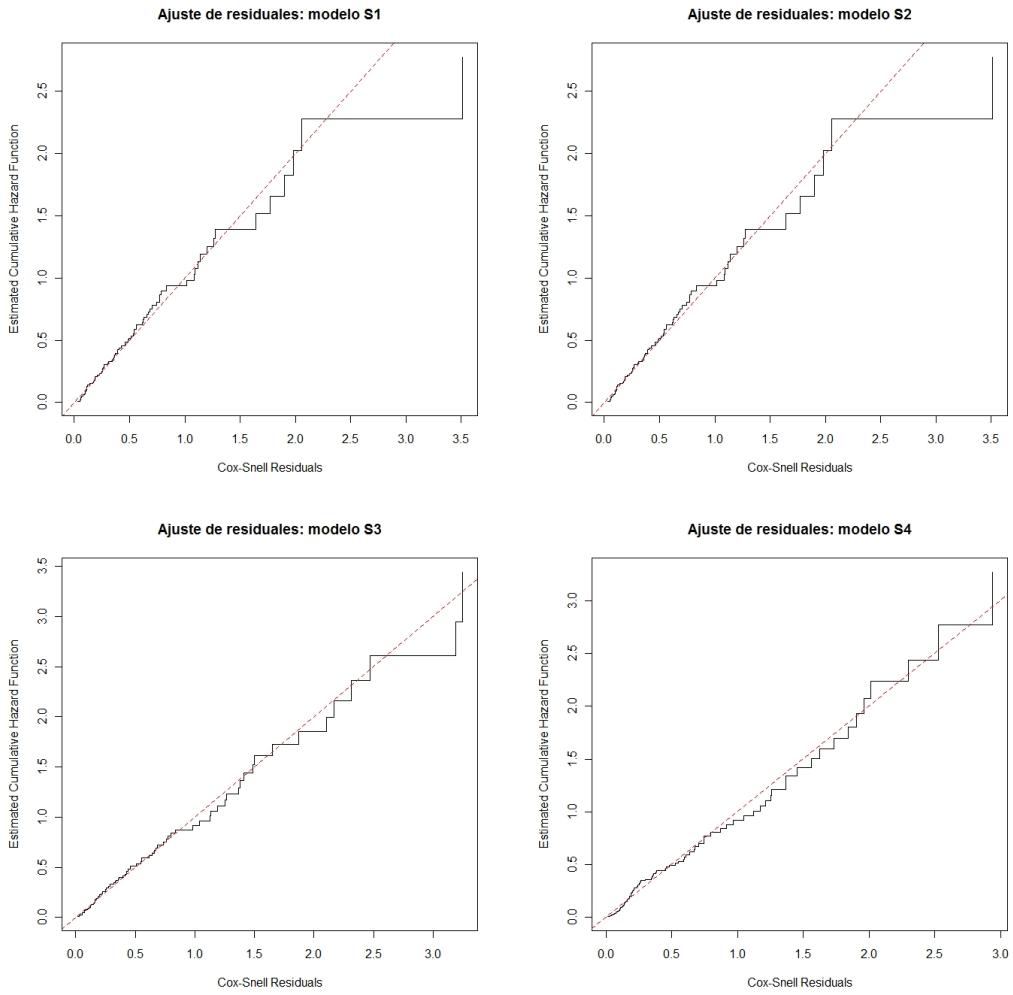

Figura 4.14: Se presentan los residuales de los modelos S1, S2, S3, y S4. Se observa que para todos los casos. La gráfica tiene un desapego en la diagonal de 45° para valores de t_i grandes, pero en general las gráficas muestran que los residuales siguen la función de riesgo acumulada de la exponencial con parámetro 1.

En la figura 4.15 se presentan para los modelos S5,...,S8. De la misma manera se observa que las gráficas de los residuales siguen la diagonal de 45° y presentan desapego a valores grandes de t_i. Los residuales siguen la función de riesgo acumulada de la exponencial con parámetro 1.

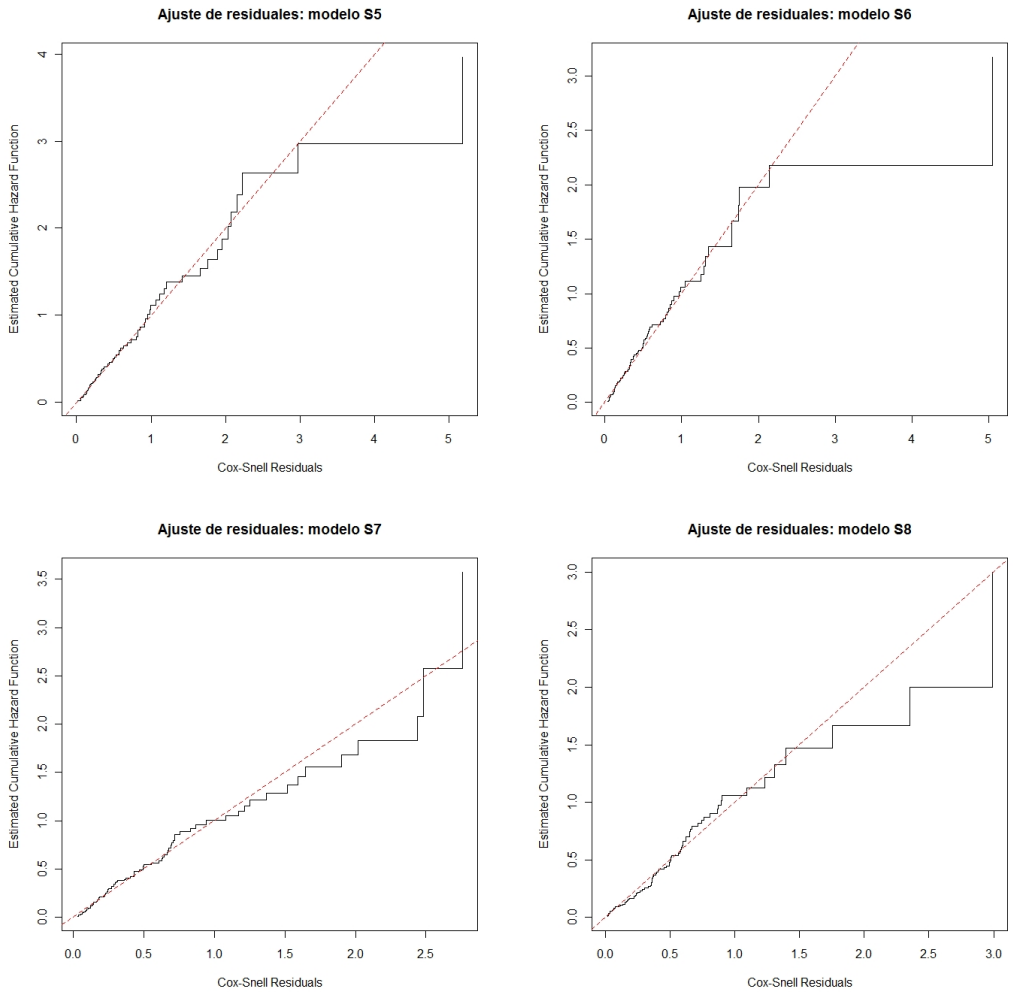

Figura 4.15: Se presentan los residuales de los modelos S5, S6, S7, y S8. Se observa que para todos los casos. La gráfica tiene un desapego de la diagonal de 45° para valores de t_i grandes, pero en general las gráficas muestran que los residuales siguen la función de riesgo acumulada de la exponencial con parámetro 1.

4.5. Conclusiones

Este capítulo sirvió para poder ajustar los modelos de clasificación y el modelo de regresión de Cox con dos bases de datos. Para la aplicación del crédito se resume lo siguiente:

- El modelo saturado fue uno de los mejores, presentó tasa de clasificación correcta del 76.7 %. Aunque el discriminante cuadrático para el modelo saturado presentó mejores tasas de clasificación para los malos créditos 57.33 %. Sin embargo, fue aún pésima, pues tuvo una tasa de clasificación correcta en los créditos buenos de 78.71 % y el modelo saturado fue de 89.57 %.

- Al quitar variables del modelo saturado las tasas de clasificación se degradaban. Se incluyeron interacciones entre variables, tomando aquellas que aparecieron en todos los modelos, se pudo mejorar muy poco la tasa correcta de los rechazados. Sin embargo, se está de acuerdo que no en todos los casos valdría la pena indagar sobre la regla cuadrática para mejorar un clasificador.

- Se encontró un modelo a partir del saturado con interacciones. Este modelo pudo mantener la tasa de aceptados mejorando la de rechazados. Este modelo es mejor que el saturado en el sentido que tiene mejores tasas de clasificación de los rechazados, aunque su diferencia sea por pocos puntos porcentuales.

- El modelo en donde se incluyen todas las interacciones es el mejor ya que presenta las mejores tasas sobre todos los modelos. Pero este modelo no sería bueno ya que contiene variables e interacciones no significativas y las tasas de clasificación con el submodelo 1.4.1 no difieren por mucho.

- Las interacciones entre variables son un poco difíciles de interpretar. Pero pueden aportar información adicional. Por ejemplo, en la interacción de *laufzeit* y *moral* se vio que no siempre los que han pagado bien todas sus deudas actuales significa que te paguen el crédito cuando la duración del crédito es alta. Este patrón en la práctica sucede a menudo. Porque si el solicitante puede pagar la deuda en un corto plazo, a la entidad financiera no le convendría tener un crédito a largo plazo, pues sería más riesgoso ya que si ocurren cambios en el escenario político-económico o incluso en el mismo cliente pueden influir en el incumplimiento. Por ejemplo, devaluación económica, el cliente se enfermó o perdió el trabajo, etc.

En la base de transplante de médula ósea se destacan los siguientes puntos:

- Se observó que al ajustar la función empírica de Kapplan-Meier, el 54.74 % de los pacientes mueren en los primeros 781 días post-operación, después de esa fecha registran pacientes que sobreviven y 15 fallas más.

- Para la regresión logística se obtuvieron modelos con buen poder predictivo, también se encontraron modelos con interacciones de variables con tasas regulares alrededor del 80 % por lo cúal son buenos clasificadores. Las variables que resaltan son C2, P, Z8 ya que se repiten en interacciones. Al no considerar C2 en el modelo (caso 2), las tasas de clasificación correcta de los muertos se degradan. Y aún más se degrada la tasa de clasificación correcta de vivos, al excluir las medidas post-operativas (caso 3).

- Para la regresión de Cox se destaca que la variable C2 cambia de signo positivo a signo negativo en los modelos con interacciones. No todas las interacciones aportan gran información lo cual puede hacer un poco largo el proceso de selección de una a una si se incluyen todas las combinaciones posibles. Otro dato interesante son las interacciones que se encuentran en S4 pues las variables A, P no son directamente significativas y las variables que se repiten de dos a más veces son A, P, C2, Z8.

- Al no considerar C2 en el modelo saturado se llega a un modelo distinto, las variables significativas son C, P, Z1, Z8. En el ejemplo del libro (caso 3) también hay variables significativas.

- Al analizar los residuales de Cox-Snell se observó que los modelos siguen la función de riesgo acumulada de la exponencial $\lambda = 1$. Sin embargo, para seleccionar un modelo no sólo se debe de confiar en eso, pues se vio que en el modelo S3 hubo un problema de ajuste en un coeficiente, esto se debe a la presencia de multicolinealidad antes mencionada. Por lo tanto, el análisis de residuales no resultó lo bastante confiable.

- Otra cuestión es que al considerar la recaída como una variable explicativa puede hacer que el ajuste sea bastante tendencioso. En el libro el modelo que usan en el ejemplo consideran la clase incluyendo la recaída C2 como una causa de falla del transplante.

- En la regresión de Cox se puede apreciar que el ajuste de la función de supervivencia es casi idéntica cuando las variables son evaluadas en la media y es difícil ver cual de los modelos es el mejor, pues por lo mencionado anteriormente los residuales de Cox-Snell no se puede ver un patrón, vemos que en los ajustes de algunos modelos de regresión algunas variables cambian de signo. Mientras que en en la regresión logística, se puede apreciar una diferencia en los modelos por las tasas de clasificación.

Resumiendo se puede apreciar que la regresión logística y la regresión de Cox pueden arrojar modelos similares, pero eso no indica que un modelo de regresión logística ajuste tan bien como un modelo de regresión de Cox o viceversa. Por ejemplo, las variables del modelo 3 de la regresión logística al ajustarlas en la regresión de Cox (modelo S5), se pierde información ya que se excluyen variables significativas. Sin embargo, en los residuales de Cox-Snell no hubo indicios de carencia de ajuste del modelo S5.

Capítulo 5

Comparación de Reglas de Clasificación

5.1. Introducción

En el proceso de clasificación se asigna una observación a una clase basado en la información observada acerca de ella. Sabiendo que el proceso de asignación no es perfecto, se cometen errores. Por esa imperfección, en el *credit scoring* está implicíto un costo para la entidad financiera y se necesita evaluar el rendimiento del clasificador. Evaluar el proceso permitiría decidir si es lo suficientemente bueno para el propósito, intentar mejorarlo o reemplazarlo por otro procedimieto de clasificación.

Para el *credit scoring* y los numerosos problemas de clasificación el objetivo principal es construir una función de puntaje $S(\boldsymbol{X})$, en la cual los miembros de las dos clases tengan distintos conjuntos de puntaje, así permitiendo que las clases queden perfectamente distinguidas.

Una manera intuitiva de resumir los resultados de un modelo de clasificación es por medio de las tasas de clasificación errónea. A partir de dichas tasas se crea la matriz de confusión en donde muestra los resultados de clasificar las observaciones por el método de clasificación y compararlas con la variable respuesta original. Al obtener esta variable dicotómica se está asignando un punto de corte k. Esto se habló en la sección 2.2 en el tema sobre el análisis de discriminante. En la regresión logística muchas veces el punto de corte es de 0.5, esto quiere decir que se asigna a un grupo en este caso el de los malos si la probabilidad estimada es menor a 0.5. La apariencia desde este punto de aproximación para modelar la medición viene de la relación cercana entre la regresión logística y el análisis de discriminante cuando la distribución de las variables explicativas es multinormal entre las 2 poblaciones (Hosmer and Lemeshow 2000, pág. 156).

En este punto de vista, las probabilidades estimadas son usadas para poder predecir el grupo al que pertenece la observación. Presumiblemente la idea es que si el modelo predice la pertenencia con precisión, entonces debería de arrojar evidencia de que el modelo ajusta

bien. Desafortunadamente, éste puede ser o no el caso, como se vio esta relación de manera práctica en la sección 4.3.1 de la base de datos de crédito, donde se observa que no hay evidencia que indique falta de ajuste en el modelo, pero las tasas de clasificación en el modelo ajustado de la población de los malos andan cercanas al 50 %.

En la sección 5.2 se introduce el concepto de los términos empleados para la construcción de la curva ROC, se define el AUC una medida de comparación entre curvas ROC. En la sección 5.3 se grafican las curvas ROC de la base de crédito alemán y son comparadas, seguida de un análisis de la muestra de créditos malos en el programa Ggobi 5.3.2. En la sección 5.4 se grafican y se comparan las curvas ROC de la regresión logística para la base de transplantes y en la sección 5.5 se concluye este capítulo.

5.2. Curvas ROC

La curva ROC (*Receiver Operating Characteristic curve*) es una de las técnicas en la estadística que ahora es utilizada en diversos campos. Su nombre proviene del uso de dichas curvas en el campo de la teoría de detección de señales, donde el objetivo es detectar la pre-sencia de una señal en particular, dejando pasar pocas ocurrencias genuinas como sea posible y simultáneamente minimizar las falsas alarmas. En la teoría de detección de señales el objetivo es asignar a cada evento una categoría, ya sea la de señal genuina o la de falsa alarma.

Para poder definir la curva ROC es necesario definir dos clases, sean P y N. En la mayoría de los casos existe una asimetría en las poblaciones de estudio. Por ejemplo, en tratar de asignar a las clases "curará/no curará, si es tratado ", en el *credit scoring* "pagará/ no pagará el crédito". La P y la N representan esta asimetría, donde la P son los positivos ("son casos"; personas con la enfermedad, transacción fraudulenta, tendrá problemas con el crédito, morirá, etc.) y la N representa a los negativos("no son casos"; normal, transacción legítima, persona saludable, no va a morir, etc.). Normalmente uno trata de identificar los casos correctamente.

Para el grupo de los "positivos", P, se da una distribución de los puntajes $Pr(s|P)$, y para los objetos del grupo de "negativos", N, la distribución $Pr(s|N)$. La clasificación de los grupos está dada por comparar los puntajes con el umbral T. Si se encuentra el umbral ideal tal que todos los miembros de la clase P tienen puntajes mayores que t, y todos los miembros de la clase N tienen puntajes menores o iguales que t, entonces es posible tener una clasificación perfecta. Pero no es usual este hecho. Normalmente, los dos conjuntos se traslapan haciendo que los de la clase N tengan puntajes altos y P tengan puntajes bajos, haciendo la clasificación perfecta imposible. Entonces la evaluación se basa en detectar cuales son los valores de P que tienden a dar puntajes altos y cuales de N tienen la tendencia a dar puntajes bajos.

La escencia de la curva ROC está basada en la tabla de 2×2 de clasificación, donde sus

resultados son a partir de la clasificación cruzada de la verdadera clase de las observaciones contra su clase pronosticada. Un conjunto de observaciones, son evaluadas empíricamente por las probabilidades conjuntas $Pr(s > t, P)$, $Pr(s > t, N)$, $Pr(s \leq t, P)$, $Pr(s \leq t, N)$.

Convenientemente se resumen las cuatro probabilidades de arriba en términos de dos probabilidades condicionales y una probabilidad marginal:

1. La probabilidad de que una observación de la clase N arroje un score mayor que t: $Pr(s > t|N)$; ésta es la tasa de falsos positivos, denotado como fpr.

2. La probabilidad de que un observación de la clase P arroje un score mayor que t: $Pr(s > t|P)$; ésta es la tasa de verdaderos positivos o simplemente positivos, denotado tpr.

3. La probabilidad marginal de que una observación pertenezca a la clase P: $Pr(P)$.

Un falso positivo se presenta cuando un objeto el cual realmente pertence a la clase N es incorrectamente asignado a la clase P, porque su puntaje cae por encima del umbral t. Un verdadero positivo se presenta cuando un objeto el cual pertenece a la clase P es correctamente asignado a la clase P, esto es porque también cae arriba del umbral t.

Análogamente se define para la población de negativos

1. La tasa de verdaderos negativos o simplemente negativos, $Pr(s \leq t|N)$, la proporción de objetos de la clase N los cuales son correctamente clasificados en N, igual a $1 - fpr$. Denotado como tnr.

2. La tasa de falsos negativos, $Pr(s \leq t|P)$, la proporción de objetos de la clase P los cuales son incorrectamente clasificados en N, igual a $1 - tpr$. Denotado fnr.

3. La probabilidad marginal de que un objeto pertenezca a la clase N: $Pr(N) = 1 - Pr(P)$.

Estas tasas pueden verse en la matriz de confusión mostrada en el cuadro 5.1. Si los eventos de interés (positivos) es el de los créditos malos o el evento de muerte.

	Malos/Muertos	Buenos/Vivos
Malos/Muertos	Positivos	Falsos Negativos
Buenos/Vivos	Falsos positivos	Negativos

Cuadro 5.1: Tasas de la curva ROC en la matriz de confusión de 2×2.

Los epidemiólogos algunas veces usan el término de Sensibilidad (*Sensivity*), denotado por Se, para decir que es la tasa de los positivos tpr, y Especificidad (*Specificity*), denotado Sp, para decir que es la tasa de verdaderos negativos tnr. También se usa el término de Prevalencia refiriendose a la proporción de la población la cual tiene una enfermedad, $Pr(P)$ si es "positivo" corresponde al estado de enfermedad. En los términos del *credit scoring* pensando

que los malos son los "casos" es lo que se está midiendo, entonces la sensibilidad (Se) es la proporción de malos que son pronosticados como malos, mientras que la especificidad (Sp) es la proporción de buenos que son pronosticados como buenos.

La curva ROC es obtenida al variar t, pero usando sólo la tasa falsa positiva y las tasa de los positivos (fpr,tpr), o bien $(1 - Sp, Se)$ como puntos en el espacio bidimensional. La tasa de falsos positivos fpr son los valores correspondientes al eje de las abscisas y la tasa de los positivos tpr son los valores del eje de las ordenadas. Se puede pensar también que es una representación completa del desempeño de la regla de clasificación, la cual varía con la elección del umbral t (Krzanowski and Hand, 2009).

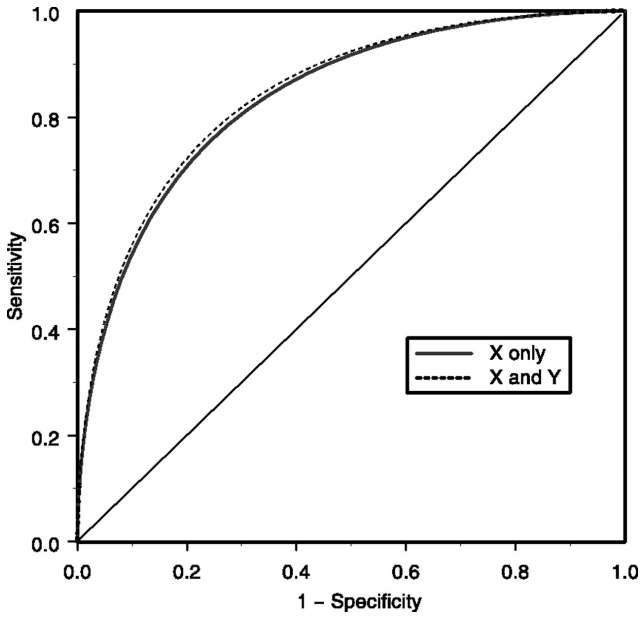

Figura 5.1: Figura hipotética de una curva ROC, la imagen puede ser encontrada en: *http://www.clinchem.org/content/vol54/issue1/images/large/zcy0010887080001.jpeg*

Entonces para ver si un clasificador es al menos bueno, se traza una diagonal uniendo las coordenadas $(0, 0)$ y $(1, 1)$. Se está en el caso cuando tpr es igual a fpr, para cualquier t del umbral. Esta línea es llamada la diagonal de chance, pues representa la asignación de los individuos de las dos poblaciones de manera aleatoria. Es decir, $Pr(s|P) = Pr(s|N) = Pr(s)$ (Krzanowski and Hand 2009, pág. 19). En este caso la probabilidad de asignar un individuo a la poblacion de positivos es la misma que a la de asignar a la población de negativos.

En la curva ROC cuando hay completa separación de $Pr(s|P)$ y $Pr(s|N)$, habrá al menos un valor en los que la asignación perfecta de cada individuo se logra, por lo que para tal t se tiene que $tpr = 1$ y $fpr = 0$. Por otro lado, ya que la curva ROC se centra únicamente en las probabilidades en donde $s > t$ en las dos poblaciones, entonces, para todos los valores menores de t se tiene $tpr = 1$ mientras que fpr varía de 0 a 1, y para todos los valores más

grandes de t, se tendría $fpr = 0$, mientras que tpr, varía de 1 a 0. Entonces la curva ROC estaría dibujada de (0, 0) a (0, 1) seguido por una linea de (0, 1) a (1, 1).

En la práctica, la curva ROC será una curva continua entre estos dos extremos, y estará por arriba del triángulo superior de la gráfica. Entre más se pegue la curva a la esquina superior izquierda, se estará cerca de una separación completa de las poblaciones, y por consiguiente tendrá una buena regla de clasificación.

Una medida usualmente utilizada es el AUC, ésta se define como el área bajo la curva ROC. Una interpretación del AUC es que es un promedio de la tasa positiva, tomados uniformemente sobre todas las posibles tasas de falsos positivos en el rango (0,1). También otra interpretación es que cuando se tienen dos curvas ROC y una está por debajo de la otra, entonces $AUC_1 \leq AUC_2$. Pero no puede darse la implicación inversa pues las curvas pueden intersectarse una con otra (Krzanowski and Hand 2009, pág. 26).

5.3. Base de Crédito Alemán

5.3.1. Comparación de los Modelos de Análisis de Discriminante y Regresión Logística

Se consideran los modelos de regresión logística y análisis de discriminante lineal del modelo 1 y la regresión logística del submodelo E.4. En la figura 5.2 se observa que las curvas ROC del modelo 1 de la regresión logística y el análisis de discriminante se encuentran bastante pegadas y se intersectan en varios puntos lo cual dificulta ver cual de estos dos clasificadores es el mejor. En la figura 5.3 se observan los modelos 1 y E.4 bajo la regresión logística. De la misma manera las curvas que describen estos dos clasificadores se encuentran bastante cerca.

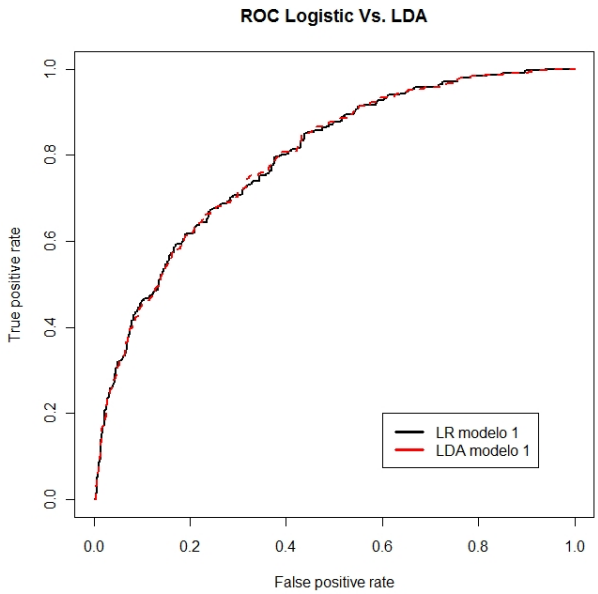

Figura 5.2: Curvas ROC del LDA y RL del modelo 1 de la base de la base de crédito.

En el cuadro 5.4 se presentan los coeficientes obtenidos al ajustar la regresión logística y el análisis de discriminante lineal.

Cuadro 5.2: Coeficientes obtenidos al ajustar el modelo 1 con la regresión logística en R y el discriminante lineal en R y SAS.

	Funciones en SAS				
Variable	d_0	d_1	$d_0 - d_1$	DA (R)	LR (R)
Laufkont	1.1731	1.8091	-0.636	-0.5562126	-0.614226
Laufzeit	0.1736	0.1255	0.0481	0.0419859	0.043016
Moral	1.9045	2.3092	-0.4047	-0.3539886	-0.391583
FamgesA1	3.0541	2.9414	0.1127	0.09853317	0.112572
FamgesA3	2.3656	2.826	-0.4604	-0.4026699	-0.46042
FamgesA4	3.4464	3.7315	-0.2851	-0.2493626	-0.209135
verwA1	1.9129	3.268	-1.3551	-1.18508549	-1.406909
verwA2	4.6976	5.2873	-0.5897	-0.51572442	-0.576205
verwA3	3.4664	4.2265	-0.7601	-0.66472353	-0.744492
verwA4	3.7284	3.3252	0.4032	0.35267125	0.329088
verwA5	6.1303	7.6602	-1.5299	-1.33793073	-1.686947
verwA6	2.8975	3.4723	-0.5748	-0.5027129	-0.538666
verwA7	3.1253	4.0464	-0.9211	-0.80559994	-0.928702
constant	-8.6349	-10.01	1.3751	-	1.386673

En las curvas que describen estos tres modelos (figura 5.3) no es posible detectar cual clasificador es mejor que otro. Por lo tanto al área bajo la curva (AUC) mostrado en el cuadro 5.3 muestra esta similitud entre ellos. En este caso no es posible determinar que clasificador

es mejor que otro, lo que resta es tratar de explicar como esta compuesta la población de malos créditos.

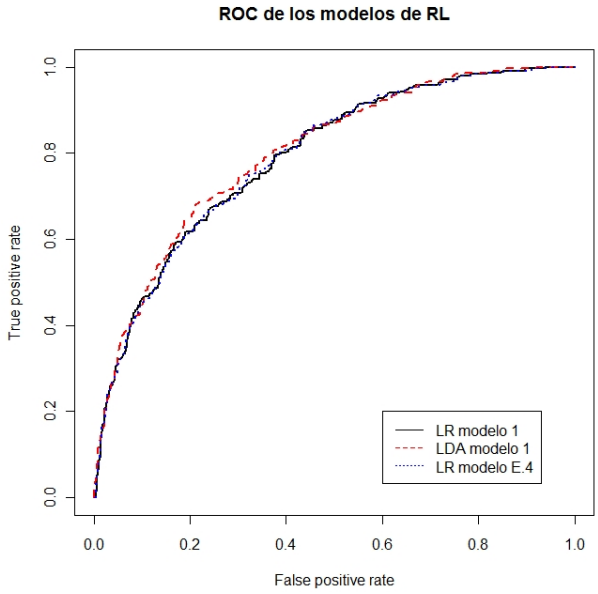

Figura 5.3: Curvas ROC de RL del modelo 1 y submodelo E.4 de la base de crédito alemán.

Modelo	AUC
LR modelo 1	0.790281
LDA modelo 1	0.7992119
LR modelo E.4	0.7902714

Cuadro 5.3: Área bajo la curva ROC de los modelos(AUC).

Como ejemplo en la figura 5.4 se muestra la regresión logística del modelo 1 contra el discriminante lineal quitando la variable *laufkont* y considerando las probabilidades apriori equiprobables, es decir que la probabilidad de ocurrencia de ambas poblaciones es igual a 0.5. Aunque con esa premisa puede afectar bastante en la práctica, ya que la estimación de las probabilidades debería reflejar el tamaño de la muestra. No obstante es un buen ejemplo ya que es un clasificador deficiente para las dos poblaciones. Sus tasas de clasificación se muestran en el cuadro 5.5

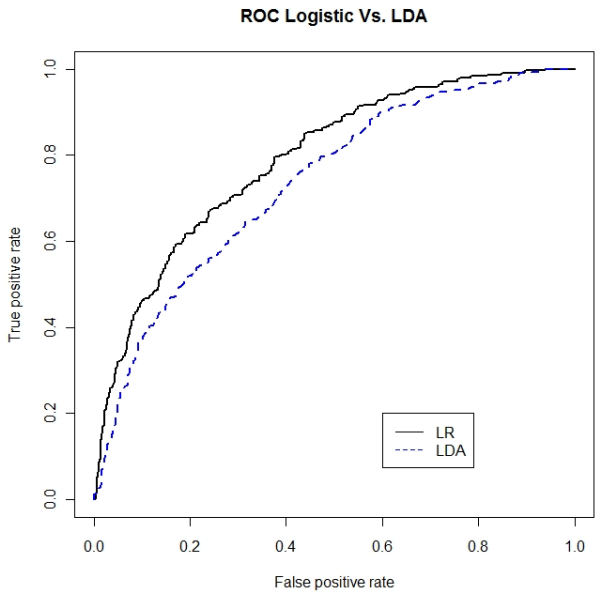

Figura 5.4: Modelo 1 de regresión logística *vs* LDA del modelo (probabilidades a priori iguales y exluyendo *laufkont*).

Cuadro 5.4: Coeficientes obtenidos al ajustar el modelo 1 con la regresión logística en R, el discriminante lineal considerando las probabilidades a priori iguales y excluyendo *laufkont* en R.

Variable	LD	LR modelo 1
Laufkont	-	-0.614226
Laufzeit	0.05549691	0.043016
Moral	-0.56376722	-0.391583
FamgesA1	0.2604557	0.112572
FamgesA3	-0.51515122	-0.460420
FamgesA4	-0.22342425	-0.209135
verwA1	-1.70046648	-1.406909
verwA2	-0.57545413	-0.576205
verwA3	-1.03985814	-0.744492
verwA4	0.30338655	0.329088
verwA5	-1.86028361	-1.686947
verwA6	-0.86651531	-0.538666
verwA7	-0.68593661	-0.928702
constant		1.386673

clase	A	R
A	67.57	32.42
R	35.00	65.00

Cuadro 5.5: Tasas de clasificación de LDA modelo 1 con probabilidades a priori iguales y sin la variable laufkont.

5.3.2. Exploración de los Datos en Ggobi

En este apartado se implementarán técnicas de visualización en Ggobi para poder determinar si existe una estructura en la base de datos, y poder explicar quiénes son los mal clasificados para la población de malos, ya que no se ha podido obtener una regla tal que pueda distinguirlos de manera correcta. Se filtraron solamente las observaciones de los malos créditos, la razón es que no se observó nada con la base entera. Después se obtuvieron las clases pronosticadas del modelo 1 de los clasificadores del análisis de discriminante, agrupando las clases pronosticadas con las variables del modelo saturado.

En la figura 5.5 se muestra el *tour* 2D realizado en dicho software, en donde se toman los clasificadores de análisis de discriminante del modelo 1. Los de color azul son aquellos rechazados que fueron bien clasificados por los dos clasificadores (LDA, QDA). Los de color rojo son aquellos en donde los dos clasificadores no los clasificaron bien. Los de color verde son aquellos en los que solamente el LDA los clasificó bien y los de color naranja son aquellos que solamente el QDA los clasificó bien. Se obtuvo esta gráfica de manera aleatoria.

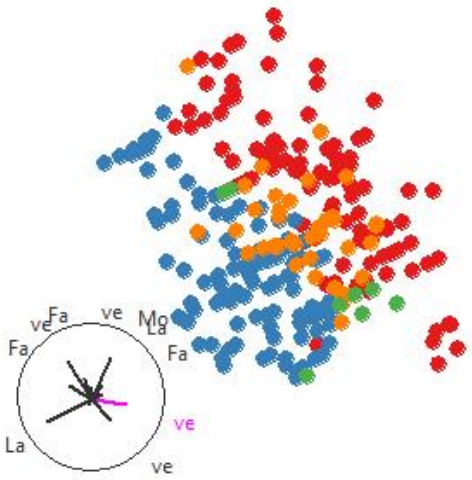

Figura 5.5: Población de malos LDA *vs* QDA.

En la figura 5.6 se observa que los rechazados tienden a agruparse. En el círculo en donde se presentan los vectores de la proyección sobresalen las variables *moral*, *laufzeit* y *laufkont*.

Haciendo la proyección con la variable *laufkont* (figura 5.6), se observa que conservan una estructura en donde se pueden separar los puntos rojos y los puntos azules. Se observa que las categorías 3 y 4 de esta variable se encuentran conformadas de puntos rojos y anaranjados.

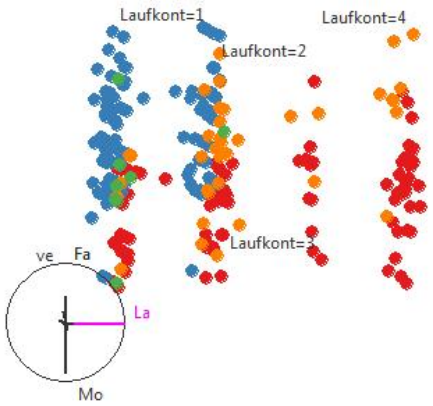

Figura 5.6: Proyección de los datos con la variable *laufkont*.

Al proyectar (figura 5.7) las variables *moral*, *laufzeit* y *laufkont* en el espacio tridimensional se aprecia que se puede separar el conjunto de observaciones para los bien clasificados y los mal clasificados. La manipulación del *tour* fue realizada en base a la variable *laufzeit* indicando que adquiere valores más grandes hacia la izquierda de la gráfica. Por lo que a pequeños valores de la variable *laufzeit* en las categorías 1 y 2 de *laufkont* se puede decir que está compuesta por los rechazados mal clasificados.

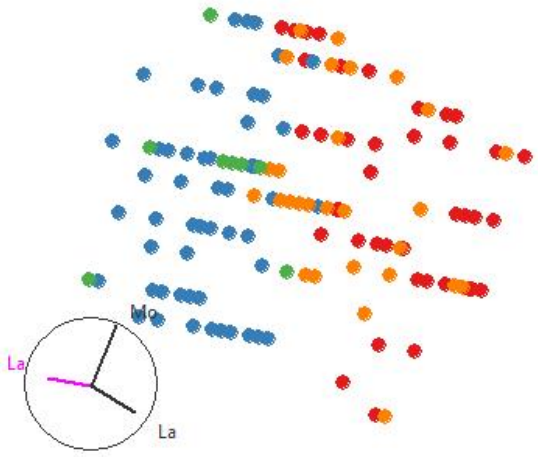

Figura 5.7: Proyección de los datos con la variables *laufkont*, *laufzeit* y *moral*.

Por otro lado, se comenta que se realizó un análisis de componentes principales, de la

misma manera se hizo un *tour* agregando los componentes principales y se hicieron las gráficas de los primeros componentes y de los últimos componentes y no se detectó ningún patrón.

5.4. Base de transplante de Médula Ósea

Se consideran los modelos de regresión logística saturado, intermedio, afinado y el modelo 3. En la figura 5.8 se presentan las curvas ROC de los modelos y se observa que las curvas ROC de los modelos intermedio, afinado y el modelo 3 tienden a parecerse conformando siluetas similares. Hay ciertos puntos en los cuales los modelos se intersectan y otros en donde se separan. Para los modelos 4, 5 y 6 (figura 5.9) se observa exactamente el mismo comportamiento, se intersectan y la curva ROC del modelo 6 esta por debajo de los otros dos modelos. Al calcular el AUC de todos los modelos (cuadro 5.6), se observa que el área bajo la curva anda alrededor de 0.8, el más grande es el modelo saturado y el más pequeño es el modelo 6.

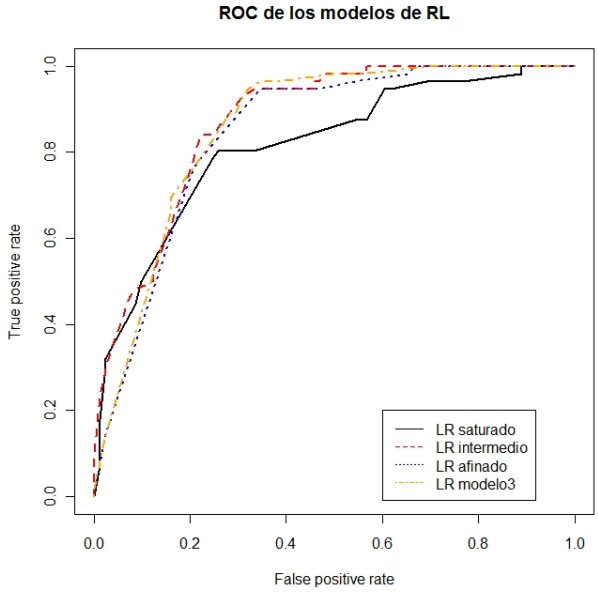

Figura 5.8: Curva ROC de los modelos saturado, intermedio, afinado y el modelo 3 de la base de datos de transplantes.

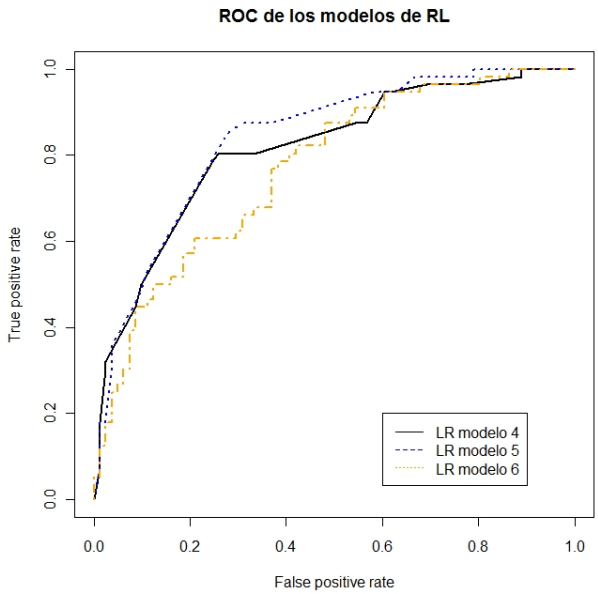

Figura 5.9: Curva ROC de los modelos 4,5 6 de la base de datos de transplantes.

Modelo	AUC
saturado	0.9142416
intermedio	0.8709215
afinado	0.8444665
modelo 3	0.8592372
modelo 4	0.8141534
modelo 5	0.8382937
modelo 6	0.768739

Cuadro 5.6: Área bajo la curva ROC de los modelos(AUC) de la base de transplantes.

5.4.1. Comparación de un Modelo de Regresión de Cox *vs* un Modelo de Regresión Logística

Comparar un modelo de clasificación con un modelo de regresión de Cox no es un asunto trivial ya que ellos arrojan resultados no equivalentes. En los métodos de *credit scoring* requieren de un horizonte temporal bien definido y ellos sólo pueden proveer las estimaciones de probabilidad de que un solicitante sea del grupo de buenos créditos o de malos créditos para ese determinado tiempo.

Stepanohva and Thomas (2002) ajustaron los modelos de regresión logística y de regresión de Cox y encontraron que el modelo de regresión de Cox fue competitivo con una regresión logística cuando modelaron incumplimiento y pago anticipado para préstamos personales. Y además con un método de riesgos competitivos y variables dependientes del tiempo fueron utilizadas para poder dar más precisión al modelo de regresión. Y se utilizaron las curvas ROC para poder ver el comportamiento del método de clasificación. Ellos al ajustar la regresión logística fijaron los modelos en un periodo particular de tiempo y los compararon con la supervivencia estimada del modelo durante ese perido de tiempo.

Eso implica que para poder ver la probabilidad de fallar en un intervalo de tiempo se fijaron en la función de supervivencia determinada sobre ese periodo.

$$P(falla\ en\ el\ periodo\ de\ tiempo\ T < t) = 1 - S(t)$$

donde S(t) es la función de supervivencia estimada del modelo de regresión de Cox.

En el caso de este trabajo y para esta base de datos se trata de clasificar quienes mueren y quienes viven. En la construcción de la regresión logística se obtuvo una función de puntaje para el tiempo en el que fue realizado el estudio. Sin embargo, la regresión logística se sitúa al final del periodo de tiempo. Por otro lado con la regresión de Cox se estima una función de supervivencia para dicho periodo de tiempo.

Para la base de transplantes se hace énfasis en todo el periodo de tiempo. En SPSS se toma la función base de la regresión para el ejemplo del libro de Klein and Moeschberger (2003). También externamente se guardan los valores de la suma ponderada evaluada en cada observación. Al final se construye la probabilidad de muerte en el periodo de tiempo, i.e P(fallar $T < 2204$)=$1 - \hat{S}(2204)$, donde $\hat{S}$ es la función de supervivencia estimada por el modelo de Cox. Para poder guardar la función base en SPSS, revise el código en el anexo B.3.

En la figura 5.10 se muestra la curva ROC de la regresión logística y la regresión de Cox para el ejemplo del libro de Klein and Moeschberger (2003) en la sección 11.1.

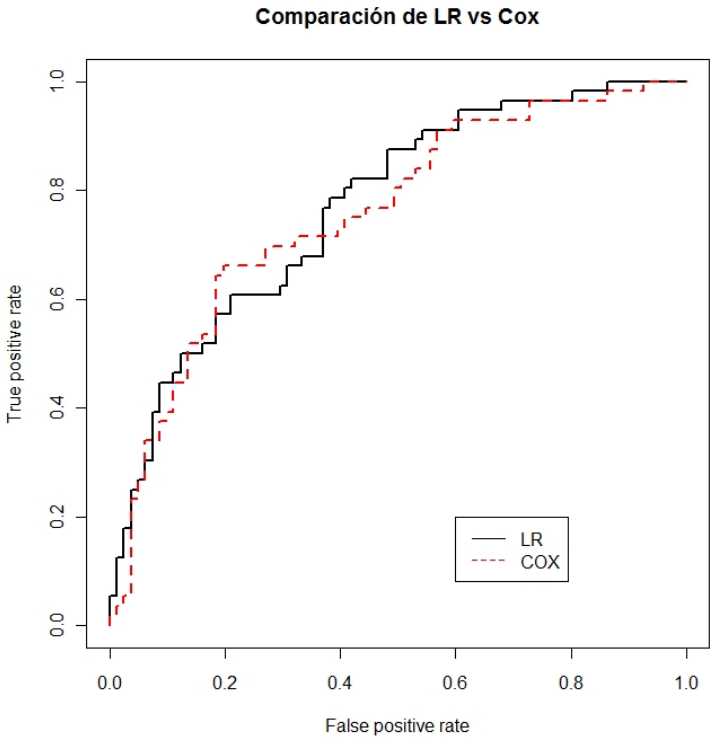

Figura 5.10: Curva ROC de la regresión logística y la regresión de Cox, considerando las variables del ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 de médula ósea.

El área bajo la curva de la curva ROC de la regresión logísitca es $AUC_{rl} = 0{,}768$. Mientras que el de la regresión de Cox es $AUC_{cox} = 0{,}757$. Las tasas de clasificación cuando se considera el punto de corte 0.5 para ambos modelos se presenta en el cuadro 5.7.

Modelo	M-M	M-V	V-M	V-V	tgcb
Regresión Logística	81.48	18.52	44.64	55.36	70.80
Regresión de Cox	90.12	9.88	60.71	39.29	69.34

Cuadro 5.7: Tasas de clasificación de los modelos de regresión logística y de regresión de Cox, considerando las variables del ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 transplantes de médula ósea.

En el cuadro 5.8 se presentan los coeficientes ajustados por los modelos de regresión considerando el ejemplo del libro.

Cuadro 5.8: Coeficientes obtenidos al ajustar la regresión logística y de Cox para el ejemplo 11.1 del libro de Klein and Moeschberger (2003) de la base de datos de 137 transplantes de médula ósea.

Variable	Cox	Logística
Intercept	-	0.2086
gA2	-1.0620	-1.5456
gA3	-0.3766	-0.4291
Z1c	-0.0018	-0.0025
Z2c	0.0208	0.0154
I(Z1c * Z2c)	0.0024	0.0082
Z7c	-0.0065	-0.0107
Z8	0.7609	1.3335
Z10	0.2645	0.2362

5.5. Conclusiones

Para la base de datos alemana se observó que las curvas ROC son muy similares. Esto se debe a que las tasas de clasificación de los modelos presentados en la sección 4.3.1 son muy similares. Para la institución financiera debiera de ser importante tal vez destacar qué tipo de cliente es el que está reflejando el modelo. En Ggobi al incluir ambas poblaciones buenos y malos, no se encontró un patrón que indique quienes son los mal clasificados. Si se filtran sólo los datos de la población de los malos, entonces, los mal clasificados por los métodos son aquellos que tienen pequeños valores de la variable *laufzeit*, en las categorías 1 y 2 de *laufkont*. Esta información puede servir cuando se presenten solicitudes de crédito con estas características, pues se analizaría al cliente con más detalle.

Para la base de datos de transplantes las curvas ROC de los modelos logísticos se intersectan y los AUC al incluir a C2 como una variable explicativa y además algunas medidas post-operativas resulta ser un buen clasificador. Para el ejemplo del libro se observó una curva ROC y un AUC inferior todas las demás curvas.

En el ejemplo del libro al comparar los modelos de regresión logística y de regresión de Cox, obtuvimos unas curvas ROC pegadas, cruzándose en distintos puntos. Por consiguiente se obtuvieron AUC similares. El modelo del libro para clasificar es malo y desde esta aproximación el modelo de Cox también lo es.

Conclusión y Discusión

Conclusiones

En los problemas de clasificación binaria se encuentran dos grupos, normalmente uno de ellos es evento de interés del investigador. Generalmente en los eventos existe una heterogeneidad o asimetría, por ejemplo, en el *credit scoring* "pagará/ no pagará el crédito", en el tratamiento de enfermedades "libre/ no libre de enfermedad". En los problemas de clasificación existen dos tareas, discriminar y clasificar, en donde dependiendo del problema pueden o no llevarse a cabo las dos tareas al mismo tiempo. Los modelos mayormente empleados para poder efectuar estas dos tareas son la regresión logística y el análisis de discriminante descritos en los capítulos 1 y 2.

Comparando la regresión logística y el análisis de discriminante (sección 2.4) se vio que son diferentes entre sí pero pueden guardar una relación muy estrecha cuando los datos tienen una distribución normal multivariada con la misma matriz de varianzas-covarianzas[1]. La pregunta sería qué modelo implementar. De acuerdo a esto, aunque no se cumpla el supuesto de la distribución de normal multivariada, el análisis de discriminante puede ser realizado, en comparación la regresión logística no asume una distribución lo cual se puede decir que es una ventaja sobre el análisis de discriminante. Sin embargo, si los datos siguen una normal multivariada u otra ley de probabilidad, entonces el discriminante es mucho más preciso.

En el ejemplo de crédito (sección 4.3) la regresión logística y el análisis de discriminante lineal presentan una similitud muy estrecha en cuanto a tasas de clasificación, se compararon los coeficientes arrojados por ambos métodos y resultaron muy similares en el paquete R. En SAS las funciones de *score* lineal son divididos de acuerdo al grupo, esto implica que SAS se basa en la regla de asignar la observación en el máximo de $ln(P_k f_k(x))$, para encontrar las reglas de clasificación de R_k se tiene que comparar las funciones (ecuación (2.45)) 2 a 2 para poder encontrar las funciones de discriminante, así la región R_k queda expresada por $(g-1)$ funciones discriminantes.

Siguiendo con el mismo ejemplo de la sección 4.3 se encontró una serie de modelos muy similares, los que tienen interacciones fueron de cierto interés, ya que mejoraron en cierto modo la interpretabilidad. Aunque se debe de tener cuidado cuando las interacciones son entre

[1]Hosmer and Lemeshow (2000) sección 2.6, Cox and Snell (1989), sección 4.4.

categóricas-categóricas y continuas-categóricas ya que puede haber pérdida de interpretabilidad, lo usual es dicotomizar las variables categóricas. Por otro lado, en cuanto a los ajustes del modelo en términos de las tasas de clasificación, se obtuvieron tasas de clasificación errónea alrededor del 50 % para la población de créditos malos, tasas de clasificación errónea cercanas al 10 % en los créditos buenos y las tasas de clasificación global se encontraron entre el 72 % y 77 %. Las curvas ROC fueron presentadas para poder comparar los modelos de clasificación. Se encontró que las curvas se encuentran muy cercanas y se cruzan en varios puntos. Esto se debe a la estrecha similitud de modelos. Además en la sección 5.3.2, se implementaron algunas técnicas de visualización donde se resaltó cuáles son los individuos erróneamente clasificados entre los modelos de regresión logística y análisis de discriminante. Se encontró que los mal clasificados pertenecen a pequeños valores de la variable *laufzeit* (duración del crédito) en las categorías 1 y 2 de la variable *laufkont* (cuenta corriente).

La aplicación con la base de transplantes de médula ósea (sección 4.4), se usó para poder discutir la diferencia entre el modelo de regresión logística y el modelo de regresión de Cox, compararlos no es un asunto trivial, esto se debe a que en el contexto producen dos resultados muy distintos. La regresión de Cox se hace un estimado de una curva de supervivencia o una curva para la función de riesgo. Por otro lado, la regresión logística calcula la probabilidad de caer en una de las categorías de la variable respuesta. La interpretación de ambos modelos puede ser similar en la construcción de razones (razón de momios, razón de riesgos) y tienen una interpretación similar a la del riesgo relativo, pero su contexto es diferente. Una ventaja es que la *hazard ratio* es una medida de la supervivencia que puede verse a lo largo del horizonte de tiempo t. Mientras que la razón de momios es una medida de comparación del evento de ocurrencia sólo al final del tiempo de estudio.

En cuanto al ajuste de modelos se puede notar que ambos modelos pueden conducir a modelos similares. Sin embargo, un modelo bajo la regresión logística no puede ser igual de significativo en la regresión de Cox, un ejemplo es el modelo 3 presentado en la sección 4.4.1. Este conjunto de variables es presentado en el modelo S5 aplicando la regresión de Cox en la sección 4.4.2.

Otro problema que se enfrentó con la base de datos de transplantes es cómo modelar el evento de interés. De manera intuitiva, la variable que indica recaída C2 se propuso que es una causa de muerte del paciente. Lo cual C2 ingresa como una variable explicativa. Por otro lado hay otras medidas post-operativas las cuales se atribuyen al desarrollo de dos tipos de complicaciones y otro a la recuperación del nivel de plaquetas en la sangre. Muchas veces va a depender de cual es el propósito de la investigación. Pero revisando algunos resultados de los modelos presentados, se observa que hay una rareza en los coeficientes estimados y en las varianzas de los coeficientes. Esto es principalmente un reflejo de que las variables consideradas en el modelo, pueden presentar algún tipo de multicolinealidad o singularidad.

Aunque no se revisaron los métodos iterativos, los paquetes estadísticos utilizan estos métodos para el ajuste de los coeficientes. Usualmente el método más utilizado es el de

Newton-Raphson, y este método utiliza la inversión de una matriz. Si la multicolinealidad es un hecho, los coeficientes ajustados no serían los valores verdaderos. Como consecuencia directa estos modelos deben de ser descartados.

Regresando al ajuste de los modelos para la base de transplantes es que, en auscencia de C2 (caso 2) las tasas de clasificación se degradan y en el ejemplo del libro (caso 3) la clasificación de los individuos baja. En la regresión de Cox se revisa el ajuste general presentando los residuales de Cox-Snell. Se observa que algunos modelos no presentan una falta de ajuste ya que siguen la hazard de la exponencial con $\lambda = 1$. De la misma manera en las curvas ROC, es difícil ver que clasificador es mejor, pues las curvas se intersectan y el AUC llega a parecerse, el caso 3 sería el único en donde se puede apreciar que una curva ROC menor a los modelos del caso 1 y 2.

Para poder comparar la regresión logística y la regresión de Cox es necesario escribir la probabilidad de fallar en términos de la curva de supervivencia. Al hacer la comparación es posible ver un modelo de clasificación y así determinar una curva ROC. Se construyó una curva ROC para el ejemplo del libro (caso 3) para los modelos de regresión logística y de Cox. Se encontró que los modelos son similares en términos de clasificación, pero la clasificación es ineficiente para poder detectar al grupo de los pacientes que vivirán.

Discusión

Por una parte el modelo de regresión logística fue utilizado para poder reducir el conjunto de variables, y a ese subconjunto se implementó el análisis de discriminante. Esa fue una forma para comparar ambos métodos. Un camino diferente hubiera sido considerar una selección automática de discriminantes, ofrecida en SAS con la función STEPDISC, aunque la discusión primordial de los métodos automáticos es que el conjunto seleccionado por el método no podría ser el mejor de todos. Otro camino es utilizar otras técnicas de análisis multivariado como los componentes principales, por mencionar una de ellas. En donde su uso recae principalmente en reducir la dimensionalidad de los datos y graficarlos resulta útil para encontrar estructuras o datos atípicos.

Por el lado de la regresión de Cox hay muchas extensiones que se pueden considerar, desde un modelo estratificado, hasta un modelo en donde se prueba la proporcionalidad del mismo, una forma es agregando variables que son dependientes con el tiempo. La censura primordialmente se supone por derecha, pero puede suceder que los individuos no fueron registrados al inicio, sino que fueron entrando en distintos momentos del estudio. Otro análisis es por parte de la teoría de riesgos competitivos, en donde se tienen múltiples causas de falla. Por ejemplo, en el caso de crédito además del incumpliento, otro evento considerado como falla es la de pago anticipado. El pago anticipado se considera como falla ya que la liquidación del crédito anticipadamente genera pérdidas a la institución financiera, pues se pierde parte de los intereses contemplados dentro del plan original de crédito. Se puede modelar el tiempo

falla del crédito considerando el evento que suceda primero, es decir, $T = min(T_1, T_2)$. No obstante también existen diversos modelos de regresión para datos de supervivencia.

Desde luego la motivación de esta tesis fue el comparar los métodos de clasificación contra un modelo de supervivencia. En el *credit scoring* usualmente se utilizan los métodos tradicionales de clasificación para poder obtener un puntaje sobre un nuevo solicitante y determinar si el cliente será bueno o malo. Pero es de suma imporancia para la entidad financiera el considerar cuanto tiempo toma en que un cliente caiga en cartera vencida. Y considerando la censura de que los clientes que inician el estudio es diferente, por eso sería conveniente utilizar análisis de supervivencia. Stepanova and Thomas (2002) comparan la regresión logística contra el modelo de regresión de Cox, para contrastarlos ven el modelo de Cox como un modelo de clasificación y graficaron las curvas ROC para ambos modelos, en donde encuentran que con base en las tablas de clasificación y las curvas ROC, los modelos en análisis de supervivencia pueden ser tan competitivos como la regresión logística cuando se trata de clasificar a los solicitantes en dos grupos.

Para poder construir la curva ROC a partir de un modelo de Cox, la tarea no es fácil pues principalmente se debe limitar el horizonte del estudio, Stepanova and Thomas (2002) dividieron el estudio en el primer año y en el segundo año. Otra razón es que en diversos paquetes estadísticos, por ejemplo, en R no se puede graficar la curva ROC de una regresión de Cox de manera directa. La decisión de asignación de las probabilidades aposteriori se debe de hacer manualmente y programarse computacionalmente. Compararlos de esta manera no es del todo satisfactorio, ya que esto requeriría hacer una regresión logística por cada horizonte de tiempo la cual se requiere comparar la supervivencia. Pero esto ayudará a comparar estos distintos modelos, pero su limitación principal es que son ajustes empíricos por lo que su continua revisión es necesaria.

En la realización de este trabajo fue posible aprender acerca de los paquetes estadísticos R, SPSS y SAS, usando la ayuda y de ejemplos que ofrecen algunos libros e internet. Fue posible determinar algunas ventajas y desventajas de cada paquete estadístico, en lo personal a veces es necesario contar con distintos paquetes ya que la carencia de uno puede ser sustituido por otro. Su uso dependerá del gusto del usuario o de la disponibilidad de ellos en los lugares de trabajo. Finalmente, modelar no es una tarea sencilla, esto además requiere una previa preparación de los datos, es decir, el tratamiento y recodificicación de variables, muestreo etc. que generalmente puede demandar más tiempo que el ajuste, interpretación y presentación de los modelos estadísticos.

Anexo A

Algunas Operaciones con Matrices

Una matriz $\boldsymbol{A}$ de $n \times m$ es un arreglo de números reales

$$\boldsymbol{A}= \begin{pmatrix} a_{11} & a_{12} & ... & a_{1m} \\ a_{21} & a_{22} & ... & a_{2m} \\ \vdots & & \ddots & \vdots \\ a_{n1} & a_{2p} & ... & a_{nm} \end{pmatrix} \tag{A.1}$$

En donde se denota cada elemento de la matriz (a_{ij}), para $i = 1, ..., n$ el número de renglón y $j = 1, ..., m$ la columna. La suma de dos matrices con el mismo número de columnas y renglones se define como

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij}) + (b_{ij}) \tag{A.2}$$

El producto de un escalar λ por una matriz se define como

$$\lambda \boldsymbol{A} = \boldsymbol{A}\lambda = \lambda(a_{ij}) \tag{A.3}$$

Otras propiedades que se derivan de las expresiones mencionadas arriba son las siguientes

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A} \tag{A.4}$$

$$(\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C} = \boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) \tag{A.5}$$

$$\boldsymbol{A} + (-1)\boldsymbol{A} = (0) \tag{A.6}$$

$$(\lambda + \mu)\boldsymbol{A} = \lambda \boldsymbol{A} + \mu \boldsymbol{A} \tag{A.7}$$

$$\lambda(\mu \boldsymbol{A}) = (\lambda\mu)\boldsymbol{A} \tag{A.8}$$

Si $\boldsymbol{A}_{l\times m}$ y $\boldsymbol{B}_{m\times n}$, entonces el producto de matrices se define como

$$\boldsymbol{AB} = (d_{ik}) = \sum_{j=1}^{m} a_{ij}b_{jk} i = 1, ..., l; k = 1, ..., n \tag{A.9}$$

$\boldsymbol{AB}$ es una matriz de dimensión $l \times n$. Otras propiedades del producto de matrices son las siguientes

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \tag{A.10}$$

$$(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC}) \tag{A.11}$$

$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC} \tag{A.12}$$

$$(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC} \tag{A.13}$$

El producto $\boldsymbol{AB}$ no es necesariamente igual a $\boldsymbol{BA}$. La matriz traspuesta de $\boldsymbol{A}_{l\times m}$ se define como $\boldsymbol{A}^T_{m\times l}$ donde cada elemento del j-ésimo renglón con la í-ésima columna de $\boldsymbol{A}^T$ es elemento del í-ésimo renglón y j-ésima columna de $\boldsymbol{A}$. Las propiedades de la matriz traspuesta son las siguientes

$$(\boldsymbol{A}^T)^T = \boldsymbol{A} \tag{A.14}$$

$$(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T \tag{A.15}$$

$$(\boldsymbol{AB})^T = \boldsymbol{B}^T\boldsymbol{A}^T \tag{A.16}$$

La matriz simétrica cumple con la propiedad de que $\boldsymbol{A} = \boldsymbol{A}^T$. Una matriz simétrica en particular es la matriz identidad $\boldsymbol{I}$. La matriz identidad se define como $\boldsymbol{I} = (\delta_{ij})$ donde $\delta_{ij} = 0$ si $i \neq j$ y 1 si $i = j$. Además cumple con la siguiente propiedad

$$\boldsymbol{AI} = \boldsymbol{IA} = \boldsymbol{A} \tag{A.17}$$

El determinante de una matriz existe sólo si esta contiene el mismo número de renglones y columnas. El determinante se define como sigue

$$det(\boldsymbol{A}) = |\boldsymbol{A}| = (-1)^{f(j_1,...,j_p)} \prod_{i=1}^{p} a_{ij_i} \tag{A.18}$$

donde la suma se realiza sobre las permutaciones $(j_1, ..., j_p)$ del conjunto de enteros $(1, ..., p)$ y $f(j_1, ..., j_p)$ es el número de trasposiciones que se necesita para cambiar $(1, ..., p)$ en $(j_1, ..., j_p)$.

Si el determinante $|\boldsymbol{A}| \neq 0$, entonces existe una única matriz $\boldsymbol{B}$ tal que $\boldsymbol{AB} = I$. Entonces $\boldsymbol{B}$ es llamada la inversa de $\boldsymbol{A}$, denotado $\boldsymbol{A}^{-1}$. Sea a_{kh} elemento de la matriz inversa

$$a_{kh} = \frac{\boldsymbol{A}_{hk}}{|\boldsymbol{A}|} = (-1)^{f(j_1,\ldots,j_p)} \prod_{i=1}^{p} a_{ij_i} \tag{A.19}$$

donde $\boldsymbol{A}_{hk}$ es la matriz cuyos elementos estan formados por el menor de $\boldsymbol{A}$. El menor de $\boldsymbol{A}$ es $(-1)^{h+k}$ el determinante de la submatriz al eliminar la fila k y la columna h de la matriz original.

La matriz cuyo determinante es cero es llamada matriz singular. Las matrices singulares no poseen inversa.

Anexo B

Código Empleado en R, SPSS y SAS de los Modelos Presentados

Aquí es presentado el código implementado en R, SPSS y SAS para los modelos discutidos en esta tesis.

B.1. Código Usado en la Regresión Logística

El ajuste de una regresión logística binaria en R

```
z1 <- glm(clase ~ Laufkont+Laufzeit+Moral+
FamgesA1+FamgesA3+FamgesA4+
verwA1+
verwA2+
verwA3+
verwA4+
verwA5+
verwA6+
verwA7
, data=datos, family=binomial(link="logit") )
summary(z1)
prediccion <- predict(z2,type="response")
prediccion <- round(prediccion,0)
tab <- table(clase,prediccion)
tab/rowSums(tab)
```

La prueba de Hosmer-Lemeshow es realizado con el siguiente código.

```
hosmerlem <-
function (y, yhat, g = 10)
```

```
{
    cutyhat <- cut(yhat, breaks = quantile(yhat,
    probs = seq(0,1, 1/g)), include.lowest = T)
    obs <- xtabs(cbind(1 - y, y) ~ cutyhat)
    expect <- xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
    chisq <- sum((obs - expect)^2/expect)
    P <- 1 - pchisq(chisq, g - 2)
    c("X^2" = chisq, Df = g - 2, "P(>Chi)" = P)
}

hosmerlem(clase,prediccion)
```

La prueba de razón de verosimilitudes se realiza ejecutando el siguiente código, es la misma para el modelo de regresiónde Cox.

```
#Likelihood ratio TEST for nested models
library(epicalc)
library(foreign)
library(survival)
library(splines)
lrtest (z2, z.step)
```

B.2. Código Usado en el Análisis de Discriminante

El ajuste del análisis de discriminante lineal en R es el siguiente:

```
z2 <- lda(datos[,-1],clase,prior=c(0.5,0.5)) # Ajuste del lda
lda.pred<-predict(z2)
tab <- table(clase,lda.pred$class)
tab/rowSums(tab)
```

Para obtener el cuadrático se cambia *lda* por *qda*.

En SAS se presenta de igual forma el código empleado para obetener el discriminante lineal:

```
*MODELO 1.2;
*discriminante lineal;
proc discrim data=agre out=ldaout1 outstat=ldaout2
method=normal pool=yes manova;
priors proportional;
class X1;
var X2 X3 X4 X7 X8 X9 X10 X11 X12 X17;
```

```
run;
*coeficientes lda;
data coefficients;
set ldaout2;
run;
proc print data=coefficients;
run;
*probabilidades aposteriori;
data coefficients2;
set ldaout1;
run;
proc print data=coefficients2;
run;
```

Para obtener el discriminante cuadrático se cambia *pool=no*, para utilizar la prueba para probar la igual de las matrices de varianzas covarianzas se utiliza *pool=test*.

B.3. Código Usado en la Regresión de Cox

Ajuste del modelo de riesgos proporcionales de Cox.

```
coxfit1<- coxph(Surv(T1,clase) ~ gA2+gA3+A+C+P+Z1+Z2+Z3+
Z4+Z5+Z6+Z7+Z8+Z9A2+Z9A3+Z9A4+Z10,
 data=datos)
 summary(coxfit1)
surv1 <- survfit(coxfit1)    #grafica de supervivencia
plot(surv1, , xlab='t',
ylab='S(t)', main='Función Estimada de Supervivencia')

coxsnellres=clase-resid(coxfitS1,type="martingale")
fitres=survfit(coxph(Surv(coxsnellres,clase)~1,
method='breslow'),type='aalen')
plot(fitres$time,-log(fitres$surv),type='s',
  xlab='Cox-Snell Residuals',
ylab='Estimated Cumulative Hazard Function',
main='Ajuste de residuales: modelo S1')
abline(0,1,col='red',lty=2)
```

Código de SPSS implementado en la sección 5.4.1

```
COXREG T1
  /STATUS=C1(1)
  /METHOD=ENTER gA2 gA3 Z1c Z2c Z1c*Z2c Z7c Z8 Z10
```

```
/OUTFILE=TABLE('C:\Users\Richard\Desktop\Bases de datos\
Survival Analysis\Transplante de medula Osea\Cox regression\Cox.sav')
  /PLOT SURVIVAL
  /SAVE=SURVIVAL HAZARD XBETA
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

B.4. Código Usado en las Curvas ROC

Obtención de una curva ROC y su gráfica en R

```
z1<- glm(clase ~ Laufkont+Laufzeit+Moral+FamgesA1+
FamgesA3+FamgesA4+verwA1+
verwA2+verwA3+verwA4+verwA5+
verwA6+verwA7, data=datos,
family=binomial(link="logit") )

prediccion.scores <- predict(z1,type="response")
pred <- prediction(prediccion.scores, datos$clase)
perf <- performance(pred, "tpr", "fpr")

#RL modelo 1
plot(perf,col='black',lty=1,lwd=2, main='ROC de los modelos de RL')
#LDA modelo 1
plot(perf2, col='red',lty=2,lwd=2,add=TRUE)
#modelo E.4
plot(perf3, col='blue',lty=3,lwd=2,add=TRUE)
legend(0.6,0.2,c('LR modelo 1','LDA modelo 1',
'LR modelo E.4'),col=c('black','red','blue'),lty=c(1,2,3))

#AUC
auc1<-performance(pred,"auc")
```

Bibliografía

1. Agresti A. (2002). *Categorical data analysis*, Second edition. John Wiley & Sons Inc.
2. Anderson T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, Third edition. John Wiley & Sons Inc.
3. Cox D. R. and Oakes D. (1984). *Analysis of Survival Data*. Chapman & Hall book.
4. Cox D. R. and Snell E. J. (1989)*Analysis of Binary Data*. Chapman & Hall book.
5. Hand D. J. (2009). Mining the past to determine the future: problems and possibilities. *International Journal of Forecasting*, **25**, 441-451.
6. Hand D. J., Krzanowski W. J. and Crowder M. J. (2007). Optimal predictive partitioning. *Statistics and Computing*, **17**, 11-21.
7. Hand D. J. and Henley W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: A Review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **160**(3), 523-541.
8. Hand D. J., Mcconway K. J., and Stanghellini E. (1997). Graphical models of applicants for credit. *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 143-155.
9. Hosmer D. W. and Lemeshow S. (2000). *Applied Logistic Regression*, Second Edition. John Wiley & Sons Inc.
10. Hosmer D. W. and Lemeshow S. (1999). *Applied Survival Analysis, regression modeling of time to event data*. John Wiley & Sons Inc.
11. Johnson R. A. and Wichern D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall.
12. Klein J. P. and Moeschberger M. L (2003). *Survival Analysis, techniques for censored and truncated data*, Second Edition. Springer.
13. Krzanowski W. J. and Hand D. J. (2009). *ROC curves for continuous data*. Chapman & Hall book.
14. Lee E. T. and Wang J. W. (2003). *Statistical Methods for Survival Data Analysis*, Third Edition. John Wiley & Sons Inc.
15. Price S. (2009). Comments on "Mining the past to determine the future: problems and possibilities". *International Journal of Forecasting*, **25**, 452-455.

16. Rencher A. C. (2002). *Methods of multivariate analysis*, Second edition. Wiley Series in probability and mathematical statistics.

17. Thomas L. C. (2000). A survey of credit and behavioral scoring: forecasting financial risk of lending to costumers . *International Journal of Forecasting*, **16**, 149-172.

18. Thomas L. C., Edelman D. B. and Crook J. N. (2002). *Credit Scoring and its applications*, Monographs on mathematical modelind and computation, SIAM.

19. Stepanova M. and Thomas L. C. (2002). Survival analysis methods for personal loan data. *Operations Research*, **50**(2), 277-289.

20. Venables W. N. and Ripley B. D.(2002). *Modern applied statistics with S* . Springer.

99