UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

USO DE LA REGRESIÓN LOGÍSTICA MULTINOMIAL EN PROBLEMAS DE APLICACIÓN

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIO

P R E S E N T A:

ROBERTO SUÁREZ HERNÁNDEZ

DIRECTOR DE TESIS:
DRA. MARÍA DEL PILAR ALONSO REYES
2015

Ciudad Universitaria, D. F.

Índice general

Introducción

El presente trabajo busca realizar una aplicación de segmentación en un club deportivo en el que se desea clasificar a los clientes, según su perfil, en tres categorías:

- Cliente Leal: Es el que está siempre en nuestra ventana de información[1].
- Cliente Intermitente: Es el que abandonó el club al menos una vez y regresó.
- Cliente Desertor: Es el que abandonó el club una sola vez y jamás regresó.

Esto con el objetivo de detectar los posibles abandonos con anticipación y poder actuar de forma adecuada según el criterio del negocio.

Para llevar a cabo la clasificación se empleará la *Regresión Logística Multinomial*, se supondrá que los clientes tienen una distribución multinomial y se buscará explicar su clasificación con base en una serie de variables otorgadas por el club deportivo.

El primer capítulo tratará acerca de los estudios de mercado y los análisis estadísticos que se llevan a cabo en una consultora.

En el segundo capítulo se dará una breve explicación de la técnica estadística que se

[1] A partir de ahora, se llamará *ventana de información* a los 28 meses de información que se tienen disponibles para los clientes en la base de datos.

utilizará y algunos conceptos previos para su mejor entendimiento.

En el tercer capítulo se realizará la aplicación de la técnica seleccionada y se obtendrá la clasificación de clientes.

Por último se incluyen las conclusiones y los anexos referidos a lo largo del documento, también el anexo incluye el código utilizado, así como las salidas de éste.

Capítulo I

Estudios de mercado

Un estudio o investigación de mercado, .[es] el proceso de planificar, recopilar, analizar y comunicar datos relevantes acerca del tamaño, poder de compra y perfil de los consumidores y por otro lado la disponibilidad de los distribuidores, con la finalidad de ayudar a los responsables de mercadeo a tomar decisiones y a controlar las acciones en una situación de mercado específica".(4)[Thompson Ivan, El Estudio de Mercado].

Generalmente el objetivo de un estudio de mercado es incrementar las ventas del negocio, ya sea por la introducción a un mercado nuevo o por profundizar en un mercado en el que ya se está inmerso, se puede estudiar de forma directa a los clientes o la forma en que interactúan con la empresa teniendo como intención encontrar una relación entre sus preferencias y comportamientos con los productos o servicios que ofrece la entidad que los estudia comprendiendo así las necesidades que falta cubrir o los aspectos a mejorar en la relación con los consumidores.

I.1. Tipos de estudios de mercado

Existen tres tipos básicos de estudios de mercado teniendo en cuenta los objetivos que persiguen las investigaciones.(5)[Jáuregui Alejandro, 7 Elementos Básicos en Metodología de investigación de Mercados].

Tipos de estudio o investigación:

1. Descriptiva
2. De causa y
3. De predicción

Investigación descriptiva

Es aquella que busca definir claramente un objeto, el cual puede ser un mercado, una industria, una competencia, puntos fuertes o débiles de empresas, algún tipo de medio de publicidad o un problema simple de mercado.

En una investigación descriptiva, el equipo de trabajo buscará establecer el «Qué» y el «Dónde», sin preocuparse por el «por qué». Es el tipo de investigación que genera datos de primera mano para realizar después un análisis general y presentar un panorama del problema.

Investigación de causa

Es aquella que busca explicar las relaciones entre las diferentes variables de un problema de mercado, es el tipo de investigación que busca llegar a los nudos críticos y buscará identificar claramente fortalezas y debilidades explicando el «por qué» y

el «cómo» suceden las cosas.

La investigación de causa, generalmente se aplica para identificar fallas en algún elemento de mercadeo, como el diseño de un empaque, algún elemento en las preferencias de los consumidores que genere alguna ventaja competitiva, alguna característica de los productos que no guste a los consumidores, etcétera.

Investigación de predicción

Es aquella que busca proyectar valores a futuro; buscará predecir variaciones en la demanda de un bien, niveles de crecimiento en las ventas, potencial de mercados a futuro, número de usuarios en algún intervalo de tiempo, comportamiento de la competencia, etcétera. En cualquier estudio predictivo, generalmente se deberán tener en cuenta elementos como el comportamiento histórico de la demanda, cambios en las estructuras de mercado, aumento o disminución del nivel de ingresos.

La investigación predictiva, es la más complicada e interesante y es la que realmente puede hacer diferencia entre el éxito y el fracaso de empresas en el largo plazo, acertar en el comportamiento de un mercado a futuro, es claramente la mejor manera de garantizar estabilidad.

I.1.1. Metodologías de investigación

Sin importar el tipo, generalmente la metodología de trabajo es igual para cualquier investigación y puede resumirse en siete pasos, los cuales son:(5)[Jáuregui Alejandro, 7 Elementos Básicos en Metodología de investigación de Mercados].

1. **Captación de datos:** El primer paso será siempre la recolección de información primaria que pueda servir como base de análisis. Existen diferentes tipos de fuentes: Encuestas propias, estudios históricos, registros de empresas, cámaras de comercio, investigaciones de campos, datos internos de la empresa, historiales de venta etcétera, el tipo de información a recolectar dependerá de los objetivos que persigue la investigación.

2. **Muestreo:** En ocasiones en que no es posible o conveniente realizar un censo (analizar a todos los elementos de una población), se selecciona una muestra, entendiendo por tal una parte representativa de la población.

   El muestreo es una herramienta de la investigación científica, cuya función básica es determinar que parte de una población debe examinarse, con la finalidad de hacer inferencias sobre esta.

   La muestra debe lograr una representación adecuada, en la que se reproduzca de la mejor manera los rasgos esenciales de la población que son importantes para la investigación. Para que una muestra sea representativa, y por lo tanto útil, debe reflejar las similitudes y diferencias encontradas en la población, es decir ejemplificar las características de ésta.(18)[Anónimo, http://www.estadistica.mat.uson.mx/ Material/ elmuestreo.pdf].

3. **Experimentación:** Consiste en manejar uno o varios elementos (precio, cantidad, calidad, publicidad) con el fin de generar datos acerca de reacciones del mercado, busca identificar el impacto de cada variable sobre el comportamiento de éste. Un ejemplo se da cuando se hacen promociones especiales en algunas zonas, para saber si el impacto es positivo o negativo para el mercado y la empresa y dados los resultados aplicar dichas promociones en general o no hacerlas.

4. **Análisis del comportamiento del consumidor:** Investiga el «por qué» las personas varían sus preferencias, aceptan o rechazan determinados productos o algunas marcas. Generalmente estas investigaciones se basan en factores de conducta y psicológicos.

5. **Análisis de la información:** Es aplicar técnicas matemáticas para estimar las relaciones existentes, con base en datos preliminares o variables aisladas. Éste análsis es primordial para la toma de decisiones con un fundamento estadístico, que respalde asi mismo la experiencia en el negocio y juntos puedan convertirse en una herramienta de gran utilidad para la empresa.

6. **Predicción o informe:** Consiste en estimar valores (investigación descriptiva), o predecir valores (investigación predictiva), que serán los resultados de la investigación y la base para obtener conclusiones.

7. **Simulación:** Consiste en modelar los resultados de mercado para producir datos artificiales y evaluar diversas alternativas. Las nuevas tecnologías han llegado incluso a simular mercados por medios virtuales.

Cabe señalar que la utilización de estos siete pasos depende del tipo de investigación que se quiera realizar, pues pueden existir situaciones en las que no sea necesario o de gran relevancia el realizar algunos de ellos o en otro caso deba ponerse un especial énfasis en alguno pudiendo convertirse incluso en el objetivo principal de la investigación.

I.2. Segmentación de clientes

La segmentación investiga el mercado con objeto de encontrar la existencia de conjuntos de consumidores homogéneos y facilita el desarrollo de las actividades de

mercadeo.

En la segmentación del mercado incide más de un criterio y los consumidores responden a un perfil que aglutina una serie de características por lo que un segmento estará definido por más de una característica. El problema consiste en encontrar un segmento óptimo, resultante del cruce de varios criterios, que mejor discrimine el comportamiento de los consumidores. (1)[Lazzari Luisa, La Segmentación de Mercados Mediante la Aplicación de Teoría de Afinidad].

Los clientes son diferentes entre sí, tienen necesidades diferentes y el valor de unos y otros es diferente, por lo tanto es muy importante establecer una segmentación utilizando técnicas estadísticas que permitan analizar sus datos y con esto poder realizar un esfuerzo específico para cada clase de clientes.

Hay diversos tipos de segmentación:

Según el objetivo

- **Segmentación táctica**
- **Segmentación estratégica**

Estos tipos de segmentación se explican ampliamente en la sección I.1.2.

Según la dimensión de cliente

- **Dimensión del valor contra necesidad:** Cada cliente presenta un valor actual –sus compras– y unas necesidades –o valor potencial–, estimado a través de estudios de mercado, encuestas sectoriales, sociodemografía, estadio de vida del cliente.

- **Valor de vida del cliente:** Es la proyección del valor de cliente a futuro, en función de su ciclo de vida.
- **Dimensión geográfica:** En los negocios basados en redes de establecimientos, es clara la importancia de la relación espacial entre el cliente y el punto de venta.
- **Dimensión comportamental:** Se entiende como el análisis de las pautas de navegación en el ámbito del comercio electrónico, con objeto de conocer al cliente y personalizar la relación con él.
- **Dimensión relacional:** Las interacciones entre la empresa y los clientes, más allá de las propias de la prestación del servicio, son claves a la hora de generar vínculos entre ambos.
- **Dimensión social:** El auge de las redes sociales online ha sacado a la luz realidades sociales conocidas, pero poco explotadas desde la segmentación de clientes; el mejor prescriptor de un producto es un amigo, pariente, alguien en las redes sociales. Existen personas con alta capacidad de prescribir, influir en su red social. Igualmente, las hay que tienen gran cantidad de relaciones sociales.

I.2.1. Segmentación táctica

Se puede definir como toda tarea de análisis de características y comportamientos de clientes, orientadas a la solución de un problema único y concreto.

Aun cuando no es la aplicación única, la gran mayoría de segmentaciones tácticas de clientes se enfocan en la optimización de campañas de mercadeo directo o relacional.

Las técnicas analíticas permiten la optimización de las campañas y existen cinco grandes tipos(3)[Córdoba Guillermo, Segmentación de clientes. Una propuesta de clasificación (II). La segmentación táctica]:

- **Retención:** Identificación de los clientes más rentables, estimación de la cuota del cliente, simulación de sendas de abandono y alertas ante eventos de riesgo de abandono.

- **Recuperación de desertores:** Son campañas altamente dependientes del motivo del abandono, y a menudo requieren una investigación de estas motivaciones de los clientes perdidos. Es clave conocer el valor de vida o valor futuro previsto del cliente, para dimensionar la oferta de recuperación, y actuar inmediatamente tras la deserción. Obviamente, siempre es preferible trabajar en la retención de un cliente que tener que hacerlo en su recuperación.

- **Venta cruzada:** Es definitivo el análisis de potencial de demanda por categoría. En el mercado minorista[1] los análisis de asociación permiten generar conjuntos de compra y patrones secuenciales de compra. Los motores de recomendación suponen una variante de venta cruzada donde la campaña se lanza en línea, durante el proceso de compra. Son campañas muy rentables en compañías o grupos altamente diversificados.

- **Mejora de ventas:** De nuevo es clave estimar correctamente la demanda total del cliente en la categoría, buscando maximizar la cuota de éste. En distribución minorista, suelen dividirse en acciones de incremento de factura media (aumento del ingreso por venta unitaria) y acciones de incremento de

[1]Cualquier mercado destinado específicamente a la realización de transacciones de valores de un tamaño relativamente pequeño y, por consiguiente, orientado al inversor individual. http://www.economia48.com/spa/d/mercado-minorista/mercado-minorista.htm

frecuencia (aumento en el número de ventas), a menudo a partir de análisis RFM –Recencia, Frecuencia, valor Monetario-

- **Captación de nuevos clientes:** El potencial de demanda se estima mediante la búsqueda de «gemelos» –clientes similares a los que resultan más rentables- o modelación sociodemográfica –modelos predictivos de demanda basados en características sociodemográficas, generalmente provenientes de fuentes públicas como censos, padrones, estudios sectoriales-. Este potencial de demanda indicará el valor esperado de cada cliente potencial y, por tanto, los recursos que conviene invertir en su captación.

I.2.2. Segmentación estratégica

Consiste en la reducción de toda la complejidad de los datos de los clientes, que pueden ser miles o millones de casos con cientos de datos, en un resumen donde:

- Los clientes se agrupan en un número reducido de segmentos
- Las variables se reducen a una sola etiqueta descriptiva del segmento, por ejemplo, de industria o geográficas, estas etiquetas sintetizan la gran riqueza de datos que han configurado el segmento, por ejemplo, cada una de las transacciones del cliente con todo detalle, sus características sociodemográficas, su distancia al punto de venta, los canales de compra usados, etcétera.

La segmentación estratégica permite hacer realidad la visión centrada en el cliente, fijando objetivos exclusivamente en éste. Si, por ejemplo, una entidad financiera busca «reducir un 5 % la tasa de abandono de los clientes históricos altamente vinculados en pasivo» entonces es necesaria una segmentación entratégica de clientes, no lo será tanto si los objetivos se plantean en términos de producto, por ejemplo

«incrementar el pasivo captado en un 5 %»

Una segmentación estratégica alcanza el éxito cuando es aceptada y utilizada por los responsables de negocio, convirtiéndose en el mapa conceptual de clientes que toda la organización acepta como punto de partida.

Para lograrlo, debe alinearse con los objetivos corporativos y la visión del problema de estos decisores de negocio. Además, requiere por supuesto capacidades de análisis avanzado de datos, como técnicas de minería de datos. Una metodología estándar sería:

1. **Análisis del negocio:** Reuniones con los responsables de negocio y futuros usuarios del modelo -dirección comercial, territorial, de mercadeo- en torno a:
    - Consenso de objetivos de la segmentación.
    - Exploración de las tipologías de clientes, conscientes o intuitivas, que la organización maneja.
2. **Definición de variables:** Se seleccionan por su relevancia, el conjunto definitivo se filtrará por criterios técnicos en la fase de modelación.
3. **Presentación al departamento técnico:** Se presenta el conjunto de variables, acuerdo sobre el modo de entrega y las tareas de las partes.
4. **Modelación:** Es la fase de análisis estadístico propiamente dicho.
5. **Validación de negocio:** Se analizan los resultados prelimirares entre los analistas y los responsables de negocio.
6. **Resultados:** Se presentan los resultados y se realiza la validación definitiva.

7. **Explotación:** La segmentación estratégica se plasma en objetivos estratégicos a través del *plan de mercadeo relacional* y el *plan de contactos*. El primero constituye el plan de mercadeo centrado en clientes, fijando los objetivos y líneas básicas de relación con cada segmento. El segundo apunta las tácticas de relación, contactos a realizar con cada segmento.

I.2.3. Técnicas utilizadas en la segmentación estratégica

Se puede dividir las técnicas utilizadas en tres grandes grupos:(2)[Rodríguez Jorge, Segmentación de Clientes]:

- Análisis de conglomerados
- Análisis de factores
- Análisis basado en respuestas

A continuación se explica cada una de las técnicas de forma específica:

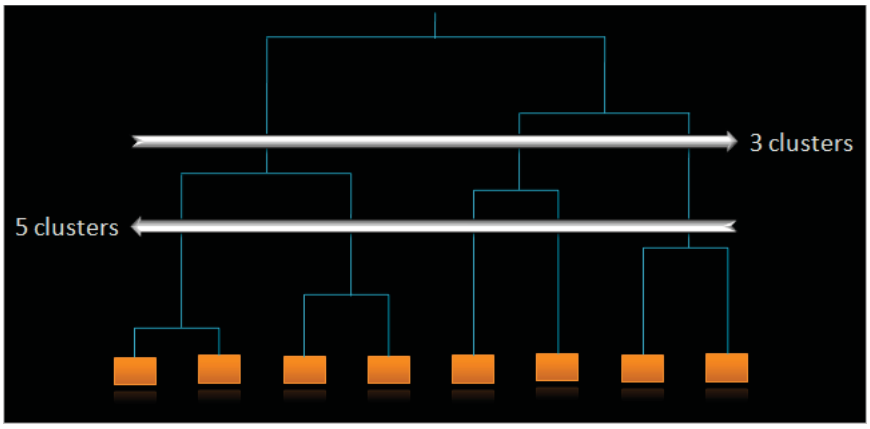
Análisis de conglomerados

Es una técnica de análisis multivariado que tiene como objetivo identificar los grupos de observaciones con características similares que existen en un conjunto, de al menos dos grupos. En mercadotecnia, por ejemplo, se puede tener una muestra de clientes o consumidores con distintas características que esté formada por un pequeño número de grupos donde los sujetos dentro de cada uno tengan propiedades semejantes. Esto podría ser importante para emplear una estrategia de mercado adecuada para cada tipo de consumidor. Este no es el único caso en el que se emplea el análisis de conglomerados, existe una infinidad de ellos y áreas de aplicación en los que resulta útil.

- **Métodos Jerárquicos:** Estos métodos tienen por objetivo unir grupos para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.(10)[UG, España, Ampliación de Análisis de Datos Multivariantes]. Existen dos tipos:

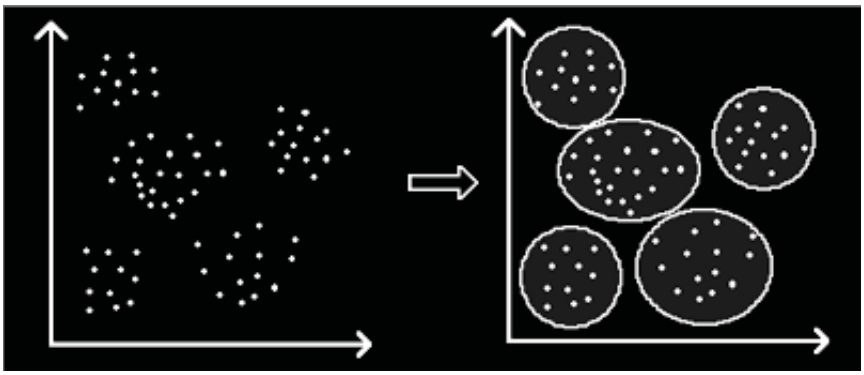
  - Aglomerativos: Se inicia asumiendo que cada observación representa un grupo, agrupando después los más cercanos hasta hacer un único grupo.
  - Divisivos: Este caso es contrario al anterior y se inicia con un único grupo separando a los sujetos más lejanos entre ellos.

  En ambos se requiere conocimiento del área de aplicación para decidir con base en la experiencia un número aproximado de grupos que se deben o desean encontrar.



- **Métodos No Jerárquicos:** En estos métodos los datos se dividen en n particiones o grupos donde cada partición representa un conjunto. Opuestamente a los métodos jerárquicos el número de conjuntos debe conocerse a priori. (11)[Rincón, Mendoza, Morán, Vega. Tecnología de Apoyo a la Logística]. En este caso se utiliza algoritmos para la segmentación como:

- K medias: Se definen puntos centrales para cada uno de los grupos definidos y mediante un proceso iterativo estos centros van cambiando, dejando fuera o incorporando las observaciones en cada iteración.



Análisis de factores

- **Análisis Factorial:** Es una técnica de reducción de datos que sirve para encontrar grupos homogéneos de variables a partir de un conjunto numeroso de variables. Esos grupos homogéneos se forman con las variables que correlacionan mucho entre sí y procurando, inicialmente, que unos grupos sean independientes de otros.(6)[UCM, España, Análisis Factorial].

  Un factor se puede definir como una variable que no puede ser medida directamente,sin embargo, sí a través de un conjunto de variables observadas correlacionadas entre sí, las cuales lo conforman.

  Se busca explicar la estructura de la correlación que existe en un conjunto de variables de tal forma que aquellas que esten más correlacionadas entre sí puedan separarse en grupos, esperando que el número de grupos sea mucho menor que el número de variables originales.

- **Análisis de Componentes Principales:** Es una técnica estadística de síntesis de la información, o reducción de la dimensión. Es decir, ante un banco de

datos con muchas variables, el objetivo será reducirlas a un menor número perdiendo la menor cantidad de información posible.(7)[Terrádez Manuel, Análisis de Componentes Principales].

El análisis de componentes principales funciona particularmente en un conjunto de variables correlacionadas transformando éste en un conjunto de menor número de variables que además no están correlacionadas.

El número de componentes dependerá del poncentaje de varianza que se decida conservar, el cubrir este requerimiento depende directamente de la persona que realiza el análisis.

Análisis basado en respuetas

- **Análisis Discriminante:** Es una técnica estadística que ayuda a identificar las características que diferencian a dos o más grupos y a crear una función capaz de distinguir con la mayor precisión posible a los miembros de uno u otro grupo.(8)[UCM, España, Análisis Discriminante].

  Los grupos deben estar definidos previamente y esta técnica funciona tanto para *discriminar* como para *clasificar*.

  Discriminar implica describir las características que diferencian a los grupos de una población para poder con esto encontrar los factores discriminantes que permitan separar los grupos de la mejor forma posible.

  Clasificar significa que los sujetos de la población serán asignados a los grupos, creando con el análisis, reglas que permitan hacerlo de forma óptima.

- **Regresión Logística:** Es un tipo especial de regresión que se utiliza para explicar y predecir una variable categórica en función de varias variables independientes que a su vez pueden ser cuantitativas o cualitativas.(9)[Castrejón Sandoval Osiris, Diseño y Análisis de Experimentos con Statistix].

La variable categórica representa los grupos para los cuáles se pretende discriminar utilizando las variables independientes para calcular la probabilidad de pertenencia de cada individuo a cada uno de los grupos, asignando a éste al grupo cuya probabilidad resulte ser mayor.

En el siguiente capítulo se profundiza en esta técnica ya que es justamente la empleada en este análisis.

Capítulo II

Regresión logística

En el análisis de regresión generalmente se busca modelar el comportamiento de una variable númerica que representa algún valor de tipo continuo, por ejemplo el precio futuro de un instrumento bursátil, el crecimiento en la producción de una empresa, etcétera. Cuando la variable que se está modelando no puede describirse como una de tipo continuo o simplemente los valores que toma no son necesariamente numéricos se deben emplear otro tipo de técnicas estadísticas, una de ellas es la regresión logística que modela la probabilidad de pertenencia a un grupo o categoría determinado con base a las relaciones lineales que presentan las variables que buscan explicar la heterogeneidad de estas clases.

II.1. Conceptos previos

II.1.1. Tipos de variables

Se llama variable a aquel elemento de un algoritmo, fórmula o proposición que representa un elemento no especificado de un determinado conjunto(18)[Anónimo http://www.tiposde.org/general/35-tipos-de-variables/].

- **Variable dependiente:** Es aquella cuyo valor depende de otras. Por ejemplo: El dinero que se debe pagar en un estacionamiento *depende* del tiempo que se haya estado en él.

- **Variable independiente:** Es aquella cuyo valor no depende de ninguna otra. Por ejemplo: En el caso del estacionamiento el tiempo pasa de forma *independiente*.

- **Variable cuantitativa:** Se expresan mediante un número, el cual, representa una cantidad. Esta variable tiene dos tipos:

  - Continua: Puede tomar cualquier valor real, tiene un dominio *teóricamente*[1] infinito. Por ejemplo: medidas como peso y estatura.

  - Discreta: Puede tomar solamente un número finito de valores. Por ejemplo: número de hermanos, número de hijos.

- **Variables cualitativas:** Representan cualidades y su medición no es numéricamente. Estas variables son de dos tipos:

  - Ordinal: A pesar de que no se tienen números se puede asignar un orden para los valores. Por ejemplo: Al opinar sobre la calidad de un producto, las opciones podrían ser malo, regular y bueno sabiendo que de algún modo malo es menor a regular y éste es menor a bueno.

  - Nominal: No existe un orden ni una escala objetiva de comparación. Por ejemplo: Al registrar el color favorito de dos personas, una contesta azul y la otra rojo, en este caso no se puede decir que uno sea mejor que otro.

[1] Esto es porque para poder realizar la captura de los datos se deben redondear estos valores lo cual los discretiza además de que sería imposible tener instrumentos de medición que obtengan los datos con una precisión infinita.

II.1.2. Momios

Los momios son una medida de asociación entre la probabilidad de que ocurra un determinado evento y de que este no suceda(12)[Szumilas, Magdalena. Explaining Odds Ratios. http://www.ncbi.nlm.nih.gov /pmc/articles/PMC2938757/]. Está representado de esta forma:

$$m(x)=\frac{p(x)}{1-p(x)}$$

Donde x es un evento, $m(x)$ es el valor del momio y $p(x)$ es la probabilidad de ocurrencia. $m(x)$ es una función continua que depende de $p(x)$ que pertenece al intervalo [0,1] se puede observar que:

$$\lim_{p(x)\to 0} \frac{p(x)}{1-p(x)} = 0$$

y

$$\lim_{p(x)\to 1^-} \frac{p(x)}{1-p(x)} = \infty$$

Por lo que el rango que toma este cociente es $[0,\infty)$ donde si $m(x)<1$ significa que el evento x tiene poca probabilidad de ocurrir con respecto a otros, $m(x)=1$ quiere decir que tiene la misma probabilidad de ocurrir que los demás y por último, si $m(x)>1$ significa que tiene mayor probabilidad de ocurrir con respecto a otros.

II.1.3. Modelos lineales generalizados

El modelo de regresión lineal simple, plantea que:

$$Y=\beta_0+\beta_1 X+\varepsilon$$

Donde Y es la variable dependiente, ésta es cuantitativa continua, β_0 es el intercepto, β_1 representa a los coeficientes de X que es la variable explicativa y por último $\varepsilon \sim N(0,\sigma^2)$.[2]

Un modelo lineal generalizado es aquel que vincula la variable respuesta con las variables explicativas y para ello se deben considerar tres componentes:

- **La componente aleatoria:** Y es una variable aleatoria con observaciones independientes e idénticamente distribuidas que pertenecen a una distribución de la familia exponencial, por ejemplo: Normal, Poisson, Gamma, Binomial.
- **La componente sistemática:** Es la parte no aleatoria del modelo, involucra a las covariables o variables explicativas y describe la relación que existe entre ellas y la de respuesta, en este caso es lineal.
- **La función *liga*:** Es la que vincula la esperanza de la distribución con la componente sistemática.

En general, tomando en cuenta estos tres componentes, el modelo se ve de esta forma:

$$g(\mu)=\sum_{i=0}^{n} \beta_i x_i$$

Y despejando se tiene:

$$\mu=g^{-1}(\sum_{i=0}^{n} \beta_i x_i)$$

Existen diversas funciones *liga* que generan distintos modelos de regresión, por ejemplo (18)[Anónimo http://academic.uprm.edu/rmacchia/agro6998/Modeloslinealesgeneralizados.pdf]:

[2]Épsilon es el ruido aleatorio

Distribución	***Liga***	**Función**
Normal	Identidad	$g(\mu)=\mu$
Binomial	Logarítmica	$g(\mu)=\ln(\frac{\mu}{1-\mu})$
Poisson	Logarítmica	$g(\mu)=\ln(\mu)$
Exponencial	Recíproca	$g(\mu)=\frac{1}{\mu}$
Normal	Probit	$g(\mu)=\Phi^{-1}[\mathbb{E}(\mu)]$

Lo que se busca con los *modelos lineales generalizados* es tratar de ampliar esa restricción sobre la variable respuesta donde ésta es una función de la media[3], en estos modelos se encuentra la *regresión logística* que se explicará con mayor detalle en la siguiente sección.

II.2. ¿Qué es la regresión logística?

El modelo logit es el complemento natural del de regresión en el caso en que las variables son categóricas. Cuando estas variables se presentan en la parte explicativas se puede lidiar fácilmente con ella mediante la introducción de variables dummy[4] (0,1) pero cuando es la explicada la que pertenece a este tipo, el modelo de regresión simple no funciona correctamente.(14)[Cramer, J. S. The logit model].

Como en cada uno de los métodos de regresión, que buscan describir la relacion que existe entre una respuesta y una o más variables explicativas, la regresión logística es utilizada para explicar una variable categórica binaria o politómica con base en variables explicativas no necesariamente continuas.

[3] $\mathbb{E}[Y]=\mu$

[4] También conocidas como variables indicadoras, sirven para expresar de forma numérica la presencia o ausencia de alguna cualidad que se quiera considerar en el modelo

En el modelo bivariado representa un experimento en el que cada uno de los sujetos tiene dos resultados posibles, sobrevivir (1) o morir (0), ganar (1) o perder (0), etcétera, debido a que la variable estudiada es categórica se busca encontrar la probabilidad de que el sujeto presente (1) o (0) en la característica estudiada.

Se puede escribir el modelo binario de la siguiente forma:

$$g(\mu)=\beta_0+\beta_1 X$$

Donde g es la función a la que se llama *liga*[5] que en el caso de la regresión logística es:

$$g(\mu)=\ln(\tfrac{\mu}{1-\mu})$$

por lo que despejando se tiene:

$$e^{\beta_0+\beta_1 X} = \tfrac{\mu}{1-\mu}$$

$$\mu = (1-\mu)e^{\beta_0+\beta_1 X}$$

$$\mu + \mu e^{\beta_0+\beta_1 X} = e^{\beta_0+\beta_1 X}$$

$$\mu(1+e^{\beta_0+\beta_1 X}) = e^{\beta_0+\beta_1 X}$$

$$\mu=\frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

que es la forma específica del modelo de regresión logística.

La principal objeción del modelo de regresión logística es que no hay restricciones para que el rango de la función sea solamente (0,1) sin embargo esta forma del

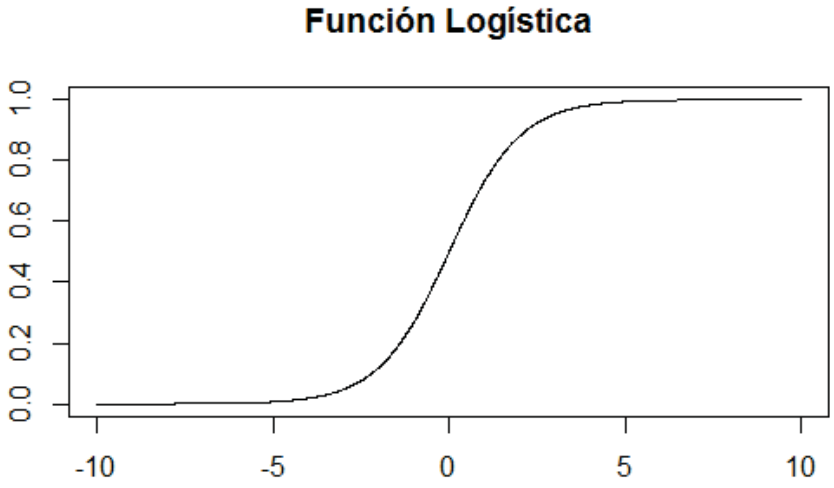
[5]En el caso de la regresión lineal simple, esta función es la identidad.

modelo cumple con este requerimiento.

No hay una justificación intuitiva directa para utilizar esta función sin embargo una explicación es que la función inversa es el cociente de momios además de que considerando $p(X)=\mu$ y $q(X)=1-\mu$ la derivada con respecto a X es:

$$\frac{\mathrm{d}\mu}{\mathrm{d}X} = p(X)q(X)\beta$$

Que tiene el mismo signo que β si β es positivo crece monótonamente de cero a uno de la misma forma que X en toda la recta real que es precisamente lo que se necesita en una función de probabilidad, la cual tiene un punto de inflexión en cero y el valor de la función en 0 es 0.5 lo cual habla de que existe simetría con respecto al punto medio.



Hay ciertas características que se deben resaltar:

- La variable respuesta será continua y acotada en el intervalo (0,1), propiamente lo que se obtiene del modelo es una probabilidad de pertenencia a un grupo y la más alta es la que define la clasificación.
- La función *liga* es el logarítmo natural del *momio* de la categoría.
- Puede ser utilizado como un cuantificador de riesgo[6]

II.2.1. Estimación por máxima verosimilitud

Se considera una sucesión independiente de observaciones cuya distribución conjunta es:

$$\prod_{j=1}^{N} P(X_j = 1|\theta)^{X_j} P(X_j = 0|\theta)^{1-X_j}$$

Donde X representa la variable aleatoria binaria y θ es un vector de parámetros desconocidos, éste es un producto conocido como función de verosimilitud y aunque esta es su forma para facilitar su uso suele aplicarse logarítmo natural de ambos lados de la igualdad y utilizando las propiedades de ésta función se obtiene la llamada función de log-verosimilitud que es la siguiente:

$$logL(\theta) = \sum_{s\in 1} \ln(p(X_s, \theta)) + \sum_{s\in 0} \ln(q(X_s, \theta))$$

Ésta ecuación se deriva e iguala a cero, resolviendo mediante métodos numéricos como *Newton-Raphson* se obtienen los valores adecuados de θ.

[6]Si se calculan *momios* se pueden comparar algunas características presentes en las categorías propuestas.

II.3. Regresión logística multinomial

Ésta es una generalización del modelo binario, se considera ahora una variable dependiente con más de dos categorias.

Los modelos multinomiales aplican para cualquier número de estados, se consideran J posibles grupos con índices *j=1,2,3,...,J* éstos son disjuntos y exhaustivos, es decir, cubren todas las posibilidades incluso si es necesario puede incluirse un espacio residual en donde haya estados difíciles de clasificar.
Sea J el número de categorías de la variable Y y $\{p_1, p_2, ..., p_J\}$ las probabilidades de las respectivas respuestas.
Obviamente se requiere que:

$$p_j(X, \theta) \geq 0$$

$$\sum_{j \in \{1,2,..J\}} p_j(X, \theta) = 1$$

La distribución de probabilidad del número de observaciones de las J categorías, con n observaciones independientes, sigue una distribución multinomial que modela la probabilidad de cada una de las posibles maneras en que n observaciones pueden repartirse entre las J categorías. Al ser una escala de medida nominal, el orden entre las categorías es irrelevante.(13)[Marín Diazaraque, Juan Miguel. http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/Categor/Tema5Cate].

Se toma una categoría como respuesta base, por ejemplo la última categoría (J) y se define un modelo, idéntico al bivariado, con respecto a ella:

$$\ln(\frac{p_j}{p_J}) = \beta_{0,j} + \beta_{1,j} X$$

donde *j=1,...,J-1*

El modelo resultande tiene *J-1* ecuaciones con sus propios parámetros, donde la probabilidad de forma explícita para cada individuo i se calcula así:

$$\pi_{ij} = \frac{e^{x_i\beta_j}}{\sum_{j=1}^{J} e^{x_i\beta_j}}$$

Al tomar como base la categoría J se supone $\beta_J = 0$ lo que da como resultado:

Para $0 < j < J$

$$P[y_i = j|x_i] = \frac{e^{x_i\beta_j}}{1+\sum_{j=1}^{J-1} e^{x_i\beta_j}}$$

Y para $j = J$

$$yP[y_i = 1|x_i] = \frac{1}{1+\sum_{j=1}^{J-1} e^{x_i\beta_j}}$$

Para un modelo con k variables independentes se tiene un total de $(k+1)(J-1)$ parámetros que deben ser estimados y cabe aclarar que cuando J=2 el modelo es equivalente al caso binario de la regresión logística.

II.4. Prueba de la significancia en la regresión logística

Una de las pruebas para los coeficientes de la regresión logística es la de Wald, la cual se obtiene al comparar el estimador por máxima verosimilitud del coeficiente $\hat{\beta}_i$ con la estimación de su error estándar.(15)[Hosmer, David W.; Lemeshow, Stanley. Applied logistic regression].

El cociente resultante, bajo la hipótesis de que $\beta_i = 0$ se distribuye normal estándar:

$$W = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$$

La prueba arroja un *valor p* que representa:

$$\mathrm{P}[|\mathrm{z}| > W]$$

El paquete R utiliza una versión alternativa de la prueba de Wald, la cual se aproxima a una distribución F o *Ji-cuadrada* (16)[Lumley, Thomas. regTermTest http://cran.r-project.org/web/packages/survey/survey.pdf].

Capítulo III

Aplicación

En éste capítulo se realizará una aplicación para los modelos estudiados, en específico para el de regresión logística multinomial.

El ejemplo consiste en segmentar una tabla que incluye información de 280 clientes en un club deportivo, se busca hacer diferencia entre los clientes que han durado más de 1 año, entre 6 y 12 y menos de 6 meses, para los cuales se realiza la siguiente clasificación:

- Cliente leal (2) 99 clientes.
- Cliente intermitente (1) 86 clientes.
- Cliente desertor (0) 99 clientes.

III.1. Variables

La tabla incluye los siguientes datos:

- Número de socio

- Sexo
- Edad
- Costo de inscripción
- Mes
- Año
- Forma de pago
- Tipo de membresía
- Tipo de pago
- Monto de pago

En esta tabla se tiene un registro de socios y cada uno de sus pagos quedando registrado únicamente el mes y el año de la realización de éste.

Con base en esto se puede crear una tabla consolidada en la que se incluya como elementos extra:

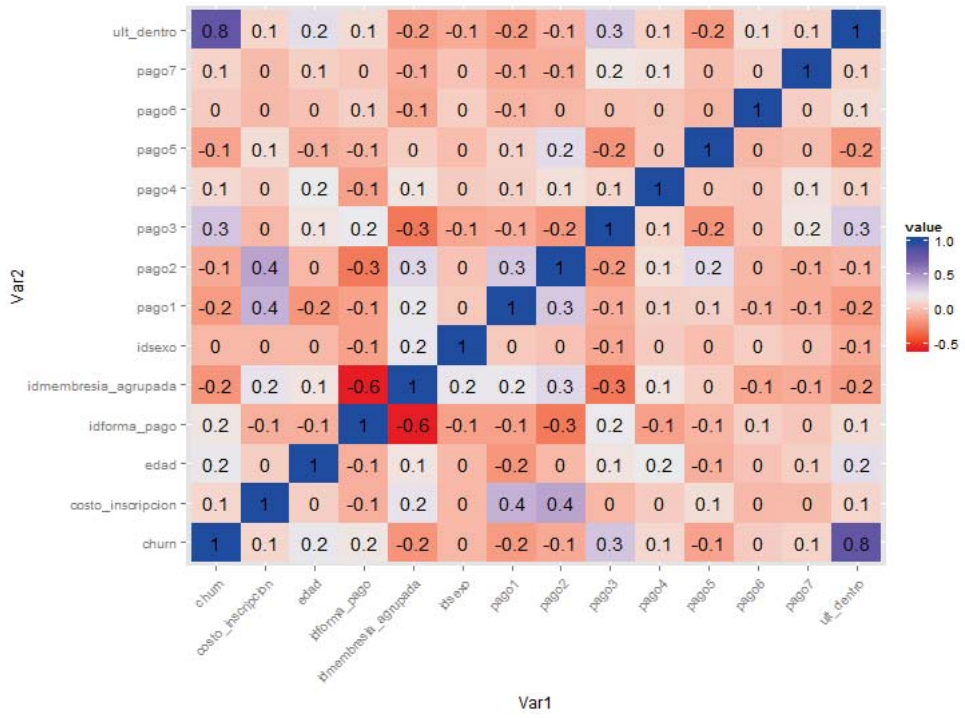
- Gasto mensual para cada tipo de pago
    - Pago1. Monto de inscripción
    - Pago2. Mensualidad
    - Pago3. Anualidad
    - Pago4. Recargos
    - Pago5. Reingreso
    - Pago6. Casillero

- Pago7. Clases extras

- Duración del último ingreso en el club

- Tipo de cliente

Se presenta las correlaciones de las variables:



Se puede observar que todas las variables presentan un grado de correlación con tipo de cliente de entre -20 y 20 % a excepción de *Pago6* y *Sexo*, la primera será conservada a pesar de su correlación por recomendación de las personas que otorgaron la información, mientras que la segunda será eliminada, se intentará aplicar el modelo de regresión multinomial para estos datos.

III.2. Comandos en R

Se utilizará esta fórmula en R:

```
formula <- Tipo_de_cliente ~ 0 + costo_inscripcion + forma_pago +
membresia + edad + pago1 + pago2 + pago3 +
pago4 + pago5 + pago6 + pago7 + ult_duracion
```

Se utiliza una intersección en cero pues un cliente que tiene duración 0 nunca ha estado en el club.

Se aplica el modelo:

```
test <- multinom(formula,data=BASE)
```

Ejecutando este comando en el paquete R se tiene como resultado los coeficientes para cada una de las variables y para los valores 1 y 2 de tipo de cliente, esto debido a que se tomo como base la categoría 0, para consultar los resultados con mayor detalle se puede revisar el apéndice A en donde se encuentra el resultado completo.

El modelo que resulta es:

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_1 * x_1 + ... + \beta_1 2 * x_1 2$$

$$logit(Prob(ClienteDesertor)) = 0{,}0001346 * costo\_inscripcion$$

$-0{,}0004327 * idforma\_pago$

$-0{,}0678837 * idmembresia\_agrupada$

$-0{,}0180553 * edad$

$-0{,}0006617 * pago1$

$-0{,}0005128 * pago2$

$+0{,}0000545 * pago3$

$+0{,}0003451 * pago4$

$+0{,}0006417 * pago5$

$-0{,}0015358 * pago6$

$-0{,}0001458 * pago7$

$+0{,}1406310 * ult\_dentro$

Para la categoría 1:

Categoría 1	**Coeficiente estimado**	**Error estándar**	**W**	$P > \|z\|$
costo_inscripcion	0.0001346	0.0000706	3.6357910	0.0569010
idforma_pago	-0.0004327	0.0064794	0.0012526	0.9717800
idmembresia_agrupada	-0.0678837	0.0415259	0.0001086	0.9916900
edad	-0.0180553	0.0111082	5.5233660	0.0190010
pago1	-0.0006617	0.0001701	159314.70	0.0000002
pago2	-0.0005128	0.0001949	1437383.00	0.0000002
pago3	0.0000545	0.0000182	980220.80	0.0000002
pago4	0.0003451	0.0002631	76486.41	0.0000002
pago5	0.0006417	0.0001687	15.37964	0.0000953
pago6	-0.0015358	0.0009743	0.0177448	0.8940600
pago7	-0.0001458	0.0001949	6.9242130	0.0086641
ult_dentro	0.1406310	0.0274631	0.0005366	0.9815200

Para la categoría 2:

Categoría 2	**Coeficiente estimado**	**Error estándar**	**W**	$P > \|z\|$
costo_inscripcion	0.0002293	0.0000878	3.6357910	0.0569010
idforma_pago	-0.0261062	0.0084956	0.0012526	0.9717800
idmembresia_agrupada	-0.2337075	0.0231485	0.0001086	0.9916900
edad	-0.0727727	0.0146938	5.5233660	0.0190010
pago1	-0.0001298	0.0001150	159314.70	0.0000002
pago2	-0.0006362	0.0002329	1437383.00	0.0000002
pago3	0.0000191	0.0000216	980220.80	0.0000002
pago4	0.0005732	0.0003555	76486.41	0.0000002
pago5	-0.0005818	0.0003733	15.37964	0.0000953
pago6	-0.0023575	0.0010040	0.0177448	0.8940600
pago7	-0.0002038	0.0002350	6.9242130	0.0086641
ult_dentro	0.4629434	0.0152635	0.0005366	0.9815200

Desviación Residual	278.3561
AIC	326.3561

Al calcular la probabilidad utilizando los valores mínimos, medios y máximos, se obtienen las siguientes probabilidades:

	Máximo	**Promedio**	**Mínimo**
costo_inscripcion	20521.7	2718.17	0
idforma_pago	130	40	1
idmembresia_agrupada	12	5	1
edad	73	40.67857	12
pago1	19503	1151.027	0
pago2	6727.769	1232.641	0
pago3	78020.33	10868.36	0
pago4	7923	248.6322	0
pago5	11948	471.1738	0
pago6	2500	33.20749	0
pago7	11025.5	458.7153	0
ult_dentro	28	12.85	1
Probabilidad 1	0.93	0.25	0.29
Probabilidad 2	0.07	0.42	0.32
Probabilidad 3	0.00	0.33	0.39
Grupo	1	2	3

A continuación se aplica la prueba de Wald para cada uno de los coeficientes:

```
apply(variables,1,function(x) regTermTest(test,x,method=c("Wald")))
```

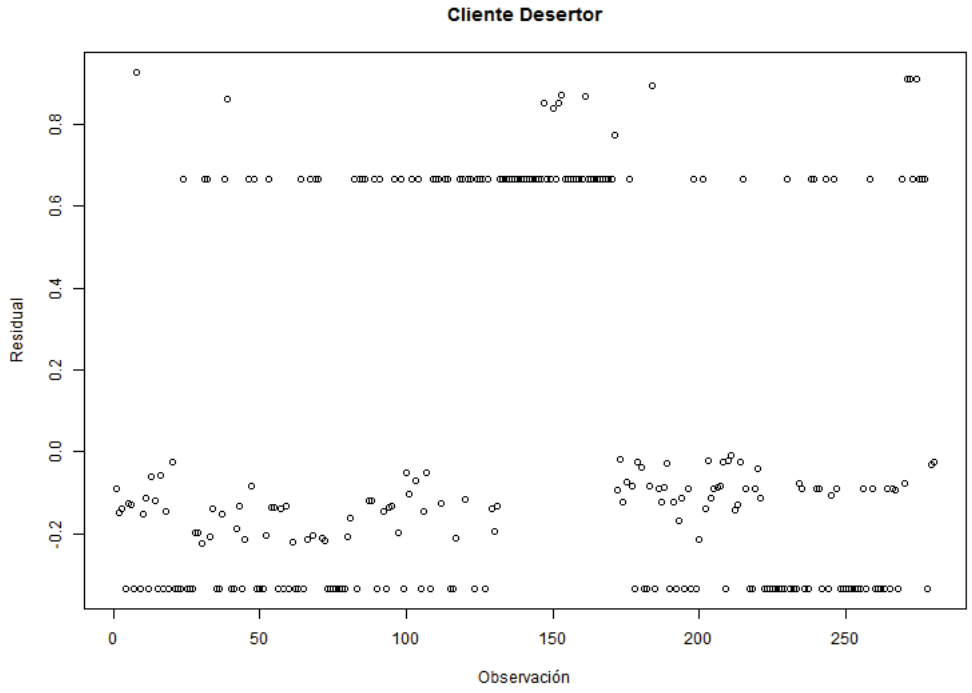
En esta línea se está utilizando la función *apply* para realizar el cálculo de la prueba de Wald para ver si se puede hacer una reducción en las variabes del modelo, el resultado completo puede consultarse en el apéndice B, cabe señalar que en esta prueba las variables significativas son aquellas que reflejan un estadístico F *grande* o equivalentemente un p-value pequeño, en este caso se considera una significancia del 10 % lo cual significa que las variables con p-value menor a .1 seguirán en el modelo.

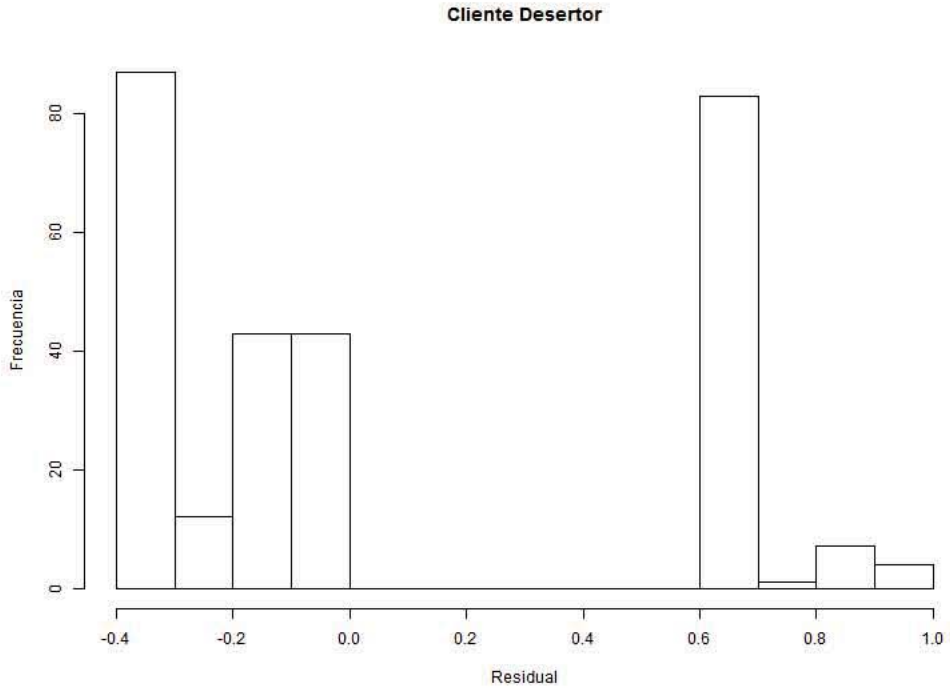
III.3. Resultados

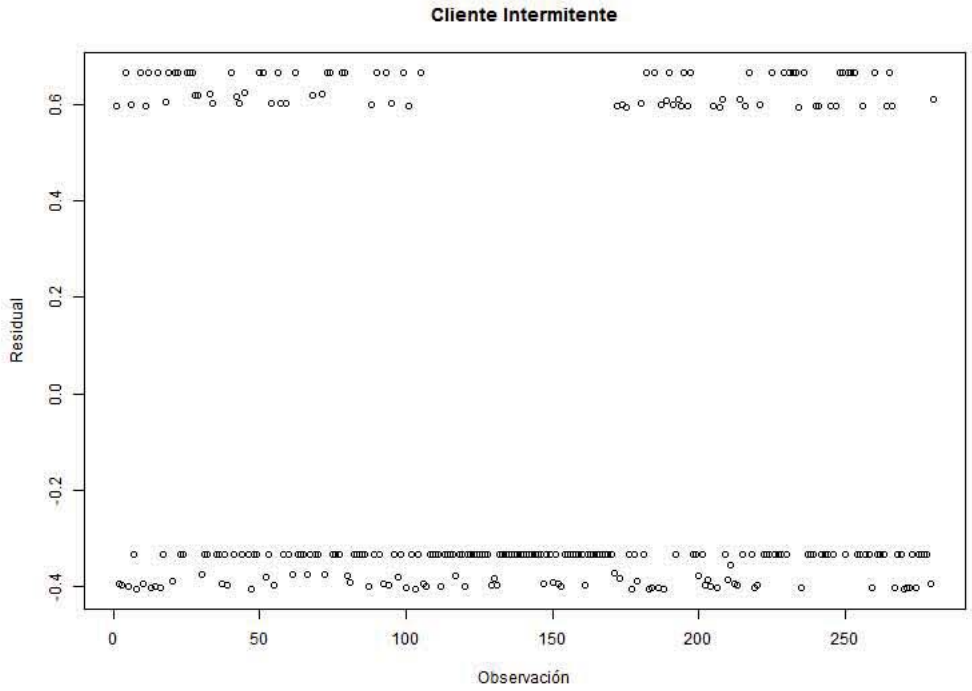
Los resultados que arroja la prueba señalan que la variable *pago3* es la única variable significativa del modelo, por lo que se debe volver a correr el modelo pero únicamente utilizando esta variable:

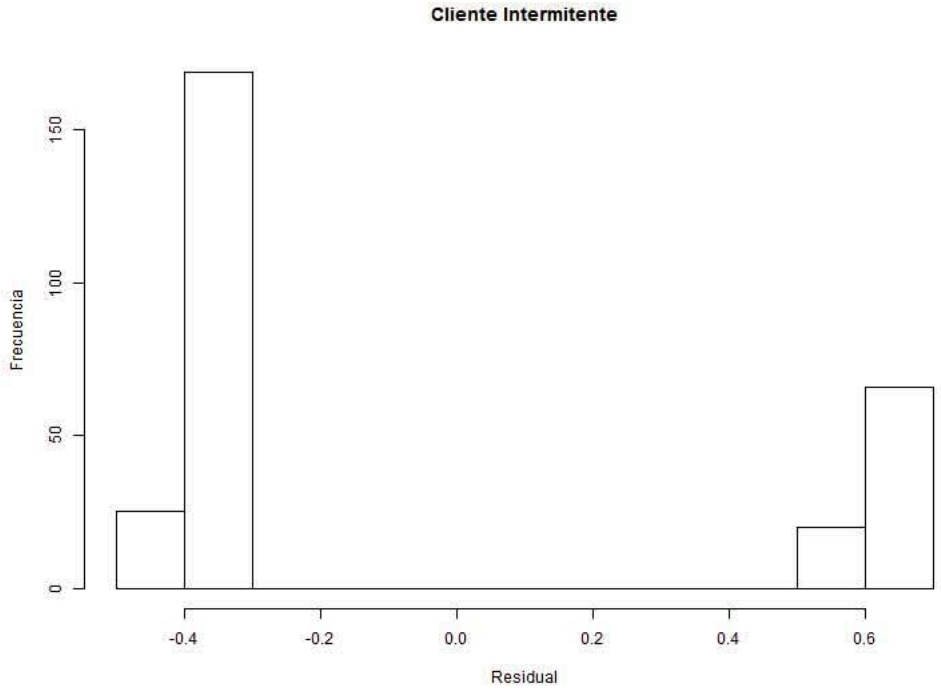
```
formula <- Tipo_de_cliente ~ 0 + pago3

test <- multinom(formula,data=BASE)
```

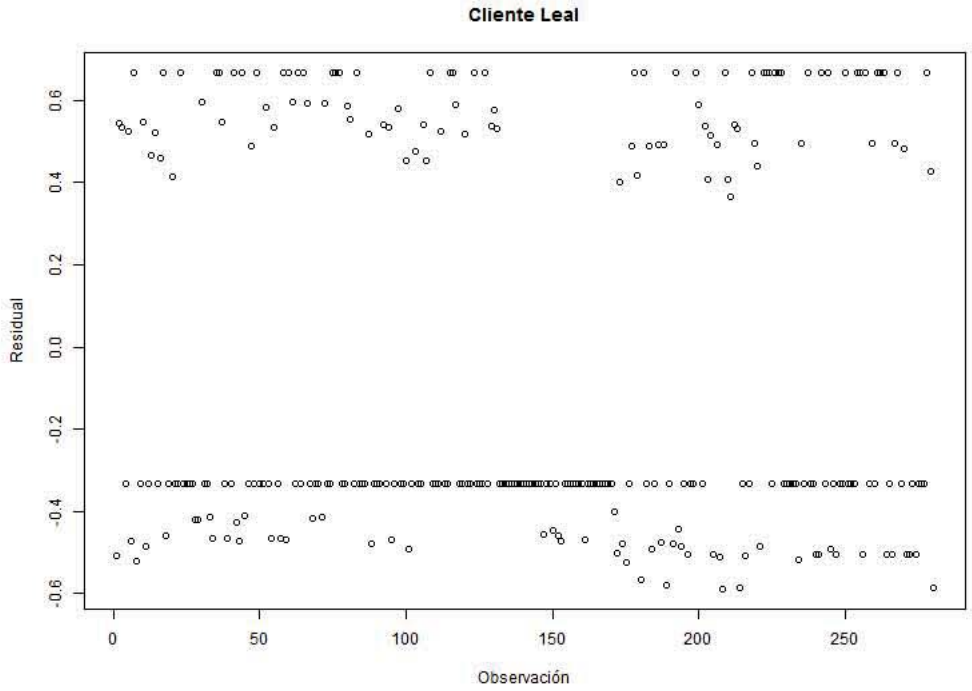
Al aplicar la prueba de Wald, este modelo resulta significativo, sin embargo al graficar los residuales la hipótesis de normalidad parece no cumplirse.

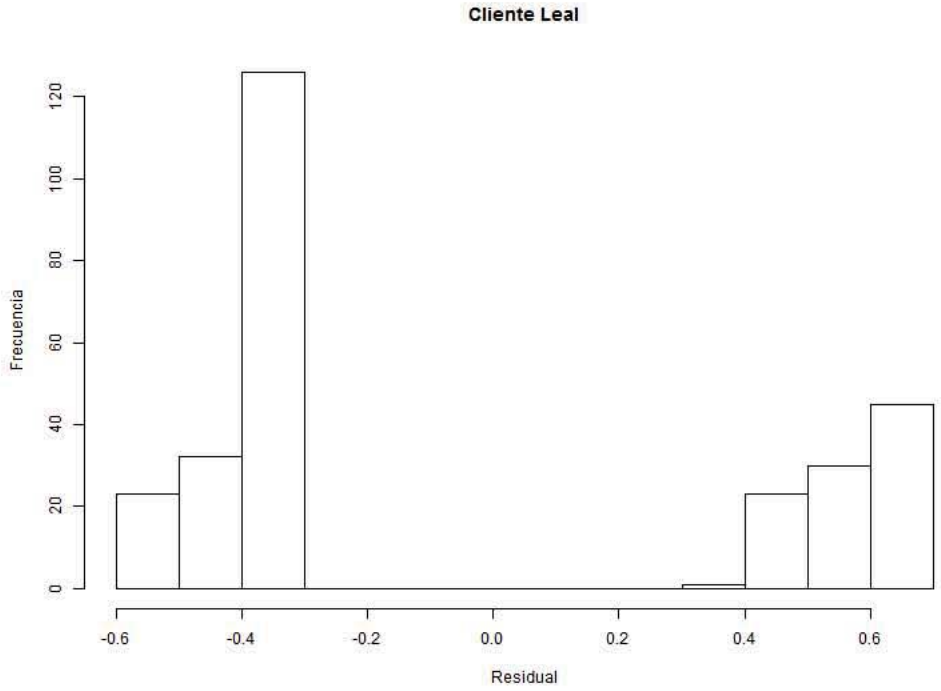


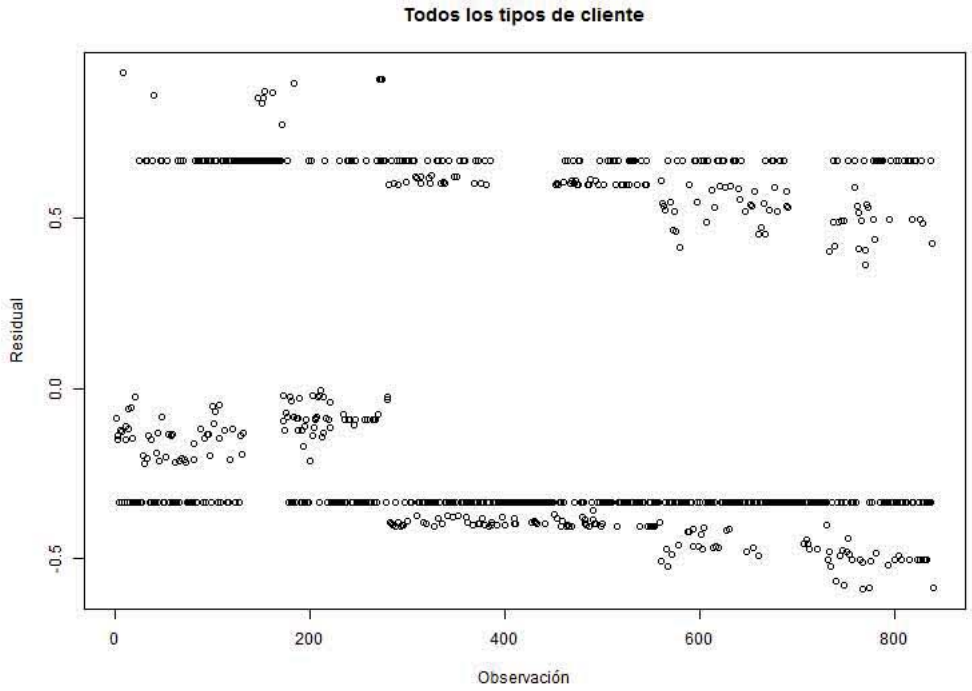

Cliente Desertor
Frecuencia
0
20
40
60
80
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
Residual

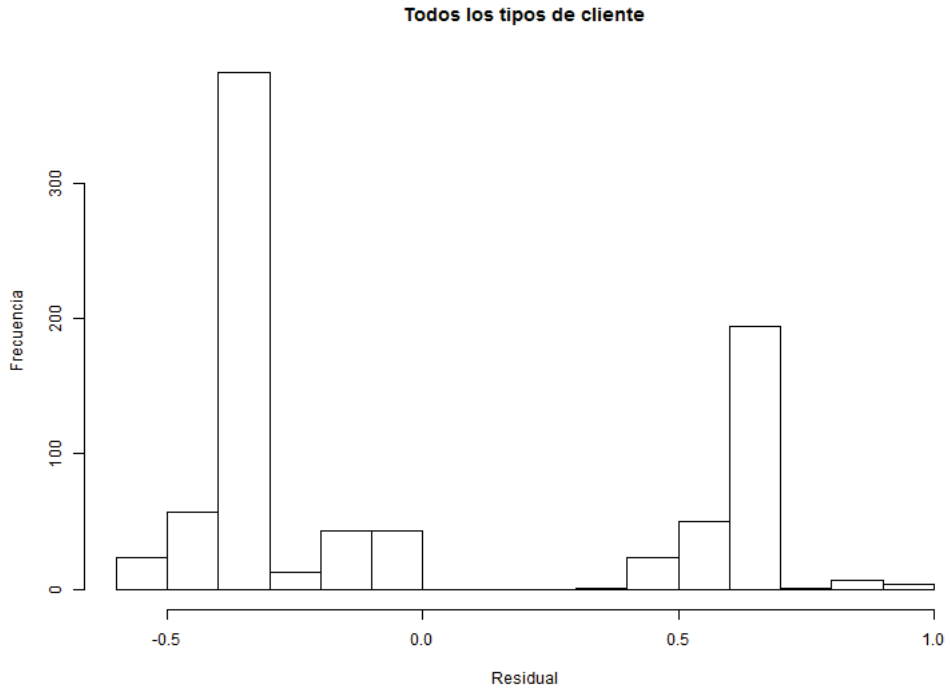

Cliente Intermitente
Residual
-0.4
-0.2
0.0
0.2
0.4
0.6
0
50
100
150
200
250
Observación


Cliente Intermitente
Frecuencia
0
50
100
150
-0.4
-0.2
0.0
0.2
0.4
0.6
Residual


Cliente Leal
Residual
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0
50
100
150
200
250
Observación


Cliente Leal
Frecuencia
0
20
40
60
80
100
120
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
Residual


Todos los tipos de cliente
Residual
-0.5
0.0
0.5
0
200
400
600
800
Observación



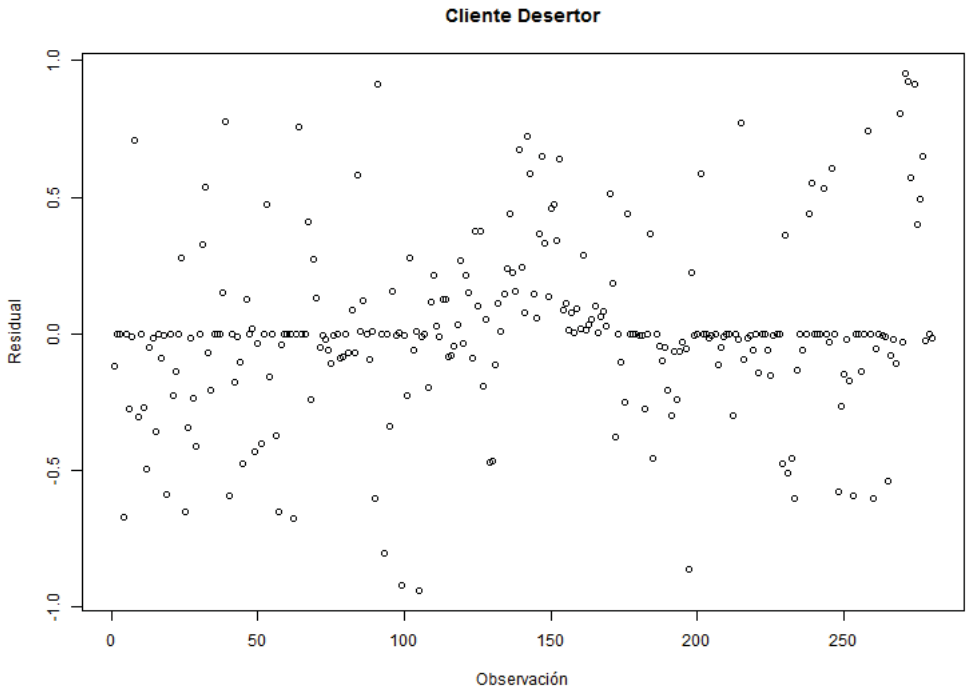
Otro factor importante a tomar en cuenta es que el error se incrementa notablemente:

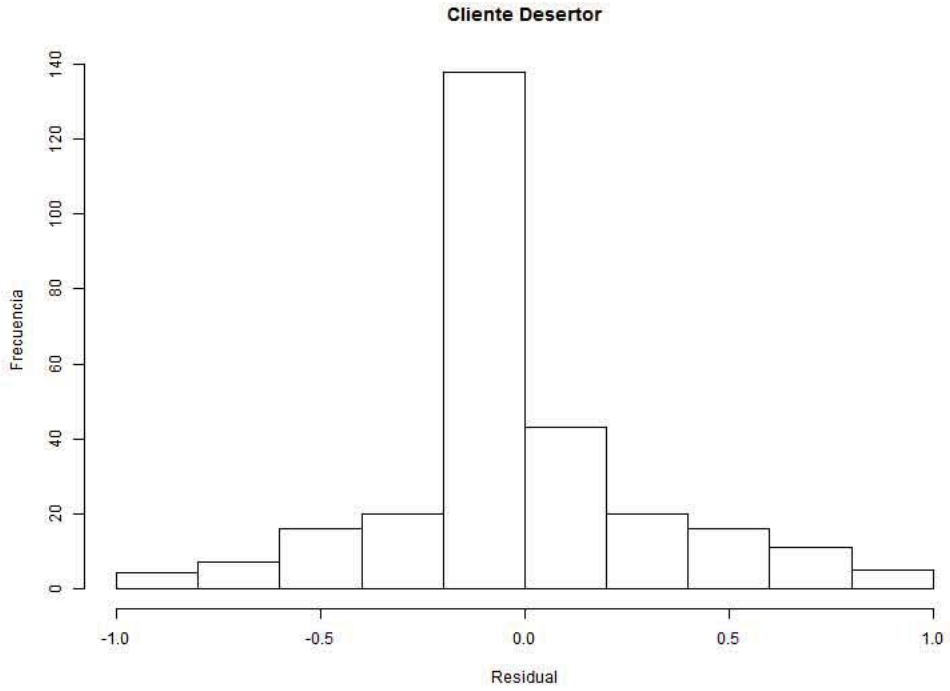
- Al considerar todas las variables disponibles para aplicar el modelo se tiene[1]:
  - Error general: 22.5 %
  - Error cliente leal: 24.2 %
  - Error cliente intermitente: 34.9 %
  - Error cliente desertor: 10.1 %
- Mientras que al considerar únicamente las variables indicadas por el test de Wald[2]:
  - Error general: 64.6 %

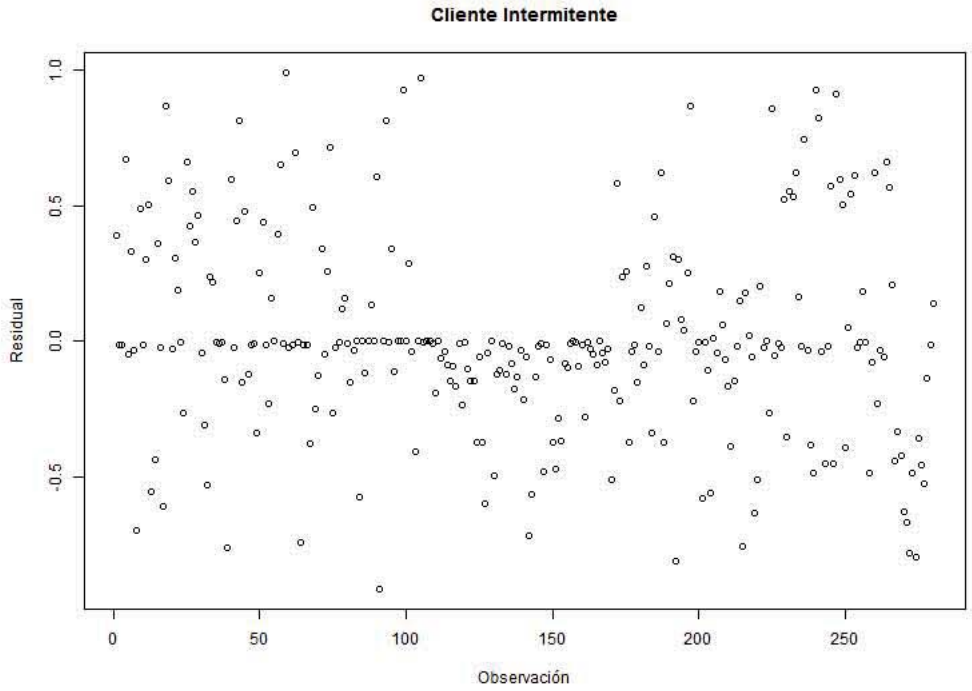
[1]Para consultar la salida de R, consultar anexos
[2]Para consultar la salida de R, consultar anexos

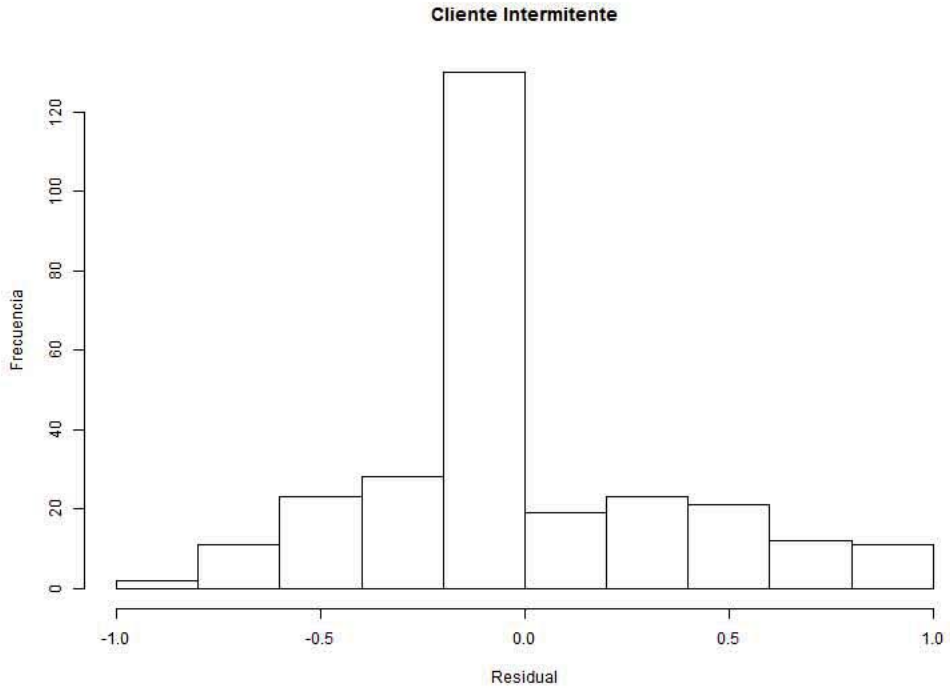
- Error cliente leal: 71.6 %
- Error cliente intermitente: 84.9 %
- Error cliente desertor:27.3 %

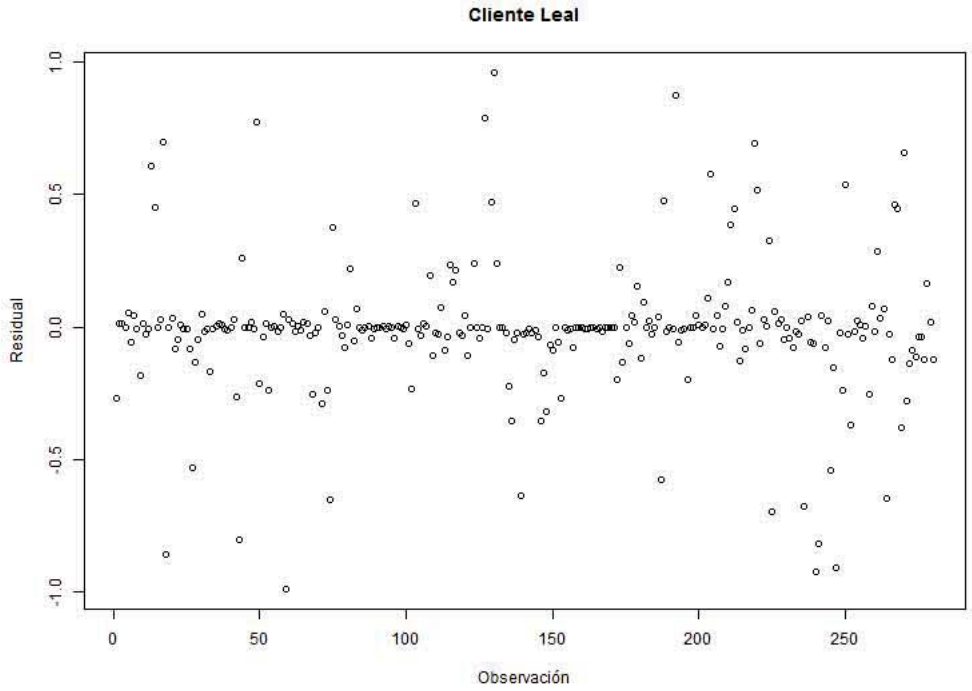
Es por eso que se decide regresar al modelo original sin hacer reducción de variables, conservando la normalidad en los residuales y con una mejora significativa en el error.

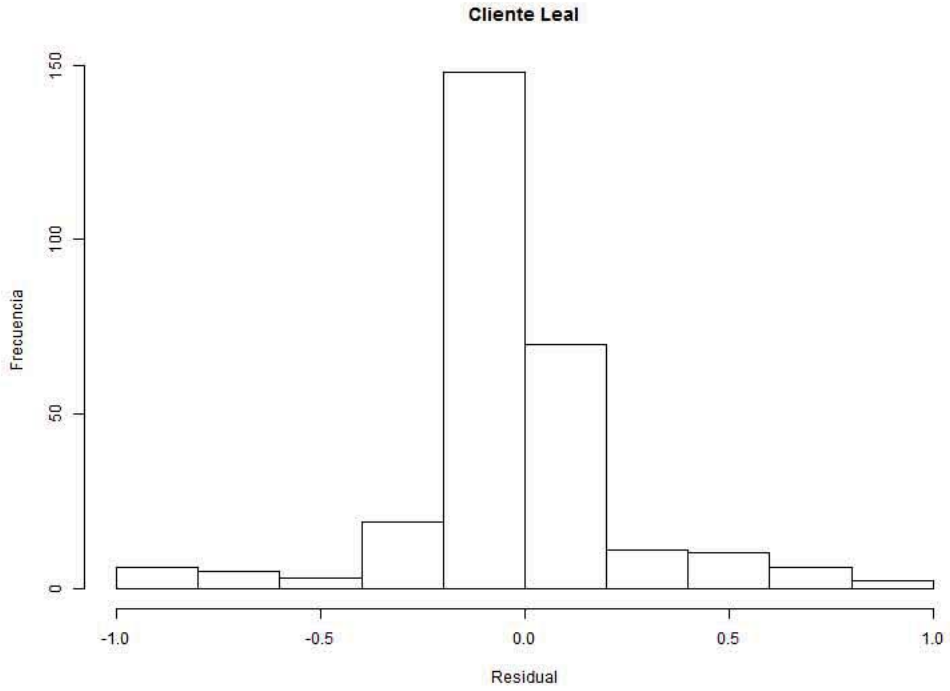


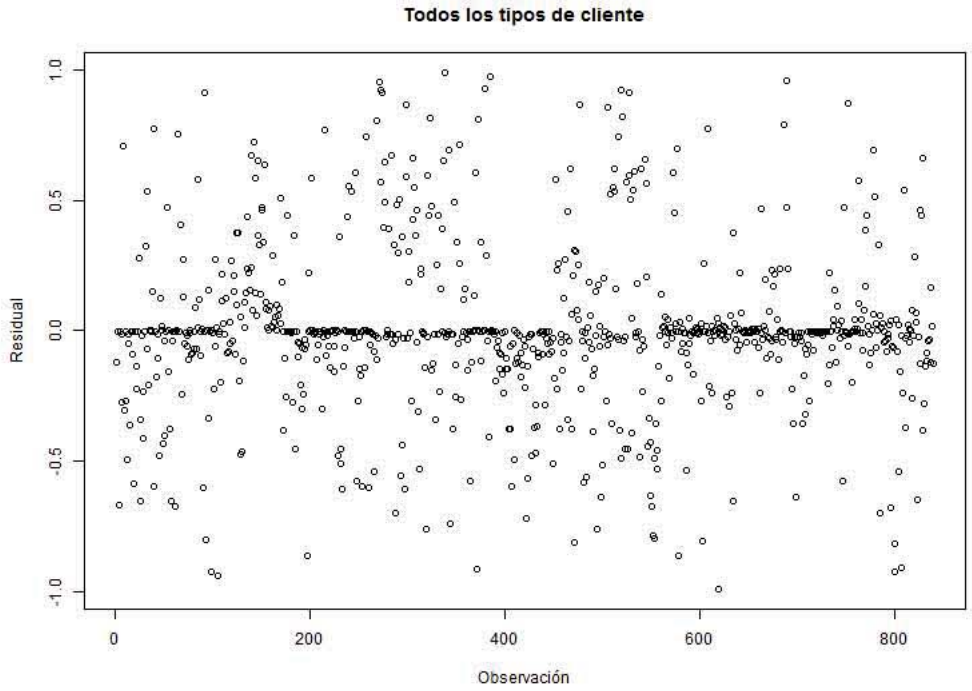

Cliente Desertor
Frecuencia
0
20
40
60
80
100
120
140
-1.0
-0.5
0.0
0.5
1.0
Residual

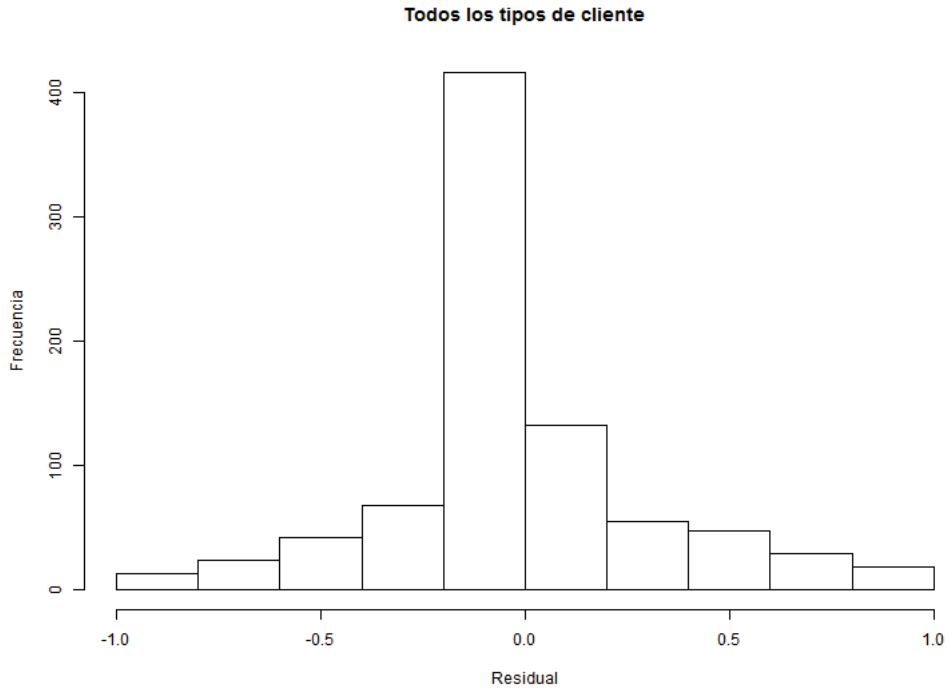

Cliente Intermitente
Residual
-0.5
0.0
0.5
1.0
0
50
100
150
200
250
Observación


Cliente Intermitente
Frecuencia
0
20
40
60
80
100
120
-1.0
-0.5
0.0
0.5
1.0
Residual


Cliente Leal
Residual
-1.0
-0.5
0.0
0.5
1.0
0
50
100
150
200
250
Observación


Cliente Leal
Frecuencia
0
50
100
150
-1.0
-0.5
0.0
0.5
1.0
Residual


Todos los tipos de cliente
Residual
-1.0
-0.5
0.0
0.5
1.0
0
200
400
600
800
Observación



III.4. Casos de sensibilidad

En esta sección ser busca mostrar algunos ejemplos con los cuales será un poco más claro cómo afectan las variables al cálculo de probabilidades para la asignación de una categoria.

- Caso 1: De cliente desertor a cliente leal

  El cliente presenta las siguientes variables:

  ```
  BASE_mult[53,]
  ```

costo_inscripcion	**idforma_pago**	**idmembresia_agrupada**	**edad**	**pago1**	**pago2**
10087	8	5	39	0	3636.727
pago4	**pago5**	**pago6**	**pago7**	**ult_dentro**	**churn**
0	0	0	3621	9	0

  Y presenta estas probabilidades:

```
 predict(test, newdata = BASE_mult[53,],
"probs")
```

0	1	2
0.5271638	0.2329826	0.2398536

Si hubiera pagado $3000 menos inscripción podría haber permanecido dentro durante más tiempo como lo muestran las probabilidades:

```
predict(test, newdata = BASE_mult[53,]+
c(0,0,0,0,0,-3000,0,0,0,0,0,0,0),
"probs")
```

0	1	2
0.1632287	0.3359769	0.5007943

- Caso 2: De cliente intermitente a cliente leal.

  El cliente presenta las siguientes variables:

```
BASE_mult[68,]
```

costo_inscripcion	idforma_pago	idmembresia_agrupada	edad	pago1	pago2
3991.3	3	11	28	0	1058.6
pago4	**pago5**	**pago6**	**pago7**	**ult_dentro**	**churn**
169.75	0	0	43	9	1

Y presenta estas probabilidades:

0	**1**	**2**
0.2415433	0.505112	0.2533447

Si su costo de inscripción hubiera sido más alto y le hubieran hecho descuento del 100 % quizás habría pasado de ser un cliente intermitente a un cliente leal, como lo muestran las probabilidades:

```
predict(test, newdata = BASE_mult[68,]+
c(10000,0,0,0,0,0,0,0,0,0,0,0,0),
"probs")
```

0	**1**	**2**
0.0514786	0.413617	0.5349044

- Caso 3: De cliente leal a cliente desertor.

  El cliente presenta las siguientes variables:

```
BASE_mult[212,]
```

costo_inscripcion	**idforma_pago**	**idmembresia_agrupada**	**edad**	**pago1**	**pago2**
4434.78	93	8	37	4828	0
pago4	**pago5**	**pago6**	**pago7**	**ult_dentro**	**churn**
163	0	0	1013.667	15	2

Y presenta estas probabilidades:

0	1	2
0.2981741	0.1461521	0.5556738

Si este cliente hubiera decidido en un principio, no pagar anualidad y el equivalente pagarlo de forma mensual quizás habría salido del club rápidamente, como lo muestran las probabilidades:

```
predict(test, newdata = BASE_mult[212,]+
c(0,0,0,0,0,20517,-20517,0,0,0,0,0,0),
"probs")
```

0	1	2
0.9999930	0.0000043	0.0000027

Conclusiones

Los estudios de mercado son parte esencial en el desarrollo de cualquier empresa, es por eso que emplear técnicas estadísticas en el análisis de datos es un recurso de gran utilidad ante la necesidad de un análisis que busca explicar las diferencias entre clientes para poder clasificarlos y enfocar la atención de forma específica en cada uno de los grupos presentes en el estudio.

De manera particular, en éste análisis se puede observar que conociendo el perfil de un cliente se puede asignar una probabilidad que permita conocer el probable comportamiento que éste cliente presentará bajo las condiciones en las que ingresa al club, otorgando con esto los elementos necesarios para poder desarrollar un plan específico y personalizado alrededor de ésta persona, permitiendo con eso maximizar la probabilidad de conservar al cliente.

Apéndice A

Salida del modelo con todas las varibles

```
Call:
multinom(formula = formula, data = BASE_mult)

Coefficients:
  costo_inscripcion  idforma_pago idmembresia_agrupada        edad
1      0.0001346043 -0.0004326669          -0.06788366 -0.01805526
2      0.0002293220 -0.0261062006          -0.23370754 -0.07277272
          pago1         pago2        pago3        pago4         pago5
1 -0.0006617088 -0.0005128125 5.450081e-05 0.0003451469  0.0006416549
2 -0.0001297922 -0.0006361752 1.907205e-05 0.0005731699 -0.0005817762
         pago6         pago7 ult_dentro
1 -0.001535808 -0.0001458122  0.1406310
2 -0.002357505 -0.0002038179  0.4629434
```

```
Std. Errors:
  costo_inscripcion idforma_pago idmembresia_agrupada       edad        pago1
1      7.059265e-05  0.006479404           0.04152591 0.01110815 0.0001700738
2      8.779558e-05  0.008495555           0.02314850 0.01469377 0.0001149654
         pago2        pago3        pago4        pago5        pago6
1 0.0001949335 1.823651e-05 0.0002631337 0.0001687306 0.0009743452
2 0.0002329376 2.162237e-05 0.0003555060 0.0003733385 0.0010040090
         pago7 ult_dentro
1 0.0001948827 0.02746308
2 0.0002350066 0.01526353

Residual Deviance: 278.3561
AIC: 326.3561
```

Apéndice B

Test de Wald con todas las variables

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for costo_inscripcion
 in multinom(formula = formula, data = BASE_mult)
F =  3.635791  on  1  and  816  df: p= 0.056901

[[2]]
Wald test for idforma_pago
 in multinom(formula = formula, data = BASE_mult)
F =  0.001252625  on  1  and  816  df: p= 0.97178

[[3]]
Wald test for idmembresia_agrupada
 in multinom(formula = formula, data = BASE_mult)
```

```
F =  0.0001085598  on  1  and  816  df: p= 0.99169

[[4]]
Wald test for edad
 in multinom(formula = formula, data = BASE_mult)
F =  5.523366  on  1  and  816  df: p= 0.019001

[[5]]
Wald test for pago1
 in multinom(formula = formula, data = BASE_mult)
F =  159314.7  on  1  and  816  df: p= < 2.22e-16

[[6]]
Wald test for pago2
 in multinom(formula = formula, data = BASE_mult)
F =  1437383  on  1  and  816  df: p= < 2.22e-16

[[7]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  980220.8  on  1  and  816  df: p= < 2.22e-16

[[8]]
Wald test for pago4
 in multinom(formula = formula, data = BASE_mult)
F =  76486.41  on  1  and  816  df: p= < 2.22e-16
```

```
[[9]]
Wald test for pago5
 in multinom(formula = formula, data = BASE_mult)
F =  15.37964  on  1  and  816  df: p= 9.5322e-05

[[10]]
Wald test for pago6
 in multinom(formula = formula, data = BASE_mult)
F =  0.01774482  on  1  and  816  df: p= 0.89406

[[11]]
Wald test for pago7
 in multinom(formula = formula, data = BASE_mult)
F =  6.924213  on  1  and  816  df: p= 0.0086641

[[12]]
Wald test for ult_dentro
 in multinom(formula = formula, data = BASE_mult)
F =  0.0005366054  on  1  and  816  df: p= 0.98152
```

Salen las variables *forma_pago*,*membresia_agrupada*,*pago6* y *ult_dentro*

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for costo_inscripcion
 in multinom(formula = formula, data = BASE_mult)
```

```
F =  5.376223  on  1  and  824  df: p= 0.020657

[[2]]
Wald test for edad
 in multinom(formula = formula, data = BASE_mult)
F =  0.0007677757  on  1  and  824  df: p= 0.9779

[[3]]
Wald test for pago1
 in multinom(formula = formula, data = BASE_mult)
F =  3138.57  on  1  and  824  df: p= < 2.22e-16

[[4]]
Wald test for pago2
 in multinom(formula = formula, data = BASE_mult)
F =  272.9676  on  1  and  824  df: p= < 2.22e-16

[[5]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  1782.498  on  1  and  824  df: p= < 2.22e-16

[[6]]
Wald test for pago4
 in multinom(formula = formula, data = BASE_mult)
F =  0.7121015  on  1  and  824  df: p= 0.39899
```

```
[[7]]
Wald test for pago5
 in multinom(formula = formula, data = BASE_mult)
F =  13.07875  on  1  and  824  df: p= 0.00031683

[[8]]
Wald test for pago7
 in multinom(formula = formula, data = BASE_mult)
F =  1.525276  on  1  and  824  df: p= 0.21717
```

Salen las variables *edad*,*pago4* y *pago7*

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for costo_inscripcion
 in multinom(formula = formula, data = BASE_mult)
F =  4.493179  on  1  and  830  df: p= 0.034327

[[2]]
Wald test for pago1
 in multinom(formula = formula, data = BASE_mult)
F =  1.164085  on  1  and  830  df: p= 0.28093

[[3]]
Wald test for pago2
 in multinom(formula = formula, data = BASE_mult)
```

```
F =  18.67314  on  1  and  830  df: p= 1.7396e-05

[[4]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  274.39  on  1  and  830  df: p= < 2.22e-16

[[5]]
Wald test for pago5
 in multinom(formula = formula, data = BASE_mult)
F =  15.20021  on  1  and  830  df: p= 0.0001045
```

Sale la variable *pago1*

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for costo_inscripcion
 in multinom(formula = formula, data = BASE_mult)
F =  0.141034  on  1  and  832  df: p= 0.70735

[[2]]
Wald test for pago2
 in multinom(formula = formula, data = BASE_mult)
F =  0.330505  on  1  and  832  df: p= 0.56552

[[3]]
Wald test for pago3
```

```
 in multinom(formula = formula, data = BASE_mult)
F =  2526.144  on  1  and  832  df: p= < 2.22e-16


[[4]]
Wald test for pago5
 in multinom(formula = formula, data = BASE_mult)
F =  4.414151  on  1  and  832  df: p= 0.035942
```

Salen lan variablen *costo_inscripcion* y *pago2*

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  14.28643  on  1  and  836  df: p= 0.00016815


[[2]]
Wald test for pago5
 in multinom(formula = formula, data = BASE_mult)
F =  0.3156144  on  1  and  836  df: p= 0.57441
```

Sale la variable *pago5*

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  14.76365  on  1  and  838  df: p= 0.00013107
```

Apéndice C

Cálculo de error con todas las variables

```
> sum(predict(test)!=BASE_mult$churn)/length(predict(test))
[1] 0.225
> sum(predict(test)[BASE_mult$churn==0]!=BASE_mult$churn[BASE_mult$churn==0])/
length(predict(test)[BASE_mult$churn==0])
[1] 0.2421053
> sum(predict(test)[BASE_mult$churn==1]!=BASE_mult$churn[BASE_mult$churn==1])/
length(predict(test)[BASE_mult$churn==1])
[1] 0.3488372
> sum(predict(test)[BASE_mult$churn==2]!=BASE_mult$churn[BASE_mult$churn==2])/
length(predict(test)[BASE_mult$churn==2])
[1] 0.1010101
```

Apéndice D

Aplicación con una variable

```
Call:
multinom(formula = formula, data = BASE_mult)

Coefficients:
         pago3
1 4.939461e-05
2 5.682568e-05

Std. Errors:
         pago3
1 1.285531e-05
2 1.266420e-05

Residual Deviance: 582.9733
AIC: 586.9733
```

Apéndice E

Test de Wald con una variable

```
> apply(lista,1,function(x) regTermTest(test,x,method=c("Wald")))
[[1]]
Wald test for pago3
 in multinom(formula = formula, data = BASE_mult)
F =  14.76365  on  1  and  838  df: p= 0.00013107
```

Apéndice F

Cálculo de error con una variable

```
> sum(predict(test)!=BASE_mult$churn)/length(predict(test))
[1] 0.575
> sum(predict(test)[BASE_mult$churn==0]!=BASE_mult$churn[BASE_mult$churn==0])/
length(predict(test)[BASE_mult$churn==0])
[1] 0.7052632
> sum(predict(test)[BASE_mult$churn==1]!=BASE_mult$churn[BASE_mult$churn==1])/
length(predict(test)[BASE_mult$churn==1])
[1] 0.8604651
> sum(predict(test)[BASE_mult$churn==2]!=BASE_mult$churn[BASE_mult$churn==2])/
length(predict(test)[BASE_mult$churn==2])
[1] 0.3131313
```

Bibliografía

[1] Lazzari Luisa, La Segmentación de Mercados Mediante la Aplicación de Teoría de Afinidad. http://www.econ.uba.ar/www/institutos/matematica/cimbage/cuaderno02/2%20TEORIA%20DE%20AFINIDAD.pdf

[2] Rodríguez Jorge, Segmentación de Clientes. http://es.slideshare.net/jrodriguezm/dm-segmentacion-declientes

[3] Córdoba Guillermo, Segmentación de clientes. Una propuesta de clasificación (II). La segmentación táctica. http://es.slideshare.net/guillermocordoba/segmentacin-de-clientes-una-propuesta-de-clasificacin-por-objetivos-dimensiones-y-modos-de-aplicacin-2699141

[4] Thompson Ivan, El Estudio de Mercado. http://www.promonegocios.net/mercado/estudios-mercados.html

[5] Jáuregui Alejandro, 7 Elementos Básicos en Metodología de investigación de Mercados. http://www.uv.mx/personal/joacosta/files/2010/08/Prediccion-de-demanda.pdf

[6] UCM, España, Análisis Factorial.

http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento
/materiales/analisis_datosyMultivariable/20factor_SPSS.pdf

[7] Terrádez Manuel, Análisis de Componentes Principales. http://www.uoc.edu/in3/emath/docs/Componentes_principales.pdf

[8] UCM, España, Análisis Discriminante. http://pendientedemigracion.ucm.es/info/socivmyt/paginas/D_departamento
/materiales/analisis_datosyMultivariable/23discr_SPSS.pdf

[9] Castrejón Sandoval Osiris, Diseño y Análisis de Experimentos con Statitix. http://www.uru.edu/fondoeditorial/libros/pdf/manualdestatistix/occompleto.pdf

[10] UG, España, Ampliación de Análisis de Datos Multivariantes. http://www.ugr.es/ gallardo/pdf/cluster-3.pdf

[11] Rincón, Mendoza, Morán, Vega. Tecnología de Apoyo a la Logística. http://es.slideshare.net/guest83cad74/cluster-no-jerarquico

[12] Szumilas, Magdalena. Explaining Odds Ratios. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/

[13] Marín Diazaraque,Juan Miguel. http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/C

[14] Cramer, J. S. The logit model

[15] Hosmer, David W.; Lemeshow, Stanley. Applied logistic regression

[16] Lumley, Thomas. regTermTest http://cran.r-project.org/web/packages/survey/survey.pdf

[17]

[18] Anónimo http://www.estadistica.mat.uson.mx/Material/elmuestreo.pdf
http://www.tiposde.org/general/35-tipos-de-variables/
ttp://academic.uprm.edu/rmacchia/agro6998/Modeloslinealesgeneralizados.pdf