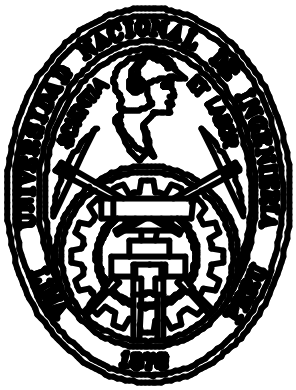
UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ECONÓMICA Y CIENCIAS SOCIALES

ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA



FACTORES QUE DETERMINAN LA CONTINUIDAD DE LAS EMPRESAS EXPORTADORAS DE CONFECCIONES DEL PERÚ

TESIS

PARA OPTAR EL TÍTULO PROFESIONAL DE

LICENCIADO EN ESTADÍSTICA

PRESENTADA POR:

JAVIER LUIS REBATTA NIETO.

LIMA – PERÚ

2006.

Dedicatoria:

A Dios, por su protección y por las fuerzas para continuar cuando las cosas parecen difíciles.

A mi hijo Javier Adrián, por las alegrías de todos los días, por hacer notar siempre su presencia y por enseñarme a ser padre.

A mi querida Cristina, por todo el amor que me brinda, por el empuje que me dá para seguir adelante y por el proyecto de vida que tenemos juntos.

A mis padres, Lucia y Javier, por estar siempre presentes, por su cariño y apoyo inagotables, por la gran confianza que me tienen y el ejemplo que siempre me dieron.

A mis hermanos, Sheila y Rogelio, por todo lo que compartimos, para que siempre alcancen sus sueños y cuando logren alcanzar una meta apunten siempre a la siguiente.

A toda mi familia por lo importante que son para mi, por los engreimientos a mi abuelita Vilma, y a todos mis tíos en especial a Ivón y Raúl, los más jóvenes y quienes me apoyaron siempre sin dudar.

Agradecimientos:

A mi asesor y amigo Lic. Luis Huamanchumo por los Invalorables consejos y las horas invertidas para elaborar este trabajo.

Al Sr. Miguel Gálvez, Gerente de Inteligencia de Mercados de PROMPEX, por el apoyo incondicional y la confianza brindada en este largo camino.

Al Sr. Juan Luis Reus, del Proyecto BID 1442/OC-PE y su equipo, por el apoyo brindado para participar en el XI Escuela de Series Temporales y Econometría, realizado en Brasil en el 2005, que fué la primera piedra para el desarrollo de este trabajo.

ÍNDICE

Índice de tablas

Índice de figuras:

Resumen

La presente investigación intenta encontrar los principales factores que determinan la continuidad de las empresas de confecciones del Perú en la actividad exportadora. El uso de la regresión logística fue inevitable debido a las ventajas que ofrece en el tratamiento de variables dicotómicas como la variable respuesta y ante la presencia de covariables que pueden ser también dicotómicas o continuas con pocos valores. Por este motivo, la aplicación de análisis discriminante u otro método similar no fue posible.

El modelo *logit* ajustado presenta interacciones significativas las cuales fueron ampliamente analizadas en este trabajo, asimismo, se realizó un exhaustivo análisis de residuos y bondad de ajuste y, finalmente, se realizó el pronóstico de continuidad para las empresas de confecciones en el año 2006.

Palabras claves: Regresión logística, continuidad, exportadores peruanos, Prueba de Hosmer y Lemeshow, Prueba de Wald, Prueba del Score, método Stepwise, confecciones, validación externa, pronósticos.

Summary

This work tries to find the different factors that describe the export continuity of Peruvian exporters of apparel and garment products. Using logistic regression was very important because the outcome variable is dichotomous and covariables are continuous with a few values or dichotomous too, then it was impossible to apply discriminant analysis or other method.

The logit model fitted has significant interactions which are deeply analized in this work and it was carried out an exhaustive residual analysis, goodness-of-fit and finally, the forecast of continuity was made for 2006.

Keywords: Logistic regression, continuity, Peruvian exporters, Hosmer and Lemeshow test, Wald test, Stepwise methods, garments and apparel, external validity, forecast.

INTRODUCCIÓN.

En la actualidad es común escuchar comentarios acerca del denominado “Boom exportador” de nuestra economía debido a las altas tasas de crecimiento obtenidas en los último años, sin embargo, no se debe olvidar que la base para que este crecimiento sea sostenido en el tiempo radica en la continuidad de las empresas exportadoras.

Por ese motivo, el apoyo que brinda la entidad de promoción de exportaciones (PROMPEX) y los gremios vinculados al comercio exterior (ADEX, SNI, Cámara de Comercio de Lima, etc.) tiene gran importancia para lograr este objetivo de mantener la continuidad, pero, ¿Cómo se puede identificar a las empresas que se encuentran a punto de abandonar la actividad exportadora?.

Este trabajo pretende dar respuesta a esta interrogante y su importancia radica en la necesidad de cubrir un vacío y poder contar, en base a la información disponible, con

un modelo que permita discriminar entre las empresas que continuarán en la actividad exportadora y las que no podrán con la mayor anticipación posible.

Se asume, en la presente investigación, que la información que las empresas exportadoras de confecciones registran en la Declaración Única de Aduanas (DUA) es suficiente para obtener un modelo que discrimine satisfactoriamente entre las empresas que continuarán o no en la actividad exportadora al año siguiente.

En ese sentido, se plantea la hipótesis que las empresas exportadoras del sector confecciones que continúan con sus operaciones de exportación deben ésta continuidad principalmente al número de mercados destino a los que van dirigidas sus ventas, a la cantidad de productos que comercian, a su participación en actividades de importación, al tamaño de las mismas, a la exportación de productos de otros sectores, y al factor de continuidad, entre otros.

El modelamiento de la información se realiza por medio de la regresión logística debido a la particularidad que presentan los datos, una variable respuesta dicotómica (empresa continua o saliente) y variables independientes que pueden ser también dicotómicas o continuas con pocos valores o ambas a la vez.

El principal resultado obtenido en esta investigación es la verificación que la información registrada por los exportadores de confecciones en la DUA es suficiente para determinar su continuidad en la actividad. Los factores incluidos en el modelo de regresión logística fueron: la cantidad de meses en un año que la empresa exporta, el incremento en sus envíos respecto al año previo, el nivel de exportaciones (tamaño de la empresa), el número de mercados destino de sus productos y la cantidad de años en la actividad exportadora.

Este trabajo cuenta con cinco capítulos. En el primero, se realiza una breve descripción del sector exportador dando énfasis en la evolución del sector confecciones que es la base de nuestro estudio. Aquí se puede obtener una visión panorámica de la evolución de las confecciones peruanas en el mercado internacional. El segundo capítulo, incluye el marco teórico que respalda nuestro análisis. En las distintas secciones se incluye los supuestos y limitaciones que rigen en la regresión logística, las estrategias para la construcción del modelo, la interpretación de los parámetros obtenidos en el ajuste y la determinación de la bondad de ajuste. La presentación de las hipótesis y definiciones operativas para el modelamiento de la continuidad exportadora de las empresas de confecciones se presenta en el capítulo III, mientras que el modelamiento en sí se realiza en el capítulo IV, donde, además se realiza un minucioso análisis de residuos y validación

externa. La interpretación de los resultados, *odds ratios* y los pronósticos obtenidos para la continuidad de las empresas en el 2006 se presenta en el capítulo V.

Asimismo, se incluyen en el documento las principales conclusiones obtenidas en el presente trabajo y se ensaya la posibilidad de brindar recomendaciones a las entidades promotoras de las exportaciones con la finalidad de disminuir la tasa de salidas de la actividad exportadora. De otro lado, se incluye cuatro anexos con tablas estadísticas, gráficos de residuales, sintaxis del SPSS y las consultas en SQL utilizadas para la obtención de la información.

ANTECEDENTES.

Las primeras investigaciones en regresión logística se realizaron en los años treinta, sin embargo, es a partir de 1970 en que comienza el auge de esta técnica especialmente en el campo de la salud. Entre las referencias más importantes en este tema se puede mencionar a Hosmer y Lemeshow (2000) con *Applied Logistic Regression* que presenta un exhaustivo desarrollo de todo el marco teórico necesario para realizar un modelamiento adecuado y es la base para la mayoría de investigaciones consultadas.

Otros desarrollos importantes son los presentados por Kleinbaum y Klein (2002), Christensen (1997) y Dobson (1983), mientras que aspectos mucho más prácticos se pueden encontrar en Catena (2003), Peña (2002), Gujarati (2003) y Acuña (2006), ésta última no ha sido publicada aún.

En nuestro país se puede encontrar los trabajos realizados para obtener la Licenciatura presentados por Corasma (2002) que modela los factores que se asocian con el bajo peso al nacer de los recién nacidos; Flores (2002) que realiza un análisis estadístico de los factores de riesgo que influyen en la enfermedad angina de pecho y Salcedo (2002) que realiza una estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación de autopartes y es lo más cercano al tipo de estudio que se desarrolla por ser también vinculado al comercio internacional.

Investigaciones en el ámbito internacional se pueden mencionar las realizadas por Komarek (2004) que presenta en una tesis doctoral el marco para realizar minería de datos para regresión logística con gran cantidad de variables independientes; Figueiredo (2003) que realiza una comparación entre los métodos de redes neuronales, regresión logística y análisis discriminante para el pronóstico de insolvencia de empresas brasileñas; Caballer (2001) que analiza la intención de donación de órganos en España con regresión logística multinivel y Vinterbo (1999) con un modelo predictivo en medicina.

Respecto a la continuidad de empresas exportadoras que tratamos en esta investigación no se encontró desarrollo alguno en comercio internacional o exportaciones, sin embargo, se pudieron encontrar algunas experiencias en otras

áreas como las realizadas por Mariaca (2002) que realiza un modelamiento para predecir problemas de crisis y continuidad en empresas bancarias y Pontual (2005) que analiza la distribución y dinámica de las empresas industriales en Brasil basándose en la denominada Ley de Gibrat y la Regresión Cuantílica para determinar que el crecimiento de las empresas no depende del tamaño de las mismas.

OBJETIVOS

1.- Objetivo general:

Determinar los factores que permiten a las empresas del sector Confecciones del Perú mantener la Continuidad en sus actividades de exportación definitiva y las principales características de la Continuidad de las empresas entrantes.

2.- Objetivos específicos:

Obtener una metodología adecuada que clasifique a las empresas entrantes según la probabilidad que continúen en la actividad exportadora del sector.

Desarrollar un modelo que permita describir la continuidad de las empresas exportadoras del sector Confecciones del Perú en sus ventas al exterior.

CAPÍTULO I

EVOLUCIÓN DE LAS CONFECCIONES DE EXPORTACIÓN EN EL PERÚ.

En el presente capítulo se hará una breve descripción del proceso de exportación y se analizará la importancia de las confecciones de nuestro país en las ventas al exterior, la diversificación de productos y empresas, así como los principales mercados destino.

1.1.- El proceso de Exportación definitiva.

La exportación definitiva es el proceso en el que las mercancías salen del territorio nacional con destino a otro para ser consumidos o utilizados finalmente generando divisas para el país que los produce. El Perú exportó en el 2005 más de US$ 17 mil millones aumentando 34% respecto al año previo.

Las exportaciones Peruanas fueron principalmente productos tradicionales los que representan el 75% del total y se encuentran concentradas en productos como: cobre, oro, molibdeno, harina de pescado, entre otros.

De otro lado, las exportaciones no tradicionales tienen al sector Textil y Confecciones como el más importante, seguido por la Agroindustria, los productos Químicos, y el sector Sidero Metalúrgico. Los productos más destacados durante el 2005 fueron: polos de algodón para adulto, espárrago fresco, alambre de cobre, camisas de algodón, páprika, pota y los artículos de joyería.

Los destinos a nivel mundial alcanzaron a ser 174 en el último año con una variedad de 3267 productos y 5411 empresas. Los mercados que más demandan nuestras exportaciones fueron: Estados Unidos, China, Chile, Canadá y Suiza.

1.2.- Las exportaciones de Confecciones del Perú.

Se mencionó en la sección anterior que el sector Textil y Confecciones fue el más importante en el rubro No Tradicional al lograr en el 2005 ventas internacionales por un total de US$ 1 274 millones lo que significó un incremento de 17% respecto al año

anterior, concentrando el 30% de las exportaciones al interior del rubro, y el 7% del total nacional.

El presente estudio se limita a las empresas exportadoras de Confecciones, es decir, aquellos que producen artículos que se encuentran clasificados en los capítulos 61 y 62 del Arancel de Aduanas. El concepto de capítulo se desprende de las partidas arancelarias que se definen a continuación.

La sub-partida nacional es una codificación y clasificación de productos que se utiliza para identificar las mercancías en el comercio internacional. Para el caso del Perú está formado por 10 dígitos de los cuales los seis primeros corresponden a una convención internacional denominado Sistema Armonizado de Designación y Codificación de Mercancías (*SA*) del cual se obtiene la nandina.

Para entender mejor la clasificación arancelaria se incluye la tabla 1.1 en la cual se muestra el significado de cada par de dígitos en la subpatida nacional. Esta tabla se obtuvo del Arancel de Aduanas del Perú[1].

[1] Publicado el 29 de diciembre del 2001 en el diario oficial El Peruano.

Considere el caso en que la partida de un producto es 0904200000, por tanto, este pertenece al capítulo 09, partida del sistema armonizado 0904 y subpartida del sistema armonizado 090420. A este nivel arancelario se debe hacer la comparación con los aranceles de otros países con la finalidad de buscar información estadística o referente a la aplicación de tasas a la importación de bienes.

Tabla 1.1: Descripción de los dígitos de una subpartida nacional.

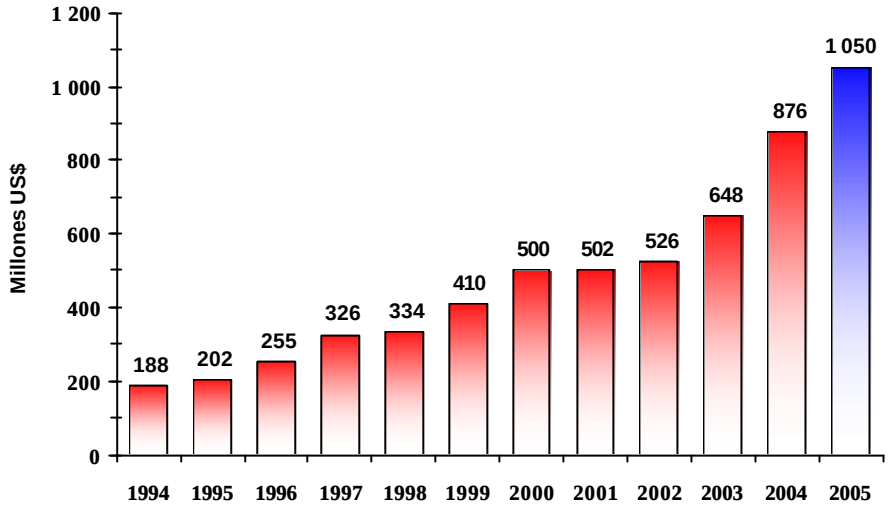
Dígitos										Denominación
1º	**2o**	**3o**	**4o**	**5o**	**6o**	**7o**	**8o**	**9o**	**10o**	
1º	2o									Capítulo
1º	2o	3º	4o							Partida del Sistema Armonizado
1º	2o	3º	4o	5o	6o					Subpartida del Sistema Armonizado
1º	2o	3º	4o	5o	6o	7o	8o			Nandina
1o	2o	3º	4o	5o	6o	7o	8o	9o	10o	Subpartida nacional

Las Confecciones representan el 82% de los envíos del Sector Textil y Confecciones peruano y obtuvo en el 2005 exportaciones por US$ 1 049 millones lo que significó un incremento de 20% respecto al año pasado y 312% en los últimos diez años con una tasa de crecimiento promedio anual en la década pasada (1996 – 2005) de 17%.

Los mercados más importantes como destino de las confecciones son: Estados Unidos, Venezuela, Chile y España, mientras que los productos que tienen mayor demanda en este sector son los polos de algodón, las camisas, blusas, y los suéteres del mismo material.

Figura 1.1: Exportaciones de Confecciones del Perú

1994 - 2005



Fuente: SUNAT. Elaboración: Gerencia de Inteligencia de Mercados - PROMPEX.

1.3.- Distribución y dinámica de las empresas exportadoras.

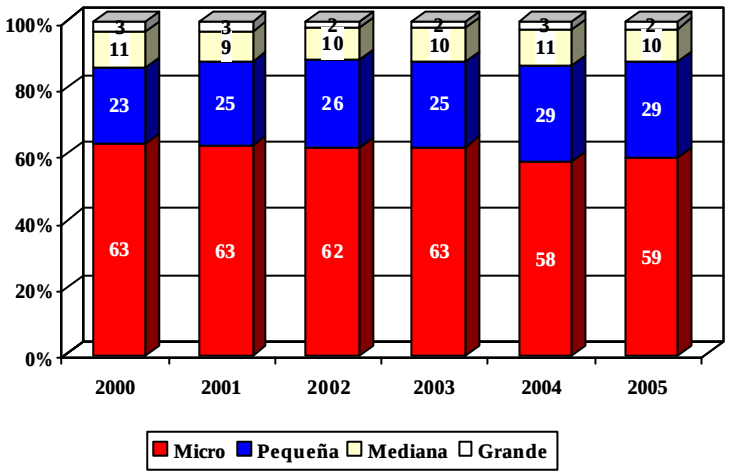
El número de empresas exportadoras de confecciones ascendió en el 2005 a 977 y presentó un incremento de 26% respecto al año previo. La cantidad de empresas que exportaron en los dos últimos años (continuas[2]) fueron el 59%, mientras que el resto fueron entrantes[3]. El 26% de las empresas que registraron ventas al exterior en el 2004 no lo lograron un año después.

[2] Ver clasificación en la sub-sección 3.1.3.a.
[3] Ver clasificación en la sub-sección 3.1.3.a.

La mayor cantidad de empresas son los denominados micros[4] que representan el 59% del universo y las pequeñas[5] tienen una participación de 29%. La situación es diferente cuando se habla de valores de exportación, en lugar de número de empresas, debido a que la mayor parte la negocian las empresas grandes[6].

Figura 1.2: Empresas de confecciones según nivel de exportación

2000 – 2005 (%)



Fuente: SUNAT. Elaboración: Gerencia de Inteligencia de Mercados - PROMPEX.

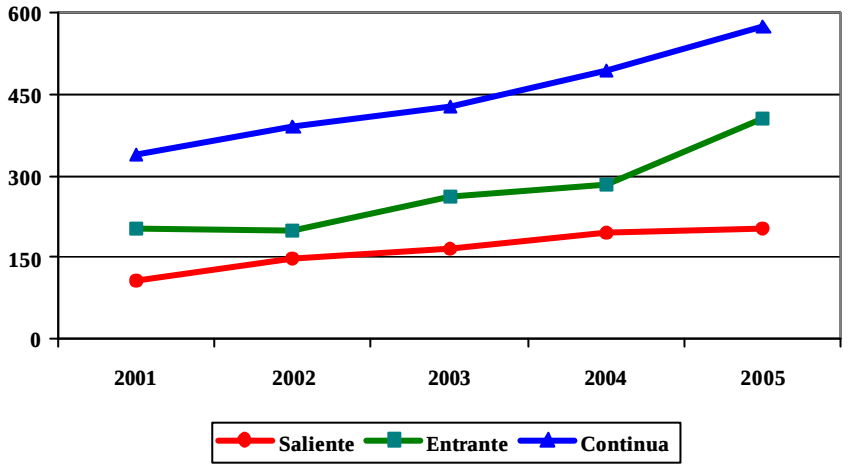
Así, el 67% de las ventas al exterior de estos productos corresponden a 21 empresas cuyos valores de envíos anuales sobrepasan los US$ 10 millones, y las medianas [7]

[4] Exportaciones FOB se encuentran entre 5 y 100 mil dólares.
[5] Exportaciones FOB se encuentran entre 100 mil y un millón de dólares.
[6] Exportaciones FOB superior a 10 millones de dólares.
[7] Exportaciones FOB se encuentran entre uno y diez millones de dólares.

concentran el 24% del total. Las empresas micro que representan la gran mayoría en número alcanzan a exportar el 2% en valor.

Figura 1.3: Número de Empresas de Confecciones según estado

2001 – 2005



Fuente: SUNAT. Elaboración: Gerencia de Inteligencia de Mercados - PROMPEX.

Todos los niveles de empresas por lo menos duplicaron en el 2005 las ventas que generaban en el año 2000, mientras que en cantidad la duplicación únicamente no se cumple en las empresas grandes que aumentaron 50% en ese periodo.

Alrededor del 28% de las empresas exportadoras dejarán la actividad en el año inmediato siguiente, aunque en el año 2005 la tasa de empresas salientes disminuyó a 26%. De otro lado, las empresas que exportan en dos periodos consecutivos

(continuas) ascendieron el último año al 59%, disminuyendo 5 puntos respecto al 2004 debido al importante incremento de la cantidad de empresas nuevas o que reingresan a esta actividad (entrantes).

El número de empresas exportadoras que continúan en la actividad exportadora del sector confecciones aumentó a través de los años y en el 2005 fueron 573, es decir, 79 empresas más que en el 2004, y 147 adicionales al año 2003. De otro lado, la dinámica de las empresas salientes[8] disminuyó significativamente superando ligeramente las 200 empresas. El 73.8% de las exportadoras de confecciones (empresas) del 2004 continuaron en esta actividad al año siguiente.

Otro importante resultado es el siguiente, de las 977 empresas del 2005, el 41% fueron entrantes, es decir, nuevas o que se reincorporaban a la actividad. Este fue el año en el cual se registró la participación más alta de empresas entrantes que no superaba anteriormente el 38%.

Mayor cantidad de empresas incrementaron sus ventas al exterior en este sector durante el 2005 alcanzando el 60% del total. Aquí no se considera a las empresas entrantes que no registraron ventas en un periodo anterior.

[8] Ver clasificación en la sub-sección 3.1.3.a.

De los 776 casos analizados en el 2005 con respecto al año previo se observó que el 15.1% registró importación[9] de productos ubicados en los mismos capítulos que sus exportaciones en el 2004, el 70.1% de las empresas no exporta otros productos que no sea confecciones, mientras que el 56.1% tiene dos años de experiencia en esta actividad y alrededor del 27.0% alcanza los cinco años.

En las mismas condiciones que el párrafo anterior, el 67.8% de las empresas exportaron a un solo mercado en el año previo, y el 83.0% no exporta a través de más de tres subpartidas nacionales.

[9] Régimen definitivo en el cual la mercancía ingresa al territorio aduanero con fines de nacionalización y consumo siendo la contraparte en el comercio exterior de las exportaciones.

CAPÍTULO II

REGRESIÓN LOGÍSTICA MULTIVARIADA.

En este capítulo se realiza una descripción de la regresión logística como herramienta efectiva para el análisis de datos categóricos, los supuestos que debe cumplir y las limitaciones que posee. Asimismo, se resalta las ventajas que brinda al exigir menos supuestos que otras aplicaciones que pretenden encontrar una relación que clasifique a elementos de una misma población en dos grupos basados en las variables independientes, como el análisis discriminantes o la misma regresión lineal multivariada.

2.1.- Introducción.

Los investigadores se encuentran cada vez mas preocupados en modelar los fenómenos que descubren en los diversos campos del quehacer humano, como por

ejemplo, modelar la relación existente entre la presencia de una enfermedad coronaria y los hábitos o rutina del paciente, sea sedentario o no, o fumador, o si tiene parientes que han sufrido estas enfermedades con anterioridad, entre otros factores.

Existen muchos casos en donde la utilización de la Regresión Lineal es suficiente, pero, si se presenta la variable dependiente como categórica y todas las independiente son continuas se puede proceder con la metodología del análisis discriminante, sin embargo, ¿Qué sucede cuando las variables independientes son binarias o dicotómicas?. La respuesta es sencilla, no se puede aplicar el método discriminante debido a que no se cumple uno de los principales requisitos, la normalidad multivariada.

Los otros dos requerimientos son: la homogeneidad de varianzas-covarianzas entre grupos y la relación lineal entre las predictoras. La linealidad se recupera mediante la utilización de logaritmos y los demás requisitos necesarios para el análisis discriminante no lo son para la regresión logística.

2.2.- Limitaciones y supuestos.

En la sección anterior se manifestó que la regresión logística y el análisis discriminante persiguen los mismos fines, y aunque el primero es menos restringido, el incumplimiento de ciertos supuestos puede traer consecuencias nefastas para el ajuste del modelo. Catena, Ramos y Trujillo (2003, pp 348-350) las mencionan, explican y dan pautas para eliminar esos problemas como se describe a continuación:

a).- Linealidad de la función logit: El cumplimiento de este supuesto está relacionado a la posibilidad de presencia de potencias en las variables predictoras que hace necesaria la presencia de un término no lineal significativo en el modelo. Este problema se puede solucionar mediante una transformación.

b).- Independencia de los errores: El valor de la variable de agrupamiento de un elemento no puede depender o ser predicho a partir del valor de otro caso. Esto puede ocurrir de dos formas: cuando los sujetos de análisis han sido medidos en la variable dependiente de manera secuencial o cuando los grupos han sido igualados en variables relevantes. La solución a este problema será el cambio de la estrategia de análisis, inclusive considerando la variable de agrupamiento como un intrasujeto.

c).- Multicolinealidad: Este fenómeno sucede cuando las variables predictoras, categóricas o métricas, se encuentran correlacionadas entre sí, lo que implica que hay predictoras redundantes las cuales deben ser identificadas y eliminar las que contribuyan menos a la capacidad predictiva del modelo.

d).- Números de variables y números de casos insuficientes: Cuando esto sucede puede presentarse el caso de problemas en la estimación de parámetros, principalmente en las variables categóricas, dado que es posible que las condiciones definidas por su combinación no contengan observaciones. Esto produce distorsión en las estimaciones de los parámetros y también en el error estándar.

e).- Puntos extremos: el efecto inmediato de este caso es que el modelo presenta baja capacidad predictiva, la identificación y eliminación es el paso más acertado, pero debe seguir exigentes criterios para efectuarse o sino se puede incrementar artificialmente la predicción.

2.3.- Ajuste del modelo.

Sea el conjunto de *p* variables independientes denotadas por el vector $x' = (x_1, x_2, \ldots\ldots, x_p)$ y sea la probabilidad condicional de la variable respuesta *Y* denotada por $P(Y = 1 / x) = \pi(x)$, entonces, el modelo de regresión logística será:

$$p(x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} \tag{2.1}$$

Sea el *odds ratio* (OR) la razón entre $\pi(x)$ y su complemento $1 - \pi(x)$, donde $\pi(x)$ es como en (2.1). Se aplica al OR una transformación que es el punto medular del estudio de la regresión logística y se representa en términos de p(x), así:

$$g(x) = \ln\left[\frac{\pi(x)}{1 - \pi(x)}\right]; \tag{2.2}$$

y reemplazando la ecuación (2.1) en (2.2), se obtiene:

$$\begin{aligned} g(x) &= x'\beta \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p \end{aligned} \tag{2.3}$$

la cual es conocida como la ecuación *logit*.

La importancia de la transformación es que g(x) tiene muchas de las propiedades deseables de un modelo de regresión lineal, es decir, la *logit* es lineal en sus

parámetros, puede ser continua y puede tomar el rango de **-8** a **8** dependiendo del rango de las variables independientes.

Hosmer y Lemeshow (2000) y Peña (2002) presentan una diferencia importante entre la regresión lineal y la logística se centra en la distribución condicional de la variable respuesta de tal modo que el error, expresado como la desviación de la observación de la media condicional, se distribuye como una normal con media 0 y varianza s^2, mientras que la logística sigue una binomial con media cero y varianza *p(1-p)*, donde *p* es una estimación de p.

Las variables independientes que son escala nominal no son apropiados para ser incluidos en el modelo como si lo son aquellas de escala de intervalo, debido a que el número utilizado para representar los niveles de las variables de escala nominal son únicamente identificadores y no tiene un significado numérico.

La solución a este problema se encuentra en la utilización de las variables diseño o dummy de tal manera que si la variable a ser transformada tiene *k* valores posibles, entonces se crearán necesariamente *k-1* variables diseño.

Sea la *j*-ésima variable independiente x_j que posee k_j niveles, entonces, las k_j-*1* variables diseño serán denotados como D_{jl} y el coeficiente para estas variables diseño serán denotados como $ß_{jl}$, para l=1,2,…, k_j-*1*. Así, la *logit* para un modelo con *p* variables que incluye la *j*-ésima discreta será:

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p \qquad (2.4)$$

La presencia de la sumatoria y los dobles sub-índices será generalmente suprimida excepto cuando se discuta la estrategia de modelamiento al ser necesario usar el valor específico de los coeficientes para alguna variable diseño en el modelo.

2.3.1.- Estimación de parámetros.

El método utilizado para estimar los parámetros en la regresión logística será el de máxima-verosimilitud. Este es un método iterativo que parte de un conjunto de coeficientes iniciales y determina el ajuste de las predicciones con respecto a la variable dependiente o de agrupamiento, se calculan los residuos, y se repite el proceso modificando los coeficientes y comparando los residuos obtenidos hasta que el ajuste no se pueda mejorar.

Considere una muestra de tamaño *n* para las *p* variables independientes (o covariantes) y la variable respuesta *Y*, luego para ajustar el modelo se requiere que se obtenga una estimación del vector de parámetros ß' = (ß$_0$, ß$_1$, …., ß$_p$) y, asumiendo que las observaciones son independientes, la función de verosimilitud es obtenida como el producto de los términos dados en la siguiente expresión:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{2.5}$$

donde: p(x) es como en (2.1).

El principio de máxima verosimilitud que utilizamos para estimar ß debe maximizar la ecuación anterior, por lo cual primero se aplica logaritmos:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{2.6}$$

y luego, para hallar el ß que maximice L(ß), se procede a obtener las derivadas respecto a los parámetros, y el conjunto de expresiones resultantes se igualan a cero y se les conoce como las ecuaciones de verosimilitud:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0, \tag{2.7}$$

y

$$\sum_{i=1}^{n} x_{ij} [y_j - \pi(x_i)] = 0 \tag{2.8}$$

para $j = 1,2,\ldots,p$.

Las ecuaciones obtenidas son no lineales en los parámetros y estos requieren métodos especiales para su solución que son iterativos y se puede utilizar cualquier programa comercial para su cálculo. El valor de ß dado por la solución de las ecuaciones anteriores será llamado estimador de máxima verosimilitud y será denotado como β.

Sea $\hat{\pi}(x)$ una estimación máximo-verosímil de p(x) y como tal representa una predicción del valor o ajuste del modelo de regresión logística que permite obtener la siguiente consecuencia:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{2.9}$$

esto es, la suma de los valores observados de *y* es igual a la suma de los valores predichos o esperados. Este resultado será de mucha utilidad cuando se discuta el tema de ajuste del modelo.

El método de estimación de las varianzas y covarianzas de los coeficientes estimados se realizan mediante máxima-verosimilitud. La teoría indica que los estimadores son obtenidos de la segunda derivada parcial de la función de verosimilitud la cual tiene la siguiente forma:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i (1-\pi_i) \tag{2.10}$$

y

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1-\pi_i) \qquad (2.11)$$

para *j, l* = 0, 1, 2, …, *p* donde π_i denota $\pi(x_i)$. Sea la matriz *(p+1)x(p+1)* que contiene los valores negativos de los términos dados en la ecuación (2.10) y (2.11) y es denotado como I(ß), matriz de información observada. Las varianzas y covarianzas de los coeficiente estimados son obtenidos de la inversa de esta matriz lo cual se denota como:

$$\text{Var}(ß) = I^{-1}(ß) \qquad (2.12)$$

Dado que en casos muy especiales no es posible presentar una expresión explícita para los elementos de la matriz se utilizará la notación Var($ß_j$) para denotar el j-ésimo elemento diagonal de esta matriz la cual es la varianza de $\hat{\beta}_j$, y Cov($ß_j$, $ß_l$) denota un elemento arbitrario fuera de la diagonal que representa la covarianza de $\hat{\beta}_j$ y $\hat{\beta}_l$ donde j,l = 1, 2,…., p.

Mayormente solo se utilizará el error estándar estimado lo cual se denota de la siguiente manera:

$$\hat{SE}(\hat{\beta}_j) = [\hat{Var}(\hat{\beta}_j)]^{1/2}, \qquad (2.13)$$

para *j* = 0, 1, 2, …., *p*. Esta notación será utilizada en el desarrollo de métodos para pruebas de coeficientes y estimación de intervalos de confianza.

La matriz de información será utilizada en la sección que discuta el ajuste del modelo y la determinación del mismo. Una formulación de esta es como sigue: $\hat{I}(\hat{\beta}) = X'VX$, donde *X* es una matriz *n* por *p+1* que contiene la data de cada sujeto de estudio, y *V* es una matriz diagonal *nxn* con elementos generales $\hat{\pi}_i(1-\hat{\pi}_i)$.

2.3.2.- Pruebas de significancia del modelo.

Una vez que se ajusta un modelo de regresión logística multivariable particular, se inicia el proceso de determinación de la significancia del modelo.

El primer paso en ese proceso es determinar la significancia individual de las variables en el modelo, así la prueba de razón de verosimilitud se aplica sobre cada uno de los *p* coeficientes de las variables independientes en el modelo. La hipótesis nula es: los *p* coeficientes para las covariables en el modelo son iguales

a cero y la distribución del estadístico *G* será chi-cuadrado con *p* grados de libertad.

2.3.2.a.-Prueba de razón de verosimilitud.

Esta prueba tiene por finalidad determinar la significancia de la inclusión de una variable en el modelo comparándolo con un modelo previo que no la incluye. El modelo inicial incluye únicamente a la constante.

De este modo, si se quiere determinar la bondad del ajuste del modelo más complejo con la inclusión de una nueva variable en el modelo *r*-ésimo se tendrá:

$$\begin{aligned} G &= -2\ln\left(\frac{l_{r-1}}{l_r}\right) \\ &= -2[\ln(l_{r-1}) - \log(l_r)] \\ &= -2(L_{r-1} - L_r) \end{aligned} \qquad (2.14)$$

donde: $r<p$, L_r es el logaritmo de la función de verosimilitud (l_r) incluyendo la variable en el modelo y L_{r-1} no la incluye.

El estadístico *G* se χ^2 distribuye como con *p* grados de libertad y la significancia nos permitirá asegurar que al menos uno o los *p* valores

de los coeficientes son diferentes de cero cuando se compara el modelo completo contra el inicial.

2.3.2.b.- Prueba de Wald.

Esta prueba se debe aplicar luego de la razón de verosimilitud para poder concluir cuales de los coeficientes son diferentes de cero. El estadístico de prueba es:

$$W_j = \hat{\beta}_j \Big/ \hat{SE}(\hat{\beta}_j) \tag{2.15}$$

Bajo la hipótesis nula que un coeficiente individual es cero, estas estadísticas seguirán una distribución normal estándar. Los valores de *p* para un nivel de significancia a permitirá determinar las variables que serán incluidas o no en el modelo, si no se rechaza la hipótesis nula el coeficiente sería cero y no ejercería ninguna influencia en el mismo.

El análogo multivariado de la prueba de Wald es obtenido del siguiente cálculo matricial:

$$\begin{aligned} W &= \hat{\beta}'[\hat{Var}(\hat{\beta})]^{-1}\hat{\beta} \\ &= \hat{\beta}'(X'VX)\hat{\beta}, \end{aligned} \tag{2.16}$$

los cuales se distribuyen como una chi-cuadrado con *p+1* grados de libertad bajo la hipótesis que cada uno de los *p+1* coeficientes es igual a cero. Prueba para exactamente los *p* coeficientes de las variables independientes son obtenidas por la eliminación de $\hat{\beta}_0$ de $\hat{\beta}$ y la fila relevante (la primera), así como la primera columna de $(X'VX)$.

Puesto que la evaluación de esta prueba requiere la capacidad de realizar operaciones matriciales para obtener $\hat{\beta}$ no hay ventaja sobre la prueba de razón de verosimilitud de la significancia del modelo.

2.3.2.c.- Prueba del Score.

La estadística *Score* es una forma cuadrática basada en el vector de derivadas parciales de la función de *log*-verosimilitud con respecto a los parámetros de interés, evaluado en los valores postulados por la hipótesis nula.

Una ventaja importante de la prueba del *Score* respecto a la de Wald o de Razón de verosimilitud es su eficiencia computacional dado que la

prueba del Score no requiere el cálculo de los estimadores de máxima verosimilitud de los parámetros del modelo.

Para el caso univariado esta prueba se basa en la distribución condicional de la derivada en la ecuación (2.8) dada la derivada en la ecuación (2.7). En este caso, se puede calcular usando $\beta_0 = \ln(n_1 / n_0)$, donde: $n_1 = \sum y_i$, $n_0 = \sum(1 - y_i)$, $\beta_1 = 0$ y $\pi = n_1 / n = \bar{y}$. Por lo tanto, el lado izquierdo de la ecuación (2.8) se convierte en $\sum x_i(y_i - \bar{y})$, y se puede mostrar que la varianza estimada es: $\bar{y}(1-\bar{y})\sum(x_i - \bar{x})^2$. El estadístico para la prueba de *Score* (ST) es:

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}} \qquad (2.17)$$

2.3.3.- Estimación de intervalos de confianza.

Los métodos utilizados para estimar intervalos de confianza para un modelo de variable multivariado son esencialmente los mismos que en el caso univariado, por lo cual, se partirá de ellos para luego generalizarlo.

La construcción de intervalos de confianza para los estimadores sigue la misma base estadística que la utilizada para la formulación de pruebas en la determinación de la significancia del modelo, siendo un caso particular la utilización de los fundamentos de la prueba de Wald para la elaboración de estos intervalos para el intercepto y la pendiente.

Así, para el caso univariado se tiene que los límites al 100(1-a)% del intervalo de confianza para el coeficiente de la pendiente son:

$$\hat{\beta}_1 \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_1) \qquad (2.18)$$

y para el intercepto son los siguientes:

$$\hat{\beta}_0 \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_0) \qquad (2.19)$$

donde: $z_{1-\alpha/2}$ es el valor de la distribución normal estándar en el punto superior 100(1-a/2)% y $\hat{SE}(\hat{\beta}_i)$ es el estimado del error estándar del parámetro estimado.

El estimador de la *logit* es:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.20}$$

con lo cual se puede obtener el estimador de la varianza de la *logit*:

$$\hat{Var}[\hat{g}(x)] = \hat{Var}(\hat{\beta}_0) + x^2 \hat{Var}(\hat{\beta}_1) + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \tag{2.21}$$

Los límites al nivel de 100(1-a)% del intervalo de confianza basado en la prueba de Wald para la *logit* son:

$$\hat{g}(x) \pm z_{1-\alpha/2} \hat{SE}[\hat{g}(x)] \tag{2.22}$$

donde: $\hat{SE}[\hat{g}(x)]$ es la raíz cuadrada positiva del estimador de la varianza obtenido en (2.21).

El estimador de la *logit* y su intervalo de confianza proporcionan las bases para obtener el estimado del valor ajustado de la probabilidad logística y su intervalo de confianza asociado:

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} \tag{2.23}$$

y los límites de confianza para la estimación de la probabilidad logística se obtienen de la siguiente expresión:

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\,\hat{SE}[\hat{g}(x)]}}{1+e^{\hat{g}(x)\pm z_{1-\alpha/2}\,\hat{SE}[\hat{g}(x)]}} \tag{2.24}$$

El intervalo de confianza para la *logit* multivariada es un poco más complicado de calcular para el caso multivariado que el presentado en (2.22) debido a que ahora más términos participan en la sumatoria. Continuando de (2.1) una expresión general para el estimador de la *logit* para un modelo que contiene *p* covariables es:

$$\hat{g}(x)=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\ldots+\hat{\beta}_px_p \tag{2.25}$$

Otra forma de expresar la ecuación (2.25) es a través del uso de la notación vectorial: $\hat{g}(x)=x'\hat{\beta}$, donde $\beta'=(\beta_0,\beta_1,\beta_2,\ldots,\beta_p)$ es el vector que denota el estimador de los *p+1* coeficientes y el vector $x'=(x_0,x_1,x_2,\ldots,x_p)$ representa el conjunto de valores de las *p* covariables en donde x_0=1.

De (2.21) se tiene una expresión para el estimador de la varianza del estimador de la *logit* presentada en (2.25):

$$\hat{Var}[\hat{g}(x)] = \sum_{j=0}^{p} x_j^2 \hat{Var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k) \tag{2.26}$$

La expresión (2.26) se puede llevar a una forma matricial equivalente para el estimado de la varianza de los coeficientes estimados, de modo que a partir de la matriz de información observada, se tiene:

$$\hat{Var}(\hat{\beta}) = (X' VX)^{-1} \tag{2.27}$$

y, de este modo se obtiene a partir de la aplicación de varianza a la ecuación (2.25) en su forma matricial y de (2.27), el siguiente resultado:

$$\begin{aligned} \hat{Var}[(\hat{g}(x))] &= \hat{Var}(x'\beta) \\ &= x' \hat{Var}(\hat{\beta}) x \\ &= x' (X' VX)^{-1} x. \end{aligned} \tag{2.28}$$

La importancia de esta última ecuación radica en la generalización del cálculo de la matriz de varianzas de la *logit* para el caso multivariado y que permite una mejor aplicación en el cálculo de los intervalos de confianza.

2.4.- Interpretación del modelo.

Al iniciar esta parte se considera que el modelo fue bien ajustado y que las variables en el modelo son significativas.

Para realizar una adecuada interpretación se debe tener la capacidad de determinar el aporte que se espera de las covariables (variables independientes) en la descripción del problema en estudio cuyos coeficientes estimados forman la pendiente (razón de cambio) de la *logit*.

2.4.1.- Interpretación según tipo de variable independiente.

El problema de la interpretación tiene dos fases: determinar la relación funcional entre la variable respuesta (dependiente) y las independientes, y definir apropiadamente la unidad de cambio para las variables independientes que en regresión logística univariada se representa de la siguiente forma: $\beta_1 = g(x+1) - g(x)$, es decir, el cambio en la *logit* correspondiente a la variación en una unidad en la variable independiente.

En esta sección se analizará la interpretación para el caso particular univariado en diferentes escalas de medida y la generalización se verá en la sección 2.4.2.

La discusión se inicia para el caso en que el coeficiente que se interpreta corresponde a una variable de escala nominal y dicotómica (dos posibles medidas), y que servirá de base para los demás casos.

2.4.1.a.- Variable independiente dicotómica.

Se considera que la variable independiente puede tomar únicamente dos valores, 1 y 0. La diferencia en la *logit* para un caso con *x=1* y *x=0* es: $g(1) - g(0) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$, y se presenta de esta forma para enfatizar la importancia de expresar la diferencia *logit* deseada en términos del modelo.

En la sección 2.2 se definió el *odds ratio* como una función del modelo logístico, sin embargo, es necesario ampliar el mismo para nuestro caso.

El *odds* para la respuesta en la que *x=1* es definido como: $\pi(1)/[1-\pi(1)]$ mientras que para el caso en que *x=0* es definido como: $\pi(0)/[1-\pi(0)]$. El *odds ratio* (*OR*), es definida como la razón de los *odds* para *x=1* y *x=0* como en la siguiente ecuación:

$$\mathrm{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{2.29}$$

Sustituyendo *x=1* y *x=0,* en la ecuación (2.1) para *p=1* (modelo univariado) se obtiene: $\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$ y $\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$ con sus correspondientes complementos: $1-\pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$ y $1-\pi(0) = \frac{1}{1+e^{\beta_0}}$, y reemplazando en (2.29) y simplificando la expresión se obtiene la relación entre el *odds ratio* y el coeficiente de regresión para una variable dicotómica en un modelo de regresión logístico:

$$\mathrm{OR} = e^{\beta_1} \tag{2.30}$$

Esta simple relación es la razón fundamental por la cual la regresión logística ha demostrado ser una poderosa herramienta analítica de investigación. La OR es una medida de asociación la cual ha encontrado amplio uso especialmente en bioestadística.

Considere el caso donde la variable salida es *y* (presencia o ausencia de continuidad de la empresa exportadora de confecciones) y la variable independiente es *x* (variación de las exportaciones en el año anterior), entonces, un *OR* estimado con un valor de *k* significa que una empresa que aumenta sus exportaciones tiene *k* veces más posibilidad de ser una empresa continua en la actividad exportadora que una que no aumentó sus ventas al exterior, para *k>1*. Cuando *k<1*, la posibilidad es menor, para *k=1* es indiferente.

Tabla 2.1 : Sistematización del caso de continuidad de empresas exportadoras de confecciones según variación.

Continuidad de empresa de confecciones	Variación de exportaciones		Total
	Incremento (1)	Decrecimiento (0)	
Presencia (1)	a	b	a + b
Ausencia (0)	c	d	c + d
Total	a + c	b + d	N

Si se da el caso que $\pi(x)$ es pequeño para *x*=1 y *0*, entonces, $[1-\pi(0)]/[1-\pi(1)]$ tiende a ser igual a uno y *OR* quedará reducido a $[\pi(1)/\pi(0)]$ que es denominado riesgo relativo (*RR*). En la tabla (2.1) se

sistematiza el caso de continuidad de las empresas exportadoras utilizado en esta sección.

En la ecuación 2.30 se mostró la relación entre el *OR* y el coeficiente obtenido por el método de máxima verosimilitud, este mismo resultado se puede obtener de una tabla cruzada como en 2.1 de la siguiente forma:

$$OR = \frac{a / c}{b / d} \tag{2.31}$$

el cual era igual a *k*, entonces, el coeficiente β_1 será igual a ln(*k*).

Considere la estimación del error estándar del coeficiente estimado como la raíz cuadrada de la sumatoria de la inversa del valor de las celdas de una tabla cruzada como 2.1:

$$\hat{SE}(\beta_1) = \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]^{1/2} \tag{2.32}$$

luego del cual, asumiendo que el tamaño de la muestra es lo suficientemente grande como para que *OR* sea normal, se tiene la posibilidad de estimar los límites de un intervalo de confianza al nivel 100(1-a)% para los *odds ratio*:

$$\exp\left[\hat{\beta}_j \pm Z_{1-\alpha/2}\,\hat{SE}(\hat{\beta}_j)\right] \tag{2.33}$$

Se debe considerar la importancia del efecto que tiene la codificación en el cálculo del *OR*. La ecuación 2.30 es correcta cuando la variable independiente es clasificado como 0 y 1, otra codificación requiere el cálculo del valor de la diferencia logit para la codificación utilizada y luego aplicar la transformación exponencial para obtener el *OR*.

Sea alguna variable independiente con dos niveles *x=a* y *x=b*, donde: *a?b*, el *log* de los *odds ratio* es la diferencia entre los *logits* estimados calculados en estos valores:

$$\begin{aligned}\ln[\hat{OR}(a,b)] &= \hat{g}(x=a) - \hat{g}(x=b)\\ &= (\hat{\beta}_0 + \hat{\beta}_1 a) - (\hat{\beta}_0 + \hat{\beta}_1 b)\\ &= \hat{\beta}_1(a-b)\end{aligned} \tag{2.34}$$

Aplicando la exponencial se obtiene:

$$\hat{OR}(a,b) = \exp[\hat{\beta}_1(a-b)] \tag{2.35}$$

el cual es idéntico a la expresión (2.30) cuando (a-b)=1.

En (2.34) y (2.35) la expresión $\hat{OR}_{(a,b)}$ es usada para representar el odds ratio:

$$\hat{OR}(a,b) = \frac{\hat{\pi}(x=a)/[1-\hat{\pi}(x=a)]}{\hat{\pi}(x=b)/[1-\hat{\pi}(x=b)]} \tag{2.36}$$

y cuando *a=1* y *b=0* se tiene $\hat{OR} = \hat{OR}_{(1,0)}$.

2.4.1.b.- Variable independiente policotómica.

Suponga que la variable independiente tiene más de dos categorías, es decir, un número fijo de valores discreto, por lo tanto, se debe formar un conjunto de variables diseño para representar las categorías de la variable. Aquí se hará una revisión de la creación de las variables diseño para el caso policotómico con *k* categorías.

Considere el caso de continuidad de las empresas exportadoras de confecciones y que el tamaño de la empresa (con *k=4* categorías) es la variable independiente. Esta información se presenta en la tabla 2.2.

Tabla 2.2 : Sistematización del caso de continuidad de empresas

Exportadora de confecciones según tamaño

Continuidad de empresa de confecciones	Tamaño de la empresa				Total
	Grande (4)	Mediana (3)	Pequeña (2)	Micro (1)	
Presencia (1)	a	b	c	d	a + b + c +d
Ausencia (0)	e	f	g	h	e +f +g +h
Total	a + e	b + f	c + g	d + h	n

Sin perdida de generalidad, debido a que ampliar el valor de *k* sigue el mismo procedimiento, se usará el tamaño de la empresa grande como referencia. Se calcula el *odds ratio* para el tamaño pequeño *OR*(pequeño) de la siguiente manera:

$$\mathrm{OR(Pequeño)} = \frac{c \times e}{a \times g} \tag{2.37}$$

y donde el grupo de referencia (tamaño de empresa grande) tendrá *OR*=1. Estas mismas estimaciones pueden ser obtenidas de un programa de regresión logística con una apropiada selección de las variables diseño.

El método especifica que todas las $k - 1$ variables creadas tendrán el valor de cero para el grupo de referencia y luego se crea una única variable diseño con el valor igual a 1 para el resto de variables como se ilustra en la tabla 2.3. A este método se le denomina celda de referencia.

No se debe olvidar que el logaritmo de *OR* nos permite obtener como resultado el coeficiente de la categoría que se compara con el grupo de referencia. De esta manera se calcula, el Ln[*OR*(mediana,grande)]= β_1, Ln[*OR*(pequeña,grande)]= β_2, Ln[*OR*(micro,grande)]= β_3.

Tabla 2.3: Creación de variables de diseño para el tamaño de la empresa exportadora de confecciones.

Tamaño de la empresa (Código)	Variables de diseños		
	Tam1	Tam2	Tam3
Grande (4)	0	0	0
Mediana (3)	1	0	0
Pequeña (2)	0	1	0
Micro (1)	0	0	1

Para verificar que esto no ocurrió por coincidencia se calcula la diferencia de la *logit* que muestra que esto es debido al diseño. La comparación del tamaño mediano con grande es como se presenta a continuación:

$$\begin{aligned}\ln[\hat{OR}(\text{mediano}, \text{grande})] &= \hat{g}(\text{mediano}) - \hat{g}(\text{grande}) \\ &= [\hat{\beta}_0 + \hat{\beta}_1(\text{Tam1} = 1) + \hat{\beta}_2(\text{Tam2} = 0) + \hat{\beta}_3(\text{Tam3} = 0)] \\ &\quad -[\hat{\beta}_0 + \hat{\beta}_1(\text{Tam1} = 0) + \hat{\beta}_2(\text{Tam2} = 0) + \hat{\beta}_3(\text{Tam3} = 0)] \\ &= \hat{\beta}_1. \end{aligned} \tag{2.38}$$

Similares cálculos se pueden hacer para los demás coeficiente de la regresión logística. Se intenta bosquejar a continuación, sin perdida de generalidad, para el caso en el cual *k=4* una expresión del error estándar estimado a partir de la tabla 2.2 para el coeficiente estimado de la variable *Tam1*.

$$\hat{SE}(\hat{\beta}_1) = \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{e} + \frac{1}{f}\right]^{1/2} \tag{2.39}$$

Esta estimación permitirá el cálculo de los límites de confianza para los *odds ratio* que es similar al desarrollo de las variables dicotómicas. Los coeficientes de la regresión logística (*log odds ratio*) tienen los siguientes límites de confianza al nivel 100(1-a)%:

$$\hat{\beta}_j \pm Z_{1-\alpha/2} \hat{SE}(\hat{\beta}_j) \qquad (2.40)$$

y los correspondientes límites de los *odds ratio* se obtienen aplicando la exponencial a la ecuación anterior:

$$\exp\left[\hat{\beta}_j \pm Z_{1-\alpha/2} \hat{SE}(\hat{\beta}_j)\right] \qquad (2.41)$$

2.4.1.c.- Variable independiente continua.

La interpretación del coeficiente estimado para este tipo de variables depende de como ingresa ésta al modelo logístico y de las unidades de la variable.

Bajo la asunción que la *logit* es lineal en la covariable continua *x*, la ecuación para la *logit* es como en la ecuación 2.3 para *p*=1, donde la pendiente es $ß_1$, entrega el cambio en los *odds ratio* para el incremento de una unidad en *x*, esto es, $ß_1 = g(x + 1) - g(x)$ para algún valor de *x*.

Se debe considerar el caso en que la variación sea en verdad relevante para el estudio y no pueda ser considerado como muy pequeño o grande.

Así, para proveer una información útil para covariables de escala continua se requiere desarrollar un método para estimación puntual e interválica en donde el cambio sea arbitrario e igual a *c* unidades en la covariable.

El *log* de los *odds ratio* para una variación de *c* unidades en *x* es obtenida de la diferencia de *logits* $g(x + c) - g(x) = c\beta_1$ y el *odds ratio* asociado es obtenido mediante la aplicación de la exponencial a este resultado con lo cual se tiene: $OR(c) = OR(x + c, x) = \exp(c\beta_1)$. Un estimado puede ser obtenido reemplazando β_1 por su estimador máximo-verosímil $\hat{\beta}_1$.

Un estimado del error estándar es obtenido multiplicando el error estándar estimado de $\hat{\beta}_1$ por *c*. Por lo tanto, los límites de confianza estimados a un nivel de 100(1-a)% de *OR(c*) son:

$$\exp\left[c\hat{\beta}_j \pm cZ_{1-\alpha/2}\hat{SE}(\hat{\beta}_j)\right] \qquad (2.42)$$

Debido a que la estimación puntual y la interválica dependen de la selección de *c*, este valor deberá ser claramente especificado en todas las tablas y cálculos. La naturaleza de la arbitrariedad en la elección de *c* puede ser molesta para algunos.

2.4.2.- El modelo multivariable.

En la sección previa se presentó la interpretación de los coeficientes estimados de un modelo de regresión logística en el caso cuando hay una variable en el modelo lo cual usualmente es muy difícil de encontrar debido a que una variable independiente puede asociarse con otra y puede generar diferentes distribuciones en los niveles de la variable respuesta.

La situación multivariada que se examina es aquella en la cual el modelo contiene dos variable independientes, una dicotómica y la otra continua y el principal interés está focalizado en el efecto de la variable dicotómica.

Por cuestiones metodológicas asuma que se quiere comparar el valor de exportaciones promedio de dos grupos de empresas de confecciones y que este valor es una variable asociada a muchos factores entre ellos el número de productos exportados. Asimismo, asuma que todos los demás factores asociados tienen sobre los dos grupos las mismas distribuciones.

Para el caso en que los dos grupos tengan una distribución aproximadamente similar de su número de productos, entonces, el análisis univariado sería

suficiente para hacer esta comparación y esta proveería una estimación correcta de la diferencia del valor de exportaciones entre los dos grupos.

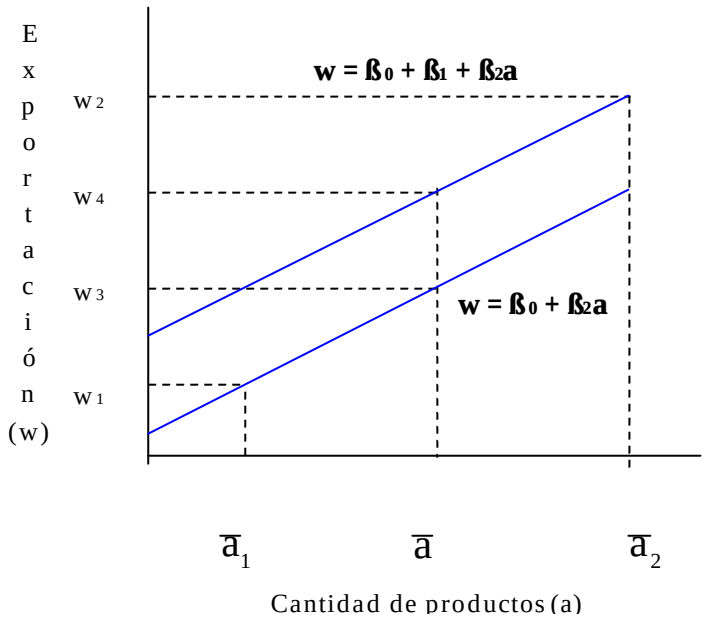
Sin embargo, para el caso en que uno de los grupos tuviera menos productos que el otro, entonces, una comparación de los grupos no tendría sentido puesto que por lo menos una porción de cualquier diferencia observada sería probablemente debido a la diferencia en la cantidad de productos. No sería posible determinar el efecto de cada grupo sin antes eliminar la discrepancia de cantidad de productos entre los grupos.

Esta situación es descrita gráficamente en la figura 2.1. Se asume que la relación entre la cantidad de productos y el valor de exportaciones es lineal con la misma pendiente significativa, diferente de cero en cada grupo.

El modelo estadístico que describe la situación en la figura 2.1 muestra que las exportaciones puede ser expresado como $w = \beta_0 + \beta_1 x + \beta_2 a$, donde *x*= 0 para el grupo 1 y *x=1* para el grupo 2 y *a* denota el número de productos exportados.
En este modelo, el parámetro β_1 representa la verdadera diferencia en las exportaciones de los dos grupos y β_2 es la razón de cambio en las exportaciones por producto.

Suponga que la media de la cantidad de productos del grupo 1 es $\bar{a}_1$ y del grupo 2 es $\bar{a}_2$. La comparación entre la diferencia de las medias y los valores de exportación se presenta en la siguiente ecuación: $(w_2 - w_1) = ß_1 + ß_2(\bar{a}_2 - \bar{a}_1)$.

Figura 2.1: Comparación del valor de exportación de dos grupos con diferente distribución de cantidad de productos.



De esa expresión se puede concluir que no solo se obtiene la diferencia entre los grupos mediante $ß_1$, sino que $ß_2(\bar{a}_2 - \bar{a}_1)$ representa la diferencia entre la cantidad de productos de los grupos.

El proceso de ajuste estadístico para la cantidad de productos envuelve comparaciones de los dos grupos en algún valor común de esta cantidad, la cual es la media de ambos grupos denotado por $\bar{a}$.

En términos del modelo esto produce una comparación de w_4 a w_3, de la siguiente manera: $(w_4 - w_3) = ß_1 + ß_2(\bar{a} - \bar{a}) = ß_1$, la verdadera diferencia entre los dos grupos. Teóricamente algún valor común de la cantidad de productos debió ser usado, y produciría la misma diferencia entre las dos líneas.

La selección de las medias tiene sentido por dos razones: mantiene correlación con el conocimiento del problema en cuestión y se encuentra en el rango en donde se cree que la asociación entre la cantidad de productos y el valor de exportaciones es lineal y constante en cada grupo.

Considere la misma situación mostrada en la figura 2.1 pero en lugar de tener las exportaciones como variable dependiente se tiene una dicotómica y que el eje vertical denota la *logit*. Bajo el modelo la *logit* es dada por la ecuación $g(x,a) = \beta_0 + \beta_1 x + \beta_2 a \cdot$

Una comparación univariada obtenida del resultado y del grupo de una tabla de clasificación cruzada 2x2 producirían *log odds ratio* aproximadamente igual a $ß_1 + ß_2(\overline{a}_2 - \overline{a})$. Esto sería estimar incorrectamente el efecto de grupo debido a la diferencia en la distribución de la cantidad de productos. Para ajustar esta diferencia se incluye la cantidad de productos en el modelo y se calcula la diferencia de *logits* en un valor común de la variable como la media combinada $\overline{a}$.

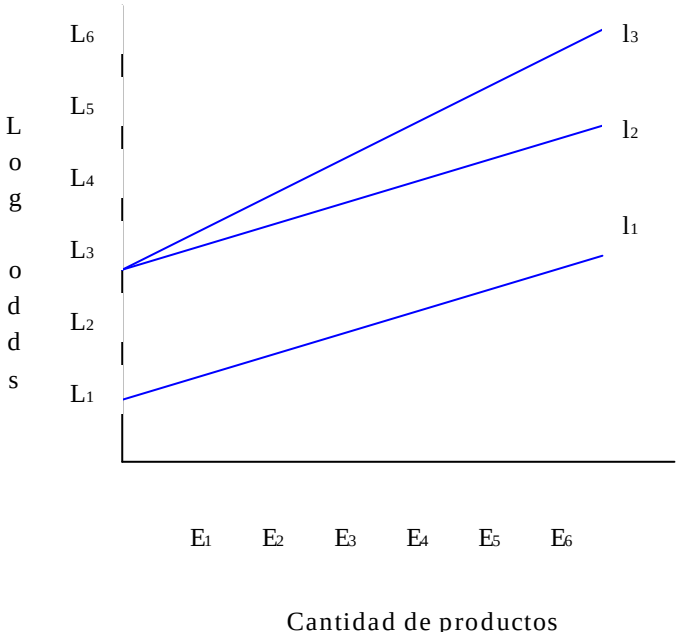
La diferencia de *logit* es $g(x = 1, \overline{a}) - g(x=0, \overline{a}) = ß_1$, así, el coeficiente $ß_1$ es el *log odds ratio* que se esperaría obtener de una comparación univariada si los dos grupos tienen la misma distribución en el número de productos.

El método de ajuste cuando todas las variables son dicotómicas, policotómicas, continuas o una mezcla de estas es idéntico al ajuste descrito en el caso de variables dicotómica-continua.

2.4.3.- Interacción y confusión.

El término variable de confusión es utilizado para describir a la covariable que está asociada con la variable respuesta y una independiente básica o factor de riesgo. Cuando ambas asociaciones están presentes, entonces, la relación entre el factor de riesgo y la variable salida es llamada confusión.

Figura 2.2: *Logits* bajo tres modelos diferentes para mostrar presencia o ausencia de interacción



Se iniciará este análisis describiendo la situación cuando la interacción no se encuentra presente. Considere un modelo que contiene el factor de riesgo de

una variable dicotómica y una covariable continua como en el ejemplo discutido en la sección anterior. Si la asociación entre la variable salida y la covariable es la misma en cada nivel del factor de riesgo (grupo), entonces, no existe interacción entre estas variables.

Gráficamente la ausencia de interacción produce un modelo con dos líneas paralelas una para cada nivel de las variables del factor de riesgo, y en general, esta ausencia es caracterizada por un modelo que no contiene términos de segundo orden o superior envolviendo dos o más variables.

Cuando la interacción está presente, la asociación entre la variable salida y el factor de riesgo difiere o depende en algunos casos del nivel de la covariable, es decir, la covariable modifica el efecto del factor de riego. En algunos casos se utiliza el término modificador de efecto para describir una variable que interactúa con un factor de riesgo.

El modelo más simple y comúnmente usado para incluir interacciones es uno en el cual la *logit* es también lineal en la variable de confusión para el segundo grupo pero con una pendiente diferente. Modelos alternativos pueden ser

formulados, los cuales permitirían una relación no lineal entre la *logit* y las variables en el modelo dentro de cada grupo.

En algunos modelos la interacción es incorporada por la inclusión de términos de orden superior apropiados. En esta sección, se asumirá que la interacción está presente mientras que la determinación de su existencia se revisará en la sección 2.5.

La figura 2.2 presenta tres diferentes *logits*. Considere el caso donde la variable respuesta es la presencia o ausencia de continuidad exportadora, el factor de riesgo es la variación en las exportaciones de la empresa en el periodo anterior y la covariable es la cantidad de productos exportados en el mismo periodo.

Suponga que l_1 corresponde a la *logit* para decrecimiento de exportación como una función de la cantidad de productos exportados, y l_2 representa la *logit* para el incremento de las exportaciones de la empresa, las cuales son paralelas indicando que la relación entre la cantidad de productos y la continuidad es la misma para las empresas que incrementaron o no sus ventas al exterior.

En este caso no hay interacción y el *log odds ratio* para la variación de las exportaciones (incremento contra decrecimiento), controlados por la cantidad de productos exportados, es dado por la diferencia entre las líneas l_2 y l_1. Esta diferencia es igual a la distancia vertical entre las dos líneas la cual es la misma para todas las cantidades de productos.

Ahora, suponga el caso en que la línea l_3 es la *logit* para el incremento en las exportaciones de la empresa en lugar de l_2. Esta recta no es paralela a l_1 indicando que la relación entre la cantidad de productos exportados y la cantidad de empresas continuas que incrementaron sus exportaciones no es la misma para aquellas que decrecieron. Cuando esto ocurre se puede decir que existe interacción entre cantidad de productos y la variación en las ventas internacionales de las empresas.

La estimación de los *log odds ratio* para la variable variación de exportaciones (incremento contra decrecimiento) controlados por la cantidad de productos exportados es dado por la distancia vertical entre las líneas l_3 y l_1 pero esta depende ahora de la cantidad de productos en la cual se hace la comparación, por lo tanto, no se puede calcular el *odds ratio* para la variación de las

exportaciones sin especificar en primer lugar la cantidad de productos en el cual se realiza, es decir, la cantidad de productos es modificador de efecto.

Determinar si la covariable X es una modificador de efecto y/o una variable de confusión envuelve muchos detalles. La figura 2.2 muestra que determinar el estado de modificación del efecto envuelve la estructura paramétrica de la *logit*, mientras que determinar el estado de la variable de confusión implica dos aspectos. En primer lugar, la covariable debe estar asociada con la variable salida, lo cual implica que la *logit* debe tener una pendiente diferente de cero en la covariable, y segundo, la covariable debe estar asociada con el factor de riesgo.

En el ejemplo, esto es caracterizado por una diferencia en el promedio de cantidad de productos entre las empresas que incrementan o no sus ventas al exterior. Sin embargo, la asociación puede ser mas compleja que una simple diferencia en la media. La esencia es que se tiene incomparabilidad en los grupos del factor de riesgo. Esta incomparabilidad se debe considerar en el modelo si se desea obtener una correcta, sin confusión, estimación del efecto para el factor de riesgo.

Un método práctico para comprobar el estado de confusión de una covariable es comparar el coeficiente estimado para la variable factor de riesgo del modelo que contiene y no contiene la covariable. Algún cambio en el coeficiente estimado para el factor de riesgo sugiere que la covariable es una variable de confusión y debería ser incluida en el modelo aún a costa de la significancia estadística de su coeficiente estimado.

De otro lado, una covariable es un modificador de efecto solo cuando el término de interacción incluido en el modelo es estadísticamente significativo y tiene significado para la investigación. Cuando una covariable es un modificador de efecto su estado como una variable de confusión es de importancia secundaria donde el estimado del efecto del factor de riesgo depende del valor específico de la covariable.

2.4.4.- Estimación de *odds ratio* en la presencia de interacción.

La presencia de interacción en el modelo nos obliga a re-plantear el proceso de estimación de los *odds ratio* los cuales no podrán ser calculados con la aplicación de la exponencial a los coeficientes estimados.

Una aproximación que siempre producirá la estimación correcta basada en el modelo se obtiene de la siguiente forma:

(1) Anotar la expresión para las *logit* en los dos niveles en que el factor de riesgo está siendo comparado,

(2) Simplificar algebraicamente la diferencia entre las dos *logits* y calcular su valor,

(3) Exponenciar el valor obtenido en el paso 2.

Considere el caso en que un modelo incluye solo dos variables y su interacción. En este modelo se denota el factor de riesgo como F, la covariable como X y su interacción FX. La *logit* para este modelo evaluado en $F=f$ y $X=x$ es:

$$g(f, x) = ß_0 + ß_1 f + ß_2 x + ß_3 fx \qquad (2.43)$$

Asuma que se requiere los *odds ratio* comparando los dos niveles de F, $F=f_1$ contra $F=f_0$, para $X=x$. Se sigue los tres pasos mencionados anteriormente, y se reemplaza los valores en 2.43, obteniéndose en primer término:

$$g(f_1, x) = ß_0 + ß_1 f_1 + ß_2 x + ß_3 f_1 x$$

y

$$g(f_0, x) = ß_0 + ß_1 f_0 + ß_2 x + ß_3 f_0 x.$$

Luego, se simplifica y calcula su diferencia para obtener los *log odds ratios* produciéndose:

$$\begin{aligned}\ln[OR(F=f_1,F=f_0,X=x)] &= g(f_1,x)-g(f_0,x)\\ &= (\beta_0+\beta_1 f_1+\beta_2 x+\beta_3 f_1 x)-(\beta_0+\beta_1 f_0+\beta_2 x+\beta_3 f_0 x)\\ &= \beta_1(f_1-f_0)+\beta_3 x(f_1-f_0)\end{aligned} \tag{2.44}$$

En tercer lugar, se obtiene los odds ratio aplicando la exponencial a la diferencia obtenida en el paso anterior:

$$OR = \exp[ß_1(f_1 - f_0) + ß_3 x(f_1 - f_0)] \tag{2.45}$$

Se debe notar que la expresión (2.44) no simplifica a un coeficiente único, en su lugar, envuelve dos coeficientes, la diferencia en el valor del factor de riego y la variable de interacción. El estimador de los *log odds ratios* es obtenido por reemplazo de los parámetros en (2.44) y (2.45) con sus estimadores.

Los límites de los intervalos de confianza para el estimador se calculan de la misma manera cuando la interacción está presente, primero se calcula los límites para los *log odds ratio* y luego se aplica la exponencial a los puntos extremos. Considere que la varianza para el estimador del *odds ratio* en (2.44) es:

$$\operatorname{Var}\{\ln[\widehat{OR}(F=f_1,F=f_0,X=x)]\} =$$
$$= (f_1-f_0)^2 \operatorname{Var}(\hat{\beta}_1) + [x(f_1-f_0)]^2 \operatorname{Var}(\hat{\beta}_3) + 2x(f_1-f_0)^2 \operatorname{Cov}(\hat{\beta}_1,\hat{\beta}_3) \quad (2.46)$$

cuya raíz positiva es la que se utiliza para el cálculo de los límites:

$$[\hat{\beta}_1(f_1-f_0)+\hat{\beta}_3 x(f_1-f_0)] \pm z_{1-\alpha/2}\, \widehat{SE}\{\ln[\widehat{OR}(F=f_1,F=f_0,X=x)]\} \quad (2.47)$$

2.5.- Estrategias para la construcción del modelo.

En las secciones previas el análisis se centró sobre la estimación, pruebas estadísticas y la interpretación de los coeficientes de regresión logística en el modelo. Lo mas frecuente es la existencia de un gran número de variables y se puede dar la posibilidad de incluir gran cantidad de estas en el modelo, entonces, es necesario desarrollar una estrategia y métodos asociados a nuestro alcance para poder simplificar estas situaciones.

La meta de cualquier método es seleccionar las variables que dan como resultado el mejor modelo dentro del contexto del problema. Para alcanzar estas metas, es necesario tener a disposición lo siguiente:

1) Un plan básico de selección de las variables para el modelo, y
2) Un conjunto de métodos para determinar la suficiencia del modelo, ambos en términos de sus variables individuales y su ajuste total.

2.5.1.- Selección de variables.

El criterio para incluir una variable en un modelo puede variar de un problema a otro y la aproximación tradicional para construir un modelo estadístico consiste en buscar aquel que sea más parsimonioso al explicar los datos.

La racionalidad para minimizar el número de variables en el modelo es aquel resultado numéricamente más estable y más fácilmente generalizable. Mayor número de variables trae consigo mayor error estándar estimado y mayor dependencia sobre los datos observados.

Algunos especialistas sugieren incluir todas las variables relevantes de manera intuitiva y cueste lo que cueste en aspecto de significancia estadística, sin embargo, la racionalidad para esta aproximación es proveer un control de confusión como sea posible dentro de los datos.

La base de esto es la posibilidad que las variables individuales no muestren fuerte confusión, pero si puede ser considerable cuando se toma colectivamente. El mayor problema con esta aproximación es que el modelo puede ser sobre ajustado produciendo estimados numéricamente inestables.

La sobre-estimación es típicamente caracterizado por irreales coeficientes estimados y/o errores estimados grandes. Esto puede ser especialmente molesto en problemas donde el número de variables en el modelo es grande relativo al número de sujetos y/o cuando la proporción total correspondiente (y=1) es cercano a cualquiera, 0 o 1.

Hosmer y Lemeshow (2000), señala que en regresión logística, hay cinco pasos que uno debe seguir en la selección de variables para el modelo los cuales son muy similares a los usados en regresión lineal.

1) El proceso de selección deberá iniciarse con un cuidadoso análisis univariado. Las variables nominales, ordinales y continuas con pocos valores enteros deberán ser analizados con tablas de contingencia de la variable de salida (y=0,1) contra los *k* niveles de la variable independiente.

La prueba chi-cuadrado de razón de verosimilitud con *k-1* grados de libertad es exactamente igual al valor de la prueba de razón de verosimilitud para la significancia de los coeficientes para las *k-1* variables de diseño en un modelo de regresión logística univariado que contenga esa variable independiente simple. Siendo la prueba chi-cuadrado de Pearson asintóticamente equivalente a la chi-cuadrado de razón de verosimilitud este puede también ser usado.

Atención especial se debe tener cuando la tabla de contingencia tiene una celda cero ya que esto produce una estimación puntual para uno de los *odds ratio* igual a cero o infinito. Incluir tales variables en los cálculos del modelo logístico puede darnos salidas numéricas indeseables.

Una estrategia para manejar la celda cero incluye las siguiente acciones: Colapsar las categorías de la variable independiente en algún sentido para eliminar la celda cero, eliminando la categoría completamente, o, si la variable es escala ordinal, modelando la variable como si fuera continua.

Para una variable continua, el análisis univariado más deseable envuelve el ajuste de un modelo de regresión logística univariado para obtener el

estimado del coeficiente, el error estándar estimado, la prueba de razón de verosimilitud para la significancia del coeficiente, y la estadística univariada de Wald.

Un análisis alternativo, lo cual es equivalente al nivel univariado puede ser basado en una prueba *t* de dos muestras. Las estadísticas descriptivas que generalmente forman parte de este tipo de pruebas son las medias grupales, desviaciones estándar, el estadístico *t* y su valor de probabilidad *p*.

La similaridad de esta aproximación al análisis de regresión logística viene del hecho que la función estimada lineal discriminante univariada de los coeficientes de regresión logística es:

$$\frac{(\bar{x}_1 - \bar{x}_0)}{s_p^2} = \frac{t}{s_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_0}} \qquad (2.48)$$

y que la función discriminante y el estimado de máxima verosimilitud del coeficiente de regresión logística son usualmente bastante cercanos cuando la variable independiente es aproximadamente distribuida como una normal con cada uno de los grupo de respuestas, *y=0,1*.

Así, el análisis univariado basado en la prueba *t* deberá ser útil para determinar si una variable puede ser incluida en el modelo, donde el valor *p*

debe ser del mismo orden de magnitud que el estadístico de Wald, la prueba del *Score*, y la de razón de verosimilitud de la regresión logística.

Para covariables continuas, se puede desear complementar la evaluación del ajuste logístico univariado con alguna clase de *scatterplot* alisado. Este es útil no solamente en comprobar la importancia potencial de la variable y posible presencia y efecto de observaciones extremas, grandes o pequeñas, pero a una escala apropiada.

2) Sobre la culminación de análisis univariado, se seleccionan las variables a ser consideradas en el análisis multivariado. Una variable que en la prueba univariada tiene un valor de $p<0.25$ es candidato para el modelo multivariado y el primer modelo debe contener todas las variables seleccionadas por cumplir esta condición.

La recomendación de considerar un nivel de *p* en 0.25 es basado sobre diversos trabajos en regresión lineal y logística. Bendel y Afifi (1977) y Mickey y Greenland (1989) muestran que el uso del tradicional nivel de *p* 0.05 a menudo falla en identificar variables que se conocen son importantes en el modelo.

Usar un nivel más alto tiene la desventaja de incluir variables que son de cuestionable importancia en la etapa de construcción del modelo, por eso es importante revisar las variables incluidas al modelo críticamente antes que se logre una decisión con respecto al modelo final.

Un problema importante con una aproximación univariada es que se ignora la posibilidad que un conjunto de variables, cada una de las cuales es débilmente asociada a la variable respuesta, produzca un importante predictor en conjunto.

Si lo mencionado en el párrafo anterior es posible, entonces se debe seleccionar un nivel de significancia lo suficientemente grande como para permitir que variables de las cuales tenemos cierta sospecha sean consideradas candidatas para ser incluidas en el modelo multivariado.

3) La importancia de cada variable incluida en el modelo multivariado deberá ser verificado, para lo cual se debe hacer lo siguiente: en primer lugar, examinar la estadística de Wald para cada variable y luego, una comparación de cada coeficiente estimado con el coeficiente del modelo conteniendo solo esta variable.

Las variables que no contribuyan al modelo basado sobre estos criterios deberán ser eliminadas y un nuevo modelo se ajustará. El nuevo modelo deberá ser comparado con el anterior, y más grande, usando la prueba de razón de verosimilitud.

Los coeficientes estimados para las variables restantes deberán ser comparados a aquellos del modelo completo, en particular, se debe analizar los coeficientes de las variables que tuvieron un marcado cambio. Esto indica que uno o más de las variables excluidas fueron importantes en el sentido que proporcionan el ajuste necesario del efecto de la variable que permanece en el modelo.

El proceso de eliminar, reajustar y verificar continua hasta que se observe que todas las variables importantes son incluidas en el modelo.

4) Cuando se tiene el modelo que contiene las variables esenciales es el momento en que las debemos observar más exhaustivamente. Las interrogantes sobre las categorías apropiadas para las variables discretas se deberán resolver en la etapa univariada. Para variables continuas se deberá chequear la asunción de linealidad en la *logit*. Esta asunción en la etapa de

selección de variables es común y consistente con la anotación de poder determinar si una variable particular debería ser incluida en el modelo.

Una vez que la variable es identificada como importante para el modelo, entonces, se puede obtener la correcta relación paramétrica o escala en la etapa de refinamiento del modelo. La excepción a esto sería cuando la función es en forma de U.

Al terminar este paso podemos llamar al modelo como el de efectos principales, entonces un importante paso en la refinación del efecto principal es determinar si el modelo es lineal en la *logit* para variables continuas.

Antes de continuar con el paso 5 se hace necesario presentar dos métodos para tratar el problema expresado en el párrafo anterior: Variables de diseño y polinomios fraccionales.

a) Variables de diseño

El procedimiento se basa en lo siguiente: La diferencia entre las *logits* para dos diferentes grupos es igual al valor de un coeficiente estimado de un modelo de regresión logística ajustado que trata a las variables agrupadas como categóricas.

Primero, se debe obtener los cuantiles de la distribución de la variable, seguido, crear una variable categórica con cuatro niveles usando tres puntos de corte basados sobre los cuantiles, luego, ajustar el modelo multivariado reemplazando la variable continua con la variable categórica de cuatro niveles. Al hacer esto, tres variables diseño deben ser usadas con el menor cuantil sirviendo como grupo de referencia.

Posterior al ajuste del modelo, se grafica los coeficientes estimados contra los puntos medianos de los grupos. En el primer cuantil se debe tomar un coeficiente igual a cero. Para ayudar en la interpretación se conecta los cuatro puntos y mediante una inspección visual se selecciona la forma paramétrica mas lógica para la escala de la variable.

El siguiente paso es reajustar el modelo usando la forma paramétrica sugerida por el gráfico y seleccionar uno que es significativamente diferente del modelo lineal y que tenga sentido para la investigación.

Es posible que dos o más parametrizaciones de la covariable pueda producir modelos similares considerando únicamente que deben ser significativamente diferentes del modelo lineal. Sin embargo, la experiencia de los investigadores permite afirmar que siempre un modelo será mas atractivo que el otro, con lo cual será más fácil interpretar los parámetros estimados.

b) Método de polinomios fraccionales

Es un método más analítico desarrollado por Royston y Altman (1994) que sugiere transformaciones. Lo que se desea es determinar que valores de x^p producen el mejor modelo para las covariables. Teóricamente se puede incluir la potencia como un parámetro adicional en el proceso de estimación pero esto agrandaría enormemente la complejidad en el proceso.

Los investigadores propusieron reemplazar totalmente la estimación máximo-verosímil de la potencia por una búsqueda a través de un pequeño pero razonable conjunto de posibles valores.

Este método puede ser usado con un modelo de regresión logística multivariada pero para simplificar se describe el proceso únicamente para una variable continua. La *logit* que es lineal en la covariable, es:

$$g(x,\beta) = \beta_0 + x\beta_1. \tag{2.49}$$

donde: β denota el vector de coeficientes del modelo. Un camino para generalizar la expresión es especificarla como sigue:

$$g(x,\beta) = \beta_0 + \sum_{j=1}^{J} F_j(x)\beta_j. \tag{2.50}$$

Las funciones $F_j(x)$ son un tipo especial de función potencia. El valor de la primera función es: $F_1(x) = x^{p_1}$. Los investigadores propusieron restricciones a la potencia que se encuentran en el conjunto $\wp = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ donde el valor inicial de p_0=0. Las funciones restantes son definidas como:

$$F_j(x) = \begin{cases} x^{p_j}, & p_j \neq p_{j-1} \\ F_{j-1}(x)\ln(x), & p_j = p_{j-1} \end{cases} \tag{2.51}$$

para j=1, ….., J y valores de potencia restringidos a $\wp$.

Considere como ejemplo de desarrollo de (2.50) y (2.51), el caso para J=2 con p_1 y p_2 igual a 2. Los pasos a realizar para obtener la *logit* son los siguiente:

Para j=1: Como p_1 es igual a dos y p_0 es cero, es decir, $p_1 ? p_0$, entonces, $F_1(x) = x^{p_1} = x^2 \cdot$

Para j=2: Dado que p_2 es igual a p_1, es decir, $p_2 = p_1 = 2$, entonces, $F_2(x) = x^2 \ln(x)$, donde: x^2 es $F_{j-1}(x) \cdot$

De este modo, se obtiene la logit siguiente:

$$g(x,\beta) = \beta_0 + \beta_1 x^2 + \beta_2 x^2 \ln(x) \cdot$$

Se debe considerar que para el caso de J=1, se ajustan 8 modelos cuando p_1 pertenece a $\wp$ y el mejor modelo será el que tiene el mayor valor del *log* de la verosimilitud. Para el caso de J=2 el proceso se repite para 36 modelos obtenidos de distintos pares de potencias $(p_1, p_2) \in \wp x \wp$ y el mejor modelo será también el que tiene mayor *log* de verosimilitud.

No se puede dejar de observar que aunque se obtuvo estos dos mejores modelos, para los valores de *J*, es necesario determinar si son significativamente mejores que el modelo lineal.

Sea *L(1)* la *log* verosimilitud para el modelo lineal, es decir, para J=1 y p_1=1, y además se denota como $L(p_1)$ a la *log* verosimilitud del mejor modelo para J=1 y $L(p_1,p_2)$ denota la *log* verosimilitud para el mejor modelo cuando J=2.

Royston y Altman (1994) sugirieron y verificaron con simulaciones que cada término en el polinomio fraccional contribuía aproximadamente con 2 grados de libertad al modelo, uno para la potencia y otro para el coeficiente. Así, la prueba de razón de verosimilitud parcial comparando el modelo lineal al mejor modelo para J=1 es:

$$G(1, p_1) = -2\{L(1) - L(p_1)\} \tag{2.52}$$

el cual es distribuido aproximadamente como chi-cuadrado con 1 grado de libertad bajo la hipótesis nula de linealidad en *x*.

Asimismo, la prueba de razón de verosimilitud parcial comparando el mejor modelo para J=1 contra el mejor modelo obtenido para J=2 es el siguiente:

$$G[p_1,(p_1, p_2)] = -2\{L(p_1) - L(p_1, p_2)\} \tag{2.53}$$

que es distribuida aproximadamente como chi-cuadrado con dos grados de libertad bajo la hipótesis nula que la segunda función es igual a cero.

En forma similar se debe trabajar para contrastar el modelo lineal contra el mejor modelo para J=2 que también es distribuido aproximadamente como una chi-cuadrado con 3 grados de libertad.

5) Una vez que se tenga refinado el modelo de efectos principales y comprobar que cada una de las variables continuas tiene la escala correcta, se chequea las interacciones entre las variables en el modelo. La existencia de interacción entre dos variables revela que el efecto de una de las variables no es constante sobre los niveles de la otra.

   La decisión final de considerar un término de interacción en el modelo se basa en el estadístico, así como en consideraciones prácticas y esta inclusión debe tener sentido desde la perspectiva del problema analizado. En ese sentido, se debe crear una lista de posibles pares de variables en el modelo que tenga alguna base científica de interactuar con otra.

La interacción de variables es creada como un producto aritmético de pares de variables de efectos principales las cuales son adicionadas una a la vez en el modelo conteniendo todos los efectos principales y determinando su significancia usando la prueba de razón de verosimilitud. La inclusión de este término que no fuera significativo provocaría un aumento en el error estándar estimado sin variar la estimación de los puntos estimados.

En general, para que un término interactivo altere las estimaciones puntuales y de intervalo, los coeficientes estimados para el término de interacción debe ser estadísticamente significativo.

Al concluir este paso el modelo se puede considerar como preliminar y se debe determinar su adecuación y chequear su ajuste lo cual se discute en la sección 2.6.

2.5.2.- Método *Stepwise.*

Este método es ampliamente usado en regresión lineal y todos los programas de análisis estadísticos desarrollan este tipo de análisis. Empleando este método se

puede obtener en forma rápida y efectiva un gran número de estadísticos ajustando ecuaciones de regresión logística simultáneamente.

El proceso que selecciona y elimina las variables del modelo se basa en un algoritmo estadístico que chequea la importancia de las variables y cada inclusión y exclusión se realiza sobre la base de una regla de decisión ajustada. La importancia de una variable es definida en términos de una medida de la significancia estadística del coeficiente para la variable.

En la regresión lineal, la estadística usada depende de la asunción del modelo, en el caso que los errores sean distribuidos como normal estándar se utiliza la prueba *F*. Por otro lado, en regresión logística se asume que los errores siguen una distribución binomial y su significancia es determinada vía la prueba chi-cuadrado de razón de verosimilitud. Entonces, un paso en el procedimiento es aquel que se considera más importante en términos estadísticos, es decir, aquel que produce el mayor cambio en la *log*-verosimilitud relativa al modelo que no contiene la variable.

A continuación se describe el algoritmo para una selección *forward* seguido por una eliminación *backward* en la regresión logística *stepwise*.

Paso inicial (0): Se tiene disponible un total de *p* variables independientes las cuales tienen la posibilidad de ser consideradas como importantes para el estudio de la variable respuesta. Se inicia con la inclusión únicamente del intercepto y se evalúa su *log*-verosimilitud, L_0. Seguido, se ajusta cada uno de los *p* posibles modelos de regresión logística univariados y se compara sus respectivas *log*-verosimilitudes. El valor obtenido de la *log*-verosimilitud para el modelo que contiene la variable x_j en el paso cero será denotado $L_j^{(0)}$, donde el sub-índice se refiere a la variable que ha sido adicionada y el superíndice se refiere al paso.

El valor de la prueba de razón de verosimilitud para el modelo que contiene únicamente la variable x_j y el intercepto se denotará por $G_j^{(0)} = -2(L_0 - L_j^{(0)})$, y su valor *p* será denotado por $p_j^{(0)}$. Por lo tanto, este valor *p* es determinado por la probabilidad $Pr[\chi^2(v) > G_j^{(0)}] = p_j^{(0)}$, donde: v=1, si x_j es continuo, y v=*k-1*, si x_j es policotómico con *k* categorías.

La variable más importante es aquella con el menor valor de *p*, y si se denota esta variable por x_{e_1} entonces $p_{e_1}^{(0)} = \min(p_j^{(0)})$; donde, *min* significa seleccionar el valor mínimo de las cantidades al interior de los paréntesis y el subíndice e_1 es

usado para denotar que la variable es un candidato para ingresar al modelo en el paso 1.

Hosmer y Lemeshow (2000), afirman y verifican con un ejemplo que la variable mas importante no tiene garantizado que es estadísticamente significativo.

La selección del nivel para juzgar la importancia de una variable es un aspecto crucial usando la regresión logística *stepwise*. Sea p_E la selección donde E es el soporte de entrada, la cual determinará cuantas variables son eventualmente incluidas en el modelo.

Muchos investigadores discutieron sobre el nivel que debe tener este indicador que separa las variables que son importantes de las que no lo son. Uno de los principales resultados a los que llegaron fue que el nivel de $p_E = 0.05$ es muy riguroso y que frecuentemente excluía variables importantes del modelo por lo cual recomiendan seleccionar un valor de p_E entre 0.15 y 0.20.

A veces la meta del análisis puede ser más amplia, y los modelos que contienen más variables se intentan para proporcionar un cuadro más completo de

modelos posibles. En estos casos, el uso de p_E = 0.25 o mayor puede ser razonablemente seleccionado.

Una variable es lo suficientemente importante como para ser incluida en el modelo si el valor de p para G es menor que p_E. Por lo tanto, el programa procede al paso (1) si $p_{e_1}^{(0)} < p_E$, o sino el proceso se detiene.

Paso (1): Este paso inicia con el ajuste del modelo de regresión logística conteniendo x_{e_1} y sea $L_{e_1}^{(1)}$ la log-verosimilitud del modelo. Para determinar si algunas de las restantes *p-1* variables son importantes dado que la variable x_{e_1} está en el modelo, se ajusta los *p-1* modelos de regresión logística que contengan x_{e_1} y x_j para *j* = 1, 2,…, *p y j ? e1*. Sea $L_{e_1 j}^{(1)}$ la log-verosimilitud para el modelo que contiene las variables x_{e_1} y x_j, entonces la estadística chi-cuadrado de razón de verosimilitud de este modelo versus el que contiene solo x_{e_1} se denota por:

$$G_j^{(1)} = -2(L_{e_1}^{(1)} - L_{e_1 j}^{(1)}) \qquad (2.54)$$

El valor de *p* para este estadístico es denotado por $p_j^{(1)}$. Sea x_{e_2} la variable con el menor valor de *p* en el paso 1, es decir, $p_{e_2}^{(1)} = \min(p_j^{(1)})$ y si este valor es menor que p_E, entonces, se procede al paso 2, sino se termina el proceso.

Paso (2): Este paso se inicia con el ajuste del modelo que contiene x_{e_1} y x_{e_2}. Es posible que únicamente x_{e_2} debiera ser incluido en el modelo y que x_{e_1} no es mucho más importante que éste, por lo tanto, el paso 2 incluye una revisión por eliminación *backward* que realiza el ajuste del modelo eliminando una de las variables adicionadas en el paso previo y determina la importancia de su continuidad.

Sea $L_{-e_j}^{(2)}$ la *log*-verosimilitud del modelo con x_{e_j} removido, de manera similar sea la prueba de razón de verosimilitud de este modelo en comparación al modelo completo en el paso 2:

$$G_{-e_j}^{(2)} = -2(L_{-e_j}^{(2)} - L_{e_1 e_2}^{(2)}) \qquad (2.55)$$

y $p_{-e_j}^{(2)}$ es su correspondiente valor *p*.

Para verificar si una variable debe ser eliminada del modelo el programa selecciona la variable que al ser removida produce el máximo valor *p*. Si se denota esta variable como x_{r_2}, entonces $p_{r_2}^{(2)} = \max(p_{-e_j}^{(2)}, p_{-e_2}^{(2)})$. Para decidir si

x_{r_2} debe ser removido, el programa compara $p^{(2)}_{r_2}$ a un segundo nivel pre-seleccionado, p_R , el cual indica algún nivel mínimo para continuar contribuyendo al modelo donde *R* es el soporte para la eliminación. Cualquier valor que se seleccione para p_R deberá ser superior al valor de p_E para evitar la posibilidad que el programa ingrese y remueva la misma variable en pasos sucesivos.

Si el máximo valor de *p* para remover $p^{(2)}_{r_2}$ excede a p_R , entonces x_{r_2} es removido del modelo, por el contrario, si $p^{(2)}_{r_2}$ es menor que p_R , entonces x_{r_2} permanece en el modelo. En otro caso, el programa continua con la fase de selección de variables.

En la fase de selección *forward* cada una de los *p-2* modelos de regresión logística son ajustados conteniendo x_{e_1} , x_{e_2} y x_j para *j* = 1, 2, …, *p*; donde j ? e_1, e_2. El programa evalúa la *log*-verosimilitud para cada modelo, calcula la prueba de razón de verosimilitud contra el modelo que contiene sólo a x_{e_1} y x_{e_2} y determina el correspondiente valor *p*. Sea x_{e_3} la variable con el mínimo valor *p*,

es decir, $p_{e_3}^{(2)} = \min(p_j^{(2)})$ y si este valor *p* es menor que p_E, $p_{e_3}^{(2)} < p_E$, entonces, el programa procede al paso (3), en otro caso, se detiene.

Paso (3): El procedimiento para el paso 3 es idéntico al paso 2, el programa ajusta el modelo incluyendo la variable seleccionada durante el paso previo, desarrolla la eliminación *backward* seguido por la selección *forward*. El proceso continua de esta manera hasta el último paso, el paso (*S*).

Paso (*S*): Este paso ocurre cuando todas las *p* variables han ingresado al modelo o todas las variables en el modelo tienen valores de *p* para remover aquellos que son menores que p_R, y las variables no incluidas en el modelo tienen valores de *p* para ingresar que excede p_E. El modelo en este paso contiene aquellas variables que son importantes con relación al criterio de p_E y p_R, las cuales pueden o no ser las variables reportadas en el modelo final. Por ejemplo, si los valores seleccionados de p_E y p_R corresponden a nuestra creencia para la significancia estadística, entonces el modelo en el paso *S* puede bien contener las variables significativas. Sin embargo, si se ha utilizado valores para p_E y p_R los cuales son menos rigurosos, entonces, se deberá seleccionar las variables para un modelo final de una tabla que resuma los resultados del procedimiento *stepwise*.

Existen dos métodos que pueden ser utilizados para seleccionar variables de una tabla resumen los cuales son comparables a los métodos comúnmente usados en la regresión lineal *stepwise*. El primero de ellos es basado en el valor *p* para ingresar al modelo en cada paso, mientras el segundo es basado en la prueba de razón de verosimilitud del modelo en el paso actual comparando con el modelo del paso anterior.

Sea *q* un paso arbitrario en el procedimiento. En el primer método se compara $p_{e_q}^{(q-1)}$ a un nivel de significancia pre-seleccionado *a* =0.15 y si es menor, entonces, el proceso avanza al paso *q*. Se detendrá en el paso cuando $p_{e_q}^{(q-1)}$ es superior a *a* y se considera el modelo obtenido en el paso previo para continuar con el análisis. En este método, el criterio de inclusión es basado en la prueba de la significancia del coeficiente para x_{e_q} condicional sobre $x_{e_1}, x_{e_2}, \ldots, x_{e_{q-1}}$ presentes en el modelo. Los grados de libertad para la prueba son 1 o *k*-1 dependiendo de si x_{e_q} es continuo o policotómicos con *k* categorías.

En el segundo método, se compara el modelo en el paso actual, paso *q*, no con el modelo en el paso previo *(q-1)* sino al modelo en el último paso, paso (*S*). Se evalúa el valor *p* para la prueba de razón de verosimilitud de estos dos modelos

y se procede de esta manera hasta este valor *p* que excede a a. Esto probaría que los coeficientes para las variables adicionadas al modelo del paso *q* al paso *S* son todos iguales a cero. Dado que la prueba tenía más grados de libertad que la prueba empleada en el primer método es posible que el segundo método seleccione un mayor número de variables que el primero.

Es conocido que el valor *p* calculado en el procedimiento de selección *stepwise* no son los valores *p* en el contexto de la prueba de hipótesis tradicional, sino que deberían entenderse como indicadores de la importancia relativa entre variables.

2.5.3.- Mejor sub conjunto de regresión logística.

Este procedimiento es una alternativa a la selección *stepwise* y fue diseñado en primera instancia, como casi todo lo anteriormente analizado para la regresión lineal, sin embargo, Hosmer, Jovanovic, y Lemeshow (1989) han mostrado que una generalización para regresión logística de lo presentado por Lawless y Singhal (1978) no son necesarias, razón por la cual se puede utilizar cualquier programa de regresión lineal sin problemas.

Desarrollar el mejor sub-conjunto de variables de regresión logística es más fácil usando la notación matricial. Sea *X* la matriz *n x (p + 1)* conteniendo los valores de las *p* variables independientes para cada sujeto con la primera columna conteniendo 1 para representar el término constante. Las *p* variables pueden ser el número total de variables o aquellas seleccionadas en la etapa de construcción del modelo. Sea *V* una matriz diagonal *n x n* con elementos generados de la siguiente manera: $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$, donde: $\hat{\pi}_i$ es el estimado de la probabilidad logística calculado usando el estimador de máxima verosimilitud $\hat{\beta}$ y la data para el *i*-ésimo caso, X_i.

Se puede mostrar que $\hat{\beta} = (X'VX)^{-1}X'Vz$, donde: $z = X\hat{\beta} + V^{-1}r$ y *r* es el vector de residuos $r = (y - \hat{\pi})$. Esta representación de $\hat{\beta}$ proporciona las bases para el uso de los programas de regresión lineal en donde se debe verificar que este permita ponderaciones, produzca estimados de coeficientes idénticos a $\hat{\beta}$ cuando se usó z_i como la variable dependiente y ponderación de casos, v_i, igual a los elementos de *V*.

Para reproducir los resultados del ajuste de máxima verosimilitud del programa de regresión logística usando uno de regresión lineal, se calcula para cada caso el valor de una variable dependiente como sigue:

$$
\begin{aligned}
z_i &= (1, x_i^{'})\beta + \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} \\
&= \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij} + \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} \\
&= \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)}
\end{aligned}
\tag{2.56}
$$

y la ponderación de los casos:

$$
v_i = \hat{\pi}_i(1-\hat{\pi}_i) \tag{2.57}
$$

Lo que se requiere antes de ejecutar el programa es el valor ajustado de $\hat{\pi}_i$, calcular el valor de z_i y v_i, luego se ejecuta este usando el valor de z_i como la variable dependiente, el valor de x_i para el vector de variables independientes, y el valor de v_i para los casos ponderados. Luego se puede demostrar que los residuales del ajuste son:

$$
z_i - \hat{z}_i = \frac{(y_i - \hat{\pi}_i)}{\hat{\pi}_i(1-\hat{\pi}_i)} \tag{2.58}
$$

y la suma de cuadrados de los residuales ponderados producidos por el programa es:

$$
\sum_{i=1}^{n} v_i (z_i - \hat{z}_i)^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)} \tag{2.59}
$$

la cual es X^2, una estadística chi-cuadrado de Pearson de un programa de regresión logística máximo verosímil.

La media de la suma de cuadrados residuales es: $s^2 = X^2/(n-p-1)$ y los estimados de los errores estándar de los coeficientes estimados producidos por el programa de regresión lineal son *s* veces la raíz cuadrada del elemento diagonal de la matriz $(X'VX)^{-1}$

Por lo tanto, para obtener el valor correcto como en la ecuación (2.13) se necesita dividir el estimado del error estándar producido por el programa de regresión lineal por *s*, la raíz cuadrada del error cuadrático medio (o error estándar del estimado).

La capacidad para duplicar el ajuste de máxima verosimilitud en un programa de regresión lineal derivó en la fundación del método sugerido para desarrollar el mejor sub-conjunto de regresión logística.

El sub-conjunto de variables seleccionado para ser el mejor modelo depende del criterio de selección adoptado. Entre los que fueron usados en primera instancia se encuentran: El R^2, razón de la suma de cuadrados de la regresión sobre la

suma de cuadrados del total, y el ajustado R^2, la razón del cuadrado medio de la regresión sobre el cuadrado medio del total. Debido a que un mayor número de variables proporciona un mayor R^2 no se recomienda la aplicación de ellos como indicador para determinar el mejor sub-conjunto de regresión logística.

En su lugar, se prefiere utilizar una medida desarrollada por Mallow (1973) denotado por C_q, la cual es una medida predictiva del error cuadrado. C_q de Mallow tiene la misma forma intuitiva como en la regresión lineal. En particular, se muestra que para un sub-conjunto *q* de las *p* variables:

$$C_q = \frac{X^2 + \lambda^*}{X^2/(n-p-1)} + 2(q+1) - n. \tag{2.60}$$

donde:

$$X^2 = \sum \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}, \tag{2.61}$$

el estadístico chi-cuadrado de Pearson para el modelo con *p* variables y λ^* es la estadística de la prueba multivariada de Wald para la hipótesis que los coeficientes para las *p-q* variables no incluidas en el modelo son iguales a cero.

Bajo la asunción que el modelo ajustado es uno correcto, la aproximación de los valores esperados de X^2 y λ^* son *(n-p-1)* y *p-q*, respectivamente. La sustitución de estas aproximaciones de los valores esperados en la expresión (2.60), con

$C_q=q+1$, proporcionan candidatos para el modelo óptimo. El programa seleccionará como el mejor sub-conjunto de regresión logística aquel con el menor valor de C_q.

El uso del mejor sub-conjunto de regresión logística en el programa de regresión lineal debe permitir seleccionar una base de *q* covariables importantes de las *p* posibles las cuales deben ser sometidas a una evaluación crítica.

Algunos programas utilizan la prueba del *Score* en vez de Wald para la selección del mejor sub-conjunto de regresión logística, en donde el mejor modelo será aquel con valor más grande en el estadístico de esta prueba. La lista de covariables y pruebas de *Score* de salida de los modelos óptimos en cada tamaño es especificado por el usuario. La dificultad con la que se encuentra esta metodología es que la prueba del *Score* aumenta conforme incrementa el número de variables en el modelo.

Una aproximación para la regresión logística se puede bosquejar a continuación. Primero, se asume que la chi-cuadrado de Pearson es igual a su media, es decir, $X^2 \approx (n - p - 1)$. Luego, se asume que la estadística de Wald para las *p-q* covariables excluidas puede ser aproximado por la diferencia entre los valores

de la prueba del *Score* para todos los *p* covariables y la misma prueba para *q* covariables llamada: $\lambda_q^* \approx S_p - S_q$, lo cual resulta en la siguiente aproximación, de la ecuación (2.41):

$$C_q = \frac{X^2 + \lambda^*}{X^2/(n-p-1)} + 2(q+1) - n.$$

$$\approx \frac{(n-p-1)+(S_p - S_q)}{1} + 2(q+1) - n$$

$$\approx S_p - S_q + 2q - p + 1. \tag{2.62}$$

donde: Los valores S_p y S_q son las respectivas pruebas para los modelos que contienen *p* y *q* covariables, respectivamente, y que son obtenidos de la salida del programa.

La ventaja del método propuesto es que se puede presentar más rápidamente muchos más modelos que los posibles en otras aproximaciones, sin embargo, existe una potencial desventaja que el método debe ser capaz de ajustar el modelo conteniendo todas las posibles covariables, el cual puede no ser posible en el análisis de un gran número de variables.

2.6.- Determinación de bondad de ajuste.

La finalidad de esta sección es presentar una serie de pruebas estadísticas a la que debe ser sometido el modelo obtenido por el método de máxima verosimilitud para conocer si efectivamente describe adecuadamente la variable respuesta.

Se concluye que el modelo ajusta los datos en forma adecuada si cumple lo siguiente:

1) Las medidas de resumen de distancia entre y y $\hat{y}$ son pequeñas;
2) La contribución de cada par $(y_i, \hat{y}_i)$, $i = 1,2,\ldots..,n$, a estas medidas resumen no son sistemáticos y es pequeño relativo a la estructura de error del modelo.

2.6.1.- Medidas de bondad de ajuste.

Al finalizar la etapa de construcción del modelo se puede seguir una secuencia lógica de pasos para determinar si el ajuste del modelo es adecuado. Se deberá primero calcular y evaluar todas las medidas de ajuste y luego examinar los componentes individuales de los estadísticos que puede ser gráficamente y en

tercer lugar se debe proceder con la examinación de otras medidas de la diferencia o distancia entre los componentes de y y $\hat{y}$.

Se debe considerar el efecto que tiene el ajuste del modelo en los grados de libertad disponibles para el desarrollo de la determinación del modelo. Hosmer y Lemeshow (2000) utilizan el término patrón covariable para describir un único conjunto de valores para las covariables en el modelo. Las diferentes combinaciones posibles entre ellas pueden dar origen a un patrón distinto.

Durante el desarrollo del modelo no es necesario trabajar con el número de patrones de covariables debido a que los grados de libertad para estas pruebas se basan en el número de parámetros que participan en los modelos y no sobre los patrones, sin embargo, este número puede ser importante cuando el ajuste del modelo es determinado.

La bondad del ajuste es determinado sobre el subconjunto de valores que fueron utilizados para ajustar las covariables en el modelo y no el total.

2.6.1.a.- Prueba Chi-cuadrado de Pearson y Desvianza

En regresión logística se cuenta con muchas formas de medir la diferencia de los valores ajustados y observados. Se debe enfatizar el hecho que los valores ajustados son calculados para cada patrón de covariables y dependen de la probabilidad estimada para aquel patrón. Se define el valor ajustado para el *j*-ésimo patrón $\hat{y}_j$, donde:

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}} \tag{2.63}$$

donde: $\hat{g}(x_j)$ es la *logit* estimada, y m_j es el número de elementos incluidos en el patrón. Para variables no agrupadas en patrones $m_j=1$.

Para un patrón de covariables particular la residual de Pearson es definida así:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1-\hat{\pi}_j)}} \tag{2.64}$$

El estadístico basado en estos residuales es la Chi-cuadrado de Pearson:

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \pi_j)^2 \tag{2.65}$$

Por su parte, la desvianza residual es definida de la siguiente manera:

$$d\left(y_j, \hat{\pi}_j\right) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j \hat{\pi}_j}\right) + (m_j - y_j)\ln\left(\frac{m_j - y_j}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{\frac{1}{2}} \qquad (2.66)$$

donde, los signos + o – son los mismos que en la expresión $(y_j - m_j\hat{\pi}_j)$.

De la expresión anterior, para patrones de covariable con $\hat{y}_j$=0, la desvianza residual es:

$$d\left(y_j, \hat{\pi}_j\right) = -\sqrt{2m_j\left|\ln\left(1-\hat{\pi}_j\right)\right|} \qquad (2.67)$$

y la desvianza residual cuando $y_j = m_j$ es:

$$d\left(y_j, \hat{\pi}_j\right) = \sqrt{2m_j\left|\ln\left(\hat{\pi}_j\right)\right|} \qquad (2.68)$$

De este modo se obtiene la estadística basada en la desvianza residual:

$$D = \sum_{j=1}^{J} d\left(y_j, \hat{\pi}_j\right)^2 \qquad (2.69)$$

y para el caso *J=n*, se cumple que la desvianza es:

$$D = -2\sum_{j=1}^{n}\left[y_j \ln\left(\frac{\hat{\pi}_j}{y_j}\right) + \left(1-y_j\right)\ln\left(\frac{1-\hat{\pi}_j}{1-y_j}\right)\right] \qquad (2.70)$$

La distribución de la estadística χ^2 y *D*, bajo la asunción que el modelo ajustado es correcto en todo aspecto, es supuesto que se distribuye como una Chi-cuadrado con *J-(p+1)* grados de libertad. Para la desvianza esta expresión se presenta dado que *D* es la prueba estadística de razón de verosimilitud de un modelo saturado con *J* parámetros contra el modelo ajustado con *p+1* parámetros. Similar teoría provee la distribución nula de χ^2.

2.6.1.b.- Prueba de Hosmer y Lemeshow

Los investigadores que dieron nombre a esta prueba propusieron en 1980 y 1982 formas de agrupamiento basadas en los valores de probabilidad estimados. Para simplificar la discusión suponga que se tiene el caso *J=n*, donde *n* columnas corresponden a los *n* valores de las probabilidades estimadas con la primera columna correspondiendo al menor valor y la *n*-ésima columna al mayor valor.

Las dos estrategias de agrupamiento fueron propuestas de esta manera:

1. Colapsar la tabla basado en los percentiles de la probabilidad estimada, y
2. Colapsar la tabla basado sobre los valores ajustados de la probabilidad estimada.

Con el primer método se usa un g=10 grupos resultando en el primer grupo los casos con probabilidad estimada más pequeña y el número de integrantes es $n'_1 = n / 10$, y el último grupo de igual tamaño ($n'_{10} = n'_1$) incluye los casos con los valores más grandes. Con el segundo método se usa 10 grupos con puntos de corte definidos en los valores $k / 10$, con k entre 1 y 9 y los grupos contienen todos los sujetos con probabilidad estimada entre cortes adyacentes.

Para la fila y=1, una estimación del valor esperado son obtenidos por la suma de la probabilidad estimada para todos los elementos en el grupo y para el caso de la fila y=0, el valor esperado estimado es obtenido mediante la suma de los sujetos en el grupo multiplicado por uno menos la probabilidad estimada. Para cada estrategia de agrupamiento, la estadística de bondad de ajuste de Hosmer y Lemeshow, $\hat{C}$, es obtenido

mediante el cálculo del estadístico Chi-cuadrado de Pearson de la tabla g x 2, tabla de las frecuencias observadas y esperadas estimadas.

Se puede definir a $\hat{C}$ como sigue:

$$\hat{C} = \sum_{k=1}^{g} \frac{\left(o_k - n'_k \bar{\pi}_k\right)^2}{n'_k \pi_k \left(1 - \pi_k\right)} \tag{2.71}$$

donde: n'_k es el número total de casos en el *k*-ésimo grupo, c_k denota el número de patrones de covariables en el *k*-ésimo decil,

$$o_k = \sum_{j=1}^{c_k} y_j \tag{2.72}$$

es el número de respuestas entre el patrón de covariables C_k, y

$$\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \hat{\pi}_j}{n'_k} \tag{2.73}$$

es la probabilidad estimada promedio.

Usando una extensión del conjunto de simulaciones, Hosmer y Lemeshow (1980) demostraron que cuando *J=n* y el modelo de regresión logística ajustado es el correcto, la distribución de la estadística $\hat{C}$ es debidamente aproximada por la distribución Chi-cuadrado con *g-2* grados de libertad.

En esta prueba se espera no rechazar la H_0, es decir, que el modelo ajusta satisfactoriamente los datos.

2.6.1.c.- Tabla de clasificación

Aquí se resume los resultados de la variable respuesta que corresponden al ajuste del modelo de regresión logística comparados en clasificación cruzada con la variable dicotómica observada.

La clasificación de la variable resultado se realiza comparando las probabilidades estimadas con un punto de corte *c* de tal forma que los valores mayores a *c* tomarán el valor de 1 y los otros 0. El valor más común de *c* es 0.5.

Lo atractivo de este tipo de aproximación a la determinación del modelo es dado por la cercana relación entre la regresión logística y el análisis discriminante cuando la distribución de las covariables es normal multivariada con dos posibles grupos de salida.

En esta aproximación, la probabilidad estimada es utilizada para predecir el grupo de clasificación de un individuo y, dado que esto se realiza de acuerdo a algún criterio derivado del modelo, se presume que este ajusta los datos adecuadamente. Desafortunadamente no siempre es así, se puede dar el caso en donde el modelo es el correcto y, por lo tanto, ajusta los datos adecuadamente pero es pobre en la clasificación de los mismos.

Suponga que $P(Y=1)=\theta_1$ y que $X \sim N(0,1)$ en el grupo con *Y*=0, y $X \sim N(\mu,1)$ en el grupo con *Y*=1, entonces, en el modelo de análisis discriminante la pendiente será $\beta_1=\mu$ y el intercepto es:

$$\beta_0 = \ln\left[\frac{\theta_1}{(1-\theta_1)}\right] - \frac{\mu^2}{2} \tag{2.74}$$

La probabilidad de clasificación (PC) es definida de la siguiente forma:

$$PC = \theta_1 \Phi\left\{\frac{1}{\beta_1}\ln\left[\frac{(1-\theta_1)}{\theta_1}\right] - \frac{\beta_1}{2}\right\} + (1-\theta_1)\Phi\left\{\frac{1}{\beta_1}\ln\left[\frac{\theta_1}{(1-\theta_1)}\right] - \frac{\beta_1}{2}\right\} \tag{2.75}$$

donde: F es la función de distribución acumulada de una $N(0,1)$.

Así, el error esperado es una función de la magnitud de la pendiente y no necesariamente del ajuste del modelo y la exactitud de la clasificación no

direcciona nuestro criterio de bondad de ajuste, es decir, que las distancias entre los valores observados y esperados no sean sistemáticos y dentro de la variación del modelo. Sin embargo, la tabla de clasificación puede ser útil junto a otras medidas basados más directamente en residuales.

Antes de continuar se debe definir dos conceptos fundamentales la sensitividad y especificidad relacionados a la correcta clasificación del modelo logístico en comparación a lo observado.

En la tabla 2.4 se resume los sujetos observados y los clasificados por el modelo en la cual se identifican los casilleros con las letras *a, b, c,* y *d* que se muestra a continuación:

Tabla 2.4: Tabla de Clasificación basado en un modelo de regresión logística.

Observado	**Clasificado**		
	Con característica	**Sin característica**	**Total**
Con característica	**a**	**b**	**a + b**
Sin característica	**c**	**d**	**c + d**
Total	**a + c**	**b + d**	**N**

Se define la **sensitividad** como el porcentaje de aciertos en la clasificación del modelo logístico dado que el sujeto clasificado posee la característica estudiada. Sea (a+b) el número de caso que fueron observados como poseedores de la característica y (a) el número de casos que fueron correctamente clasificados, entonces:

$$\text{Sensitividad} = \left(\frac{a}{a+b}\right)x100 \tag{2.76}$$

Asimismo, se define la **especificidad** como el porcentaje de casos de éxito en la clasificación siguiendo el modelo logístico cuando los sujetos observados no cumplen con poseer la característica estudiada. Sea (c+d) el número de casos que no poseían la característica en la muestra y d el número de casos con clasificación correcta, entonces:

$$\text{Especificidad} = \left(\frac{d}{c+d}\right)x100 \tag{2.77}$$

El porcentaje de clasificación correcta en la tabla es estimado de la siguiente forma:

$$\left(\frac{a+d}{N}\right)x100 \tag{2.78}$$

Hosmer y Lemeshow (2000) afirma que la clasificación es sensible al tamaño relativo de los dos grupos componentes y siempre favorece la clasificación hacia el grupo más grande lo cual es independiente del ajuste del modelo y puede ser observado en la expresión de PC que es una función de θ_1.

La desventaja de utilizar PC como un criterio es que reduce un modelo probabilístico, donde la salida es medida en forma continua, a un modelo dicotómico donde la variable predictora es binaria.

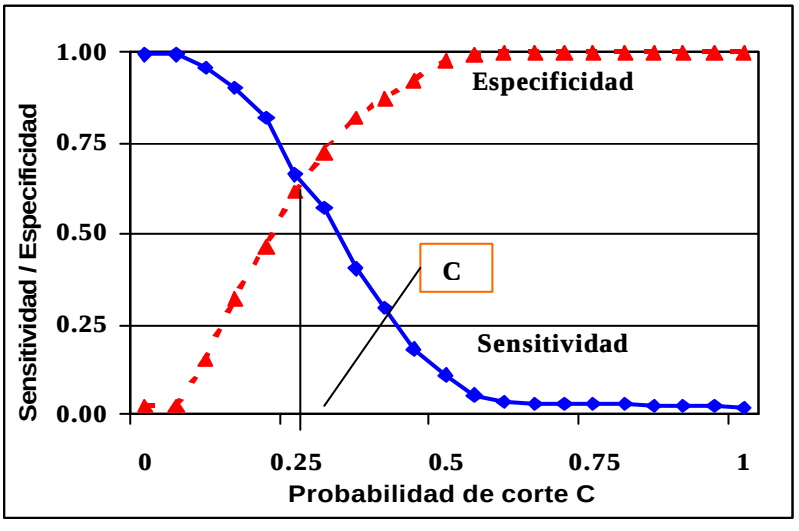
La sensitividad y especificidad no deben ser usadas como medidas de funcionamiento del modelo debido a que dependen fuertemente de la distribución de probabilidades en la muestra. Por lo tanto, si se comparan dos modelos la diferencia con respecto a la especificidad y sensitividad puede depender enteramente de una "mezcla paciente" (*patient mix*) que da la superioridad de un modelo sobre otro.

2.6.1.d.- Área bajo la curva *ROC* (*Receiver Operating Characteristic*)

Este método nos brinda una descripción más completa de la clasificación de los datos. La curva se origina del trazado de la probabilidad de detectar una señal verdadera (sensitividad) y falsa (1- especificidad) para un rango entero de posibles puntos de corte.

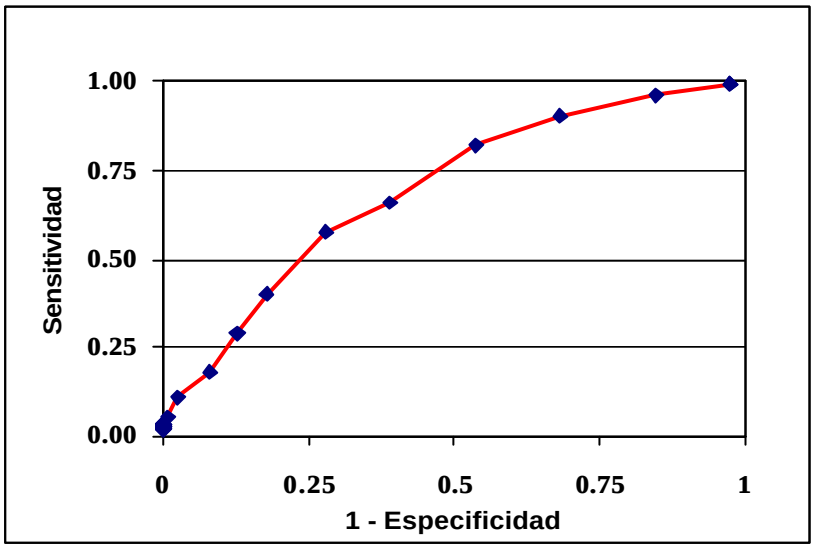
El área bajo la curva *ROC*, en el rango de cero a uno, provee una medida de capacidad del modelo para discriminar entre los sujetos que experimentan la respuesta de interés contra los que no.

Figura 2.3: Sensitividad versus Especificidad para distintos puntos de corte.



Si nuestro objetivo fuera obtener un punto de corte óptimo entonces se debe elegir aquel que maximice tanto la sensitividad como la especificidad. Esta selección es facilitada a través de un gráfico como la figura 2.3 en el que se puede apreciar que el punto óptimo es *c* donde ambas curvas se cruzan.

Figura 2.4: Curva ROC. Sensitividad y 1-Especificidad para distintos puntos de corte.



Un gráfico de la sensitividad versus (1 – especificidad) sobre todos los posibles puntos de corte es mostrado en la figura 2.4. Aquí la curva generada es llamada la curva *ROC* y el área bajo ella provee una medida de la capacidad discriminante de tal forma que un sujeto que tenga la

característica tendrá una alta probabilidad *P(Y=1)* respecto a otro que no la posea de ser clasificado correctamente.

Las figuras 2.3 y 2.4 fueron adaptadas de Hosmer y Lemeshow (2000, pp. 162-163).

Se puede adoptar como regla general que una discriminación es considerada aceptable si el valor del ROC se encuentra entre 0.7 y 0.8, como excelente para valores entre 0.8 y 0.9, y sobresaliente para mayores de 0.9. Asimismo, es inusual encontrar valores ROC por encima de 0.9 debido a que esto significaría una separación completa y en ese caso es imposible calcular los parámetros en la regresión logística.

2.6.2.- Diagnóstico de la regresión logística.

La clave para el diagnóstico de regresión logística, como en regresión lineal, son los componentes de la suma de cuadrados residuales. En la regresión lineal la asunción principal es que la varianza del error no depende de la media condicional $E(Y_j / x_j)$.

En regresión logística se tiene errores binomiales y como un resultado la varianza del error es una función de la media condicional:

$$\begin{aligned} \text{Var}\,(Y_j/x_j) &= m_j E\,(Y_j/x_j)[1 - E\,(Y_j/x_j)] \\ &= m_j \pi(x_j)[1 - \pi(x_j)] \end{aligned} \tag{2.79}$$

Por lo tanto, se inicia con residuales como en (2.64) y (2.66) los cuales han sido divididos por estimados de sus errores estándar. Sea r_j y d_j los valores como en las expresiones mencionadas y para patrones de covariables x_j se espera que si el modelo de regresión logístico es correcto estas cantidades tienen una media aproximadamente igual a cero y una varianza aproximada a uno.

Para cada patrón de covariables, existen otros valores importantes para la formación e interpretación del diagnóstico de la regresión lineal como son la matriz *hat* y los valores *leverage* que se derivan de esta.

La matriz *hat* provee los valores ajustados como una proyección de la variable respuesta en el espacio covariable. Sea X una matriz $Jx(p+1)$ que contiene los valores para todos los J patrones covariables formados de los valores observados de las p covariables, con valores de uno en la primera columna para

reflejar la presencia del intercepto en el modelo. Esta matriz es frecuentemente llamada matriz de diseño.

En regresión lineal la matriz *hat* es: $H = X(X'X)^{-1}X'$ y sea $\hat{y} = Hy$, entonces, se puede escribir los residuales de regresión $\hat{y} - y$ de la siguiente manera: $(I - H)y$ donde *I* es la matriz identidad de orden *J*.

Pregibon (1981) derivó una aproximación lineal a los valores ajustados usando mínimos cuadrados ponderados como un modelo que produjo una matriz *hat* para la regresión logística:

$$H = V^{1/2}X(X'VX)^{-1}X'V^{1/2} \qquad (2.80)$$

donde: *V* es una matriz diagonal *JxJ* con elementos de la siguiente forma:

$$v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)] \qquad (2.81)$$

Los elementos diagonales de la matriz *hat* son llamados valores *leverage* en la regresión lineal y son proporcionales a la distancia de x_j a la media de los datos. Este concepto de distancia es muy importante aquí debido a que puntos distanciados de la media pueden tener influencia considerable sobre los valores de los parámetros estimados. La extensión del concepto de *leverage* a la regresión logística requiere de algunos detalles adicionales.

Sea h_j el *j*-ésimo elemento diagonal de la matriz *H* definida en la ecuación (2.80), se puede mostrar que:

$$\begin{aligned} h_j &= m_j \pi(x_j)[1-\pi(x_j)] x_j' (X'VX)^{-1} x_j' \\ &= v_j x b_j \end{aligned} \tag{2.82}$$

donde:

$$b_j = x_j' (X'VX)^{-1} x_j' \tag{2.83}$$

y $x_j' = \left(1, x_{1j}, x_{2j}, ..., x_{pj}\right)$ es el vector de valores covariables definiendo el *j*-ésimo patrón covariable. La suma de los elementos de la diagonal de *H* es igual a *p+1*, el número de parámetros en el modelo.

En la regresión lineal, la dimensión de la matriz *hat* es usualmente *nxn* y, por lo tanto, ignora algún patrón de covariables común en la data. Con esta formulación, cualquier elemento diagonal en la matriz tiene un límite superior de 1/*k*, donde: *k* es el número de sujetos con el mismo patrón de covariables.

Si se formula la matriz *hat* para la regresión logística como una matriz *nxn*, entonces, cada elemento diagonal es limitado por $1/m_j$, donde m_j es el número total de sujetos con el mismo patrón de covariables. Cuando la matriz *hat* se

basa en datos agrupados por patrones de covariables, el límite superior para un elemento diagonal es 1.

Es recomendable que la estadística de diagnóstico sea calculado tomando en cuenta patrones de covariable, lo cual es más importante cuando el número de patrones de covariables J es mucho menor que n o en el caso que alguno de los valores de m_j son mayores que 5. Esto debido a que cuando J es menor que n existe el riesgo que se pueda errar en hallar la influencia y/o en un pobre ajuste de los patrones covariables.

Considere el patrón de covariable con m_j sujetos, y_j=0, y la probabilidad logística estimada $\hat{\pi}_j$, entonces, el residual de Pearson definido en (2.64) calculado individualmente para cada sujeto con este patrón de covariables es:

$$r_j = \frac{(0 - \hat{\pi}_j)}{\sqrt{\hat{\pi}_j(1 - \hat{\pi}_j)}}$$

$$= -\sqrt{\frac{\hat{\pi}_j}{(1 - \hat{\pi}_j)}}, \qquad (2.84)$$

mientras que los residuales de Pearson basado en todos los sujetos con este patrón de covariables es:

$$r_j = \frac{(0 - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

$$= -\sqrt{m_j}\sqrt{\frac{\hat{\pi}_j}{(1-\hat{\pi}_j)}} \qquad (2.85)$$

los cuales incrementan negativamente cuando m_j aumenta.

Si se usa la estadística de Pregibon (1981) para aproximar a la regresión lineal los residuales del *j*-ésimo patrón de covariables, $[y_j - m_j\hat{\pi}(x_j)] \approx (1 - h_j)y_j$, entonces, la varianza residual es: $m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)](1 - h_j)$, lo cual sugiere que la residual de Pearson no tiene varianza igual a 1 a menos que ellos hayan sido estandarizados. La residual de Pearson estandarizada para patrones de covariables x_j es:

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_j}} \qquad (2.86)$$

Otra estadística de diagnóstico útil es una que examina el efecto que tiene sobre el valor de coeficientes estimados y todas las medidas resumen de ajuste, como la X^2 y D, la eliminación de todos los sujetos con un patrón covariable particular.

La variación en el valor de los coeficientes estimados es análogo a la medida propuesta por Cook (1977, 1979) para regresión lineal la cual es obtenida como la diferencia estandarizada entre $\hat{\beta}$ y $\hat{\beta}_{-J}$, donde estas representan estimadores de máxima-verosimilitud calculados usando todos los J patrones covariables y excluyendo los m_j sujetos con patrón x_j, respectivamente, y estandarizando vía la matriz de covarianza estimada de $\hat{\beta}$.

Pregibon (1981) mostró mediante una aproximación lineal que para la regresión logística esta cantidad es:

$$\begin{aligned}\Delta\hat{\beta}_j &= (\beta - \beta_{-j})'(X'VX)(\beta - \beta_{-j}) \\ &= \frac{r_j^2 h_j}{(1-h_j)^2}. \\ &= \frac{r_{sj}^2 h_j}{(1-h_j)}.\end{aligned} \tag{2.87}$$

Usando una aproximación lineal similar se puede probar que el decrecimiento en el valor de la Chi-cuadrado de Pearson debido a la eliminación de los sujetos con patrón de covariables x_j es:

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)}$$

$$= r_{sj}^2. \qquad (2.88)$$

Una cantidad similar puede ser obtenida por el cambio en la desvianza,

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}. \qquad (2.89)$$

Si se reemplaza r_j^2 por d_j^2 se produce la siguiente aproximación:

$$\Delta D_j = \frac{d_j^2}{(1-h_j)}. \qquad (2.90)$$

la cual es similar en forma a la expresión en la ecuación (2.88).

Estas estadísticas de diagnóstico son conceptualmente muy atractivas debido a que ellos permiten identificar esos patrones covariables que tienen pobre ajuste (alto valor de ΔX_j^2 y/o ΔD_j), y aquellos que tienen un gran reparto de influencia sobre los valores de los parámetros estimados (grandes valores de $\Delta\hat{\beta}_j$). Después de identificar estos patrones de influencia (casos) se puede comenzar a direccionar el rol que tienen en el análisis.

Se debe recordar que se espera de la aplicación de todo lo que se mostró anteriormente en esta sección. Considere primero la medida de ajuste ΔX_j^2, esta medida es pequeña cuando y_j y $m_j\hat{\pi}(x_j)$ son cercanos. Lo más probable es que suceda cuando y_j=0 y $\hat{\pi}(x_j)$<0.1, o $y_j=m_j$ y $\hat{\pi}(x_j)$>0.9.

Tabla 2.5: Valores probables de las estadísticas de diagnóstico por valores de la probabilidad logística estimada.

$\hat{\pi}$	Estadísticas de diagnóstico		
	ΔX^2	$\Delta\hat{\beta}$	h
< 0.1	Grande o pequeño	Pequeño	Pequeño
0.1 – 0.3	Moderado	Grande	Grande
0.3 – 0.7	Moderado o pequeño	Moderado	Moderado o pequeño
0.7 – 0.9	Moderado	Grande	Grande
> 0.9	Grande o pequeño	Pequeño	Pequeño

Igualmente Δx_j^2 es grande cuando y_j está mas lejos de $m_j\hat{\pi}(x_j)$ y es más posible de ocurrir cuando el valor de y_j=0 y $\hat{\pi}(x_j)$>0.9, o con $y_j=m_j$ y $\hat{\pi}(x_j)$<0.1.

Estos mismo patrones covariables no tienen usualmente un gran $\Delta\hat{\beta}_j$, por lo tanto, cuando $\hat{\pi}(x_j)$<0.1 o $\hat{\pi}(x_j)$>0.9, $\Delta\hat{\beta}_j = \Delta X_j^2 h_j$, y h_j es próximo a cero.

El diagnóstico de influencia, $\Delta\hat{\beta}_j$, es grande cuando ambos ΔX_j^2 y h_j son al menos moderados lo cual puede ocurrir cuando los valores de $\hat{\pi}(x_j)$ se encuentren entre 0.1 y 0.3 o entre 0.7 y 0.9. Se consolida lo manifestado en la tabla 2.5 para ser usado como una guía para entender e interpretar las estadísticas de diagnósticos.

En regresión logística, al igual que en la regresión lineal, se realiza en primer lugar una determinación visual, cómo la distribución del diagnóstico bajo la hipótesis que el modelo ajustado es conocido solo en ciertos ajustes limitados.

Se definen siete estadísticos de diagnóstico los cuales pueden ser divididos en tres categorías:

1) La construcción básica de bloques, los cuales son de interés en ellos mismos pero también son usados para formar otros diagnósticos, r_j, d_j, h_j.
2) Medidas derivadas del efecto de cada patrón covariable sobre el ajuste del modelo, r_{sj}, ΔX_j^2 y ΔD_j.

3) Una medida derivada del efecto de cada patrón covariables sobre el valor de los parámetros estimados $\Delta\hat{\beta}_j$.

Un número de diferentes tipos de gráficos ha sido sugerido para usarse dirigidos a un aspecto particular del ajuste, algunos son formados de los siete diagnósticos mientras otros requieren cálculos adicionales. Se hace impracticable el intentar considerar todos los gráficos sugeridos razón por la cual se presta atención a aquellos que tenga mayor facilidad de obtener su interpretación en el análisis de regresión logística.

Los gráficos que se desarrollan son los siguientes:

1.- Gráfico ΔX_j^2 versus $\hat{\pi}_j$.

2.- Gráfico ΔD_j versus $\hat{\pi}_j$.

3.- Gráfico $\Delta\hat{\beta}_j$ versus $\hat{\pi}_j$.

Otros que son útiles en algunas ocasiones:

4.- Gráfico ΔX_j^2 versus h_j.

5.- Gráfico ΔD_j versus h_j.

6.- Gráfico $\Delta\hat{\beta}_j$ versus h_j, debido a que permite una determinación directa de la contribución de la leverage al valor de la estadística de diagnóstico. Un gráfico adicional que tiene especial utilidad es:

7.- Gráfico ΔX_j^2 versus $\hat{\pi}_j$, donde el tamaño del símbolo graficado es proporcional al tamaño de $\Delta\hat{\beta}_j$.

Se prefiere utilizar los diagnósticos ΔX^2 y ΔD graficados contra la probabilidad logística estimada en lugar de r_j y d_j contra $\hat{\pi}_j$, debido a las siguientes razones:

1..- Cuando $\tilde{n}$, mayor número de residuales positivos corresponden a los patrones covariables donde $y_j=m_j$ y residuales negativos cuando $y_j=0$. Por lo tanto, el signo del residual no es útil.

2.- Grandes residuales, sin importar el signo, corresponden a pobres ajustes de puntos.

3.- La forma de la curva permite determinar que patrones tienen y_j=0 y cuales tienen $y_j=m_j$.

Las formas de los dos primeros gráficos son similares y muestran curvas cuadráticas donde los puntos de la curva van desde lo alto del lado izquierdo a la esquina inferior del lado derecho que corresponde al patrón de covariables con $y_j=m_j$. La ordenada de estos puntos es proporcional a $(1-\hat{\pi}_j)^2$ desde $m_j=1$ para la mayoría de patrones covariables. Los puntos sobre la otra curva van de tal forma que hacen una cruz con la anterior y corresponden al patrón covariable cuando $y_j=0$. La ordenada de estos puntos es igualmente proporcional a $(0-\hat{\pi}_j)^2$. Patrones de covariable que son pobremente ajustados serán generalmente representados por puntos que se ubican en la parte superior de la esquina izquierda o derecha del gráfico. La determinación de la distancia es parcialmente basada sobre valores numéricos e impresión visual.

Una propiedad de Pearson con respecto a la desvianza asegura que el rango de ΔX^2 es mucho mayor que ΔD y siempre que se pueda elegir se prefieren estos gráficos versus π. Un ajuste razonable se obtiene cuando la mayor parte de los valores de ΔX^2 y ΔD son menores de 4 o en su defecto no mucho mayor de 4. Este valor se usa como una aproximación del percentil superior 95% de las distribuciones mencionadas, bajo las m distribuciones asintóticas, estas cantidades son distribuidas aproximadamente como $\chi^2_{(1)}$ con $\chi^2_{0.95}(1) = 3.84$.

En el tercer gráfico, el diagnóstico de influencia $\Delta\beta$ es graficado con respecto a $\hat{\pi}$ donde se podría apreciar puntos que se alejan del resto de la data. Se observa que los valores entre ellos mismo no son valores diferenciadamente grandes. La experiencia en este tipo de análisis permite afirmar que la influencia de diagnóstico debe ser mayor que 1.0 para que un patrón de covariable individual tenga efecto sobre los coeficientes estimados, sin embargo, hay siempre excepciones y es recomendable observar siempre los valores que se distancia del resto de $\Delta\beta$.

Se debe considerar que grandes valores de $\Delta\beta$ ocurren más frecuentemente cuando ΔX^2 y *leverage* son al menos moderadamente grandes, asimismo, esto puede ocurrir cuando cada componente es grande.

El gráfico número 7 por su parte presenta ΔX^2 versus $\hat{\pi}$ con el tamaño del símbolo proporcional a $\Delta\beta$. Este gráfico permite comprobar la contribución de residuales y leverage a $\Delta\beta$. Puede darse el caso en que el círculo de mayor tamaño se encuentre la parte que corresponde al mayor valor de ΔX^2 y otro círculo de gran tamaño en el menor valor de ΔX^2 pero se ubica en la región donde se espera encontrar el máximo *leverage.*

Un problema que se identifica en el diagnóstico de influencia $\Delta\beta$ es que siendo una medida de cambio sobre todos los coeficientes en el modelo simultáneamente es necesario examinar los cambios en los coeficientes individuales debido a patrones de covariables específicos identificados como influénciales.

Suponga que se tiene la situación en la cual las estadísticas de resumen indican que hay una desviación sustancial de ajuste y que se tiene evidencia que más de unos cuantos patrones de covariables, y_j difieren de $m_j\hat{\pi}_j$, es probable que al menos una de las siguientes tres cosas se están cumpliendo: (1) El modelo logístico no provee una buena aproximación a la correcta relación entre la media condicional, $E(Y/x_j)$, y x_j, (2) no se ha medido y/o incluido una importante covariable en el modelo, o (3) al menos una de las covariables en el modelo no fue ingresado en la escala correcta.

La gran conclusión de esta sección es que no se debe proceder a presentar los resultados de un modelo ajustado hasta que el ajuste mencionado no haya superado las estadísticas de resumen y las de diagnóstico.

2.6.3.- Determinación de ajuste vía validación externa.

En algunas situaciones se puede separar una parte de los datos, ajustar el modelo con los datos restantes, que son la mayoría, y luego probar el ajuste de este con los datos excluidos en primer término. En otras ocasiones puede ser posible obtener una nueva muestra de datos que determinen la bondad de ajuste del modelo desarrollado.

Este tipo de determinación es frecuentemente llamado validación y puede ser importante cuando el modelo ajustado es usado para predecir la respuesta de casos futuros. La razón para considerar este tipo de determinación de desarrollo del modelo es que el ajuste se realiza siempre de manera optimista sobre el conjunto de datos comprendido en el modelo desarrollado.

El uso de validación de datos se eleva a una determinación de bondad de ajuste donde el modelo ajustado es considerado como teóricamente conocido y no se ejecuta una estimación.

Algunos de los diagnósticos discutidos en la sección anterior ($\Delta X^2, \Delta D, \Delta\beta$) imitan esta idea por calcular para cada patrón covariable una cantidad basada

en la exclusión de un patrón particular, así, con un nuevo conjunto de datos una determinación más completa es posible.

Los métodos para la determinación de bondad de ajuste en la validación muestral, paralelos a estos, son descritos en la sección anterior para el desarrollo muestral. La mayor diferencia es que los valores de los coeficientes en el modelo son considerados como constantes fijas más que valores estimados.

Suponga que la validación muestral consiste de n_v observaciones (y_j, x_j), para *j*=1, 2, …., n_v, los cuales pueden ser agrupados en J_v patrones covariables.

En concordancia con la notación previa, sea y_j el número de respuestas positivas entre los m_j sujetos con patrón covariable $x = x_j$ para *j*=1,2,…, J_v. Además, la probabilidad logística para el *j*-ésima patrón covariable es π_j, el valor del modelo logístico estimado previamente usando el patrón x_j de la validación muestral. Las cantidades mencionadas brindan las bases para el cálculo de las medidas de ajuste, χ^2, *D* y *C* de la validación muestral.

Los cálculos de la chi-cuadrado de Pearson se obtienen de la ecuación (2.65) con obvios cambios de la validación muestral. En este caso, χ^2 es calculado como la suma de los J_v términos independientes. Si cada $m_j\pi_j$ es lo suficientemente grande como para usar la aproximación normal de la binomial, entonces, χ^2 es distribuido como $\chi^2(J_v)$ bajo la hipótesis que el modelo es correcto.

En la práctica, se espera que el número de sujetos observados en cada patrón covariable sea pequeño, la mayor parte m_j=1, por lo tanto, no se puede emplear las *m* distribuciones asintóticas. Como una alternativa se puede utilizar los resultados obtenidos por Osius y Rojek (1992) que obtienen una estadística que sigue la distribución normal estándar (Z) bajo la hipótesis que el modelo es correcto y J_v es suficientemente grande. Un proceso similar se presentó anteriormente, específicamente al calcular estadísticos estandarizados:

$$Z = \frac{X^2 - J_v}{\sigma_v}, \tag{2.91}$$

donde:

$$\sigma_v^2 = 2J_v + \sum_{j=1}^{J_v} \frac{1}{m_j\pi_j(1-\pi_j)} - 6\sum_{j=1}^{J_v} \frac{1}{m_j} \tag{2.92}$$

La prueba usa un valor de probabilidad de 2 colas basados en Z.

CAPÍTULO III

SISTEMA DE HIPÓTESIS Y DEFINICIONES OPERATIVAS.

En este capítulo se presentan las hipótesis a verificar en cumplimiento de los objetivos del estudio.

3.1.- Sistema de hipótesis.

3.1.1.- Hipótesis General

Las empresas exportadoras del sector Confecciones que continúan con sus operaciones de exportación deben esta continuidad principalmente al número de mercados destino a los cuales van dirigidas sus ventas, a la cantidad de productos que comercian, a su participación en actividades de importación, al tamaño de las mismas, a la exportación de productos de otros sectores, y al factor de continuidad.

3.1.2.- Hipótesis específicas:

Se consideran en el estudio dos hipótesis específicas.

3.1.2.a.- Hipótesis específica 1:

Las empresas exportadoras del sector Confecciones se clasifican en dos categorías mutuamente excluyentes respecto a su continuidad en las operaciones de exportación (continuas o salientes) para el periodo anual inmediato siguiente.

3.1.2.b.- Hipótesis específica 2:

Los factores que describen la continuidad de las empresas exportadoras del sector Confecciones del Perú son una función lineal que deriva en un modelo *logit*.

3.2.- Definiciones operativas:

Con la finalidad de obtener uniformidad de conceptos se incluye en esta sección las siguientes definiciones.

3.1.3.a.- Definición 1: Estado de las empresas exportadoras.

La Empresa Exportadora Continua es aquella que registra movimiento de exportación por un valor superior a los US$ 5000 en el presente año y en el inmediato anterior. Por otro lado, las empresas Entrantes y Salientes presentan exportaciones únicamente en el presente año o en el periodo anterior, respectivamente.

3.1.3.b.- Definición 2: Factor de continuidad.

El factor de continuidad está definido por el número de años, incluyendo el periodo de análisis, que la empresa ha mantenido la condición de continua, es decir, exportó sin interrupción alguna, la cual es independiente del nivel de

ventas alcanzado siempre que supere la cota de cinco mil dólares americanos impuesta para ser sujeto de estudio.

3.1.3.c.- Definición 3: Clasificación de empresas según exportaciones.

Las empresas se clasifican según su valor de exportación en tres variables *dummy* según cuartiles. X_{11} es igual a uno si las exportaciones de la empresa se encuentran entre US$ 53.7 y US$ 248.2 miles, X_{12} es igual a uno para exportaciones entre US$ 15.5 y US$ 53.7 miles y X_{13} es uno si sus ventas al exterior registran valores entre US$ 5.0 y US$ 15.5 miles. En todos los casos si no se encuentran en el rango asume el valor de cero.

CAPÍTULO IV

MODELAMIENTO DE LA CONTINUIDAD EXPORTADORA DE LAS EMPRESAS DE CONFECCIONES.

4.1.- Modelamiento y obtención de resultados.

La presente investigación se basará en la información que las empresas exportadoras de confecciones presentan en la Declaración Única de Aduana (DUA) a la Superintendencia Nacional de Administración Tributaria (SUNAT) en sus actividades de exportación desde el año 2002 para evitar distorsiones a través del tiempo y por el cambio del sistema arancelario producido en el 2001.

La aplicación del método desarrollado en el capítulo II se iniciará directamente a nivel multivariado debido a que el análisis previo se realizó en el capítulo I. Las variables consideradas en la investigación se encuentran en la tabla 4.1.

Tabla 4.1: Variables consideradas en el caso de continuidad de las empresas exportadoras de confecciones

Variable (Periodo)	Valores	Nombre en la base	Nombre breve
Año de estudio	2002 – 2005	anho	t
Cadena	Identificador	cadena	
Exportaciones FOB		fob	
Estado de la empresa (t)	Continua = 1. Saliente = 0.	estado	Y_t
Estado de la empresa (t-1)	Continua = 1. Entrante = 0.	estado	Y_0
Factor de continuidad	1, 2, ….., n	factor	X_1
Meses de exportación	1, 2, …, 12	meses	X_2
Variación de Exportación	Aumentó=1. Disminuyó=0.	incrm	X_3
Número de partidas	1, 2, ….. ,n.	part	X_4
Número de mercado	1, 2, ….. ,n.	merc	X_5
Exportación en otros sectores	Si= 1. No = 0.	otros	X_6
Mercados en otros sectores	1, 2, ….. ,n.	otrmer	X_7
Importaciones de confecciones	Si=1. No=0.	mconf	X_8
Importaciones excepto de confecciones	Si=1. No=0.	mnconf	X_9
Variable dummy 1	Si FOB entre US$ 53.7 y 248.2 miles =1. Otros = 0.	N1	X_{11}
Variable dummy 2	Si FOB entre US$ 15.5 y 53.7 miles =1. Otros = 0.	N2	X_{12}
Variable dummy 3	Si FOB entre US$ 5.0 y 15.5 miles =1. Otros = 0.	N3	X_{13}

Se debe recordar que las variables *dummy* referidas al tamaño de la empresa se obtienen del valor FOB[10] de exportación en el periodo *t*.

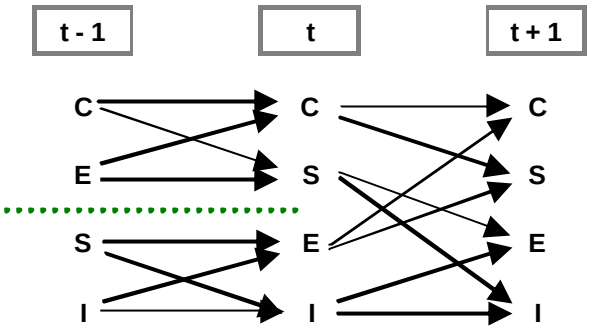
Como se distingue en la tabla 4.1, la mayoría de variables presenta un rezago de un periodo debido a que las variables de estado de las empresas en el periodo *t* dependen de la situación en que se encontraba la empresa un año antes (*t-1)* desde su propia definición de continuidad.

En la figura 4.1 se aprecia dos grupos de datos separados por la línea punteada entre los periodos *t-1* y *t*. Para efectos prácticos se trabajará con el grupo de datos superior por las razones que se expone a continuación a partir de la misma figura.

Una empresa continua (C) en el periodo *t* solo pudo tener dos posibles estados en el periodo previo *t-1*, a saber, entrante (E) o continua, no pudo ser saliente (S) o estar inactiva (I) en ese tiempo. Lo mismo sucede para las empresas salientes en el periodo *t* solo pudieron ser continuas o entrantes en el periodo *t-1*.

[10] Free on board (libre a bordo): Incoterm 2000. Término de comercio internacional que limita las responsabilidades en aspectos documentario, de entrega de carga y de seguros entre el exportador y el importador al momento en que la mercancía cruza la borda del buque en el puerto de origen.

Figura 4.1: Diagrama de posibles cambios de estado de las empresas exportadoras de confecciones en el tiempo (Caso de empresas continuas en periodo *t*)

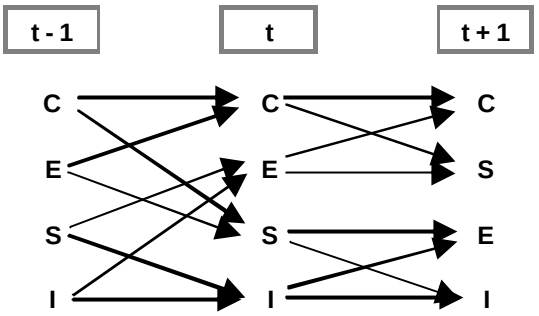


C: Continua E: Entrante S: Saliente I: Inactiva

De otro lado, las empresas entrantes en el periodo *t*, así como las inactivas, no registran movimiento de exportación en *t-1*, razón por la cual un análisis hacia atrás no tendría sentido y conviene más identificar las características un estado hacia delante, es decir, en el periodo *t+1*.

Dado que únicamente las empresas entrantes tendrían información, entonces, este es un caso especial en el que si el resto de empresas con información son continuas (no entrantes) se retornaría a la fase previa vista para *t-1*, al análisis de continuidad explicado en los párrafos anteriores repitiéndose el ciclo.

Figura 4.2: Diagrama de posibles cambios de estado de las empresas exportadoras de confecciones en el tiempo (Caso de empresas Entrantes en periodo *t*)



C: Continua E: Entrante S: Saliente I: Inactiva

Con esto se pretende decir que existen dos problemas que analizar: primero la continuidad en *t*, donde la empresa puede ser continua o entrante en *t-1;* y en segundo lugar, la continuidad en el periodo *t+1*, donde la empresa fue entrante o no entrante (continua) en *t*.

En la figura 4.2 se realiza un reordenamiento de los estados para visualizar mejor el segundo problema. Sin embargo, son uno solo debido a que el estado de la empresa en *(t-1)* del primer problema se convierte en el estado de la empresa en *t* para el segundo caso debiendo deducirse únicamente un modelo.

En los siguiente pasos se procederá a realizar el modelamiento de la continuidad de las empresas exportadoras del sector confecciones en esta actividad.

El tipo de modelo que se construirá en este trabajo será uno de regresión logística autoregresivo el cual se especifica de la siguiente forma:

$$g_t(y_{i(t-1)}, x_{i(t-1)}) = \beta_{0t} + a_{(t-1)}y_{i(t-1)} + \sum_{h=1}^{H}\beta_{ht}x_{hi(t-1)} \qquad (4.1)$$

donde: i=1, 2, …,N; h=1, 2, …, p; y g denota la función *logit*.

Se inicia el proceso con la estimación del estadístico de Wald para cada variable mediante la aplicación del ajuste de modelos unidimensionales como se observa en la tabla 4.2. Los resultados muestran que a excepción de la variable N_2 (exportaciones de la empresa se encuentran entre US$ 15.5 y 53.7 miles) todas las variables pueden ser incluidas en el modelamiento a un nivel de significancia de 0.05.

El siguiente paso es comenzar la construcción del modelo multivariado con la inclusión de todas las variables de la tabla 4.1 (Método *enter*), y se realiza la selección de las que se mantienen a partir de la significancia en la prueba de *Wald.* El procesamiento se realiza con el *software* SPSS versión 11.0.

Tabla 4.2: Estadística de Wald para todas las variables.

Variables	B	Desviación estándar	Wald	g.l	Significancia	Exp(B)	I.C 95.0% para Exp(B)	
							Inferior	Superior
ESTADOT0	1.134	0.129	77.420	1	0.000	3.107	2.414	3.999
FACTORT0	0.407	0.041	99.994	1	0.000	1.502	1.387	1.627
MESEST0	0.382	0.028	190.775	1	0.000	1.465	1.388	1.547
PART0	0.157	0.035	20.202	1	0.000	1.170	1.092	1.252
MERCT0	0.656	0.092	50.708	1	0.000	1.927	1.609	2.309
OTROST0	0.404	0.142	8.061	1	0.005	1.498	1.133	1.981
OTRMERT0	0.119	0.055	4.674	1	0.031	1.127	1.011	1.256
MNCONFT0	0.583	0.155	14.081	1	0.000	1.791	1.321	2.429
N1	0.795	0.167	22.609	1	0.000	2.215	1.596	3.074
N2	-0.022	0.141	0.025	1	0.874	0.978	0.742	1.289
N3	-1.732	0.139	155.855	1	0.000	0.177	0.135	0.232

Variables como el estado de continuidad en el periodo previo, el número de productos, la exportación de otro tipo de bienes, la importación de confecciones y otros productos, entre otras, no son incluidas en el modelo, obteniéndose el siguiente ajuste:

Modelo *logit* 1:

$$g_t(X_{i(t-1)}) = -2.018 + 0.279X_{1(t-1)} + 0.303X_{2(t-1)} + 0.979X_{3(t-1)} + 0.207X_{5(t-1)} - 0.158X_{7(t-1)} - 0.367X_{13(t-1)}$$

El nivel de sensibilidad del modelo alcanza el 88.3% y la especificidad el 51.0% logrando una clasificación efectiva del 77.8%. Sin embargo, el área bajo la curva ROC

se encuentra ligeramente bajo el nivel mínimo aceptable que es 0.7, obteniéndose 0.696. Para este modelo el nivel de la significancia para el ingreso fue 0.25 y el punto de corte c=0.5.

Repitiendo el proceso con el mismo método pero únicamente para las variables seleccionadas en el paso anterior se obtiene el siguiente modelo reajustado:

Modelo *logit* 2:

$$g_t(X_{i(t-1)}) = -1.739 + 0.273X_{1(t-1)} + 0.280X_{2(t-1)} + 1.006X_{3(t-1)} + 0.192X_{5(t-1)} - 0.780X_{7(t-1)} - 0.681X_{13(t-1)}$$

El nivel de sensibilidad es ahora 89.2%, mientras que la especificidad descendió a 47.6% que a su vez hizo disminuir la efectividad de la clasificación a 77.6%. La capacidad discriminatoria del modelo es 68.4%.

Se intenta ahora con el método de inclusión de variables *forward stepwise* luego del cual se aplicará la eliminación *backward* con los mismos puntos de corte y el nivel de significancia para la inclusión, sin embargo, en la eliminación se disminuirá la significancia a 0.10.

La primera variable en ingresar al modelo fue meses de exportación en el año previo (X_2), luego ingresó la tercera variable de diseño para el tamaño de la empresa (X_{13}), la variación en la exportación (X_3), factor de continuidad (X_1) y número de mercados destino de exportación de confecciones (X_5). Todas las variables son del periodo t-1. El modelo obtenido por el método *forward stepwise* es el siguiente:

Modelo *logit* 3:

$$g_t(X_{i(t-1)}) = -1.712 + 0.275X_{1(t-1)} + 0.276X_{2(t-1)} + 1.017X_{3(t-1)} + 0.153X_{5(t-1)} - 0.703X_{13(t-1)}$$

El nivel de sensibilidad alcanzado fue 89.2%, la especificidad 44.8% y la clasificación correcta alcanzó el 76.8%. El área bajo la curva *ROC* fue 67.0%. Si bien es cierto que se logra un incremento de la sensibilidad no sucede lo mismo con los otros indicadores.

A este modelo se aplicará la eliminación *backward* mencionada anteriormente con un nivel de significancia de 0.10. A ese nivel se elimina la variable cantidad de mercados destino (X_5) y no se eliminaría ninguna variable adicional inclusive a un nivel de significancia (a) de 0.01. El modelo obtenido se denomina **Modelo *logit* 4**.

Podría ser posible considerar un nivel de significancia menos exigente de tal forma que la variable X_5 se mantenga en el modelo lo cual ocurre con a=0.13. El **Modelo *logit* 5** resultante es el mismo que el modelo *logit* 3 y se mantienen los mismos indicadores.

El modelamiento se encargó de eliminar aquellos casos en los cuales existía correlaciones medianas entre las variables independientes (0.7<|?|<0.8), es decir, no existe redundancia de información. Las principales correlaciones fueron las observadas entre las variables diseño de las cuales solo se mantuvo en el modelo X_{13}. El otro caso fue la correlación existente entre la exportación de productos de otro sector con el número de mercados en ese rubro, ninguna de las se mantiene en el modelo.

Adicionalmente se realizan todas las posibles combinaciones que nos pueden llevar a reconocer la existencia de interacciones a partir de las cinco variables elegidas. El resultado obtenido es la significancia de las interacciones de las variables X_1*X_5, X_3*X_2, X_3*X_5 y X_5*X_{13}. El modelo resultante es el siguiente:

Modelo *logit* 6:

$$g_t(X_{i(t-1)}) = -0.867 + 0.194X_{2(t-1)} + 0.895X_{3(t-1)} - 1.108X_{13(t-1)} + 0.136X_{1(t-1)}X_{5(t-1)} + 0.171X_{3(t-1)}X_{2(t-1)} - 0.487X_{3(t-1)}X_{5(t-1)} + 0.278X_{5(t-1)}X_{13(t-1)}$$

Los indicadores resultantes del modelo fueron: sensibilidad 88.5%, especificidad 46.8%, clasificación correcta 76.8% y el área bajo la curva *ROC* fue 67.6%.

Se intentará mejorar la capacidad discriminatoria del modelo y de clasificación a través de incrementos al punto de corte *c* para 0.60 (**modelo *logit* 7**), 0.65 (**modelo *logit* 8)** y 0.70 (**modelo *logit* 9**). Conforme se aumenta el punto de corte *c* por encima de 0.5 se mejora notablemente la especificidad pasando de 46.8% a 64.6% (cuando *c*=0.6) hasta alcanzar luego el 74.4% para *c*=0.65, y esta variación repercute notablemente en desmedro de la sensibilidad quien retrocede a niveles de 82.2% y 76.0%, respectivamente.

Cuando se intenta aumentar *c* a 0.7 la especificidad supera a la sensibilidad lo cual no es lo que se requiere debido a que se busca maximizar *c* igualando ambos indicadores, por lo tanto, el valor más adecuado para *c* se encuentra entre 0.65 y 0.70.

Con estas variaciones el nivel de clasificación correcta fue 77.7% (*c*=0.6), 75.5% (*c*=0.65) y 73.3% (*c*=0.7). Asimismo, el área bajo la curva *ROC* que describe la

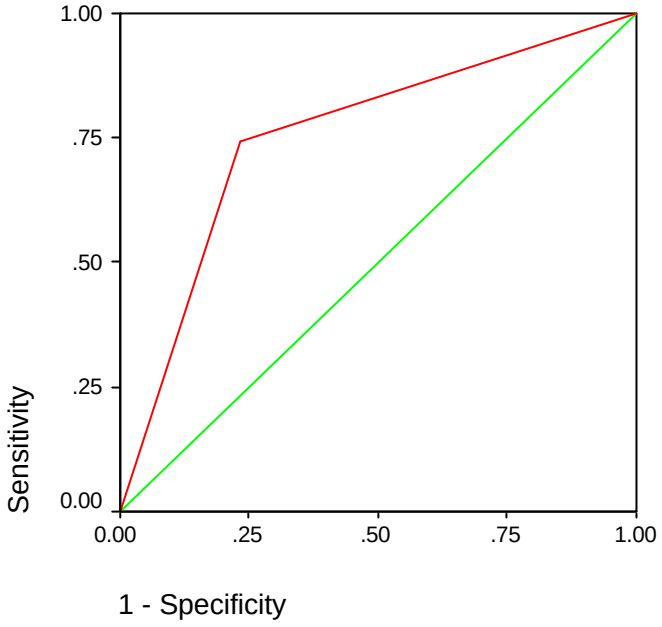
capacidad discriminatoria del modelo alcanza los niveles de 73.4%, 75.2% y 75.6%, respectivamente.

El valor de *c* que maximiza la sensibilidad y especificidad es 0.66 con el cual se alcanza los niveles de 75.8% y 74.9%, respectivamente. La clasificación correcta ascendió a 75.5% y la capacidad discriminatoria fue también 75.5%. El gráfico del mismo se presenta en la figura 4.3, mientras que el **modelo *logit* 10** resultante tiene las mismas variables y parámetros que el obtenido en el **modelo *logit* 6**.

La importancia del procedimiento realizado radica en que esta elección maximiza la clasificación para la sensibilidad y especificidad aumentando la capacidad discriminatoria representada por el área bajo la curva ROC (línea de color rojo), mientras que el área bajo la diagonal (color verde) representa el 50%, es decir, la probabilidad 0.5.

Las estadísticas de todos los modelos se pueden visualizar en el Anexo A seguido del número de modelo respectivo.

Figura 4.3: Curva ROC para Modelo *logit* 10 y punto de corte c=0.66



En el anexo A.10 se puede observar en la nota *c* al pie del cuadro el valor de -2*log likelihood* inicial que es igual a 1518.434. El dar el primer paso, es decir, incluir la primera variable representa para el modelo una variación en la verosimilitud de 312.362, el cual es altamente significativo, y es igual a decir, que la variable incluida en el modelo realiza un aporte importante a la capacidad predictiva que justifica su presencia. En la tabla 4.3 se incluye la información con las variaciones en la verosimilitud paso a paso.

Tabla 4.3: Variaciones en -2log likelihood para el modelo logit 10.

Paso	Variación	Chi-cuadrado	g.l.	Significancia
1	Paso	312.362	1	0.000
	Bloque	312.362	1	0.000
	Modelo	312.362	1	0.000
2	Paso	24.400	1	0.000
	Bloque	336.763	2	0.000
	Modelo	336.763	2	0.000
3	Paso	17.501	1	0.000
	Bloque	354.264	3	0.000
	Modelo	354.264	3	0.000
4	Paso	10.282	1	0.001
	Bloque	364.545	4	0.000
	Modelo	364.545	4	0.000
5	Paso	9.976	1	0.002
	Bloque	374.522	5	0.000
	Modelo	374.522	5	0.000
6[a]	Paso	-0.325	1	0.568
	Bloque	374.196	4	0.000
	Modelo	374.196	4	0.000
7	Paso	3.861	1	0.049
	Bloque	378.058	5	0.000
	Modelo	378.058	5	0.000
8	Paso	5.396	1	0.020
	Bloque	383.454	6	0.000
	Modelo	383.454	6	0.000
9	Paso	2.164	1	0.141
	Bloque	385.618	7	0.000
	Modelo	385.618	7	0.000

a. El valor chi-cuadrado negativo indica que el valor del estadístico ha disminuido respecto al paso previo.

En el primer paso se incluye la variable número de meses al año que la empresa exportó un año antes (X_2); con alta significancia en el paso, el bloque y como modelo. El segundo paso incluye la variable tamaño de la empresa entre US$ 5.0 y US$ 15.5

miles (X_{13}); en el tercero ingresa la interacción entre las variables factor de continuidad e incremento (X_1*X_3); en el cuarto paso, ingresa la interacción entre el factor de continuidad y el número de mercados destino (X_1*X_5), mientras que en el paso cinco, la interacción que ingresa es la existente entre el incremento y el número de meses de exportación (X_3*X_2), en todos los casos con significancia superior al 0.01.

El paso seis presenta un valor negativo de 0.325, que más allá que sea menor a cero no es significativo hasta un nivel de 0.10. Esta falta de significancia nos indica que una variable que fue incluida en el modelo en un paso anterior al interactuar con las demás variables disminuyó la importancia de su aporte y por ese motivo se le pudo retirar. Este factor es la interacción de las variables X_1*X_3 que fue introducido en el paso tres.

Tabla 4.4: Prueba de Hosmer y Lemeshow para el modelo logit 10.

Paso	Chi-cuadrado	g.l.	Significancia
1	4.462	7	0.725
2	3.195	8	0.922
3	4.036	8	0.854
4	3.084	8	0.929
5	12.168	8	0.144
6	3.084	8	0.929
7	3.258	8	0.917
8	2.209	8	0.974
9	2.507	8	0.961

En el paso siete se permite el ingreso de otra interacción, el incremento en la exportación de la empresa con el número de mercados (X_3*X_5), y en el paso ocho se incluye la variable (X_3) incremento en las exportaciones de la empresa, ambos con nivel de significancia de 0.05.

En el último paso, ingresa la interacción entre las variables número de mercados destino con el tamaño de la empresa entre US$ 5.0 y 15.5 miles (X_5*X_{13}) con una significancia de 0.141.

Se debe recordar que el nivel de significancia considerado para permitir el ingreso de una variable al modelo fue 0.25 debido a que se esperaba que la inclusión de un mayor número de variables daría una mejor descripción del fenómeno en estudio debido a la interacción con otras variables, y que el mismo modelo se encargaría de auto-ajustarse con el nivel de significancia para la salida de las variables que fue 0.05, como sucedió en el paso 6.

La prueba de Hosmer y Lemeshow (Tabla 4.4) indica que el modelo presenta problemas de significancia en el paso ocho en el cual ingresó la interacción de variables (X_5*X_{13}) que como se apreció en la Tabla 4.3 presentaba un baja significancia (0.14). En los pasos anteriores no hay significancia al incluir las variables en el

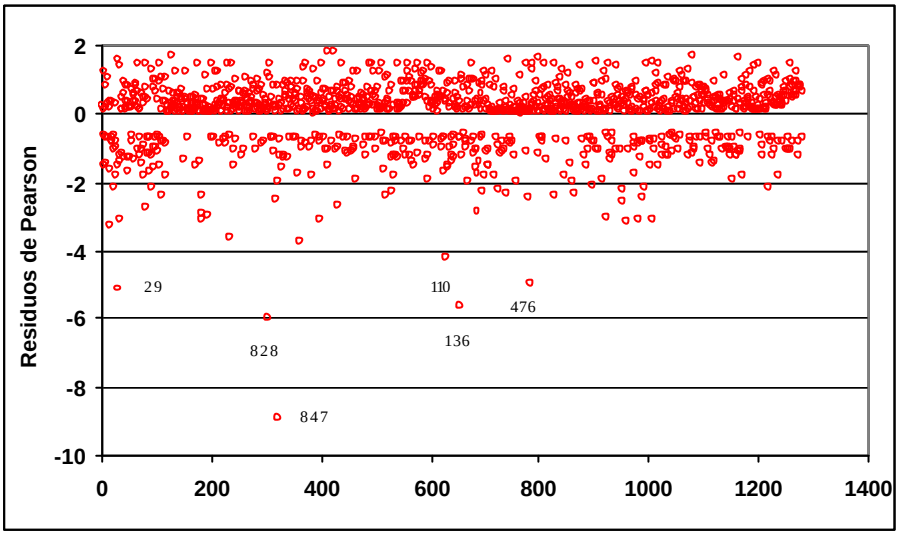
modelo y no se puede rechazar la hipótesis nula a un nivel de significancia de 0.05, es decir, el ajuste es el adecuado si se retira esta última interacción.

En el análisis de residuos se verifica que existen observaciones *outliers* que al ser eliminadas hacen más evidente la no significancia para la presencia de esta interacción en el modelo.

4.2.- Análisis de residuos

El análisis de residuos es una de las partes más importantes en la evaluación de cualquier modelo debido a que nos permite identificar la existencia de información que puede influenciar en el ajuste correcto del modelo.

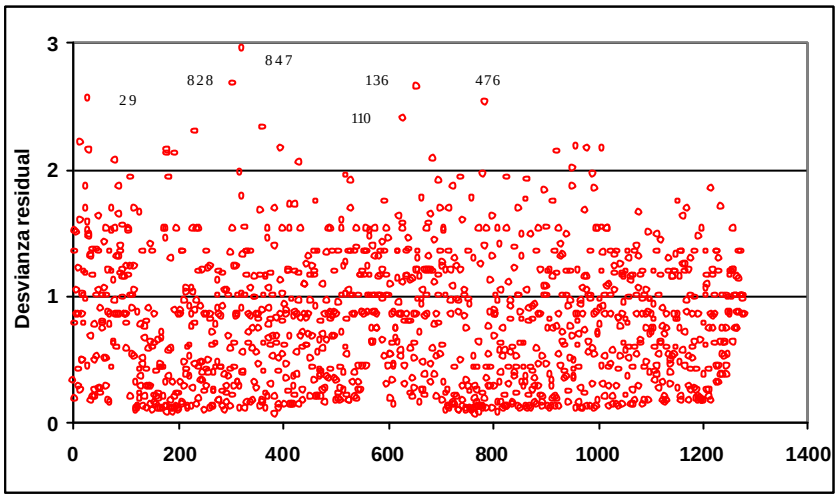
Figura 4.4: Gráfico de Dispersión entre los residuos de Pearson y las observaciones



En el anexo B se presenta la figura B.1 que presenta los residuos ordinarios para las 1280 observaciones que se utilizaron en el modelamiento. No se observa aquí ningún caso atípico al interior de la nube de puntos.

Realizándose un gráfico similar pero esta vez con los residuos de Pearson (Figura 4.4) se puede distinguir una serie de datos alejados del resto en la parte inferior de la nube de puntos los cuales son identificados con los siguientes números: 847, 828, 136, 29, 476 y 110.

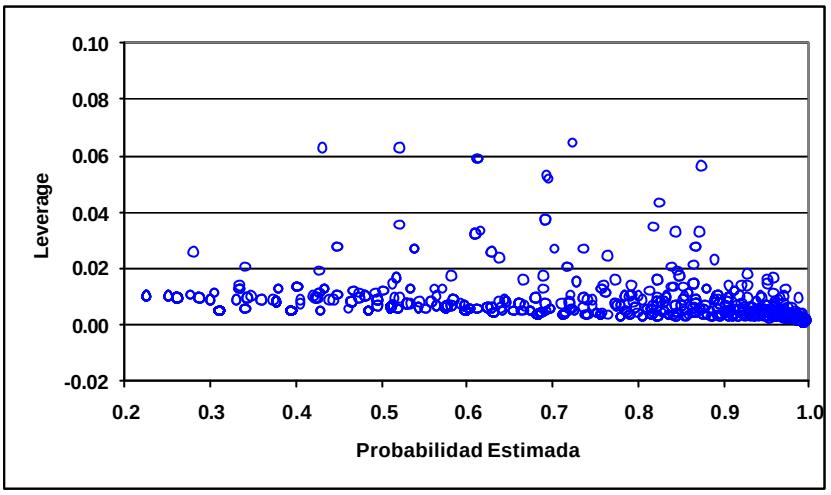
Figura 4.5: Gráfico de Dispersión entre la Desvianza residual y las observaciones



Los casos seleccionados serán comparados contra los obtenidos por otros criterios que se analiza a continuación y también con la identificación de outliers por parte del mismo modelo para aquellos casos que se alejan más de dos desviaciones estándar de la media.

Mediante una rápida revisión a la figura 4.5 se puede identificar una serie de puntos que se encuentra ligeramente alejados del resto, el indicador al que se hace referencia es la desvianza residual e identifica los mismos casos que la figura 4.4.

Figura 4.6: Gráfico de Dispersión entre los Leverage y las Probabilidades estimadas

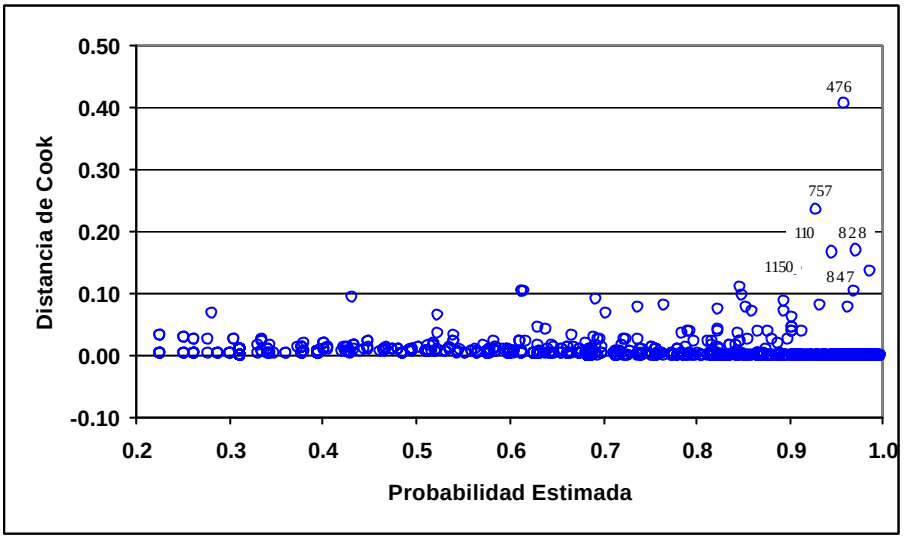


Los pequeños valores obtenidos para los valores leverage (<0.07) no permite concluir nada acerca de la existencia de datos influyentes en el ajuste del modelo según se puede observar en la figura 4.6.

Los puntos ubicados en la parte superior derecha de la figura 4.7 presentan las mayores distancias de Cook para nuestro caso que son las observaciones que presentan mayor influencia en la estimación de los parámetros del modelo en el caso

que sea removida. Los números que identifican a las principales observaciones fueron: 476, 757, 828, 110, 1150 y 847.

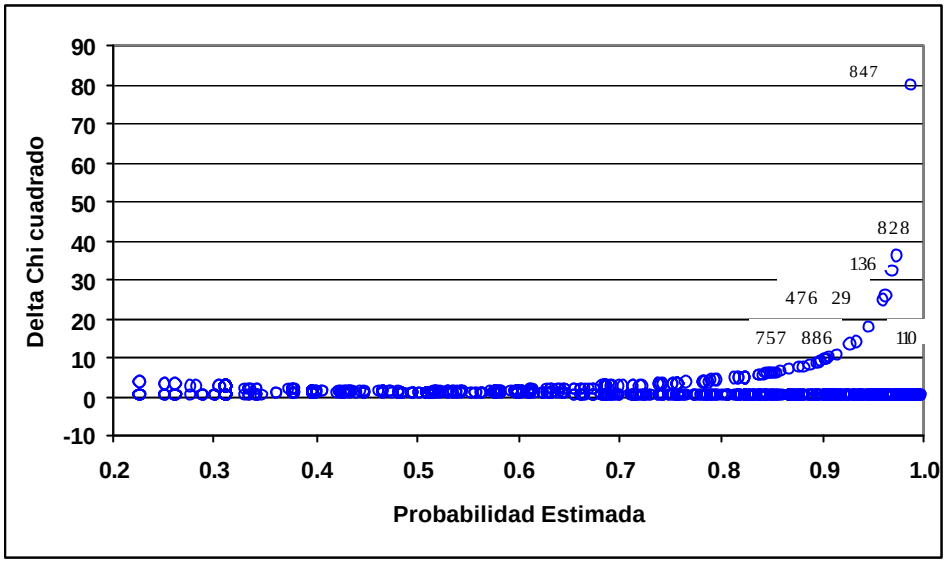
Figura 4.7: Gráfico de Dispersión entre la Distancia de Cook y las Probabilidades estimadas



En la figura 4.8 se observa dos curva que tienen una forma similar a la cuadrática, una primera que se orienta del lado inferior izquierdo del gráfico al lado superior derecho del mismo que representa a las empresas salientes (y_j=0) y una segunda desde el lado izquierdo a una posición más baja en el lado derecho en la que se presenta los casos de las empresas continuas (y_j=1).

Figura 4.8: Gráfico de Dispersión Delta Chi Cuadrado

y las Probabilidades estimadas

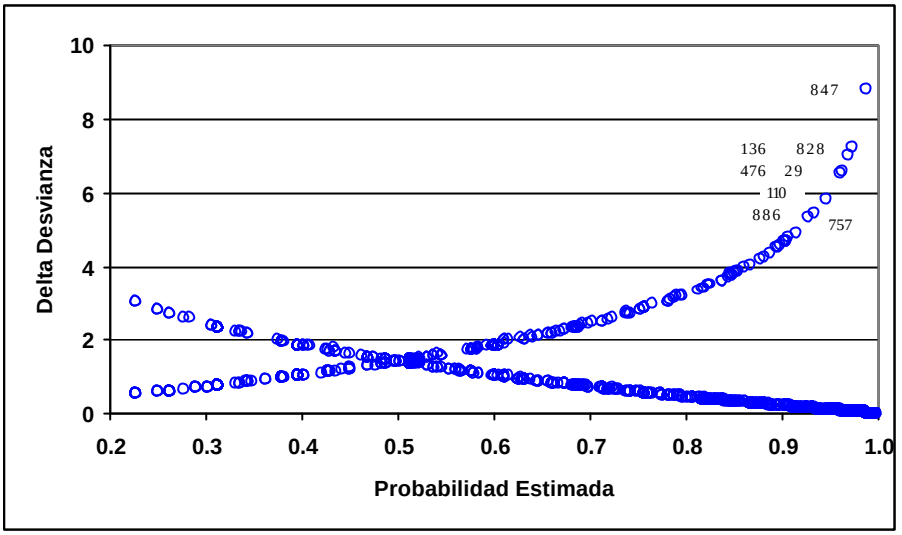


Las observaciones consideradas a ser candidatos por tener un pobre ajuste son las identificadas con los números: 847, 828, 136, 29, 476, 110, 886 y 757. Sin embargo, además de las indicadas se considerará en el análisis todas aquellas observaciones con delta chi-cuadrado mayor a 4.

En la figura 4.9 se observa una situación similar y también se obtendrá una serie de observaciones candidatos a ser analizados debido a su pobre ajuste bajo el mismo criterio mencionado en el caso anterior. El total de casos que cumplen los requisitos en ambas figuras ascienden a 22. Las demás observaciones identificadas fueron: 13, 33, 269, 370, 379, 707, 718, 921, 955, 1112, 1142, 1150, 1170, 1197.

Figura 4.9: Gráfico de Dispersión Delta Desvianza

y las Probabilidades estimadas



Los 22 casos seleccionados en los párrafos anteriores coinciden exactamente con los casos que exceden el límite de dos desviaciones estándar de la media, es decir, son identificados por este método también como *outliers* y candidatos a ser removidos del ajuste del modelo.

Los gráficos de delta chi-cuadrado, delta desvianza y distancia de Cook contra la estadística *leverage* entregan resultados similares y son incluidas en el anexo B.

Tabla 4.5: Estadísticas del modelo logit 13 con las 22 observaciones eliminadas (c=0.66)

Variables	B	Desviación estándar	Wald	g.l	Significancia	Exp(B)	I.C 95.0% para Exp(B)	
							Inferior	Superior
MESEST0	0.213	0.047	20.753	1	0.000	1.237	1.129	1.355
INCRMT0	0.991	0.382	6.717	1	0.010	2.694	1.273	5.700
N3	-0.941	0.332	8.020	1	0.005	0.390	0.203	0.748
FACTORT0*MERCT0	0.219	0.039	31.416	1	0.000	1.244	1.153	1.343
INCRMT0*MESEST0	0.280	0.073	14.634	1	0.000	1.323	1.146	1.528
INCRMT0*MERCT0	-0.547	0.172	10.073	1	0.002	0.579	0.413	0.811
MERCT0*N3	0.173	0.213	0.657	1	0.418	1.189	0.783	1.805
CONSTANTE	-1.302	0.312	17.362	1	0.000	0.272		

El re-procesamiento de la información en el modelo *logit* 13 (tabla 4.5) pretende reproducir el modelamiento obtenido para el modelo *logit* 10 pero sin considerar las 22 observaciones mencionadas en el párrafo anterior (tabla 4.6). La eliminación de estas observaciones produce los siguientes cambios al modelo *logit* 10:

1.- Se observan importantes cambios en la mayoría de los estimados de los parámetros de las variables lo cual es una muestra de la influencia que ejercían las observaciones excluidas.

2.- En términos cuantitativos el modelo *logit* 13 presentó una sensibilidad de 76.5% y la especificidad ascendió a 78.3% con un punto de corte c=0.66.

3.- El nivel de clasificación correcta fue 77.0% y la capacidad discriminatoria ascendió a 77.1%, con un límite inferior de 74.4% y superior de 80.4%.

Asimismo, el retiro de estas observaciones produce perdida de significancia para la interacción X_5*X_{13} la cual es retirada del modelo generándose el modelo *logit* 14 final que es el siguiente:

Modelo *logit* final.

$$g_t(X_{i(t-1)}) = -1.328 + 0.213X_{2(t-1)} + 0.918X_{3(t-1)} - 0.707X_{13(t-1)} + 0.224X_{1(t-1)}X_{5(t-1)} + 0.279X_{3(t-1)}X_{2(t-1)} - 0.484X_{3(t-1)}X_{5(t-1)}$$

La variación de parámetros, con respecto al modelo *logit* 13, es importante únicamente en la variable X_{13}. El nivel de sensibilidad alcanzó el 76.3% y la especificidad fue 78.9% con una clasificación correcta del 77.0%. La capacidad discriminatoria mejora ligeramente alcanzando el 77.6%.

Tabla 4.6: Observaciones outliers eliminadas de los datos en el procesamiento del modelo *logit* 13

Caso	Delta beta	Delta X2	Delta D	Prob. Estim	Y	Leverage
13	.039	10.743	4.938	0.915	1	0.004
29	.077	25.921	6.600	0.963	1	0.003
33	.039	9.380	4.692	0.903	1	0.004
110	.167	17.417	5.864	0.945	1	0.009
136	.104	32.123	7.017	0.970	1	0.003
269	.025	7.455	4.278	0.881	1	0.003
370	.089	8.541	4.540	0.894	1	0.010
379	.021	7.898	4.378	0.887	1	0.003
476	.406	24.504	6.553	0.960	1	0.016
707	.039	9.380	4.692	0.903	1	0.004
718	.072	8.647	4.556	0.896	1	0.008
757	.234	13.174	5.364	0.928	1	0.017
828	.170	35.984	7.246	0.973	1	0.005
847	.137	79.730	8.794	0.988	1	0.002
886	.080	13.980	5.434	0.933	1	0.006
921	.047	9.475	4.712	0.904	1	0.005
955	.039	7.216	4.225	0.878	1	0.005
1112	.027	8.923	4.598	0.899	1	0.003
1142	.038	6.584	4.066	0.867	1	0.006
1150	.139	9.896	4.818	0.907	1	0.014
1170	.061	9.472	4.716	0.904	1	0.006
1197	.047	9.475	4.712	0.904	1	0.005

4.3.- Validación externa.

Debido a que se desea aplicar medidas de pronóstico para el 2006 a partir de los datos obtenidos en el 2005 se debe proceder a revisar el ajuste obtenido por validación externa para los 776 casos cuando *(t-1)* es igual a 2004. Debe recordarse que para el ajuste del modelo *logit* 14 se utilizó los datos del 2002 y 2003 para *(t-1).*

En la tabla 4.7 se observa los estimados y observados para *(t-1)* igual a 2004 (t=2005), aquí se aprecia que la especificidad es 68.3%, la sensibilidad es 78.0% y una clasificación correcta que alcanza el 75.5% de los casos.

Tabla 4.7: Tabla de Clasificación para validación externa.

(*t-1*=2004)

Observado	Predicho			
	ESTADOT			Porcentaje Correcto
ESTADOT	Saliente	Continua	Total	
Saliente	136	63	199	68.3
Continua	127	450	577	78.0
Total	**263**	**513**	**776**	**75.5**

La estadística de Osius y Rojek es 1.20 que equivale a un nivel de probabilidad de la distribución normal estándar de 0.89 (prueba de dos colas) con lo cual no se puede

rechazar la hipótesis nula con un nivel de significancia de 0.05, es decir, el modelo es el correcto y ajusta adecuadamente los datos al no encontrase evidencia de diferencias entre la distribución de los datos observados y los estimados. Los valores previos para el cálculo del estadístico fueron: $X^2 = 1304.5$, $s_v = 439.3$ y $J_v = n_v = 776$.

CAPÍTULO V

ANÁLISIS DE RESULTADOS DE LA REGRESIÓN LOGÍSTICA MULTIVARIADA.

En el siguiente capítulo se presenta los resultados obtenidos aplicando la metodología desarrollada en el capítulo II y se realiza además el análisis correspondiente para los pronósticos del modelo obtenido para el año 2006.

5.1- Interpretación de *odds ratios*.

En esta sección se realizará el análisis de los estimados de los parámetros del modelo obtenido procurando tener especial cuidado cuando se trate a las estimaciones de los parámetros de las interacciones debido a que tienen un tratamiento de análisis diferente a la interpretación de las variables independientes.

El *odds ratio* para la variable dummy 3 (X_{13}) es 0.493, es decir, las empresas que registraron exportaciones de confecciones entre US$ 5 y US$ 15.5 miles en el año previo tienen menor posibilidad de ser continuas que las exportadoras de ventas superiores.

Las empresas de confecciones que incrementaron sus exportaciones en el año anterior tienen 2.504 más posibilidades de continuar en la actividad exportadora que una empresa que no incrementó sus ventas en el mismo periodo.

Considerando linealidad en la *logit*, una empresa de confecciones que exporta un mes más que otra tiene 1.237 más posibilidad de ser continua, si la diferencia es de dos meses la posibilidad se incrementa a 1.531 y para tres meses de diferencia esta ventaja se amplía a 1.895. Asimismo, diferencias de 4, 5 ó 6 meses hacen que la empresa exportadora de confecciones tenga, respectivamente, 2.344, 2.901 y 3.589 más oportunidad de ser continuas.

El aumento en la diferencia de meses incrementa la posibilidad de una empresa respecto a otra en continuar en la actividad exportadora y ésta se calcula de la siguiente manera: exp [diferencia de meses*0.213)]. Una empresa que se diferencia de

otra en once meses de exportaciones tiene 10.412 más oportunidad de continuar exportando en el siguiente periodo.

Una empresa que incrementó sus exportaciones el año previo y exportó un mes del año tiene 3.3 veces más oportunidades de ser continua que una que no incrementó sus envíos de confecciones. Si el número de meses de exportación asciende a 12 la posibilidad se incrementa a cerca de 71 veces más si la empresa incrementó sus exportaciones contra otra que no lo hizo. (Para otras variaciones ver anexo A.16).

De otro lado, si el número de mercados es mayor o igual a dos, entonces, la empresa que incrementó sus ventas el año previo tendrá menos posibilidades de continuar en las exportaciones que otra empresa que no aumentó en el mismo periodo. La única excepción es cuando las exportaciones se dirigen a un solo mercado. Cuando el número de mercados es uno la mayor proporción de empresas continuas se registran en aquellas que incrementaron sus ventas al exterior, lo mismo sucede cuando son dos mercados destino pero la tendencia cambia para tres o más mercados que es lo que explica el indicador.

El incremento del factor de continuidad en una unidad aumenta su posibilidad de ser continua en 25% si se dirige a un mercado, 57% si es a dos mercados, hasta llegar

a 283% cuando se dirige a seis destinos. Mayores incrementos en el factor de continuidad aumenta exponencialmente la posibilidad de continuar exportando.

5.2.- Análisis de pronósticos de continuidad para el año 2006.

La aplicación del modelo final de regresión logística a la información para el 2005 *(t-1)* proporciona pronósticos para el 2006 *(t)* logrando discriminar entre las empresas que mantendrán su continuidad un año más y aquellas que dejarán de realizar esta actividad.

El 70.9% de las empresas que exportaron en el 2005 continuarán haciéndolo en el siguiente año con las siguientes características:

- Las empresas que tienen mayor cantidad de mercados destino aumentan su probabilidad de continuar exportando. El 62.5% de las empresas que tienen un solo mercado de destino son continuas en el periodo siguiente, el 80.8% que registra dos mercados destino es también continua, mientras que las empresas con tres o más mercados en el 90.0% de los casos mantienen la continuidad.

- El 65.7% de las empresas continuas son aquellas que exportan por encima de US$ 15.5 miles.

- Las empresas tienen mayor posibilidad de ser pronosticadas como continuas si incrementaron su valor de exportación respecto al año previo (79.9% de los casos).

- Exportar en el año hasta en dos meses distintos es un signo de tener alta probabilidad de dejar la actividad exportadora al año siguiente (93.8% de los casos). Si una empresa exporta en 4 meses o más distintos en un año es altamente probable que continue haciéndolo el siguiente año (94.3% de los casos).

- El 54.0% de las empresas entrantes en el 2005 (factor de continuidad=1) continuarán exportando en el 2006, el 70.7% de los que exportaron en los últimos dos años (2004 y 2005) lo harán nuevamente, mientras que las empresas que tienen tres años en la actividad serán continuas en el 79.2% de los casos. Asimismo, el porcentaje de empresas continuas supera el 82.0% cuando el factor de continuidad es por lo menos de cuatro años.

La información consignada en los párrafos superiores puede ser encontrada en el anexo A.15.

CONCLUSIONES

Las principales conclusiones que se obtienen del presente trabajo fueron:

- ✓ Los datos que los exportadores de confecciones del Perú registran en la declaración única de Aduanas (DUA) proporcionan información suficiente para discriminar a las empresas en dos categorías mutuamente excluyentes según su continuidad en la actividad exportadora (continua o saliente) bajo el ajuste de un modelo *logit* multivariado con una capacidad discriminatoria aceptable (77.6%) y un porcentaje de clasificación correcta de 77.0%.

- ✓ De los doce factores analizados se pueden señalar que los más relevantes para el modelo *logit* multivariado de continuidad de las empresas del sector confecciones son los siguientes: El incremento de las exportaciones respecto al año previo, el nivel de exportaciones (tamaño de la empresa en ventas internacionales), el número de mercados destino, el número de meses en el

año que realiza embarques y la cantidad de años en la actividad exportadora (factor de continuidad).

- ✓ El incremento de las exportaciones como factor juega un rol importante debido a su presencia como covariable individual y mediante interacciones con las variables número de meses en el año que la empresa realiza embarques y el número de mercados destino.

- ✓ La variable número de mercados destino también es de gran importancia para el modelo debido a la significancia de la interacción con las variables factor de continuidad e incremento de exportaciones, aunque no ingresa al modelo como covariable individual.

- ✓ Los estadísticos de distancia de Cook, delta chi-cuadrado y delta desvianza presenta excelentes aproximaciones para la detección de valores *outliers* o puntos extremos que no fueron posible identificar mediante los *leverage* que entregaron valores muy pequeños (menores a 0.1).

- ✓ La capacidad para pronosticar valores futuros de continuidad de las empresas exportadoras del sector confecciones es plenamente comprobada

mediante la prueba de Osius y Rojek para validación del modelo *logit* de pronóstico con un alto nivel de significancia.

- ✓ La continuidad de las empresas entrantes (no entrantes = continuas), con visión de pronóstico, tienen la misma problemática de desarrollo cuando se analiza lo mismo para las empresas continuas (no continuas = entrantes) debido a que la entrada al modelo, en el ajuste, validación o pronóstico, pueden tener únicamente empresas entrantes y continuas y el resultado es continuas o salientes.

- ✓ La mayor parte de las empresas entrantes en el 2005 (54.0%) continuarán en la actividad exportadora en el 2006 (un año después) en lugar de abandonarla y la probabilidad de mantenerse en la actividad aumenta a medida que se incrementa el factor de continuidad (número de años exportando).

- ✓ La variable ESTADOT0, estado de continuidad de la empresa en un año previo, no se incluye en el modelo debido a la falta de significancia, lo cual se puede deber a la redundancia de información debido a la presencia del

factor de continuidad, que no solo cuenta la exportación un periodo atrás sino hasta seis años.

RECOMENDACIONES

Como consecuencia de los resultados obtenidos en esta investigación se plantea las siguientes recomendaciones:

- ✓ Ampliar la investigación mediante una encuesta que permita la inclusión de otras variables relevantes no disponibles en la DUA con la finalidad de mejorar la capacidad discriminatoria del modelo como por ejemplo: adquisición de nueva maquinaria, ampliación de personal, participación en alguna feria o misión comercial, existencia de un área de comercio exterior o inteligencia de negocio, entre otros.

- ✓ Incentivar la práctica y adecuación de los resultados de este sector empresarial a otros de la realidad exportadora nacional no tradicional.

- ✓ Sugerir la preparación de acciones de contingencia para evitar anticipadamente la posibilidad que una empresa exportadora pueda dejar esta actividad debido a los factores reconocidos en la investigación o en la ampliación del mismo.

BIBLIOGRAFÍA

Acuña, E. (2006). Análisis de regresión. Universidad de Puerto Rico.

Bendel, R. Afifi, A. (1977). "Comparison of stopping rules in forward stepwise regression". Journal of the American Statistical Association, **72**, 46-53.

Caballer, A. (2001). "La actitud e intención de la donación de órganos en la población Española: Análisis mediante regresión logística multinivel". Tesis doctoral. Universitat Jaume.

Catena, A. Ramos, M., Trujillo, H. (2003). Análisis multivariado, un manual para investigadores. Biblioteca Nueva.

Christensen, R. (1997). Log-linear models and logistic regression. Springer.

Cook, R. (1977). "Detection of influential observations in linear regression". Technometrics, **19**, 15-18.

Cook, R. (1979). "Influential observations in linear regression". Journal of the American Statistical Association, **74**, 169-174.

Corasma, V. (2002). "Factores que se asocian con el bajo peso del recién nacido". Tesis Monográfica. Universidad Nacional Mayor de San Marcos.

Dobson, A. (1983). An introduction to statistical modelling. Chapman and Hall.

Faraggi, D. Reiser, B. (2002). "Estimation of the área under the ROC curve". Statistics in medicine, **21**, 3093-3106.

Flores, L. (2002). "Análisis estadístico de los factores de riesgo que influyen en la enfermedad angina de pecho". Tesis Monográfica. Universidad Nacional Mayor de San Marcos.

Gujarati, D. (2003). Econometría. Mc Graw Hill.

Hogg, R. and Craig, A. (1995). Introduction to mathematical statistical. Prentice Hall.

Hosmer, D. Jovanovic, B. Lemeshow, S. (1989). "Best subsets logistic regression". Biometrics, **45**, 1265-1270.

Hosmer, D., Lemeshow, S. (2000). Applied logistic regression. Second edition. John Wiley and Son, INC.

Jaccard, J. (2001). "Interaction Effects in Logistic Regression". Serie Quantitative Applications in the social Sciences. Sage University Paper.

Kleinbaum, D. Klein, M. (2002). Logistic regression. Second edition. Springer.

Komarek, P. (2004). "Logistic regression for data mining and high dimensional classification". Tesis Doctoral. Carnegie Mellon University.

Lawless, J. Singhal, K. (1978). "Efficient screening of non-normal regression models". Biometrics, **34**, 318-327.

Mallows, C. (1973). "Some comments on Cp". Technometrics, **15**, 661-676.

Mariaca, R. (2002). "Predicción de problemas de crisis y continuidad en empresas bancarias".

Mickey, J. Greenland, S. (1989). "A study of the impact of confounder-selection criteria on effect estimation". American Journal of Epidemiology, **129**, 125-137.

Osius, G. Rojek, D. (1992). "Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom". Journal of the American Statistical Association, **87**, 1145-1152.

Peña, D. (2002). Análisis de datos multivariantes. Mc Graw Hill.

Pontual, E. (2005). "Distribução e dinâmica do tamanho de empresas industriais".

Pregibon, D. (1981). "Logistic regression diagnostics". Annal of Statistics, **9**, 705-724.

Royston, P. Altman, D. (1994). "Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling". Applied Statistics, **43**, 429-467.

Salcedo, C. (2002). "Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación". Informe profesional. Universidad Nacional Mayor de San Marcos.

Vinterbo, S. (1999) “Predictive model in medicine: some methods for construction and adaptation”.

Visauta, B. (1998). Análisis Estadístico con SPSS para Windows. Mc Graw Hill.

Zhang, Y. (2004). Notas de clase STA6938 - Modelos de Regresión Logística. Universidad Central de Florida.

ANEXOS

Anexo A.

Estadísticas para los modelos *logit*.

A.1.- Estadísticas para el modelo *logit* 1 con todas las variables

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	375.112	14	.000
	Block	375.112	14	.000
	Model	375.112	14	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1143.322	.254	.366

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	183	176	51.0
		Continua	108	812	88.3
	Overall Percentage				77.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ESTADOT0	-.082	.247	.110	1	.740	.921
	FACTORT0	.279	.073	14.849	1	.000	1.322
	MESEST0	.303	.039	61.415	1	.000	1.354
	PART0	-.017	.041	.166	1	.684	.983
	MERCT0	.207	.113	3.350	1	.067	1.229
	OTROST0	.215	.268	.644	1	.422	1.240
	OTRMERT0	-.158	.104	2.289	1	.130	.854
	MCONF0	-.021	.429	.002	1	.961	.979
	MNCONFT0	.226	.201	1.266	1	.261	1.254
	LIMAT0	-.070	.279	.064	1	.800	.932
	INCRMT0	.979	.226	18.855	1	.000	2.662
	N1	.277	.305	.826	1	.364	1.319
	N2	.349	.313	1.246	1	.264	1.417
	N3	-.367	.319	1.329	1	.249	.692
	Constant	-2.018	.530	14.481	1	.000	.133

a. Variable(s) entered on step 1: ESTADOT0, FACTORT0, MESEST0, PART0, MERCT0, OTROST0, OTRMERT0, MCONF0, MNCONFT0, LIMAT0, INCRMT0, N1, N2, N3.

A.2.- Estadísticas para el modelo logit 2 con todas las variables menos las eliminadas en el modelo logit 1 (a>0.25).

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	371.610	6	.000
	Block	371.610	6	.000
	Model	371.610	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1146.824	.252	.363

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	171	188	47.6
		Continua	99	821	89.2
	Overall Percentage				77.6

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	FACTORT0	.273	.061	19.968	1	.000	1.314
	MESEST0	.280	.032	76.061	1	.000	1.324
	MERCT0	.192	.107	3.230	1	.072	1.212
	OTRMERT0	-.078	.067	1.333	1	.248	.925
	INCRMT0	1.006	.208	23.468	1	.000	2.735
	N3	-.681	.160	18.024	1	.000	.506
	Constant	-1.739	.306	32.343	1	.000	.176

a. Variable(s) entered on step 1: FACTORT0, MESEST0, MERCT0, OTRMERT0, INCRMT0, N

A.3.- Estadísticas para el modelo *logit* 3 (a = 0.25)

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients					
			Constant	MESEST0	N3	INCRMT0	FACTORT0	MERCT0
Step 1	1	1264.492	-.325	.211				
	2	1211.502	-.639	.324				
	3	1206.163	-.770	.374				
	4	1206.071	-.789	.382				
Step 2	1	1236.041	.135	.166	-.796			
	2	1186.539	-.153	.275	-.787			
	3	1181.748	-.285	.322	-.772			
	4	1181.671	-.303	.329	-.771			
Step 3	1	1231.561	-.035	.168	-.799	.227		
	2	1179.886	-.433	.279	-.790	.366		
	3	1174.625	-.613	.330	-.775	.419		
	4	1174.532	-.638	.338	-.773	.426		
Step 4	1	1219.358	-.358	.142	-.756	.415	.112	
	2	1158.413	-1.103	.236	-.718	.779	.216	
	3	1150.701	-1.497	.282	-.693	.977	.274	
	4	1150.502	-1.569	.290	-.690	1.014	.285	
	5	1150.502	-1.571	.291	-.690	1.015	.285	
Step 5	1	1216.105	-.473	.136	-.759	.414	.102	.112
	2	1155.792	-1.249	.224	-.730	.779	.204	.154
	3	1148.328	-1.641	.268	-.706	.979	.264	.155
	4	1148.139	-1.710	.276	-.703	1.016	.275	.153
	5	1148.139	-1.712	.276	-.703	1.017	.275	.153

a. Method: Forward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1518.434

d. Estimation terminated at iteration number 4 because log-likelihood decreased by less than .010 percent.

e. Estimation terminated at iteration number 5 because log-likelihood decreased by less than .010 percent.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	7.139	1	.008
	Block	343.902	3	.000
	Model	343.902	3	.000
Step 4	Step	24.030	1	.000
	Block	367.932	4	.000
	Model	367.932	4	.000
Step 5	Step	2.363	1	.124
	Block	370.295	5	.000
	Model	370.295	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1174.532	.236	.339
4	1150.502	.250	.360
5	1148.139	.251	.362

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.462	7	.725
2	3.195	8	.922
3	6.307	8	.613
4	6.083	8	.638
5	6.253	8	.619

Classification Table[a]

Observed			Predicted		
			ESTADOT		Percentage Correct
			Saliente	Continua	
Step 1	ESTADOT	Saliente	198	161	55.2
		Continua	158	762	82.8
	Overall Percentage				75.1
Step 2	ESTADOT	Saliente	154	205	42.9
		Continua	100	820	89.1
	Overall Percentage				76.2
Step 3	ESTADOT	Saliente	152	207	42.3
		Continua	100	820	89.1
	Overall Percentage				76.0
Step 4	ESTADOT	Saliente	163	196	45.4
		Continua	105	815	88.6
	Overall Percentage				76.5
Step 5	ESTADOT	Saliente	161	198	44.8
		Continua	99	821	89.2
	Overall Percentage				76.8

a. The cut value is .500

Variables in the Equation[f]

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465	1.388	1.547
	Constant	-.789	.121	42.451	1	.000	.454		
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390	1.313	1.471
	N3	-.771	.156	24.355	1	.000	.463	.341	.628
	Constant	-.303	.156	3.766	1	.052	.739		
Step 3[c]	MESEST0	.338	.030	129.694	1	.000	1.402	1.323	1.486
	INCRMT0	.426	.159	7.149	1	.008	1.531	1.120	2.093
	N3	-.773	.157	24.283	1	.000	.462	.339	.628
	Constant	-.638	.202	10.032	1	.002	.528		
Step 4[d]	FACTORT0	.285	.061	22.121	1	.000	1.330	1.181	1.498
	MESEST0	.291	.031	87.890	1	.000	1.337	1.259	1.421
	INCRMT0	1.015	.207	24.075	1	.000	2.759	1.840	4.139
	N3	-.690	.159	18.895	1	.000	.502	.368	.685
	Constant	-1.571	.288	29.684	1	.000	.208		
Step 5[e]	FACTORT0	.275	.061	20.288	1	.000	1.317	1.168	1.484
	MESEST0	.276	.032	74.933	1	.000	1.318	1.238	1.404
	MERCT0	.153	.101	2.291	1	.130	1.165	.956	1.419
	INCRMT0	1.017	.207	24.098	1	.000	2.765	1.842	4.151
	N3	-.703	.159	19.439	1	.000	.495	.362	.677
	Constant	-1.712	.304	31.699	1	.000	.180		

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: INCRMT0.

d. Variable(s) entered on step 4: FACTORT0.

e. Variable(s) entered on step 5: MERCT0.

f. Stepwise procedure stopped because removing the least significant variable result in a previously fitted mode

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	MESEST0	-759.217	312.362	1	.000
Step 2	MESEST0	-678.565	175.460	1	.000
	N3	-603.036	24.400	1	.000
Step 3	MESEST0	-677.416	180.299	1	.000
	INCRMT0	-590.836	7.139	1	.008
	N3	-599.439	24.346	1	.000
Step 4	FACTORT0	-587.266	24.030	1	.000
	MESEST0	-630.253	110.004	1	.000
	INCRMT0	-587.867	25.231	1	.000
	N3	-584.696	18.890	1	.000
Step 5	FACTORT0	-585.012	21.886	1	.000
	MESEST0	-620.791	93.442	1	.000
	MERCT0	-575.251	2.363	1	.124
	INCRMT0	-586.693	25.246	1	.000
	N3	-583.798	19.457	1	.000

A.4.- Estadísticas para el modelo *logit* 4 (a = 0.10)

Block 1: Method = Backward Stepwise (Likelihood Ratio)

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients					
			Constant	FACTORT0	MESEST0	MERCT0	INCRMT0	N3
Step 1	1	1216.105	-.473	.102	.136	.112	.414	-.759
	2	1155.792	-1.249	.204	.224	.154	.779	-.730
	3	1148.328	-1.641	.264	.268	.155	.979	-.706
	4	1148.139	-1.710	.275	.276	.153	1.016	-.703
	5	1148.139	-1.712	.275	.276	.153	1.017	-.703
Step 2	1	1219.358	-.358	.112	.142		.415	-.756
	2	1158.413	-1.103	.216	.236		.779	-.718
	3	1150.701	-1.497	.274	.282		.977	-.693
	4	1150.502	-1.569	.285	.290		1.014	-.690
	5	1150.502	-1.571	.285	.291		1.015	-.690

a. Method: Backward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1518.434

d. Estimation terminated at iteration number 5 because log-likelihood decreased by less than .010 percent.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	370.295	5	.000
	Block	370.295	5	.000
	Model	370.295	5	.000
Step 2[a]	Step	-2.363	1	.124
	Block	367.932	4	.000
	Model	367.932	4	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1148.139	.251	.362
2	1150.502	.250	.360

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.253	8	.619
2	4.089	7	.769

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	161	198	44.8
		Continua	99	821	89.2
	Overall Percentage				76.8
Step 2	ESTADOT	Saliente	163	196	45.4
		Continua	105	815	88.6
	Overall Percentage				76.5

a. The cut value is .500

Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	FACTORT0	.275	.061	20.288	1	.000	1.317	1.168	1.484
	MESEST0	.276	.032	74.933	1	.000	1.318	1.238	1.404
	MERCT0	.153	.101	2.291	1	.130	1.165	.956	1.419
	INCRMT0	1.017	.207	24.098	1	.000	2.765	1.842	4.151
	N3	-.703	.159	19.439	1	.000	.495	.362	.677
	Constant	-1.712	.304	31.699	1	.000	.180		
Step 2[a]	FACTORT0	.285	.061	22.121	1	.000	1.330	1.181	1.498
	MESEST0	.291	.031	87.890	1	.000	1.337	1.259	1.421
	INCRMT0	1.015	.207	24.075	1	.000	2.759	1.840	4.139
	N3	-.690	.159	18.895	1	.000	.502	.368	.685
	Constant	-1.571	.288	29.684	1	.000	.208		

a. Variable(s) entered on step 1: FACTORT0, MESEST0, MERCT0, INCRMT0, N3.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	FACTORT0	-585.012	21.886	1	.000
	MESEST0	-620.791	93.442	1	.000
	MERCT0	-575.251	2.363	1	.124
	INCRMT0	-586.693	25.246	1	.000
	N3	-583.798	19.457	1	.000
Step 2	FACTORT0	-587.266	24.030	1	.000
	MESEST0	-630.253	110.004	1	.000
	INCRMT0	-587.867	25.231	1	.000
	N3	-584.696	18.890	1	.000

A.5.- Estadísticas para el modelo logit 5 (a = 0.13)

Block 1: Method = Backward Stepwise (Likelihood Ratio)

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients					
			Constant	FACTORT0	MESEST0	MERCT0	INCRMT0	N3
Step 1	1	1216.105	-.473	.102	.136	.112	.414	-.759
	2	1155.792	-1.249	.204	.224	.154	.779	-.730
	3	1148.328	-1.641	.264	.268	.155	.979	-.706
	4	1148.139	-1.710	.275	.276	.153	1.016	-.703
	5	1148.139	-1.712	.275	.276	.153	1.017	-.703

a. Method: Backward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1518.434

d. Estimation terminated at iteration number 5 because log-likelihood decreased by less than .010 percent.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	370.295	5	.000
	Block	370.295	5	.000
	Model	370.295	5	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1148.139	.251	.362

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.253	8	.619

Contingency Table for Hosmer and Lemeshow Test

		ESTADOT = Saliente		ESTADOT = Continua		
		Observed	Expected	Observed	Expected	Total
Step 1	1	91	91.208	41	40.792	132
	2	70	74.213	58	53.787	128
	3	71	65.919	71	76.081	142
	4	44	46.543	84	81.457	128
	5	39	34.375	89	93.625	128
	6	18	21.799	110	106.201	128
	7	16	12.838	111	114.162	127
	8	8	7.085	120	120.915	128
	9	1	3.671	127	124.329	128
	10	1	1.349	109	108.651	110

Classification Table[a]

			Predicted		
			ESTADOT		Percentage
	Observed		Saliente	Continua	Correct
Step 1	ESTADOT	Saliente	161	198	44.8
		Continua	99	821	89.2
	Overall Percentage				76.8

a. The cut value is .500

Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	FACTORT0	.275	.061	20.288	1	.000	1.317	1.168	1.484
	MESEST0	.276	.032	74.933	1	.000	1.318	1.238	1.404
	MERCT0	.153	.101	2.291	1	.130	1.165	.956	1.419
	INCRMT0	1.017	.207	24.098	1	.000	2.765	1.842	4.151
	N3	-.703	.159	19.439	1	.000	.495	.362	.677
	Constant	-1.712	.304	31.699	1	.000	.180		

a. Variable(s) entered on step 1: FACTORT0, MESEST0, MERCT0, INCRMT0, N3.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	FACTORT0	-585.012	21.886	1	.000
	MESEST0	-620.791	93.442	1	.000
	MERCT0	-575.251	2.363	1	.124
	INCRMT0	-586.693	25.246	1	.000
	N3	-583.798	19.457	1	.000

Correlation Matrix

		Constant	ESTADOT0	FACTORT0	MESEST0	MERCT0	OTRMERCT0	INCRMT0	N1	N2	N3	MCONF0	MNCONFT0	OTROST0	PARTO
Step	Constant	1.000	-0.146	-0.245	-0.295	-0.088	0.068	-0.492	-0.470	-0.603	-0.617	0.033	-0.102	-0.051	-0.197
1	ESTADOT0	-0.146	1.000	-0.532	-0.150	-0.021	0.068	0.379	-0.040	-0.042	0.002	0.005	-0.040	-0.051	-0.010
	FACTORT0	-0.532	1.000	-0.080	-0.081	0.008	0.282	0.019	0.004	0.023	-0.004	0.005	-0.003	0.018	
	MESEST0	-0.150	-0.080	1.000	-0.231	-0.012	-0.069	0.321	0.483	0.531	0.021	0.025	-0.094	-0.197	
	MERCT0	-0.021	-0.081	-0.231	1.000	-0.406	-0.013	-0.124	-0.125	-0.126	-0.033	0.036	0.244	-0.061	
	OTRMERCT0	0.068	0.008	-0.012	-0.406	1.000	0.049	-0.009	-0.017	-0.043	0.051	0.120	-0.758	0.006	
	INCRMT0	0.379	0.282	-0.069	-0.013	0.049	1.000	-0.002	0.013	0.037	0.000	0.018	-0.012	-0.001	
	N1	-0.470	-0.040	0.019	0.321	-0.124	-0.009	-0.002	1.000	0.745	0.751	0.032	-0.002	-0.019	
	N2	-0.603	-0.042	0.004	0.483	-0.125	-0.017	0.013	0.745	1.000	0.841	0.033	0.075	-0.014	0.015
	N3	-0.617	0.002	0.023	0.531	-0.126	-0.043	0.037	0.751	0.841	1.000	0.031	0.079	-0.021	-0.013
	MCONF0	0.033	0.005	-0.004	0.021	-0.033	0.051	0.000	-0.032	-0.033	-0.031	1.000	-0.328	-0.023	-0.014
	MNCONFT0	-0.102	-0.040	0.005	0.025	0.036	-0.120	-0.018	0.050	0.075	0.079	-0.328	1.000	0.028	0.098
	OTROST0	-0.051	-0.051	-0.003	-0.094	0.244	-0.758	-0.012	-0.002	-0.014	-0.021	-0.023	0.028	1.000	0.021
	PARTO	-0.197	-0.010	0.018	-0.197	-0.061	0.006	-0.001	-0.019	0.015	-0.013	-0.014	0.098	0.021	1.000

A.6.- Estadísticas para el modelo logit 6 con las 5 covariables e interacciones y punto de corte (c=0.5).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	198	161	55.2
		Continua	158	762	82.8
	Overall Percentage				75.1
Step 2	ESTADOT	Saliente	154	205	42.9
		Continua	100	820	89.1
	Overall Percentage				76.2
Step 3	ESTADOT	Saliente	160	199	44.6
		Continua	113	807	87.7
	Overall Percentage				75.6
Step 4	ESTADOT	Saliente	180	179	50.1
		Continua	120	800	87.0
	Overall Percentage				76.6
Step 5	ESTADOT	Saliente	182	177	50.7
		Continua	119	801	87.1
	Overall Percentage				76.9
Step 6	ESTADOT	Saliente	184	175	51.3
		Continua	118	802	87.2
	Overall Percentage				77.1
Step 7	ESTADOT	Saliente	183	176	51.0
		Continua	119	801	87.1
	Overall Percentage				76.9
Step 8	ESTADOT	Saliente	190	169	52.9
		Continua	115	805	87.5
	Overall Percentage				77.8
Step 9	ESTADOT	Saliente	168	191	46.8
		Continua	106	814	88.5
	Overall Percentage				76.8

a. The cut value is .500

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

A.7.- Estadísticas para el modelo logit 7 con las 5 covariables e interacciones (c=0.6).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	254	105	70.8
		Continua	234	686	74.6
	Overall Percentage				73.5
Step 2	ESTADOT	Saliente	240	119	66.9
		Continua	205	715	77.7
	Overall Percentage				74.7
Step 3	ESTADOT	Saliente	247	112	68.8
		Continua	200	720	78.3
	Overall Percentage				75.6
Step 4	ESTADOT	Saliente	244	115	68.0
		Continua	184	736	80.0
	Overall Percentage				76.6
Step 5	ESTADOT	Saliente	243	116	67.7
		Continua	192	728	79.1
	Overall Percentage				75.9
Step 6	ESTADOT	Saliente	241	118	67.1
		Continua	187	733	79.7
	Overall Percentage				76.2
Step 7	ESTADOT	Saliente	246	113	68.5
		Continua	190	730	79.3
	Overall Percentage				76.3
Step 8	ESTADOT	Saliente	255	104	71.0
		Continua	193	727	79.0
	Overall Percentage				76.8
Step 9	ESTADOT	Saliente	232	127	64.6
		Continua	164	756	82.2
	Overall Percentage				77.2

a. The cut value is .600

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

A.8.- Estadísticas para el modelo *logit* 8 con las 5 covariables e interacciones (c=0.65).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	254	105	70.8
		Continua	234	686	74.6
	Overall Percentage				73.5
Step 2	ESTADOT	Saliente	250	109	69.6
		Continua	218	702	76.3
	Overall Percentage				74.4
Step 3	ESTADOT	Saliente	259	100	72.1
		Continua	220	700	76.1
	Overall Percentage				75.0
Step 4	ESTADOT	Saliente	266	93	74.1
		Continua	229	691	75.1
	Overall Percentage				74.8
Step 5	ESTADOT	Saliente	263	96	73.3
		Continua	227	693	75.3
	Overall Percentage				74.7
Step 6	ESTADOT	Saliente	263	96	73.3
		Continua	225	695	75.5
	Overall Percentage				74.9
Step 7	ESTADOT	Saliente	267	92	74.4
		Continua	206	714	77.6
	Overall Percentage				76.7
Step 8	ESTADOT	Saliente	266	93	74.1
		Continua	218	702	76.3
	Overall Percentage				75.7
Step 9	ESTADOT	Saliente	267	92	74.4
		Continua	221	699	76.0
	Overall Percentage				75.5

a. The cut value is .650

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

A.9.- Estadísticas para el modelo *logit* 9 con las 5 covariables e interacciones (c=0.7).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

Observed			Predicted		
			ESTADOT		Percentage Correct
			Saliente	Continua	
Step 1	ESTADOT	Saliente	295	64	82.2
		Continua	324	596	64.8
	Overall Percentage				69.7
Step 2	ESTADOT	Saliente	279	80	77.7
		Continua	268	652	70.9
	Overall Percentage				72.8
Step 3	ESTADOT	Saliente	280	79	78.0
		Continua	272	648	70.4
	Overall Percentage				72.6
Step 4	ESTADOT	Saliente	291	68	81.1
		Continua	283	637	69.2
	Overall Percentage				72.6
Step 5	ESTADOT	Saliente	280	79	78.0
		Continua	262	658	71.5
	Overall Percentage				73.3
Step 6	ESTADOT	Saliente	283	76	78.8
		Continua	266	654	71.1
	Overall Percentage				73.3
Step 7	ESTADOT	Saliente	280	79	78.0
		Continua	264	656	71.3
	Overall Percentage				73.2
Step 8	ESTADOT	Saliente	290	69	80.8
		Continua	272	648	70.4
	Overall Percentage				73.3
Step 9	ESTADOT	Saliente	290	69	80.8
		Continua	273	647	70.3
	Overall Percentage				73.3

a. The cut value is .700

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

A.10.- Estadísticas para el modelo *logit* 10 con las 5 covariables e interacciones (c=0.66).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Iteration History[a,b,c,d,e]

Iteration		-2 Log likelihood	Coefficients								
			Constant	MESEST0	N3	FACTORT0 by INCRMT0	FACTORT0 by MERCT0	INCRMT0 by MESEST0	INCRMT0 by MERCT0	INCRMT0	MERCT0 by N3
Step 1	1	1264.492	-.325	.211							
	2	1211.502	-.639	.324							
	3	1206.163	-.770	.374							
	4	1206.071	-.789	.382							
Step 2	1	1236.041	.135	.166	-.796						
	2	1186.539	-.153	.275	-.787						
	3	1181.748	-.285	.322	-.772						
	4	1181.671	-.303	.329	-.771						
Step 3	1	1228.742	.082	.154	-.779	.072					
	2	1171.753	-.285	.252	-.750	.160					
	3	1164.385	-.482	.297	-.726	.224					
	4	1164.170	-.520	.305	-.723	.241					
	5	1164.170	-.521	.306	-.723	.241					
Step 4	1	1223.414	.048	.141	-.765	.060	.023				
	2	1163.107	-.343	.225	-.732	.144	.045				
	3	1154.256	-.567	.261	-.706	.218	.060				
	4	1153.889	-.619	.267	-.701	.242	.065				
	5	1153.888	-.622	.267	-.701	.244	.065				
Step 5	1	1219.917	.077	.115	-.778	-.006	.034	.045			
	2	1155.469	-.297	.168	-.755	.014	.069	.098			
	3	1144.469	-.525	.184	-.732	.038	.095	.142			
	4	1143.914	-.584	.186	-.728	.049	.103	.154			
	5	1143.912	-.588	.186	-.728	.050	.103	.155			
Step 6	1	1219.903	.074	.116	-.777		.033	.043			
	2	1155.553	-.289	.166	-.758		.071	.104			
	3	1144.753	-.503	.177	-.739		.099	.157			
	4	1144.239	-.556	.178	-.736		.107	.173			
	5	1144.238	-.559	.178	-.736		.108	.174			
Step 7	1	1218.871	.121	.106	-.775		.037	.058	-.064		
	2	1152.909	-.187	.145	-.754		.078	.141	-.145		
	3	1141.046	-.363	.150	-.732		.106	.215	-.213		
	4	1140.379	-.407	.150	-.728		.114	.240	-.237		
	5	1140.376	-.410	.150	-.728		.115	.242	-.239		
Step 8	1	1214.808	-.184	.131	-.769		.047	.022	-.145	.476	
	2	1147.478	-.625	.185	-.745		.097	.083	-.286	.699	
	3	1135.625	-.842	.195	-.725		.131	.146	-.375	.773	
	4	1134.983	-.891	.195	-.721		.141	.168	-.399	.781	
	5	1134.980	-.894	.195	-.721		.142	.170	-.401	.781	
Step 9	1	1209.123	-.188	.129	-1.338		.048	.022	-.221	.620	.395
	2	1144.620	-.615	.184	-1.243		.094	.085	-.384	.850	.353
	3	1133.429	-.818	.194	-1.139		.126	.148	-.467	.898	.297
	4	1132.818	-.864	.194	-1.110		.136	.170	-.486	.896	.279
	5	1132.816	-.867	.194	-1.108		.136	.171	-.487	.895	.278

a. Method: Forward Stepwise (Likelihood Ratio)

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1518.434

d. Estimation terminated at iteration number 4 because log-likelihood decreased by less than .010 percent.

e. Estimation terminated at iteration number 5 because log-likelihood decreased by less than .010 percent.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Contingency Table for Hosmer and Lemeshow Test

		ESTADOT = Saliente		ESTADOT = Continua		
		Observed	Expected	Observed	Expected	Total
Step 1	1	109	111.094	76	73.906	185
	2	89	86.600	82	84.400	171
	3	56	54.371	76	77.629	132
	4	41	42.372	90	88.628	131
	5	24	23.375	71	71.625	95
	6	20	22.670	124	121.330	144
	7	12	10.005	111	112.995	123
	8	7	4.732	118	120.268	125
	9	1	3.812	172	169.188	173
Step 2	1	73	77.977	42	37.023	115
	2	81	79.465	58	59.535	139
	3	51	50.378	55	55.622	106
	4	45	43.339	63	64.661	108
	5	31	29.097	57	58.903	88
	6	29	26.667	72	74.333	101
	7	22	25.223	115	111.777	137
	8	14	13.738	118	118.262	132
	9	8	6.260	102	103.740	110
	10	5	6.884	238	236.116	243
Step 3	1	87	88.916	45	43.084	132
	2	68	72.567	59	54.433	127
	3	56	51.237	51	55.763	107
	4	48	45.121	65	67.879	113
	5	36	38.559	87	84.441	123
	6	28	27.310	96	96.690	124
	7	19	17.576	112	113.424	131
	8	9	9.091	109	108.909	118
	9	7	5.643	121	122.357	128
	10	1	2.980	175	173.020	176
Step 4	1	84	85.311	42	40.689	126
	2	71	72.095	53	51.905	124
	3	66	65.676	71	71.324	137
	4	53	48.964	74	78.036	127
	5	37	36.640	91	91.360	128
	6	21	23.521	109	106.479	130
	7	12	13.968	117	115.032	129
	8	11	7.677	117	120.323	128
	9	3	3.844	125	124.156	128
	10	1	1.302	121	120.698	122
Step 5	1	87	87.441	40	39.559	127
	2	63	65.328	48	45.672	111
	3	62	58.446	56	59.554	118
	4	48	51.666	80	76.334	128
	5	43	38.782	85	89.218	128
	6	23	26.677	107	103.323	130
	7	22	15.891	106	112.109	128
	8	3	9.266	124	117.734	127
	9	7	3.979	124	127.021	131
	10	1	1.524	150	149.476	151
Step 6	1	84	85.311	42	40.689	126
	2	71	72.095	53	51.905	124
	3	66	65.676	71	71.324	137
	4	53	48.964	74	78.036	127
	5	37	36.640	91	91.360	128
	6	21	23.521	109	106.479	130
	7	12	13.968	117	115.032	129
	8	11	7.677	117	120.323	128
	9	3	3.844	125	124.156	128
	10	1	1.302	121	120.698	122
Step 7	1	85	88.289	43	39.711	128
	2	79	78.134	56	56.866	135
	3	62	61.239	68	68.761	130
	4	49	47.457	78	79.543	127
	5	38	35.174	90	92.826	128
	6	23	23.229	104	103.771	127
	7	14	14.484	113	112.516	127
	8	5	7.419	123	120.581	128
	9	4	2.678	123	124.322	127
	10	0	.898	122	121.102	122
Step 8	1	86	87.172	41	39.828	127
	2	75	75.306	53	52.694	128
	3	62	61.961	67	67.039	129
	4	53	52.856	86	86.144	139
	5	38	35.601	91	93.399	129
	6	20	22.703	109	106.297	129
	7	16	13.399	112	114.601	128
	8	6	6.676	122	121.324	128
	9	3	2.520	126	126.480	129
	10	0	.806	113	112.194	113
Step 9	1	87	89.523	41	38.477	128
	2	69	71.690	52	49.310	121
	3	65	62.685	66	68.315	131
	4	54	49.489	76	80.511	130
	5	35	36.026	93	91.974	128
	6	24	23.669	105	105.331	129
	7	16	14.343	112	113.657	128
	8	5	7.569	124	121.431	129
	9	3	2.934	123	123.066	126
	10	1	1.073	128	127.927	129

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	254	105	70.8
		Continua	234	686	74.6
	Overall Percentage				73.5
Step 2	ESTADOT	Saliente	250	109	69.6
		Continua	218	702	76.3
	Overall Percentage				74.4
Step 3	ESTADOT	Saliente	270	89	75.2
		Continua	241	679	73.8
	Overall Percentage				74.2
Step 4	ESTADOT	Saliente	272	87	75.8
		Continua	236	684	74.3
	Overall Percentage				74.7
Step 5	ESTADOT	Saliente	265	94	73.8
		Continua	233	687	74.7
	Overall Percentage				74.4
Step 6	ESTADOT	Saliente	263	96	73.3
		Continua	233	687	74.7
	Overall Percentage				74.3
Step 7	ESTADOT	Saliente	273	86	76.0
		Continua	228	692	75.2
	Overall Percentage				75.4
Step 8	ESTADOT	Saliente	269	90	74.9
		Continua	223	697	75.8
	Overall Percentage				75.5
Step 9	ESTADOT	Saliente	269	90	74.9
		Continua	223	697	75.8
	Overall Percentage				75.5

a. The cut value is .660

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465	1.388	1.547
	Constant	-.789	.121	42.451	1	.000	.454		
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390	1.313	1.471
	N3	-.771	.156	24.355	1	.000	.463	.341	.628
	Constant	-.303	.156	3.766	1	.052	.739		
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357	1.281	1.438
	N3	-.723	.157	21.096	1	.000	.485	.357	.661
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273	1.128	1.436
	Constant	-.521	.168	9.629	1	.002	.594		
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307	1.229	1.389
	N3	-.701	.158	19.711	1	.000	.496	.364	.676
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067	1.023	1.113
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276	1.121	1.453
	Constant	-.622	.173	12.944	1	.000	.537		
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205	1.118	1.298
	N3	-.728	.160	20.841	1	.000	.483	.353	.660
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109	1.054	1.165
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051	.885	1.249
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167	1.060	1.285
	Constant	-.588	.175	11.338	1	.001	.556		
Step 6[f]	MESEST0	.178	.035	26.268	1	.000	1.195	1.116	1.279
	N3	-.736	.159	21.467	1	.000	.479	.351	.654
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114	1.062	1.168
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190	1.111	1.275
	Constant	-.559	.167	11.234	1	.001	.572		
Step 7[g]	MESEST0	.150	.037	16.562	1	.000	1.161	1.081	1.248
	N3	-.728	.159	20.966	1	.000	.483	.353	.659
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122	1.069	1.176
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274	1.155	1.405
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788	.622	.998
	Constant	-.410	.183	5.028	1	.025	.664		
Step 8[h]	MESEST0	.195	.043	20.431	1	.000	1.215	1.117	1.322
	INCRMT0	.781	.338	5.328	1	.021	2.183	1.125	4.236
	N3	-.721	.159	20.507	1	.000	.486	.356	.664
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153	1.091	1.218
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185	1.055	1.331
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670	.512	.876
	Constant	-.894	.281	10.100	1	.001	.409		
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214	1.116	1.321
	INCRMT0	.895	.347	6.659	1	.010	2.448	1.240	4.831
	N3	-1.108	.310	12.739	1	.000	.330	.180	.607
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146	1.084	1.211
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187	1.056	1.333
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320	.906	1.924
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615	.460	.820
	Constant	-.867	.282	9.473	1	.002	.420		

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

Correlation Matrix

		Constant	MESEST0	N3	Constant	MESEST0	N3	FACTORT0 by INCRMT0	FACTORT0 by MERCT0	INCRMT0 by MESEST0	Constant	MESEST0	N3	FACTORT0 by MERCT0	INCRMT0 by MESEST0	INCRMT0 by MERCT0	INCRMT0	MERCT0 by N3
Step	Constant	1.000	-0.814															
1	MESEST0	-0.814	1.000															
Step	Constant	1.000	-0.810	-0.624														
2	MESEST0	-0.810	1.000	0.346														
	N3	-0.624	0.346	1.000														
Step	Constant				1.000	-0.720	-0.605	-0.346										
3	MESEST0				-0.720	1.000	0.339	0.104										
	N3				-0.605	0.339	1.000	0.054										
	FACTORT0 by INCRMT0				-0.346	-0.104	0.054	1.000										
Step	Constant				1.000	-0.580	-0.596	-0.364	-0.208									
4	MESEST0				-0.580	1.000	0.305	-0.108	-0.341									
	N3				-0.596	0.305	1.000	0.059	0.030									
	FACTORT0 by MERCT0				-0.208	-0.341	0.030	0.027	1.000									
	FACTORT0 by INCRMT0				-0.364	-0.108	0.059	1.000	0.027									
Step	Constant				1.000	-0.496	-0.596	-0.295	-0.163	0.022								
5	MESEST0				-0.496	1.000	0.287	0.397	-0.537	-0.579								
	N3				-0.596	0.287	1.000	0.089	-0.002	-0.056								
	FACTORT0 by MERCT0				-0.163	-0.537	-0.002	-0.324	1.000	0.458								
	FACTORT0 by INCRMT0				-0.295	0.397	0.089	1.000	-0.324	-0.696								
	INCRMT0 by MESEST0				0.022	-0.579	-0.056	-0.696	0.458	1.000								
Step	Constant										1.000	-0.436	-0.600	-0.285	-0.269			
6	MESEST0										-0.436	1.000	0.276	-0.467	-0.447			
	N3										-0.600	0.276	1.000	0.030	0.010			
	FACTORT0 by MERCT0										-0.285	-0.467	0.030	1.000	0.326			
	INCRMT0 by MESEST0										-0.269	-0.447	0.010	0.326	1.000			
Step	Constant										1.000	-0.532	-0.541	-0.188	0.098	-0.397		
7	MESEST0										-0.532	1.000	0.249	-0.480	-0.526	0.380		
	N3										-0.541	0.249	1.000	0.036	0.026	-0.021		
	FACTORT0 by MERCT0										-0.188	-0.480	0.036	1.000	0.311	-0.160		
	INCRMT0 by MESEST0										0.098	-0.526	0.026	0.311	1.000	-0.714		
	INCRMT0 by MERCT0										-0.397	0.380	-0.021	-0.160	-0.714	1.000		
Step	Constant										1.000	-0.661	-0.359	-0.445	0.473	0.159	-0.761	
8	MESEST0										-0.661	1.000	0.222	-0.168	-0.666	0.056	0.472	
	INCRMT0										-0.761	0.472	0.009	0.436	-0.536	-0.499	1.000	
	N3										-0.359	0.222	1.000	0.039	0.010	-0.024	0.009	
	FACTORT0 by MERCT0										-0.445	-0.168	0.039	1.000	0.003	-0.349	0.436	
	INCRMT0 by MESEST0										0.473	-0.666	0.010	0.003	1.000	-0.257	-0.536	
	INCRMT0 by MERCT0										0.159	0.056	-0.024	-0.349	-0.257	1.000	-0.499	
Step	Constant										1.000	-0.662	-0.232	-0.443	0.476	0.123	-0.728	0.054
9	MESEST0										-0.662	1.000	0.121	-0.170	-0.665	0.061	0.457	-0.010
	INCRMT0										-0.728	0.457	-0.192	0.395	-0.518	-0.531	1.000	0.226
	N3										-0.232	0.121	1.000	0.123	-0.014	0.333	-0.192	-0.856
	FACTORT0 by MERCT0										-0.443	-0.170	0.123	1.000	0.000	-0.284	0.395	-0.110
	INCRMT0 by MESEST0										0.476	-0.665	-0.014	0.000	1.000	-0.252	-0.518	0.024
	MERCT0 by N3										0.054	-0.100	-0.856	-0.110	0.024	-0.399	0.226	1.000
	INCRMT0 by MERCT0										0.123	0.061	0.333	-0.284	-0.252	1.000	-0.531	-0.399

Casewise List

Case	Selected Status	Observed ESTADOT	Predicted	Predicted Group	Temporary Variable Resid	Temporary Variable ZResid
13	S	S**	0.915	C	-0.915	-3.272
29	S	S**	0.963	C	-0.963	-5.084
33	S	S**	0.903	C	-0.903	-3.056
110	S	S**	0.945	C	-0.945	-4.154
136	S	S**	0.970	C	-0.970	-5.659
269	S	S**	0.881	C	-0.881	-2.726
370	S	S**	0.894	C	-0.894	-2.907
379	S	S**	0.887	C	-0.887	-2.807
476	S	S**	0.960	C	-0.960	-4.910
707	S	S**	0.903	C	-0.903	-3.056
718	S	S**	0.896	C	-0.896	-2.928
757	S	S**	0.928	C	-0.928	-3.598
828	S	S**	0.973	C	-0.973	-5.985
847	S	S**	0.988	C	-0.988	-8.922
886	S	S**	0.933	C	-0.933	-3.728
921	S	S**	0.904	C	-0.904	-3.070
955	S	S**	0.878	C	-0.878	-2.679
1112	S	S**	0.899	C	-0.899	-2.983
1142	S	S**	0.867	C	-0.867	-2.559
1150	S	S**	0.907	C	-0.907	-3.124
1170	S	S**	0.904	C	-0.904	-3.068
1197	S	S**	0.904	C	-0.904	-3.070

a S = Selected, U = Unselected cases, and ** = Misclassified cases.

b Cases with studentized residuals greater than 2.000 are listed.

A.11.- Estadísticas para el modelo *logit* 11 con las 5 covariables e interacciones (c=0.67).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	254	105	70.8
		Continua	234	686	74.6
	Overall Percentage				73.5
Step 2	ESTADOT	Saliente	279	80	77.7
		Continua	268	652	70.9
	Overall Percentage				72.8
Step 3	ESTADOT	Saliente	278	81	77.4
		Continua	262	658	71.5
	Overall Percentage				73.2
Step 4	ESTADOT	Saliente	274	85	76.3
		Continua	240	680	73.9
	Overall Percentage				74.6
Step 5	ESTADOT	Saliente	270	89	75.2
		Continua	242	678	73.7
	Overall Percentage				74.1
Step 6	ESTADOT	Saliente	265	94	73.8
		Continua	239	681	74.0
	Overall Percentage				74.0
Step 7	ESTADOT	Saliente	275	84	76.6
		Continua	238	682	74.1
	Overall Percentage				74.8
Step 8	ESTADOT	Saliente	271	88	75.5
		Continua	229	691	75.1
	Overall Percentage				75.2
Step 9	ESTADOT	Saliente	273	86	76.0
		Continua	229	691	75.1
	Overall Percentage				75.4

a. The cut value is .670

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	MESEST0	-759.217	312.362	1	.000
Step 2	MESEST0	-678.565	175.460	1	.000
	N3	-603.036	24.400	1	.000
Step 3	MESEST0	-652.463	140.755	1	.000
	N3	-592.640	21.110	1	.000
	FACTORT0 * INCRMT0	-590.836	17.501	1	.000
Step 4	MESEST0	-622.360	90.831	1	.000
	N3	-586.800	19.712	1	.000
	FACTORT0 * MERCT0	-582.085	10.282	1	.001
	FACTORT0 * INCRMT0	-584.600	15.311	1	.000
Step 5	MESEST0	-585.712	27.512	1	.000
	N3	-582.394	20.876	1	.000
	FACTORT0 * MERCT0	-581.478	19.044	1	.000
	FACTORT0 * INCRMT0	-572.119	.325	1	.568
	INCRMT0 * MESEST0	-576.944	9.976	1	.002
Step 6	MESEST0	-587.134	30.030	1	.000
	N3	-582.862	21.486	1	.000
	FACTORT0 * MERCT0	-584.304	24.371	1	.000
	INCRMT0 * MESEST0	-584.600	24.962	1	.000
Step 7	MESEST0	-579.389	18.403	1	.000
	N3	-580.676	20.975	1	.000
	FACTORT0 * MERCT0	-583.853	27.329	1	.000
	INCRMT0 * MESEST0	-582.558	24.739	1	.000
	INCRMT0 * MERCT0	-572.119	3.861	1	.049
Step 8	MESEST0	-579.383	23.786	1	.000
	INCRMT0	-570.188	5.396	1	.020
	N3	-577.754	20.528	1	.000
	FACTORT0 * MERCT0	-583.847	32.713	1	.000
	INCRMT0 * MESEST0	-571.570	8.160	1	.004
	INCRMT0 * MERCT0	-571.645	8.309	1	.004
Step 9	MESEST0	-578.227	23.639	1	.000
	INCRMT0	-569.784	6.752	1	.009
	N3	-572.997	13.178	1	.000
	FACTORT0 * MERCT0	-581.509	30.201	1	.000
	INCRMT0 * MESEST0	-570.563	8.310	1	.004
	MERCT0 * N3	-567.490	2.164	1	.141
	INCRMT0 * MERCT0	-571.628	10.439	1	.001

A.12.- Estadísticas para el modelo *logit* 12 con las 5 covariables e interacciones (c=0.68).

Block 1: Method = Forward Stepwise (Likelihood Ratio)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	312.362	1	.000
	Block	312.362	1	.000
	Model	312.362	1	.000
Step 2	Step	24.400	1	.000
	Block	336.763	2	.000
	Model	336.763	2	.000
Step 3	Step	17.501	1	.000
	Block	354.264	3	.000
	Model	354.264	3	.000
Step 4	Step	10.282	1	.001
	Block	364.545	4	.000
	Model	364.545	4	.000
Step 5	Step	9.976	1	.002
	Block	374.522	5	.000
	Model	374.522	5	.000
Step 6[a]	Step	-.325	1	.568
	Block	374.196	4	.000
	Model	374.196	4	.000
Step 7	Step	3.861	1	.049
	Block	378.058	5	.000
	Model	378.058	5	.000
Step 8	Step	5.396	1	.020
	Block	383.454	6	.000
	Model	383.454	6	.000
Step 9	Step	2.164	1	.141
	Block	385.618	7	.000
	Model	385.618	7	.000

a. A negative Chi-squares value indicates that the Chi-squares value has decreased from the previous step.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1206.071	.217	.312
2	1181.671	.231	.333
3	1164.170	.242	.348
4	1153.888	.248	.357
5	1143.912	.254	.365
6	1144.238	.254	.365
7	1140.376	.256	.368
8	1134.980	.259	.373
9	1132.816	.260	.375

Classification Table[a]

			Predicted		
			ESTADOT		Percentage Correct
	Observed		Saliente	Continua	
Step 1	ESTADOT	Saliente	295	64	82.2
		Continua	324	596	64.8
	Overall Percentage				69.7
Step 2	ESTADOT	Saliente	279	80	77.7
		Continua	268	652	70.9
	Overall Percentage				72.8
Step 3	ESTADOT	Saliente	278	81	77.4
		Continua	262	658	71.5
	Overall Percentage				73.2
Step 4	ESTADOT	Saliente	274	85	76.3
		Continua	240	680	73.9
	Overall Percentage				74.6
Step 5	ESTADOT	Saliente	273	86	76.0
		Continua	251	669	72.7
	Overall Percentage				73.7
Step 6	ESTADOT	Saliente	272	87	75.8
		Continua	251	669	72.7
	Overall Percentage				73.6
Step 7	ESTADOT	Saliente	275	84	76.6
		Continua	245	675	73.4
	Overall Percentage				74.3
Step 8	ESTADOT	Saliente	282	77	78.6
		Continua	260	660	71.7
	Overall Percentage				73.7
Step 9	ESTADOT	Saliente	275	84	76.6
		Continua	235	685	74.5
	Overall Percentage				75.1

a. The cut value is .680

Variables in the Equation[i]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	MESEST0	.382	.028	190.775	1	.000	1.465
	Constant	-.789	.121	42.451	1	.000	.454
Step 2[b]	MESEST0	.329	.029	127.859	1	.000	1.390
	N3	-.771	.156	24.355	1	.000	.463
	Constant	-.303	.156	3.766	1	.052	.739
Step 3[c]	MESEST0	.306	.030	106.994	1	.000	1.357
	N3	-.723	.157	21.096	1	.000	.485
	FACTORT0 by INCRMT0	.241	.061	15.419	1	.000	1.273
	Constant	-.521	.168	9.629	1	.002	.594
Step 4[d]	MESEST0	.267	.031	73.453	1	.000	1.307
	N3	-.701	.158	19.711	1	.000	.496
	FACTORT0 by MERCT0	.065	.022	9.123	1	.003	1.067
	FACTORT0 by INCRMT0	.244	.066	13.521	1	.000	1.276
	Constant	-.622	.173	12.944	1	.000	.537
Step 5[e]	MESEST0	.186	.038	23.971	1	.000	1.205
	N3	-.728	.160	20.841	1	.000	.483
	FACTORT0 by MERCT0	.103	.026	16.273	1	.000	1.109
	FACTORT0 by INCRMT0	.050	.088	.322	1	.571	1.051
	INCRMT0 by MESEST0	.155	.049	9.966	1	.002	1.167
	Constant	-.588	.175	11.338	1	.001	.556
Step 6[e]	MESEST0	.178	.035	26.268	1	.000	1.195
	N3	-.736	.159	21.467	1	.000	.479
	FACTORT0 by MERCT0	.108	.024	19.923	1	.000	1.114
	INCRMT0 by MESEST0	.174	.035	24.825	1	.000	1.190
	Constant	-.559	.167	11.234	1	.001	.572
Step 7[f]	MESEST0	.150	.037	16.562	1	.000	1.161
	N3	-.728	.159	20.966	1	.000	.483
	FACTORT0 by MERCT0	.115	.024	22.225	1	.000	1.122
	INCRMT0 by MESEST0	.242	.050	23.470	1	.000	1.274
	INCRMT0 by MERCT0	-.239	.121	3.915	1	.048	.788
	Constant	-.410	.183	5.028	1	.025	.664
Step 8[g]	MESEST0	.195	.043	20.431	1	.000	1.215
	INCRMT0	.781	.338	5.328	1	.021	2.183
	N3	-.721	.159	20.507	1	.000	.486
	FACTORT0 by MERCT0	.142	.028	25.275	1	.000	1.153
	INCRMT0 by MESEST0	.170	.059	8.176	1	.004	1.185
	INCRMT0 by MERCT0	-.401	.137	8.587	1	.003	.670
	Constant	-.894	.281	10.100	1	.001	.409
Step 9[h]	MESEST0	.194	.043	20.336	1	.000	1.214
	INCRMT0	.895	.347	6.659	1	.010	2.448
	N3	-1.108	.310	12.739	1	.000	.330
	FACTORT0 by MERCT0	.136	.028	23.249	1	.000	1.146
	INCRMT0 by MESEST0	.171	.059	8.325	1	.004	1.187
	MERCT0 by N3	.278	.192	2.091	1	.148	1.320
	INCRMT0 by MERCT0	-.487	.147	10.925	1	.001	.615
	Constant	-.867	.282	9.473	1	.002	.420

a. Variable(s) entered on step 1: MESEST0.

b. Variable(s) entered on step 2: N3.

c. Variable(s) entered on step 3: FACTORT0 * INCRMT0 .

d. Variable(s) entered on step 4: FACTORT0 * MERCT0 .

e. Variable(s) entered on step 5: INCRMT0 * MESEST0 .

f. Variable(s) entered on step 7: INCRMT0 * MERCT0 .

g. Variable(s) entered on step 8: INCRMT0.

h. Variable(s) entered on step 9: MERCT0 * N3 .

i. Stepwise procedure stopped because removing the least significant variable result in a previously fitted model.

A.13.- Estadísticas para el modelo *logit* 13 con las 5 covariables e interacciones sin los 22 casos observados (c=0.66).

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients							
			Constant	MESEST0	INCRMT0	N3	FACTORT0 by MERCT0	INCRMT0 by MESEST0	INCRMT0 by MERCT0	MERCT0 by N3
Step 1	1	1124.595	-.206	.135	.650	-1.396	.052	.019	-.208	.431
	2	1032.223	-.738	.199	.908	-1.256	.112	.095	-.378	.368
	3	1004.105	-1.102	.213	.980	-1.055	.174	.200	-.489	.252
	4	999.728	-1.269	.213	.990	-.958	.211	.266	-.538	.186
	5	999.601	-1.301	.213	.991	-.941	.218	.280	-.547	.173
	6	999.601	-1.302	.213	.991	-.941	.219	.280	-.547	.173

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1461.535

d. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	461.934	7	.000
	Block	461.934	7	.000
	Model	461.934	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	999.601	.308	.447

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	8.737	8	.365

Contingency Table for Hosmer and Lemeshow Test

		ESTADOT = Saliente		ESTADOT = Continua		
		Observed	Expected	Observed	Expected	Total
Step 1	1	88	92.168	39	34.832	127
	2	67	68.156	45	43.844	112
	3	61	62.859	65	63.141	126
	4	51	47.799	74	77.201	125
	5	40	34.295	93	98.705	133
	6	23	18.286	103	107.714	126
	7	7	9.290	120	117.710	127
	8	0	3.209	127	123.791	127
	9	0	.766	127	126.234	127
	10	0	.170	127	126.830	127

Classification Table[a]

			Predicted		
			ESTADOT		Percentage
	Observed		Saliente	Continua	Correct
Step 1	ESTADOT	Saliente	264	73	78.3
		Continua	216	704	76.5
	Overall Percentage				77.0

a. The cut value is .660

A.14.- Estadísticas para el modelo logit 14 (c=0.66).

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients						
			Constant	INCRMT0	N3	FACTORT0 by MERCT0	INCRMT0 by MESEST0	INCRMT0 by MERCT0	MESEST0
Step 1	1	1131.020	-.197	.489	-.780	.050	.019	-.124	.138
	2	1034.242	-.749	.743	-.744	.115	.093	-.266	.200
	3	1004.787	-1.133	.867	-.713	.181	.198	-.394	.214
	4	1000.391	-1.298	.910	-.707	.217	.265	-.469	.213
	5	1000.271	-1.327	.918	-.707	.224	.279	-.484	.213
	6	1000.271	-1.328	.918	-.707	.224	.279	-.484	.213

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 1461.535

d. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	461.264	6	.000
	Block	461.264	6	.000
	Model	461.264	6	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1000.271	.307	.447

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	7.380	8	.496

Contingency Table for Hosmer and Lemeshow Test

		ESTADOT = Saliente		ESTADOT = Continua		
		Observed	Expected	Observed	Expected	Total
Step 1	1	87	90.646	39	35.354	126
	2	67	68.217	45	43.783	112
	3	60	62.491	64	61.509	124
	4	53	48.990	73	77.010	126
	5	40	35.199	95	99.801	135
	6	22	18.305	104	107.695	126
	7	8	9.069	117	115.931	125
	8	0	3.163	126	122.837	126
	9	0	.755	127	126.245	127
	10	0	.167	130	129.833	130

Classification Table[a]

			Predicted		
			ESTADOT		Percentage
	Observed		Saliente	Continua	Correct
Step 1	ESTADOT	Saliente	266	71	78.9
		Continua	218	702	76.3
	Overall Percentage				77.0

a. The cut value is .660

Variables in the Equation

								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	INCRMT0	.918	.372	6.077	1	.014	2.504	1.207	5.194
	N3	-.707	.166	18.188	1	.000	.493	.356	.682
	FACTORT0 by MERCT0	.224	.039	33.782	1	.000	1.251	1.160	1.350
	INCRMT0 by MESEST0	.279	.073	14.521	1	.000	1.322	1.145	1.527
	INCRMT0 by MERCT0	-.484	.156	9.654	1	.002	.616	.454	.836
	MESEST0	.213	.047	20.828	1	.000	1.237	1.129	1.356
	Constant	-1.328	.311	18.192	1	.000	.265		

a. Variable(s) entered on step 1: INCRMT0, N3, FACTORT0 * MERCT0 , INCRMT0 * MESEST0 , INCRMT0 * MERCT0 , MESEST0.

Correlation Matrix

		Constant	INCRMT0	N3	FACTOR BY MERCT0	INCRMT0 BY MESEST0	INCRMT0 BY MERCT0	MESEST0
Step	Constant	1.000	-0.765	-0.335	-0.532	0.402	0.179	-0.629
1	INCRMT0	-0.765	1.000	0.008	0.475	-0.513	-0.495	0.462
	N3	-0.335	0.008	1.000	0.031	0.011	-0.033	0.215
	FACTOR BY MERCT0	-0.532	0.475	0.031	1.000	-0.013	-0.324	-0.127
	INCRMT0 BY MESEST0	0.402	-0.513	0.011	-0.013	1.000	-0.281	-0.592
	INCRMT0 BY MERCT0	0.179	-0.495	-0.033	-0.324	-0.281	1.000	0.036
	MESEST0	-0.629	0.462	0.215	-0.127	-0.592	0.036	1.000

A.15.- Estadísticas para los pronósticos de la regresión logística (c=0.66).

INCREM * ESTADO1 Crosstabulation

Count

		ESTADO1		
		Saliente	Continua	Total
INCREM	Disminuye	91	139	230
	Incrementa	193	554	747
Total		284	693	977

N3 * ESTADO1 Crosstabulation

Count

		ESTADO1		
		Saliente	Continua	Total
N3	Otro	267	455	722
	Exportó entre US$ 5.0 y 15.5 miles	17	238	255
Total		284	693	977

MERCADO * ESTADO1 Crosstabulation

Count

		ESTADO1		
		Saliente	Continua	Total
MERCADO	1	238	396	634
	2	39	164	203
	3	4	76	80
	4	1	42	43
	5	1	12	13
	6		2	2
	7	1	1	2
Total		284	693	977

MESES * ESTADO1 Crosstabulation

Count

		ESTADO1		
		Saliente	Continua	Total
MESES	1	96	4	100
	2	115	10	125
	3	37	82	119
	4	19	62	81
	5	10	77	87
	6	6	77	83
	7	1	52	53
	8		58	58
	9		39	39
	10		40	40
	11		61	61
	12		131	131
Total		284	693	977

FACTOR * ESTADO1 Crosstabulation

Count

		ESTADO1		
		Saliente	Continua	Total
FACTOR	1	186	218	404
	2	53	128	181
	3	21	80	101
	4	10	48	58
	5	2	46	48
	6	12	173	185
Total		284	693	977

Anexo A.16: Cálculos previos para el análisis de las interacciones del modelo logit 14 (c=0.66)

$ß_3$	0.918
$ß_{3,2}$	0.279

X_2	$\exp(ß_3+ß_{3,2}X_2)$
1	3.310
2	4.375
3	5.783
4	7.645
5	10.105
6	13.356
7	17.655
8	23.336
9	30.846
10	40.772
11	53.893
12	71.236

$ß_3$	0.918
$ß_{3,5}$	-0.484

X_5	$\exp(ß_3+ß_{3,5}X_5)$
1	1.543
2	0.951
3	0.586
4	0.361
5	0.223
6	0.137

$ß_1$	0.000
$ß_{1,5}$	0.224

X_5	$\exp(ß_1+ß_{1,5}X_5)$
1	1.251
2	1.565
3	1.958
4	2.450
5	3.065
6	3.834

ANEXO B.

Gráficos complementarios de análisis de residuos.

Gráfico de Dispersión entre los residuos ordinarios y las observaciones

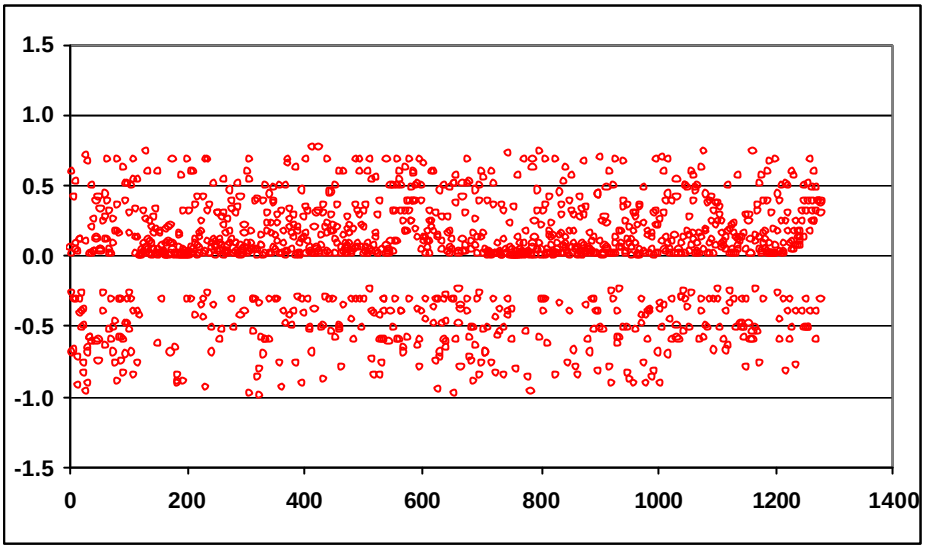


Gráfico de Dispersión entre la Distancia de Cook y los Leverage

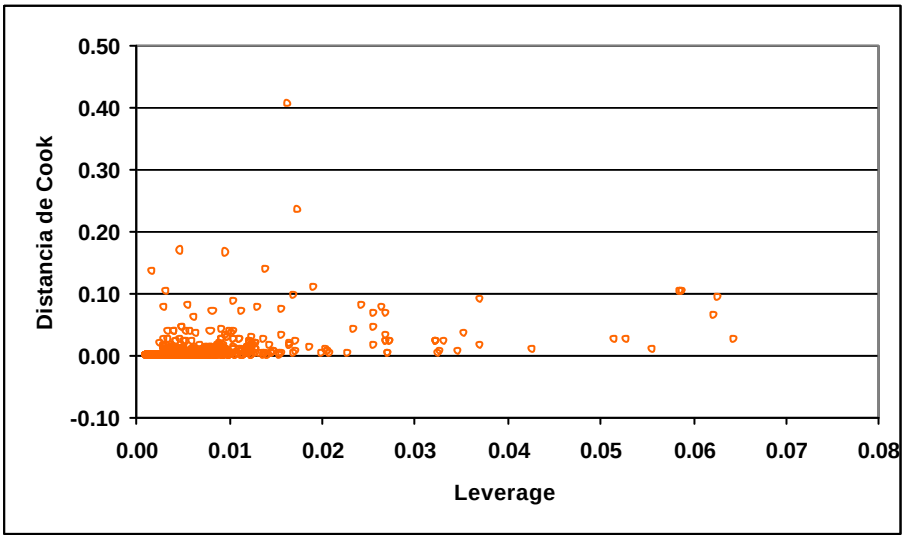


Gráfico de Dispersión Delta Chi Cuadrado

y los Leverage

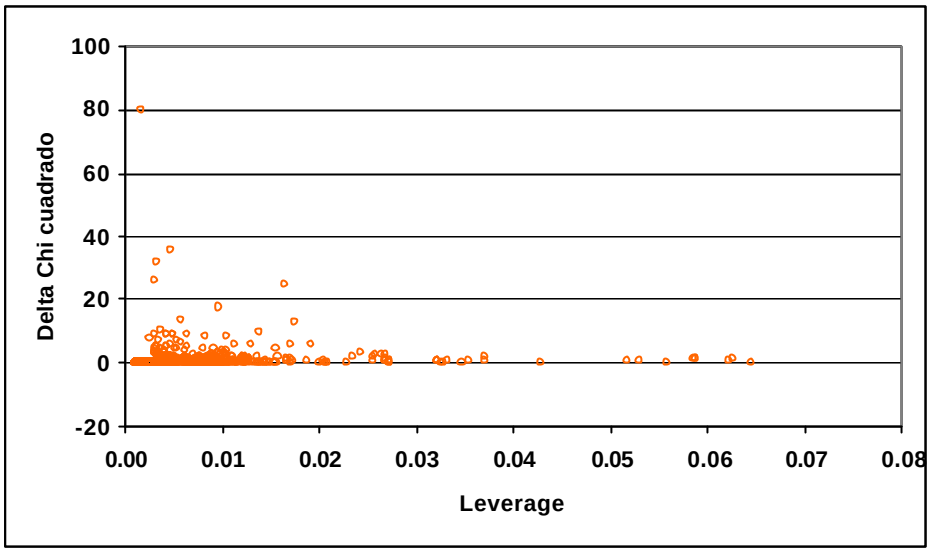
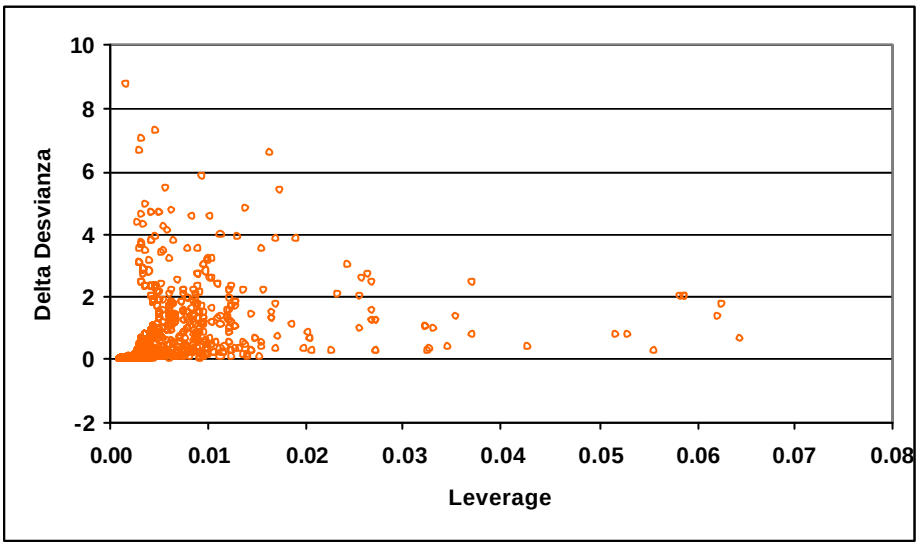


Gráfico de Dispersión Delta Desvianza

y los Leverage



ANEXO C.

Sentencias SQL utilizadas en las consultas a la base de datos de exportación de Aduanas.

Consulta C.1.

```
-- Permite obtener exportaciones de las empresas 2002 - 2005.
-- Con esto se tiene las variables: Número de mercados.
select substr(a.femb,1,4), a.ndoc, a.cpaides, sum(a.vfobserdol)
from exportacion a
where (a.part_nandi between '6100000000' and '6299999999')
        and (a.femb between 20020101 and 20051231)
group by substr(a.femb,1,4), a.ndoc, a.cpaides
```

Consulta C.2.

```
-- Establece exportaciones de las empresas 2002 - 2005.
-- Con esto se tiene las variables: número de partidas.
select substr(a.femb,1,4), a.ndoc, a.part_nandi, sum(a.vfobserdol)
from exportacion a
where (a.part_nandi between '6100000000' and '6299999999')
        and (a.femb between 20020101 and 20051231)
group by substr(a.femb,1,4), a.ndoc, a.part_nandi
```

Consulta C.3.

```
-- Obtiene partidas por empresa y mes 1994 - 2005.
-- Con esta consulta se calcula las variables número de meses de exportación.
select substr(a.femb,1,6), a.ndoc, sum(a.vfobserdol)
from exportacion a
where (a.part_nandi between '6100000000' and '6299999999')
        and (a.femb between 19940101 and 20051231)
group by substr(a.femb,1,6), a.ndoc
```

Consulta C.4.

```
-- Con esta consulta se obtiene las variables: Número de mercados y exportaciones
-- de otros sectores diferente al de confecciones.
select substr(a.femb,1,4), a.ndoc, a.cpaides, sum(a.vfobserdol)
from exportacion a
where (a.part_nandi<'6100000000' or a.part_nandi>'6299999999')
        and (a.femb between 20020101 and 20051231)
group by substr(a.femb,1,4), a.ndoc, a.cpaides
```

Consulta C.5.

```
-- Consulta para obtener las importaciones por empresa 2002 – 2005 de sectores
-- distintos a las confecciones.
select a.cod_anho, a.libr_tribu, sum(a.fob_dolpol)
from importacion a
where (a.cod_anho in ('2002','2003', '2004','2005'))
    and (a.part_nandi<'6100000000' or a.part_nandi>'6299999999')
group by a.cod_anho, a.libr_tribu
```

Consulta C.6.

```
-- Consulta para obtener las importaciones por empresa 2002 – 2005 de confecciones.
select a.cod_anho, a.libr_tribu, sum(a.fob_dolpol)
from importacion a
where (a.cod_anho in ('2002','2003','2004','2005'))
    and (a.part_nandi between 6100000000 and 6299999999)
group by a.cod_anho, a.libr_tribu
```

ANEXO D.

Sentencias SPSS utilizadas en la gestión de la base de datos de la investigación.

```
*-----------------------------------------------------
* selecciona años 2002 y 2003 en t-1
*-----------------------------------------------------
USE ALL.
COMPUTE filter_$=(anhot0<2004).
VARIABLE LABEL filter_$ 'anhot0<2004'+
' (FILTER)'.
VALUE LABELS filter_$  0 'Not Selected' 1 'Selected'.
FORMAT filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE .

*-----------------------------------------------------
*modelo con todas las variables (modelo 1)
*-----------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=ENTER estadot0 factort0 mesest0 part0 merct0 otrost0 otrmert0
  mconf0 mnconft0 limat0 incrmt0 n1 n2 n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.30) ITERATE(20) CUT(.5) .
```

```
ROC
  pgr_1  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .


*------------------------------------------------------------------------------------------------
*todas las variables menos las eliminadas en el modelo anterior (modelo 2)
*------------------------------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=ENTER factort0 mesest0 merct0 otrmert0
  incrmt0 n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.30) ITERATE(20) CUT(.5) .

ROC
  pgr_2  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```

```
*-----------------------------------------------------
* método forward (modelo 3)
*-----------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) estadot0 factort0 mesest0 part0 merct0 otrost0 otrmert0
  mconf0 mnconft0 limat0 incrmt0 n1 n2 n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.10) ITERATE(20) CUT(.5) .

ROC
  pgr_3  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*-----------------------------------------------------------------------
* método backward - significancia= 0.10 (modelo 4)
*-----------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=BSTEP(LR) factort0 mesest0 merct0 incrmt0 n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.10) ITERATE(20) CUT(.5) .
```

```
ROC
  pgr_4  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*------------------------------------------------------------------------
* método backward - significancia= 0.13 (modelo 5)
*------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=BSTEP(LR) factort0 mesest0 merct0 incrmt0 n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.13) ITERATE(20) CUT(.5) .

ROC
  pgr_5  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```

```
*------------------------------------------------------
* Correlaciones de todas las variables.
*------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=ENTER estadot0 factort0 mesest0 merct0 otrmert0 incrmt0 n1
  n2 n3 limat0 mconf0 mnconft0 otrost0 part0
  /PRINT=CORR
  /CRITERIA PIN(.25) POUT(.30) ITERATE(20) CUT(.5) .

*------------------------------------------------------------------------
* modelo con interacciones - punto c=0.5  (modelo 6).
*------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.5) .

ROC
  pgr_6  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```

```
*-------------------------------------------------------------------------
* modelo con interacciones - punto c=0.6  (modelo 7).
*-------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.6) .
ROC
  pgr_7  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*-------------------------------------------------------------------------
* modelo con interacciones - punto c=0.65  (modelo 8).
*-------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.65) .
```

```
ROC
  pgr_8  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*------------------------------------------------------------------------
* modelo con interacciones - punto c=0.7  (modelo 9).
*------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.7) .

ROC
  pgr_9  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```

```
*------------------------------------------------------------------------------
* modelo con interacciones - punto c=0.66  (modelo 10).
*------------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.66) .
ROC
  pgr_10  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*------------------------------------------------------------------------------
* modelo con interacciones - punto c=0.67  (modelo 11).
*------------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.67) .
```

```
ROC
  pgr_11  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .

*------------------------------------------------------------------------
* modelo con interacciones - punto c=0.68  (modelo 12).
*------------------------------------------------------------------------
LOGISTIC REGRESSION VAR=estadot
  /METHOD=FSTEP(LR) factort0 mesest0 merct0 incrmt0 n3 factort0*mesest0
  factort0*merct0 factort0*n3 factort0*incrmt0 merct0*mesest0
  incrmt0*mesest0 incrmt0*n3 merct0*n3 incrmt0*merct0 incrmt0*n3
  /SAVE PRED PGROUP COOK LEVER DFBETA RESID LRESID SRESID ZRESID DEV
  /CLASSPLOT /CASEWISE OUTLIER(2)
  /PRINT=GOODFIT CORR ITER(1) CI(95)
  /CRITERIA PIN(.25) POUT(.05) ITERATE(20) CUT(.68) .

ROC
  pgr_12  BY estadot (1)
  /PLOT = CURVE(REFERENCE)
  /PRINT = SE COORDINATES
  /CRITERIA = CUTOFF(INCLUDE) TESTPOS(LARGE) DISTRIBUTION(FREE) CI(95)
  /MISSING = EXCLUDE .
```