UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE LETRAS Y CIENCIAS HUMANAS

E.A.P. DE ESTADÍSTICA

Factores importantes que determinan el ingreso a la Universidad Nacional Mayor de San Marcos, 2003

MONOGRAFÍA

Para optar el Título de Licenciado en Estadística

AUTOR

José Romualdo Moina Fuentes

LIMA – PERÚ
2003

FACTORES IMPORTANTES QUE DETERMINAN EL INGRESO A LA UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS-2003

JOSÉ ROMUALDO MOINA FUENTES

Monografía presentada a consideración del Cuerpo Docente de la Facultad de Ciencias Matemáticas de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para optar el Título Profesional de Licenciado en Estadística.

Aprobado por:

Jacinto Pedro Mendoza Solís

Jurado

Olga Lidia Solano Dávila

Asesora

Lima – Perú

Diciembre 2003

FICHA CÁTALOGRÁFICA

MOINA FUENTES, JOSÉ ROMUALDO

Factores Importantes que Determinan el Ingreso a la Universidad Nacional Mayor de San Marcos-2003. (Lima) 2003.

VI, 100p., 29.7 cm, (UNMSM, Licenciado, Estadística, 2003).

Monografía Universidad Nacional Mayor de San Marcos,
Facultad de Ciencias Matemáticas 1. Estadística

I. UNMSM/FCM. Título (Serie).

A mi padre, mi madre y hermanos

RESUMEN

FACTORES IMPORTANTES QUE DETERMINAN EL INGRESO A LA UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS-2003

JOSÉ ROMUALDO MOINA FUENTES

DICIEMBRE 2003

Asesora: **Mg. Olga Lidia Solano Dávila**

Título: **Licenciado en Estadística**

==========

El presente estudio intenta determina las variables más importantes al momento de explicar mejor el éxito en el examen de Admisión (para este caso entiéndase como éxito el ingreso a la Universidad Nacional Mayor de San Marcos). A fin de cumplir con los objetivos planteados se aplicó el Modelo de Regresión Logística. Para esto se utilizó la información de la Encuesta a los Postulantes Sanmarquinos 2003 proporcionada por la Oficina Técnica del Estudiante (OTE) y la información que se obtuvo de la Ficha que el postulante llenó al momento de Inscribirse. Esta última fue proporcionada por la Comisión Ejecutiva de Admisión (CEA). De acuerdo a los resultados de este análisis, las variables que más influyen en el examen de admisión son: modalidad a la que postula, el colegio de procedencia, la cantidad de veces que ha postulado, el área académica a la que postula y el promedio general en la educación secundaria.

PALABRAS CLAVES: Encuesta de postulantes, Análisis Exploratorio, Log de Odds, Modelo de Regresión Logística Múltiple, Análisis de Influencia

ABSTRACT

IMPORTANT FACTORS WHICH DETERMINATE THE ADMISSION TO NATIONAL UNIVERSITY OF SAN MARCOS-2003

JOSÉ ROMUALDO MOINA FUENTES

DECEMBER 2003

Adviser: **Mg. Olga Lidia Solano Dávila**
Title obtained: **Degree in Statistics**

The present study attempts to determine the most important variables at the time to better explain the success in the Admission exam (for this case it must be understood as successful admission to the National University of San Marcos). In order to accomplish the proposed objectives, the Logistic Regression Model was applied. For this, we used the information of the survey to the postulants of the San Marcos 2003 provided by the Technical Office of Student (OTE) and the information obtained from the file that the postulant fills at the time of registration. This one was provided by the Executive Committee on Admission (CEA). According to the results of this analysis, the variables that most influence in the admission exam are: modality which postulate to, school of precedence, the quantity of times that has postulated, the academic area that postulates and the average general in secondary education.

KEYWORDS: Survey of Postulants, Exploratory Analysis of Log Odds, Multiple Logistic Regression Model, Analysis of Influence

Índice

CAPÍTULO I
INTRODUCCIÓN

Al revisar los componentes del proceso de enseñanza aprendizaje, esto es lo cognitivo, afectivo y psicomotor; se aprecia que son múltiples las variables que intervienen en su desarrollo. Es sabido que en la edad adolescente el alumno está sujeto a una especial sensibilidad para comprender el mundo y para entenderse a sí mismo. En este entorno, las demás personas toman una importancia especial y las propias apreciaciones y valoraciones sobre sí mismo cobran nuevas dimensiones que lo proyectan positiva o negativamente ante el mundo y sus tareas, específicamente en sus rendimientos académicos. (cf. Bloom, 1972, 1977; Rogers, 1989; Carrasco, 1993; Gardner, 1994).

Así, un punto de partida importante de este estudio consiste en entender la Educación como un proceso que intenta conducir al alumno al máximo desarrollo de sus potencialidades tanto intelectuales como afectivas y valorativas. En esta línea, el foco del presente estudio consistió en fijar su análisis en algunas variables alterables que pudieran influir en el desarrollo de los rendimientos académicos, y posteriormente en el éxito en el examen de admisión, en el ámbito de la educación secundaria, su preparación pre-universitaria y sus características generales y familiares.

Por esto podemos decir, en primer lugar, que el paso del colegio a la universidad está marcado por la preparación pre-universitaria la que tiene por finalidad capacitar al

postulante para rendir satisfactoriamente el examen de admisión. Este último es el que determina el ingreso o no a esta casa de estudios. No obstante, el postulante trae consigo otros antecedentes las cuales pudieran ser o son las causantes del éxito o fracaso a la hora de rendir el examen por la cual pasa cada uno de los que aspiran ingresar a esta universidad.

Estos antecedentes serán medidos, en lo posible, por las diferentes variables disponibles que se tiene para este estudio. Esta información fue acumulada de diferentes fuentes: 1.- De la encuesta aplicada a los postulantes a la UNMSM en el examen de admisión 2003[1]; 2.- De la ficha que el postulante llena al inscribirse para el examen de admisión[2].

Para el factor escolar se tomara en cuenta el rendimiento durante su educación secundaria (de esta medida solo tomaremos en cuenta su promedio global) y las características de su educación secundaria (tipo de colegio, etc.). Para el factor demográfico se tomará en cuenta el sexo, edad, etc. así como también características familiares, de vivienda, etc. y por último, para el factor nivel socioeconómico, se tomará como indicador en gasto per cápita[3] y el ingreso familiar. Estos factores de hecho tendrán, unos más que otros, influencia en la determinación del ingreso a la universidad.

Se intenta pues, lograr la mayor y más parsimoniosa explicación del Rendimiento (si ingresa o no a la universidad) expresado en las principales variables tomadas en cuenta en este estudio. Es decir, nos interesa detectar y aumentar intersecciones de la manera

[1] Encuesta aplicada por la Oficina Técnica del Estudiante. ***OTE-UNMSM***.
[2] Información recogida por la Comisión Ejecutiva de Admisión-***CEA-UNMSM***.
[3] Elaborado por la Oficina Técnica del Estudiante. Elaborado bajo una técnica de regresión aplicado a la encuesta del postulante 2003 en base a la ***Encuesta de Hogares del INEI (ENAHO)***.

más simple y menos complicada entre las variables o índices que dan cuenta de una mayor influencia en las posibilidades de ingresar a la UNMSM en el año 2003.

En resumen, para esta investigación se busca establecer en forma sistemática responder a las interrogantes relacionadas con los factores que determina el ingreso del postulante. Claro está que primero se estudiará y analizará dichos factores.

1.1 FORMULACIÓN DEL PROBLEMA.

Al momento de dar el examen de admisión, el postulante trae consigo antecedentes individuales, familiares y sociales las cuales, como se mencionó anteriormente, tienen un cierto grado de influencia sobre el éxito o fracaso en cuanto al ingreso a esta universidad. Ahora bien, es muy importante identificarlos, explicarlos y poder establecer la magnitud de influencia las cuales afectan o determinan el ingreso del postulante a la universidad.

Es posible encontrar algunas investigaciones referentes al tema, tales como la que hizo la Universidad Nacional de Ingeniería[4], en las cuales explican y determinan estos factores (en esta investigación solo se tomó en cuenta algunos factores).

En el Perú son muy escasos y poco sistemáticos los estudios que hayan tratado de identificarlos, menos aún en la Universidad Nacional Mayor de San Marcos.

Es por ello que se pretende responder algunas interrogantes relacionadas con los postulantes tales como: ¿son las características de la educación secundaria tales como el rendimiento, tipo de colegio, etc. las que influyen en su ingreso a la universidad? O ¿son

[4] ORDÓÑEZ MERCADO A. ***Factores Importantes en el Examen de Admisión-UNI***. Tecnia, Vol 8 N°1 1998.

las características familiares, demográficas (sexo, edad, etc.), nivel socioeconómico las que influyen más? O en todo caso ¿cuáles de estas combinaciones son las más determinantes en aumentar las posibilidades de ingresar a la universidad? Intentaremos responder a estas interrogantes.

1.2 JUSTIFICACIÓN O IMPORTANCIA DEL ESTUDIO

Los problemas que afectan el éxito en el examen de admisión tienen consecuencia en el postulante y desde luego en su familia, pues en la mayoría de casos es esta la que solventa los gastos que amerita una preparación pre universitaria y desde luego la inscripción del postulante. Más aún que en esta universidad la convocatoria a un examen de admisión se hace cada año[5]. Esto implica más gastos en la preparación pre-universitaria, una nueva inscripción al examen de admisión y algo muy importante, la pérdida de un año en la cual el alumno ya hubiera reducido su periodo universitario, un año más tendrá que esperar en su afán de conseguir un título profesional.

Por esto, una investigación de este tipo pretende conocer, identificar y medir los principales problemas u obstáculos que el postulante trae consigo a la hora de rendir el examen de admisión que posteriormente decae en el éxito o fracaso para ingresar a esta universidad. Y por supuesto, se pretende conocer cuál es el grado en que afecta cada una de las variables consideradas en este estudio. Por ello, la adecuada intersección (descriptiva o experimental) entre educación secundaria + familia + características personales + comunidad, dará mayor influencia en las posibilidades de ingreso a esta universidad.

[5] A partir del 2004 la universidad convocará dos fechas para el examen de admisión.

Por último, esta investigación deberá servir como marco referencial a partir de la cual se desarrollen nuevos estudios que profundicen en algunos aspectos que por el diseño mismo de la investigación no está siendo considerado.

1.3 OBJETIVO

1.3.1 Objetivo general.

Identificar y establecer cuáles de los factores (variables) que mejor describen y explican el ingreso del postulante a la hora de rendir el examen de admisión en la U.N.M.S.M en el año 2003.

1.3.2 Objetivos específicos.

- Determinar y Comprender los niveles de relación de algunas variables de los factores Demográficos y de Educación Secundaria con el éxito o fracaso a la hora de rendir el examen de admisión.
- Identificar los factores demográficos, escolares y familiares más importantes del postulante a la hora de rendir su examen de admisión.
- Determinar cuál o cuáles de las variables es o son las que mayor capacidad predictiva tiene o tienen sobre la posibilidad de ingresar de los postulantes.

1.4 MARCO TEÓRICO

A continuación se desarrollan los fundamentos teóricos y empíricos de las variables consideradas para este trabajo, es decir, se expone las definiciones de las variables tomadas en cuenta al tratar de explicar las posibilidades de ingreso del postulante.

1.4.1 Rendimiento en el Examen de Admisión

Expresada en este estudio como el éxito (ingreso) o fracaso (no ingreso) a la hora de dar el examen de admisión. Es pues para algunos autores, la noción relativa a que cuando se entregan a todos los alumnos las más apropiadas condiciones o ambientes de aprendizaje, las que son capaces de alcanzar un alto nivel de dominio.

El rendimiento académico es una medida de las capacidades respondientes o indicativas que manifiestan, en forma estimativa, lo que una persona ha aprendido como consecuencia de un proceso de instrucción o formación (Pizarro, 1985). Este mismo autor define, desde la perspectiva del alumno, como la capacidad respondiente de este frente a estímulos educativos, susceptibles de ser interpretados según objetivos o propósitos educativos pre-establecidos.

Este tipo de rendimiento académico puede ser entendido en relación con un grupo social que fija los niveles mínimos de aprobación ante un determinado cúmulo de conocimientos o aptitudes (Carrasco, 1985).

1.4.2 Condiciones de la Educación Secundaria

Son las variables que caracterizan individualmente a los postulantes en cuanto a su educación secundaria. La educación secundaria es una etapa en la cual los estudiantes adquieren un gran porcentaje de sus conocimientos ya sean académicamente y socialmente que de cierta forma tendrá un impacto en su vida académica y social. En esta etapa juegan muchas variables las que los caracterizan, pero mencionaremos las disponibles para este estudio:

Tipo de colegio.- Es el tipo de gestión que tuvo el colegio en donde el postulante terminó su educación secundaria. De las diferentes clases de gestión, estas se clasificaron solo en dos.

Promedio general de notas en su educación secundaria.- Es el rendimiento académico que tuvo el postulante en la secundaria.

1.4.3 Condiciones Pre Universitarias

Son las variables que caracterizan individualmente a los postulantes en cuanto a su preparación pre-universitaria. Estas son:

Tipo de preparación.- Se refiere a la manera que el postulante se preparó durante un determinado tiempo para postular al examen de admisión.

Tiempo transcurrido desde que termino su educación secundaria.- Es el tiempo en años desde que el postulante termina su educación secundaria hasta el momento en que da el examen de admisión.

Número de veces que postuló a San Marcos anteriormente.- Es la cantidad de veces que el postulante ha intentado ingresar anteriormente a la universidad sin mayores éxitos.

Área a la que postula.- Área determinada según el tipo de especialidad a la que aspira el postulante.

Modalidad de postulación.- Clasificado según la modalidad por la que postula.

1.4.4 Aptitudes Propias del Postulante

Refiere a las capacidades del postulante para enfrentar las exigencias del examen de admisión, capacidades que las adquieren durante su vida escolar y familiar. Para ser cuantificadas estas variables requiere de un análisis especial por las cuales no serán incluidas en esta monografía.

1.4.5 Características Generales y Familiares

Son las variables que caracterizan individualmente a los postulantes en sí y en cuanto a su entorno familiar. Estas son:

Edad del postulante.- Medida en años cumplidos hasta el momento de dar el examen de admisión.

Sexo del postulante.- Distinción en cuanto a la condición orgánica que hace la diferencia del hombre y la mujer.

Situación familiar.- Que se refiere a la condición del postulante en cuanto a si vive con sus padres o no.

Número de personas que viven en el hogar.- cantidad de personas bajo el mismo techo en donde vive el postulante.

Nivel educativo del padre.- Referida al último nivel de estudios alcanzado por el padre del postulante.

1.4.6 Características Económicas

Son las variables que caracterizan individualmente a los postulantes en cuanto a su condición económica. Estas son:

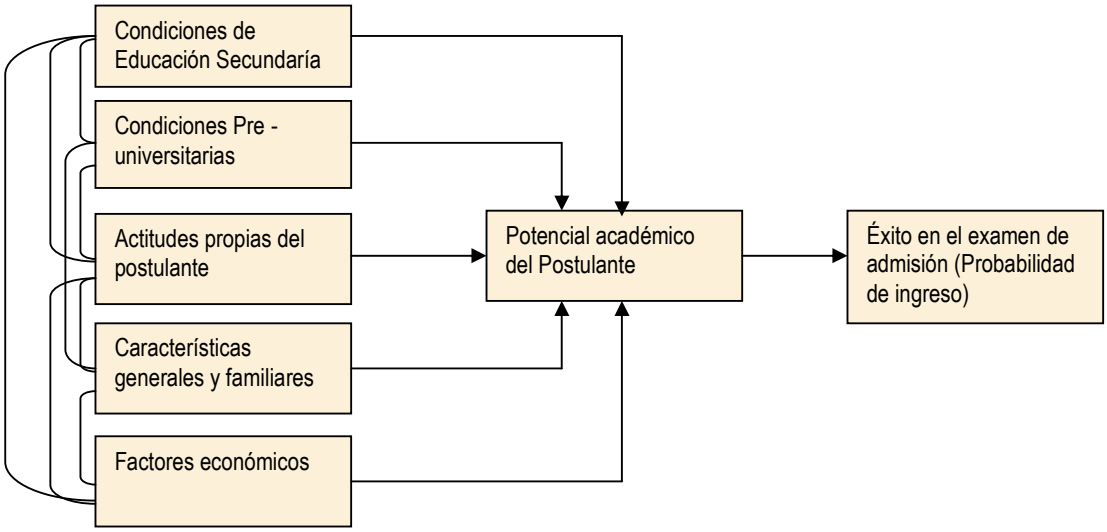
El gasto per cápita del postulante.- Gasto promedio *estimado* de los hogares por persona. Se refiere al gasto de consumo final, es decir a las compras de bienes y servicios que los hogares financian con su ingreso.

Ingreso mensual familiar.- Es la suma de los ingresos de todos los que conforman la familia.

Situación laboral del postulante.- Condición laboral actual.

Estos grupos de variables (factores) pueden ser esquematizados de la siguiente manera:

Gráfico I.1
Esquema de la interacción de las variables que caracterizan al postulante



Cabe resaltar que todas estas variables se obtuvieron del cuestionario aplicado en la encuesta y la ficha del postulante que el postulante llenó a la hora de inscribirse al examen de admisión[6]. Se observará que las posibles variables explicativas bajo la estructura descrita en la figura anterior que pueden tener un cierto grado de influencia sobre la probabilidad de ingreso en el examen de admisión.

[6] Información proporcionada por la ***CEA-UNMSM***. Esta se empató con la información recolectada en la encuesta mediante el código de cada postulante.

CAPÍTULO II
METODOLOGÍA

Teniendo en cuenta la formulación del problema y su justificación o importancia de este estudio comenzamos a plantear el diseño metodológico.

2.1 DISEÑO METODOLÓGICO

Al momento de establecer las variables, se centró la mirada en los factores del alumno, de su educación secundaria y la familia, estimados como los de mayor importancia. Así se seleccionaron variables alterables que dieran cuenta en mejor medida de resultados que influyesen directamente en el ingreso a la universidad.

Se exponen a continuación aspectos relativos al tipo de investigación realizada y al diseño muestral. La descripción de las técnicas aplicada se describirá en la siguiente sección.

2.1.1 Tipo de estudio

El estudio realizado es cuantitativo, correlacional, estimativo y multivariado por los factores (o variables independientes). Para los efectos de este trabajo es importante identificar las variables que intervienen en la elaboración de nuestro estudio, estas ya fueron definidas en la sección anterior pero que posteriormente se describirán cada una de ellas.

2.1.2 Diseño muestral

Población Objetivo.- La población objetivo estuvo compuesta por todos los postulantes que lograron inscribirse para el examen de admisión de la UNMSM - 2003.

Población Muestreada.- La población muestreada estuvo compuesta por los postulantes que lograron inscribirse para el examen de admisión de la UNMSM – 2003 en las fechas programadas.

Unidad de Análisis.- La unidad de análisis estuvo constituida por un postulante que logró inscribirse al examen de admisión en la fecha programada que comprendió desde 13 de enero y el 25 de febrero del 2003.

Tipo de Muestreo.- Muestreo sistemático y probabilístico.

Proceso de muestreo.- La muestra que se obtuvo fue de tipo probabilístico, sistemático el cual consiste en seleccionar un primer elemento al azar y luego cada k-término de una lista o dejar pasar a k-individuos y preguntar al que sigue y así sucesivamente.

El muestreo sistemático es aplicable cuando la población esta ordenada siguiendo una tendencia conocida, la cual asegura una cobertura de unidades de todos los tipos. Y teniendo en cuenta que la CEA cuenta con un cronograma de inscripciones ordenado por modalidades (prueba general, primeros puestos, deportistas calificados, etc) y en el caso de la modalidad de prueba general estaba ordenado alfabéticamente, lo cual aseguró una cobertura total de los postulantes en todas sus modalidades.

Tamaño de muestra.- Para calcular el tamaño de muestra era necesario saber el tamaño de la población, en este caso el número de postulantes a la UNMSM 2003 era desconocido. Para solucionar este inconveniente se hizo una proyección[7] en base a la

[7] La proyección se hizo utilizando El Método biparamétrico de Holt, debido a que la serie de postulantes presentaba una tendencia determinística de tipo lineal. OGPL-UNMSM.

información histórica de la cantidad de postulantes de la OGPL-UNMSM. Esta proyección fue de 51 435 postulantes para el año 2003.

Además para el cálculo del tamaño de muestra, teniendo en cuenta el diseño muestral propuesto, requiere conocer ciertos parámetros (σ^2, ρ)[8] de las variables de interés, debido a que estos parámetros se estiman usando información previa, por ejemplo datos de los postulantes al examen de admisión del año 2002. Sin embargo, no fue posible obtener dicha información, por lo que se calculó el tamaño de la muestra usando una alternativa en estos casos, es decir, utilizando la fórmula para el tamaño de muestra del MIA (muestreo irrestricto aleatorio), que no requiere conocer los parámetros antes mencionados. Esta fórmula es:

$$n = \frac{Np(1-p)}{(N-1)D + p(p-1)} \qquad II.1$$

Dónde:

N: 51 435 (Número postulantes proyectados).

n: Tamaño de muestra.

p: Valor aproximado de la proporción de alumnos que ingresan a la universidad.

$$D = \left(E / Z\right)^2 \qquad II.2$$

E: Margen de error

Z: Valor de la abscisa en la distribución normal que garantiza una confianza prefijada.

Además se tuvo las siguientes consideraciones:

- Nivel de confianza para los resultados del 95% (Z=1.96),

[8] σ^2: Varianza poblacional.
ρ : Es el llamado coeficiente de correlación intraclase que mide la correlación entre las unidades de una misma muestra sistemática y es calculada sobre todos los posibles pares de diferentes unidades dentro de cada muestra sistemática

- p y q =0.5; y
- Un error relativo del 2% (E=0.02).

Para una población proyectada de 51 435 postulantes, con un margen de error de 0.02, se obtiene una muestra de 2 294 postulantes a encuestar.

Pero como era de esperarse esta cifra proyectada no iba a ser exacta, la cantidad total de postulantes para el año 2003 fue de 46 585[9] lo que nos dio un tamaño de nuestra de 2117 postulantes. Ahora, esta cantidad ha sido reducida a 1 588 casos por los siguientes motivos: presencia de muchos casos perdidas "missing" o datos incompletos en algunas variables, especialmente en la que se pedía el promedio general de notas del colegio pues en muchos casos no recordaban sus notas, peor aún, no traían sus certificados de colegio, pues estas ya no eran obligatorias. Debido a estos inconvenientes la muestra fue reducida significativamente.

Para resolver este inconveniente tenemos que probar si esta muestra resultante después de la depuración sigue siendo representativa de la población. Para esto, utilizaremos el Test para proporción poblacional el cual prueba la siguiente hipótesis:

H_0: La diferencia de las proporciones no es significativa ($p_{población}=p_{muestra}$).

Si el p_valor es mayor al nivel de significancia de 0.05, no rechazaremos la hipótesis nula y se concluirá que esta muestra es representativa de la población. Está claro que se desea encontrar un p-valor mayor al nivel de significancia. Este "test" se aplicará a las siguientes variables: Sexo, Tipo de colegio, Número de veces que postulo anteriormente a la universidad y Área académica a la que postula. Estos resultados se obtuvieron de un programa la cual calcula la significancia del "test". Los resultados se presentan en el siguiente cuadro:

[9] Información proporcionada por la CEA-UNMSM.

Cuadro II.1

COMPARACIÓN DE LAS PROPORCIONES ENTRE LA POBLACIÓN Y LA MUESTRA REDUCIDA

(a) Postulante a la UNMSM, según tipo de colegio

	Población según CEA		Muestra reducida		
Colegio	Total	Porcentaje	Total	Porcentaje	p_valor
Estatal	35084	75.31	1177	74.12	p_valor>0.05
Particular	11501	24.69	411	25.88	p_valor>0.05
TOTAL	46585	100.00	1588	100.00	

(b) Postulante a la UNMSM, según sexo

	Población según CEA		Muestra reducida		
Sexo	Total	Porcentaje	Total	Porcentaje	p_valor
Femenino	23664	50.80	766	48.24	p_valor<0.05
Masculino	22921	49.20	822	51.76	p_valor<0.05
TOTAL	46585	100.00	1588	100.00	

(c) Postulante a la UNMSM, según número de veces que postulo anteriormente a San Marcos

	Población según CEA		Muestra reducida		
	Total	Porcentaje	Total	Porcentaje	p_valor
Ninguna	25781	55.34	876	55.16	p_valor>0.05
Una	12835	27.55	445	28.02	p_valor>0.05
Dos	5368	11.52	177	11.15	p_valor>0.05
Tres	1868	4.01	65	4.10	p_valor>0.05
Cuatro a mas	733	1.57	25	1.57	p_valor>0.05
TOTAL	46585	100	1588	100.00	

(d) Área o bloque al cual postula

	Población según CEA		Muestra reducida		
	Total	Porcentaje	Total	Porcentaje	p_valor
Ciencias de la salud	14094	30.25	477	30.04	p_valor>0.05
Humanidades	12112	26.00	443	27.90	p_valor>0.05
Cc. básicas e Ingenierías	10755	23.09	370	23.30	p_valor>0.05
Cc. económico-empresarial	9624	20.66	298	18.76	p_valor>0.05
TOTAL	46585	100	1588	100	

Vemos que solo la variable Sexo presenta una ligera diferencia con la población, pero veremos más adelante que esta variable no es un factor influyente en el rendimiento en el examen de admisión (si ingresa o no a la universidad). Concluimos que la muestra reducida por la eliminación de los datos "missing" de algunas variables es representativa de la población.

2.2 DESCRIPCIÓN DE VARIABLES

A continuación describiremos las variables según su nombre en la base de datos, el tipo, nivel de medición y categorías según sea el caso.

Cuadro II.2

Descripción de las variables que intervienen en el estudio

VARIABLE	NOMBRE EN BD	TIPO	NIVEL DE MEDICIÓN	CATEGORÍAS
1. Éxito en el examen de admisión	IND_INGR	Cualitativa	Nominal	0: No ingresó 1: Si ingresó
2. Colegio de procedencia	COLPRO	Cualitativa	Nominal	0: Estatal 1: Particular
3. Promedio de notas en el colegio	P43	Cuantitativa	Razón	No tiene
4. Modalidad de postulación	MODALIDA	Cualitativa	Nominal	0: Prueba general 1: CEPUSM 2: Primeros puestos 3: Otras modalidades
5. Tipo de preparación pre-universitaria	TIPPRE	Cualitativa	Nominal	0: Solo 1: Grupo de estudio 2: Academia 3: CEPUSM
6. Tiempo desde que termino la secundaria	TIEM_EGR	Cuantitativa continua	Razón	No tiene
7. Cuántas veces postulo anteriormente a la universidad	POSUSM	Cuantitativa discreta		
8. Área académica	ARECOD	Cualitativa	Nominal	0: Ciencias básicas e ingeniería 1: Ciencias de la Salud 2: Ciencias económico-empresarial 3: Humanidades
9. Edad	EDAPOS	Cuantitativa continua	Razón	Según la edad cumplida
10. Sexo	SEXPOS	Cualitativa	Nominal	0: Masculino 1: Femenino
11. Índice de pobreza	IND_POBR	Cuantitativa continua	Razón	No tiene
12. Nivel de estudios del padre	N_PAD_AG	Cualitativa	Ordinal	0: Analfabeto/inicial 1: Primaria 2: Secundaria/carrera corta 3: Superior no universitaria 4: Superior universitaria 5: Post-grado (Maestría, Doctorado)
13. Gasto per cápita	GASTPER	Cuantitativa continua	Razón	Según el gasto estimado

La forma de esta base de datos se muestra en el Anexo 1.

2.3 TÉCNICAS ESTADÍSTICAS USADAS

Para el análisis de los datos, usaremos el Modelo de Regresión Logística, la cual nos proporciona, según nuestro objetivos planteados, las variables o factores que influyen significativamente en los postulantes a la hora de rendir el examen de admisión lo que posteriormente determina su ingreso a la universidad. Este Modelo lo desarrollaremos en el siguiente capítulo.

CAPÍTULO III

MODELO DE REGRESIÓN LOGISTICA

3.1 DEFINICIÓN

Los modelos de regresión, como en el caso lineal, pueden usarse con dos objetivos: 1) predictivo en el que el interés del investigador es predecir lo mejor posible la variable dependiente, usando un conjunto de variables independientes y 2) estimativo en el que el interés se centra en estimar la relación de una o más variables independientes con la variable dependiente. El segundo objetivo es el más frecuente en estudios etiológicos en los que se trata de encontrar factores determinantes de una enfermedad o un proceso. Es el objetivo que sigue este trabajo.

El objetivo de esta técnica estadística es también, expresar la probabilidad de que ocurra un hecho como función de ciertas variables, supongamos que son k $(k > 0)$, que se consideran potencialmente influyentes. La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, llamémosla Y, que puede ser dicotómica o politómica (en este trabajo nos referiremos solamente al primer caso) y una o más variables independientes, llamémoslas X.

Al ser la variable *Y* dicotómica, podrá tomar el valor "*0*" si el hecho no ocurre y "*1*" si el hecho ocurre; el asignar los valores de esta manera o a la inversa es intrascendente, pero es muy importante tener en cuenta la forma en que se ha hecho llegado el momento de interpretar los resultados. Las variables independientes (también llamadas explicativas) pueden ser de cualquier naturaleza: cualitativas o cuantitativas. La probabilidad de que $Y=1$ se denotará por *p*.

Como hemos visto dada una variable dependiente binaria *Y*, y según el modelo de regresión lineal $Y=\beta_0+\beta_1X_1+...+\beta_kX_k+u$ presenta serias inconsistencias. Para evitarlas, se han desarrollado modelos no lineales, los cuales tratan de resolver el problema de que si se utilizara una regresión lineal y al estimar sus parámetros, estaríamos ajustando una recta la cual al predecir nuevos valores de *Y* pueden proporcionar valores mayores que *1* o menores que *0* (lo cual está en contradicción con la definición de probabilidad).
La idea consiste en utilizar un modelo de la forma:

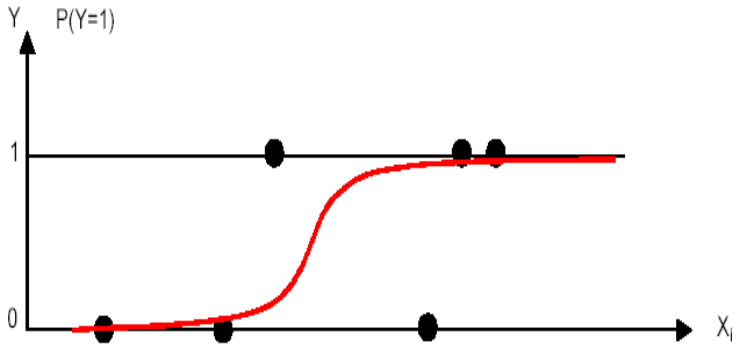
$$Y = f(\beta_0 + \beta_1 X_1 + ... + \beta_k X_k) + u \qquad \textit{III.1.1}$$

Donde f es una función real que depende de la expresión lineal $\beta_1+\beta_2X_2+...+\beta_kX_k$.
Con el nuevo modelo, y razonando de forma similar al caso del modelo lineal, se cumplirá:

$$E[Y]=P(Y=1)=f(\beta_0+\beta_1X_1+...+\beta_kX_k) \qquad \textit{III.1.2}$$

Ahora bien, ¿qué tipo de función f estamos buscando?, obviamente *f* deberá ser distinta de la función identidad. Además necesitamos una función que esté acotada por los valores *0* y *1* (puesto que su valor coincidirá con el de una probabilidad). En el gráfico siguiente se muestra una idea intuitiva del tipo de función que buscamos para construir nuestro nuevo modelo:

Gráfico III.1
Tipo de función que se busca encontrar.



Pues bien, de entre las funciones *f* que presentan una forma similar a la de la gráfica, hay dos que son las que se utilizan con mayor frecuencia: la función logística (y que da lugar a los modelos Logit), y la función de distribución de una normal estándar (asociada a los modelos Probit).

Como se mencionó, una posible solución a las inconsistencias que presentaba el modelo de probabilidad lineal -para explicar el comportamiento de una variable dependiente binaria- es usar un modelo Logit de la forma:

$$Y = f(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k) + u \qquad \textit{III.1.3}$$

Donde *f* es la función logística, esto es:

$$f(z) = \frac{\exp(z)}{1 + \exp(z)} \qquad \textit{III.1.4}$$

Por tanto, tendremos que:

$$E[Y] = P(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)} \qquad \textit{III.1.5}$$

3.2 MODELOS DE REGRESIÓN LOGÍSTICA SIMPLE

3.2.1 Introducción

Es importante entender que en la regresión logística el objetivo es el mismo que cualquier técnica estadística de construcción de modelo, esto es, el mejor y más parsimonio ajuste que describa la relación entre la variable de salida o respuesta y el grupo de variables independientes o predictores.

La diferencia con el modelo de regresión lineal es que la variable de salida en el modelo de regresión logística es binaria o dicotómica.

En cualquier problema de regresión la cantidad clave es el valor medio de la variable de salida, dado el valor de la variable independiente. Esta cantidad se llama 'la media condicional' y será expresado como '*E(Y / x)*' donde *Y* denota la variable de salida y *x* denota el valor de la variable independiente. En la regresión lineal asumimos que esta media puede ser expresada como una ecuación lineal en *x* (o algunas transformaciones de *x* o *Y*), tal como

$$E(Y/x) = \beta_0 + \beta_1 x \qquad \textit{III.2.1}$$

Esta expresión implica que es posible para E(Y/x) tomar cualquier valor de x en el rango entre $-\infty$ y $+\infty$.

Hay dos razones primarias para elegir la distribución logística. Primero, desde el punto de vista matemático, es una función extremadamente flexible y fácilmente de usar, y segundo, se presta a una interpretación clínico significativo. Una interpretación más detallada se presenta más adelante.

Para simplificar la notación, usaremos la cantidad $\pi(x) = E(Y/x)$ para representar la media condicional de *Y* dado *x* cuando se usa la distribución logística. La forma específica del modelo de la regresión logística que usaremos es:

$$P(y=1) = \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \qquad \textit{III.2.2}$$

Una transformación de *π(x)* que es fundamental en el estudio de la regresión logística es la transformación logit. Esta transformación se define en términos de *π(x)*, como:

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \qquad \textit{III.2.3}$$

La importancia de esta transformación es que *g(x)* tiene muchas de las propiedades deseables del modelo de regresión lineal. La función logit, *g(x)*, es lineal en sus parámetros, puede ser continuo, y puede extenderse de $-\infty$ a $+\infty$, dependiendo del rango de *x*.

3.2.2 Ajuste del Modelo de Regresión Logística

Supongamos que tenemos una muestra de "n" observaciones independientes del par (x_i,y_i), $i=1, 2, ..., n$, donde y_i denota el valor de la variable de salida dicotómica y x_i es el valor de la variable independiente para el *i-ésimo* sujeto. Además, asumimos que la variable de salida se ha codificado como *0* y *1*, representando la ausencia y presencia de una característica, respectivamente. Para ajustar el modelo de regresión logística en la ecuación *(III.2.2)* para un grupo de datos se requiere que estimemos los valores de β_0 y β_1, los parámetros desconocidos.

En la regresión lineal, el método usado más a menudo para estimar los parámetros desconocidos es el de mínimos cuadrados. Bajo la suposición de una regresión lineal el

método de mínimos cuadrados produce estimadores con un número de propiedades estadísticas deseables. Desafortunadamente, cuando se aplica el método de mínimos cuadrados a un modelo con una variable de salida dicotómica los estimadores no tienen estas mismas propiedades.

El método general de estimación que conduce a la función de mínimos cuadrados bajo el modelo de la regresión lineal se llama máxima verosimilitud. Este método proporcionará un fundamento para acercar la estimación con el modelo de regresión logística. En un sentido más general el método de máxima verosimilitud rinde valores para los parámetros desconocidos con la máxima probabilidad de obtener el grupo de datos observados. Para aplicar este método primero debemos construir una función, llamada la función de verosimilitud.

Una manera conveniente de expresar la contribución de la función de verosimilitud para el par (x_i,y_i) es con la expresión

$$\pi(xi)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i} \qquad \textit{III.2.4}$$

Dado que se asume que las observaciones son independientes, la función de verosimilitud se obtiene como el producto de los términos dados en la expresión *(III.2.4)* como sigue:

$$l(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{1-y_i} \qquad \textit{III.2.5}$$

El principio de máxima verosimilitud establece que los valores de β son los que maximizan la expresión de la ecuación *(III.2.5).* Sin embargo, esto es matemáticamente fácil de trabajarlo con el logaritmo de la ecuación *(III.2.6)*. Esta expresión, log de verosimilitud, se define como

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

Para encontrar los valores de β que maximicen $L(\beta)$ diferenciaremos $L(\beta)$ con respecto a β_0 y β_1, y luego igualar a cero. Estas ecuaciones, conocidas como ecuación de verosimilitud, son:

$$\sum [y_i - \pi(x_i)] = 0 \qquad \textit{III.2.7}$$

y

$$\sum x_i [y_i - \pi(x_i)] = 0 \qquad \textit{III.2.8}$$

Los valores de β dado por la solución de las ecuaciones *(III.2.7)* y *(III.2.8)* se llama estimación de máxima verosimilitud y lo denotaremos como $\hat{\beta}$. Obviamente no es fácil de estimar estos valores, para eso se utiliza métodos numéricos, como por ejemplo, el método de Newton Raphson.

3.2.3 Prueba de Significancia para los Coeficientes

Después de estimar los coeficientes, nuestra primera impresión al modelo ajustado comúnmente concierne a la evaluación de la significancia de las variables en el modelo. Esto generalmente involucra la formulación de pruebas de hipótesis para determinar si la variable independiente es significativamente relacionada con la variable respuesta.

Un acercamiento para la evaluación de la significancia de los coeficientes es comparando los valores observados de la variable respuesta predichos por los siguientes dos modelos: con y sin la variable de respuesta. Si los valores predichos con la variable respuesta en el modelo son mejores, o más exacto en el mismo sentido, que cuando la variable no está en el modelo. Entonces sentimos que la variable es ‘significativa’.

El método general para evaluar la significancia de las variables se ilustra fácilmente en el modelo de regresión lineal, y esto motiva usarlo para aprovechar su uso en la regresión logística

En la regresión lineal, la evaluación de la significancia del coeficiente *(β_1)* es una referencia para aprovechar a formar una *tabla de análisis de varianza*. Esta tabla particiona la suma total de cuadrados respecto a su media en dos partes: (1) la suma de cuadrados respecto a la línea de regresión *SSE*, (o *suma de cuadrados residual*), y (2) la suma de cuadrados de los valores predichos, basados en el modelo de regresión, acerca de la media de la variable dependiente *SSR* (o *suma de cuadrados debido a la regresión*). El *SSE* se define de la siguiente manera:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textit{III.2.9}$$

Bajo el modelo que no contiene la variable independiente, solo con β0, y $\beta 0 = \bar{y}$, la media de la variable respuesta. En este caso, $\hat{y}i = \bar{y}$ y SSE es igual a la varianza total. Cuando incluimos la variable independiente en el modelo cualquier disminución en SSE será a causa de que el coeficiente de la variable independiente no es cero. El cambio en el valor de SSE es debido a la fuente de variabilidad de la regresión, denotado como *SSR*, eso es,

$$SSR = \left[\sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 \right] - \left[\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \right] \qquad \textit{III.2.10}$$

En la regresión lineal, interesa el tamaño de *SSR*. Un valor grande sugiere que la variable independiente es importante, mientras que un valor pequeño sugiere que la variable independiente no es útil en predicción de la respuesta.

Este principio es la misma con la regresión logística: *comparar los valores observados de la variable respuesta con los valores predichos obtenidos de los modelos con y sin la variable independiente*. En la regresión logística, la comparación de los valores observados y predichos se basa en el logaritmo de la función verosimilitud definida en la ecuación *(III.2.6)*.

El test para evaluar la hipótesis de que si el coeficiente de la variable de respuesta es significativa es el test de razón de verosimilitud, esta es:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \qquad \textit{III.2.11}$$

Donde $\hat{\pi}i = \hat{\pi}(x_i)$.

La estadística *D* en la ecuación *(III.2.11)* se llama *deviance*. Esta estadística se puede expresar de la siguiente manera: del modelo ajustado

$$D = -2ln(verosimilitud\ del\ modelo\ ajustado) \qquad \textit{III.2.12}$$

Para propósitos de evaluación de significancia de una variable independiente comparamos el valor de *D* con y sin la variable independiente. El cambio en *D* debido a la inclusión de la variable independiente en el modelo se obtiene como:

$$G = D(modelo\ sin\ la\ variable) - D(modelo\ con\ la\ variable) \qquad \textit{III.2.13}$$

Esta estadística juega el mismo rol en la regresión logística como en el numerador del test F parcial de la regresión lineal. La estadística *G* se puede expresar de la siguiente manera:

$$G = -2\ln\left[\frac{(verosimilitud\_\sin\_la\_\mathrm{var}iable)}{(verosimilitud\_con\_la\_\mathrm{var}iable)}\right] \qquad \textit{III.2.14}$$

Para el caso específico de una variable independiente, es fácil mostrar que cuando la variable no está en el modelo, el estimador máximo verosimilitud de β_0 es $ln(n_1/n_0)$ donde $n_1 = \sum y_i$, $n_0 = \sum (1 - y_i)$ y el valor predicho es constante, n_1/n. En este caso el valor de G es:

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}_i^{\,y_i} \left(1 - \hat{\pi}_i\right)^{(1-y_i)}} \right] \qquad \textit{III.2.15}$$

o

$$G = 2\left\{ \sum_{i=1}^{n} \left[y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i) \right] - \left[n_i \ln(n_i) + n_0 \ln(n_0) - n \ln(n) \right] \right\} \qquad \textit{III.2.16}$$

Bajo la hipótesis que β_1 es igual a cero, la estadística G sigue una distribución Chi-cuadrado con un grado de libertad.

Se sugieren otros dos test estadísticamente equivalentes, estos son el *test de Wald* y el *test Score*. Las suposiciones que necesitan para estos dos test son los mismos del *test de razón de verosimilitud*.

El *test de Wald* se obtiene comparando la estimación máxima verosimilitud del parámetro $\hat{\beta}_1$, a la estimación de su error de estándar. El cociente que resulta, bajo la hipótesis de que $\beta_1 = 0$, sigue una distribución normal estándar. La estadística de Wald para el modelo de regresión logística es

$$W = \frac{\hat{\beta}_1}{S\hat{E}(\hat{\beta}_1)} \qquad \textit{III.2.17}$$

Que sigue una distribución normal estándar.

3.2.4 Estimación por Intervalo de Confianza

La base para la construcción de la estimación por intervalos es la misma teoría estadística usada para formular los test de significancia para el modelo. En particular, el estimador por intervalo de confianza para β_1 y el intercepto se basa en su respectiva prueba de Wald. El intervalo de confianza al $100(1-\alpha)\%$ para β_1 es

$$\hat{\beta}_1 \pm z_{1-\alpha/2} S\hat{E}(\hat{\beta}_1), \qquad \textit{III.2.18}$$

Y para el intercepto es

$$\hat{\beta}_0 \pm z_{1-\alpha/2} S\hat{E}(\hat{\beta}_0), \qquad \textit{III.2.19}$$

Donde $z_1\text{-}\alpha/2$ es bajo el $100(1-\alpha)\%$ de una distribución normal estándar y $S\hat{E}(.)$ representa el estimador del error estándar del parámetro según el modelo respectivo.

La logit es la parte lineal del modelo de regresión logística, y como tal, es más como la línea ajustada a un modelo de regresión lineal. La estimación de la logit es

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \textit{III.2.20}$$

El estimador de la varianza del estimador de la logit requiere obtener la varianza de la suma. En este caso es:

$$\hat{Var}\left[\hat{g}(x)\right] = \hat{Var}(\hat{\beta}_0) + x^2 \hat{Var}(\beta_1) + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \qquad \textit{III.2.21}$$

En general la varianza de la suma es igual a la suma de la varianza de cada término y la covarianza de cada posible par de términos formado de las componentes de la suma. El intervalo de confianza al $100(1-\alpha)\%$ según Wald para la logit es

$$\hat{g}(x) \pm z_{1-\alpha/2} S\hat{E}\left[\hat{g}(x)\right], \qquad \textit{III.2.22}$$

Donde $S\hat{E}\left[\hat{g}(x)\right]$ es el la raíz cuadrada del estimador de la varianza en la ecuación *(III.2.21)*.

3.2.5 Prueba de Hosmer - Lemeshow

Evalúa la bondad del modelo construyendo una tabla de contingencia, divide la muestra en aproximadamente 10 grupos iguales a partir de las probabilidades estimadas, para comparar las frecuencias observadas con las esperadas en cada uno de estos grupos a través de la prueba χ^2 con *j-2* grados de libertad, en donde *j* es el número de grupos formados.

Se calcula los deciles de las probabilidades estimadas p_i; $i = 1,.....,n$ y $D_i,.....,D_9$ que son los deciles observados divididos en 10 grupos dados por:

$$A_j = \left\{ i \in \{1,...,n\} / \hat{p}_i \in \left\lfloor D_{j-1}, D_j \right\rfloor \right\}, j = 1,...,10$$

Dónde: $D_0 = 0$, $D_{10} = 1$

Sea:

n_j = número de casos en A_j ; $j = 1,...10$

o_j= número de $y_i = 1$ en A_j; $j = 1,...10$

$$\overline{p}_j = \frac{1}{n_j} \sum_{i \in A_j} \hat{p}_i \text{ ; j = 1,... ,10} \qquad \textit{III.2.23}$$

Las hipótesis a contrastar son:

H_0: El modelo es adecuado

H_1: El modelo no es adecuado

Estadístico de prueba es:

$$\chi^2 = \sum_{j=1}^{10} \frac{\left(o_j - n_j \bar{p}_j\right)^2}{\bar{p}_j n_j \left(1 - \bar{p}_j\right)} \sim \chi^2{}_{\alpha, j-2} \qquad \textit{III.2.24}$$

Decisión: si $X^2 \geq \chi^2{}_{\alpha, j-2}$ rechazamos H_0 y concluimos que el modelo no es adecuado a un nivel de significancia α.

3.3 REGRESIÓN LOGÍSTICA MÚLTIPLE

3.3.1 Modelo de Regresión Logística Múltiple

Consideremos una colección de *p* variables independientes expresado por el vector $x' = (x_1, x_2, ..., x_p)$. Por el momento asumiremos que cada una de estas variables es por lo menos una escala intervalar. La probabilidad condicional esta denotado por $P(Y=1|x) = \pi(x)$. El logit del modelo de regresión logística múltiple se da por la siguiente ecuación:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p, \qquad \textit{III.3.1}$$

En este caso el modelo de regresión logística es

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}, \qquad \textit{III.3.2}$$

Si algunas de las variables independientes son discretas o de escala nominal tal como la raza, sexo, o grupos, estos son inapropiados incluirlos en el modelo como si fueran variables de escala intervalar. Los números usados para representar los varios niveles de estas variables de escala nominal son solamente identificadores, y no tienen ninguna significación numérica. En esta situación el método de selección es usar una colección de variables denominadas '*variables dummy*'.

En general, si una variable de escala nominal tiene *k* posibles valores, entonces será necesario utilizar *k-1* variables dummy. Para ilustrar esta noción supongamos que la *j*-

ésima variable independiente x_j tiene k_j niveles. Las k_j*-1* variables dummy se denotará como D_{jl} y los coeficientes para estas variables dummy serán denotadas como β_{jl}, $l=1,2,....,k_j-1$. Así, la logit para el modelo con p variables y con la *j-ésima* variable discreta será

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p \qquad \textit{III.3.3}$$

3.3.2 Ajuste del Modelo de Regresión Logística Múltiple

Asumimos que tenemos una muestra de n observaciones independientes (x_i,y_i), $i=1,2,...,n$. Como en el caso univariante, ajustar el modelo requiere que obtengamos estimadores del vector $\beta'=(\beta_0, \beta_1,..., \beta_p)$. El método de estimación en el caso multivariante será la misma como en la situación univariante (máxima verosimilitud). La función de verosimilitud es casi idéntica dado en la ecuación *(III.2.5)* con el único cambio que $\pi(x)$ es ahora definida como en la ecuación *(III.3.2)*. Allí serán $p+1$ ecuaciones de verosimilitud que se obtiene por la diferenciación de la función logaritmo de verosimilitud con respecto a los $p+1$ coeficientes. Las ecuaciones de verosimilitud que resulta pueden ser expresadas como sigue:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \qquad \textit{III.3.4}$$

Y

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0 \qquad \textit{III.3.5}$$

Para $j=1,2,...,p$.

Sea $\hat{\beta}$ que denota la solución de estas ecuaciones. Así, los valores ajustados para el modelo de regresión logística múltiple es $\hat{\pi}(x_i)$, el valor de la expresión en la ecuación *(III.3.2)* calculado usando $\hat{\beta}$, y x_i.

El método de estimación de la varianza y covarianza de los coeficientes estimados siguen desde la teoría bien desarrollada de la estimación máxima verosimilitud. Esta teoría establece que los estimadores son obtenidos de la matriz de segunda derivada parcial de la función de logaritmo de verosimilitud. Esta derivación parcial tiene la siguiente forma general:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j{}^2} = -\sum_{i=1}^{n} x^2{}_{ij}\pi_i(1-\pi_i) \qquad \textit{III.3.6}$$

Y

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij}x_{il}\pi_i(1-\pi_i) \qquad \textit{III.3.7}$$

Para *j, l=0,1,2,...,p* donde π_i representa $\pi(x_i)$. Dado la matriz *(p+1)*(p+1)* contiene los negativos de los términos dados en la ecuación *(III.3.6)* y *(III.3.7)* se denota como *I(β)*. Esta matriz se llama matriz de información observada. La varianza y covarianza de los coeficientes estimados se obtienen de la inversa de la matriz el cual se denota como *Var(β)=I-1(β)*. Usaremos la notación $Var(\beta_j)$ que indica el *j-ésimo* elemento de la diagonal de esta matriz, que es la varianza de $\hat{\beta}_j$, y $Cov(\beta_j, \beta_l)$ para indicar un elemento arbitrario de la diagonal, que es la covarianza de $\hat{\beta}_j$ y $\hat{\beta}_l$. Los estimadores de la varianza y covarianza, que denotaremos por $\hat{Var}(\hat{\beta})$, se obtiene al evaluar *Var(β)* en $\hat{\beta}$. Usaremos $\hat{Var}(\hat{\beta}_j)$ y $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_l)$, *j, l=0,1,2,...,p* para denotar los valores de esta matriz. Para más adelante tendremos la ocasión de usar solamente la estimación del error estándar de los coeficientes estimados, que denotaremos como

$$\hat{SE}(\hat{\beta}_j) = \left[\hat{Var}(\hat{\beta}_j)\right]^{1/2} \qquad \textit{III.3.8}$$

Para *j=0,1,2,...,p*. Usaremos esta notación en el desarrollo de los métodos para los test de los coeficientes y la estimación por intervalo de confianza.

3.3.3 Test para la Significancia del Modelo

Una vez que tenemos ajustado el modelo de regresión logística múltiple, comenzaremos el proceso de evaluación. Como en el caso univariante, el primer paso en este proceso generalmente es fijar la significancia de las variables en el modelo. El *test de razón de verosimilitud* para la significación total de los *p* coeficientes para las variables independientes en el modelo se realiza exactamente de la misma manera como en el caso univariante. El "test" se basa en la estadística *G* dado en la ecuación *(III.2.16)*. La única diferencia es que los valores ajustados, $\hat{\pi}$, bajo el modelo se basa sobre el vector que contiene *p+1* parámetros, $\hat{\beta}$, bajo la hipótesis nula de que los "*p*" coeficientes para las covariantes en el modelo son igual a cero, la distribución de *G* será una Chi cuadrado con *p* grados de libertad.

Antes de concluir que cualquier o todos los coeficientes son ceros, desearemos mirar el *test de Wald* univariante

$$W_j = \hat{\beta}_j / S\hat{E}(\hat{\beta}_j). \qquad \textit{III.3.9}$$

Bajo la hipótesis que un coeficiente individual es cero, esta estadística seguirá una distribución normal estándar. Si la meta es obtener el mejor modelo ajustado minimizaremos el número de parámetros, el próximo paso lógico es ajustar un modelo reducido que contenga solo las variables significativas, y comparar con el modelo que contiene todas las variables.

El test de razón de verosimilitud que compara estos dos modelos se obtiene usando la definición de *G* dado en la ecuación *(III.2.16)* este tendrá una distribución Chi-cuadrado con 2 grados de libertad bajo la hipótesis que los coeficientes para las variables excluidas son cero.

Cada vez que una variables independiente categórica es incluida (o excluida) de un modelo, todas las variables dummy deben de ser incluidas (o excluidas). A causa de los múltiples grados de libertad se debe ser cuidadoso en el uso de la estadística de Wald (*W*) para evaluar la significancia de los coeficientes.

La analogía multivariable del test de Wald es obtener del cálculo del siguiente vector-matriz

$$W = \hat{\beta}'\left[\hat{Var}(\hat{\beta})\right]^{-1}\hat{\beta}$$
$$= \hat{\beta}'(XVX)\hat{\beta} \qquad \textit{III.3.10}$$

Que tiene una distribución Chi-cuadrado con $p+1$ grados de libertad bajo la hipótesis de que cada uno de los $p+1$ coeficientes es igual a cero.

3.3.4 Estimación por Intervalos de Confianza

Hemos discutido la estimación por intervalos de confianza para los coeficientes de un modelo de regresión logística simple en la sección (3.2). El método para la estimación por intervalos de confianza para un modelo de variables múltiples es esencialmente el mismo. Un intervalo de confianza al $100(1-\alpha)\%$ para los coeficientes se obtienen desde las ecuaciones *(III.2.18)* y *(III.2.19)* para el termino constante. El estimado por intervalo de confianza para la logit es un poco más complicado para el modelo de variable múltiple que el resultado presentado en la ecuación *(III.2.22).* La idea básica es el mismo, solamente que ahora hay más términos involucrados en la

suma. Siguiendo de la ecuación *(III.3.1)* una expresión general para la estimación de la logit de un modelo que contiene *p* covariantes en:

$$\hat{g}(X) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p \qquad \textit{III.3.11}$$

Una vía alternativa para expresar la estimación de la logit en la ecuación *(III.3.11)* es el uso directo de la notación vectorial como $\hat{g}(X) = X'\hat{\beta}$, donde el vector $\hat{\beta}_0' = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ denota la estimación de los *p+1* coeficientes y el vector $X'=(x_0,x_1,x_2,...,x_p)$ representa la constante y el grupo de los valores de los *p-covariantes* en el modelo, donde $x_0=1$.

Siguiendo de la ecuación *(III.2.21)* donde una expresión para la estimación de la variance del estimador de la logit en la ecuación *(III.3.11)* es:

$$\hat{Var}\left[\hat{g}(X)\right] = \sum_{j=0}^{p} x^2{}_j \hat{Var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2x_j x_k \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k) \qquad \textit{III.3.12}$$

Podemos expresar este resultado mucho más conciso usando la matriz para la estimación de la variance del estimador de los coeficientes. De la expresión de la matriz de información observada, tenemos que:

$$\hat{Var}(\hat{\beta}) = (X'VX)^{-1} \qquad \textit{III.3.13}$$

Siguiendo de la ecuación *(III.3.13)* como una expresión equivalente para la estimación en la ecuación *(III.3.12)* es

$$\begin{aligned} \hat{Var}\left[(\hat{g}(x))\right] &= x'\hat{Var}(\hat{\beta})x \\ &= x'(X'VX)^{-1}x \end{aligned} \qquad \textit{III.3.14}$$

Afortunadamente, todos los paquetes de software que contiene la regresión logística proporciona la opción para el uso de crear una nueva variable conteniendo los valores estimados de la ecuación *(III.3.14)* o el error estándar para todos los sujetos de los datos.

3.4 MEDIDAS DE CONFIABILIDAD DEL MODELO

Las siguientes son unas medidas que cuantifican el nivel de ajuste del modelo al conjunto de datos:

3.4.1 La Desvianza

Es similar a la suma de cuadrados del error de la regresión lineal y es igual como el negativo de dos veces la función de verosimilitud maximizada más una constante

$$D = -2\sum_{i=1}^{n}\left[y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)\right] \qquad III.4.1$$

D es equivalente a la prueba de razón de verosimilitud para probar la validez del modelo logístico.

***Decisión**:*

- Si *D* es mayor que una χ^2 con *n-k* grados de libertad para un nivel de significación dado entonces el modelo logístico no es confiable.

3.4.2 El Pseudo-R2

Es similar al R^2 de la regresión lineal se define por:

$$Pseudo - R^2 = \left(1 - \frac{Devianza}{Devianza.Nula}\right)100\% \qquad III.4.2$$

3.5 ANÁLISIS DE LOS RESIDUOS PARA LA REGRESIÓN LOGÍSTICA

Existen varios tipos de residuos que permiten cotejar si una observación es discordante o no.

3.5.1 Residuos de Pearson

Definidos por:

$$r_j = \frac{y_j - m_j \hat{\pi}_j}{\sqrt{m_j \hat{\pi}_j \left(1 - \hat{\pi}_j\right)}} \qquad \textit{III.5.1}$$

Dónde:

y_i, representa el número de veces que $y = 1$ entre las m_j repeticiones de X_j. si los valores de la variable de respuesta están agrupados

El residual de Pearson es similar al residual estudentizado usado en la regresión lineal. Así un residual de Pearson en valor absoluto mayor que 2 indica un dato anormal.

La *estadística* χ^2 *de Pearson* es la suma de cuadrados de los residuales de Pearson.

$$\chi_P^2 = \sum_{i=1}^{J} r_j^2 \qquad \textit{III.5.2}$$

3.5.2 Residuos de Pearson Estandarizado

Definido por:

$$r_{ij} = \frac{r_j}{\sqrt{1 - h_j}} \qquad \textit{III.5.3}$$

Dónde:

r_j : son los residuos de Pearson.

h_j : es el leverage para la observación j-ésima. Es el elemento de la diagonal principal de la matriz H, definida en la sección 3.6.1.

3.5.3 Residuos de Desvianza

Los residuos de desvianza tienen el mismo signo que $y_j - m_j \hat{p}_j$ y están definidos por:

$$d_j = \pm \left\{ 2 \left[y_j \ln\left(\frac{y_j}{m_j \hat{p}_j} \right) + \left(m_j - y_j\right) \ln\left(\frac{m_j - y_j}{m_j \left(1 - \hat{p}_j\right)} \right) \right] \right\}^{1/2} \qquad \textit{III.5.4}$$

Si el residual de desvianza es mayor que 4 en valor absoluto entonces la observación correspondiente es anormal.

La desvianza es igual a la suma de cuadrados de los residuales desvianza:

$$\chi_D^2 = \sum_{j=1}^{J} d_j^2 \qquad \textit{III.5.5}$$

Otras medidas son:

$\chi_P^2 = \sum_{j=1}^{n} r_j^2 \overset{H_0}{\sim} \chi_{i-k}^2$ y

$\chi_D^2 = \sum_{j=1}^{n} d_i^2 \overset{H_0}{\sim} \chi_{i-k}^2$

Donde *k* es el número de parámetros estimados en el modelo logit.

3.6 MEDIDAS DE INFLUENCIA EN REGRESIÓN LOGÍSTICA

Son de gran utilidad para detectar la presencia de datos influyentes sobre el modelo de regresión logística.

3.6.1 Leverage

Son los elementos de la diagonal de la matriz de predicción ***H***.

El leverage para la observación *i-ésima* es el elemento *i-ésimo* de la diagonal principal de la matriz ***H***, h_i, y toma valores entre *0* y *1* con un valor medio de *p/n*.

$H = X_*\left(X'_* X\right)^{-1} X_*$

Donde:

$X_* = W^{1/2} X$ y $W = diag(\hat{P}_i(1-\hat{P}_i))$

Las estadísticas de influencia derivadas de h_j son:

3.6.2 La Distancia de Cook (ΔB_j):

Mide la influencia de la estimación deβ.

$$\Delta B_j = \frac{r_{ij}^2}{(1-hj)}, \qquad \textit{III.6.1}$$

Dónde:

r_{ij} : residuo de pearson estandarizado

Sí $\Delta B_j > 1$ es potencialmente influyente en los valores de los parámetros estimados.

3.6.3 Estadística Delta Chi-Cuadrado de Pearson ($\Delta\chi^2_{P(j)}$):

Reducción de las medidas en Pearson χ^2 que resulta de borrar todos los casos del *i-ésimo* patrón:

$$\Delta\chi^2_{P(j)} = \frac{r_j^2}{(1-h_j)} \qquad \textit{III.6.2}$$

Donde

r_j : son los residuos de Pearson.

h_j : es el leverage para la observación *j*-ésima. Es el elemento de la diagonal principal de la matriz *H*.

3.6.4 Estadística Delta Desvianza

Mide el cambio en la desvianza que resulta de eliminar la j-*ésima* observación.

$$\Delta\chi^2_{D(j)} = \frac{d_j^2}{(1-h_j)} \qquad \textit{III.6.3}$$

$\Delta\chi^2_{P(j)}$ y $\Delta\chi^2_{D(j)}$ Son medidas de influencia en el j-ésimo valor ajustado del Modelo de Regresión Logística. Como regla valores mayores de 4 indican un cambio significativo (porque la distribución es asintóticamente χ^2).

3.6.5 Gráficos para el Diagnóstico

Son de gran utilidad para detectar la presencia de datos influyentes.

Para tener una descripción rápida de la información proporcionada por las estadísticas descritas arriba utilizamos, según Hosmer y Lemeshow (1989) los siguientes gráficos:

1. Delta Chi-cuadrado contra la probabilidad estimada.
2. Delta desvianza contra la probabilidad estimada.
3. Distancia de Cook contra la probabilidad estimada.
4. Delta Chi-cuadrado contra leverage.
5. Delta desvianza contra leverage.
6. Distancia de Cook contra leverage.

A continuación se presenta un cuadro de resumen de estas estadísticas para la medida de influencia:

Cuadro III.1

MEDIDAS DE INFLUENCIA

ESTADISTICA	INFLUYEN EN :
Delta Chi-cuadrado	las estimaciones de la probabilidad de la variable dependiente
Delta Desvianza	las estimaciones de la probabilidad de la variable dependiente
DFFITS	las estimaciones de la probabilidad de la variable dependiente
COOK	las estimaciones de los parámetros del modelo
DFBETAS (i)	las estimaciones del parámetro que acompaña a cada variable del modelo

CAPÍTULO IV
ANÁLISIS DE LOS DATOS

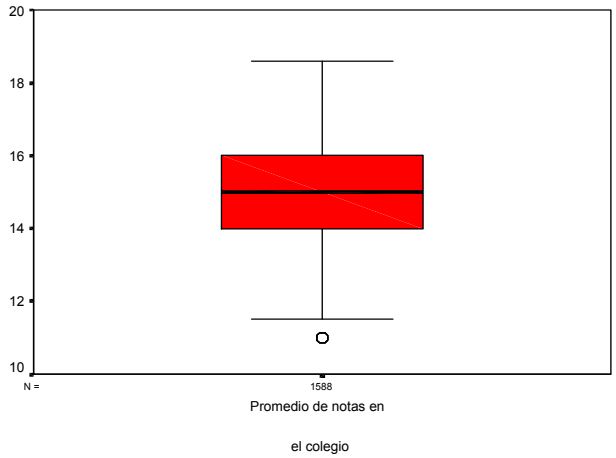
4.1 ANÁLISIS EXPLORATORIO PREVIO

La finalidad del Análisis Exploratorio de Datos (AED) es examinar los datos previamente a la aplicación de cualquier técnica estadística. De esta forma el analista consigue un entendimiento básico de sus datos y de las relaciones existentes entre las variables analizadas. Para conseguir este objetivo el AED proporciona métodos sistemáticos sencillos para organizar y preparar los datos, detectar fallas en el diseño y recogida de los mismos, identificación de casos atípicos (outliers) y comprobación de supuestos necesarios a las técnicas multivariantes. El examen previo de los datos es un paso necesario, que lleva tiempo, y que habitualmente se descuida por parte de los analistas de datos. Las tareas implícitas en dicho examen pueden parecer insignificantes y sin consecuencias a primera vista, pero son una parte esencial de cualquier análisis estadístico.

4.1.1 Análisis Exploratorio de las Variables Cuantitativas

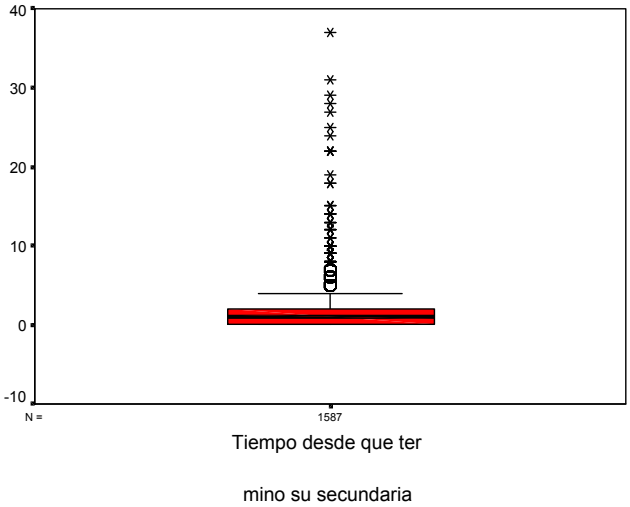
Para la variable *Promedio de notas en el colegio*, el gráfico de cajas nos dice que es aproximadamente simétrica. Presenta solo un "outlier". Como veremos más adelante esta variable es la única que tiene este comportamiento.

Gráfico IV.1
Promedio de notas en el colegio.



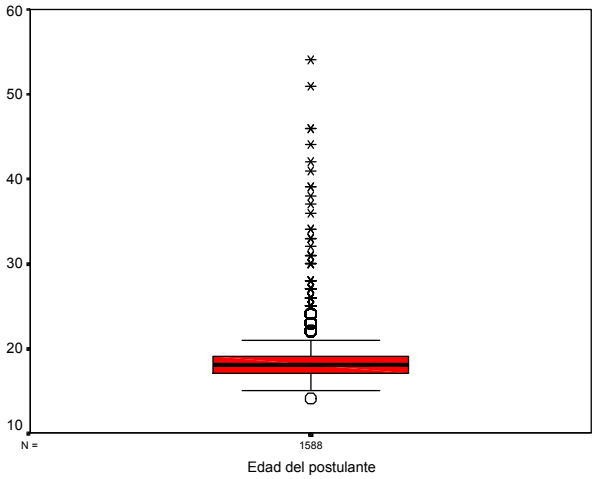
Con respecto a la variable *Tiempo desde que termino su secundaria* vemos que esta es asimétrica, que no tiene mucha dispersión pero presenta muchos datos discordantes u outliers.

Gráfico IV.2
Tiempo desde que termino su secundaria.



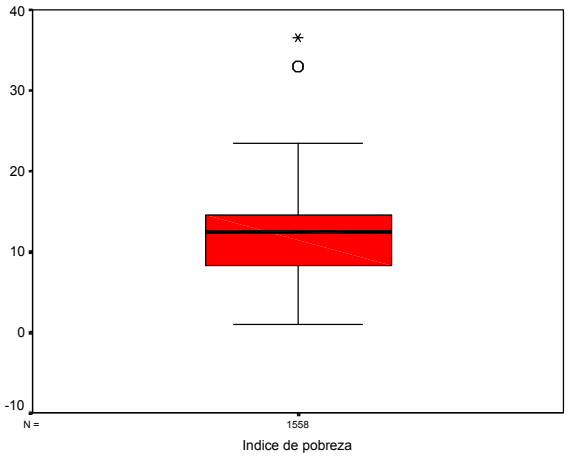
Para la variable *Edad del postulante* notamos que esta tiene una distribución simétrica dentro de su dispersión, pero que presenta, como la variable anterior, muchos datos discordantes u outliers.

Gráfico IV.3
Edad del postulante.



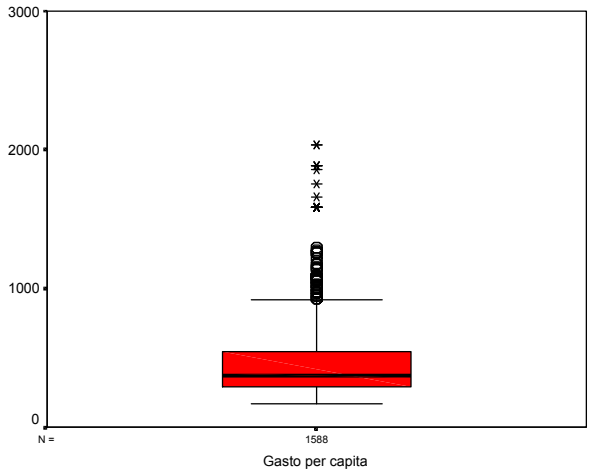
En cuanto a la variable *Índice de pobreza* notamos que es un tanto asimétrica, pero que no presenta muchos datos discordantes como las anteriores variables. Estos datos discordantes se pueden explicar dado que algunos postulantes provienen de provincias las cuales presentan un Índice de pobreza más alto que cualquier distrito de Lima.

Gráfico IV.4
Índice de pobreza.



Por último, para la variable *Gasto per cápita* es también asimétrica, la cual presenta datos discordantes como las anteriores. Cabe recalcar nuevamente que esta variable fue estimada en base a otras variables útiles para esta operación.

Gráfico IV.5
Gasto per cápita.

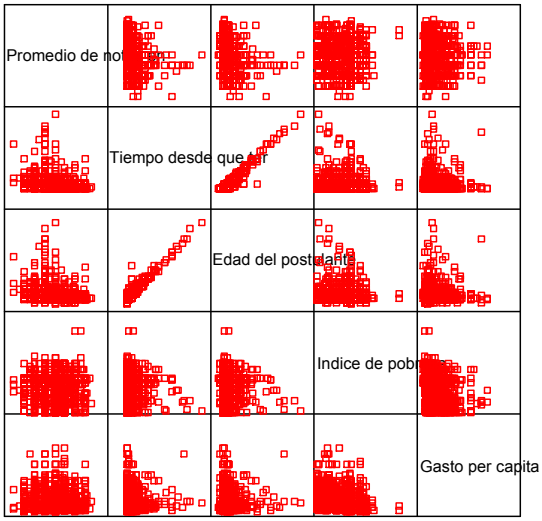


En resumen, vemos que existen muchos datos discordantes u outliers en casi todas las variables cuantitativas de este estudio. Esto podría ocasionar complicaciones a la hora de ajustar un buen modelo, para tratar de evitarlo, antes de modelarlo lo clasificaremos en categorías a fin de evitar posibles influencias en las estimaciones de los parámetros.

Avanzando con esta exploración evaluaremos la correlación entre estas variables cuantitativas, mostraremos el gráfico de la matriz de dispersión y luego una tabla en la cual se muestra los Coeficientes de Correlación Pearson que evalúa las relaciones lineales dos a dos entre estas variables.

Gráfico IV.6

Matriz de dispersión entre variables cuantitativas.



Cuadro IV.1

Coeficiente de correlación de Pearson entre variables cuantitativas.

Correlations

		Edad del postulante	Indice de pobreza	Tiempo desde que termino la secundaria	Promedio de notas en el colegio	Gasto per capita
Edad del postulante	Pearson Correlation	1.000	-.045	.960**	-.196**	-.022
	Sig. (2-tailed)	.	.076	.000	.000	.373
	N	1588	1558	1587	1588	1588
Indice de pobreza	Pearson Correlation	-.045	1.000	-.032	-.032	-.314**
	Sig. (2-tailed)	.076	.	.209	.205	.000
	N	1558	1558	1557	1558	1558
Tiempo desde que termino la secundaria	Pearson Correlation	.960**	-.032	1.000	-.178**	-.010
	Sig. (2-tailed)	.000	.209	.	.000	.695
	N	1587	1557	1587	1587	1587
Promedio de notas en el colegio	Pearson Correlation	-.196**	-.032	-.178**	1.000	.023
	Sig. (2-tailed)	.000	.205	.000	.	.361
	N	1588	1558	1587	1588	1588
Gasto per capita	Pearson Correlation	-.022	-.314**	-.010	.023	1.000
	Sig. (2-tailed)	.373	.000	.695	.361	.
	N	1588	1558	1587	1588	1588

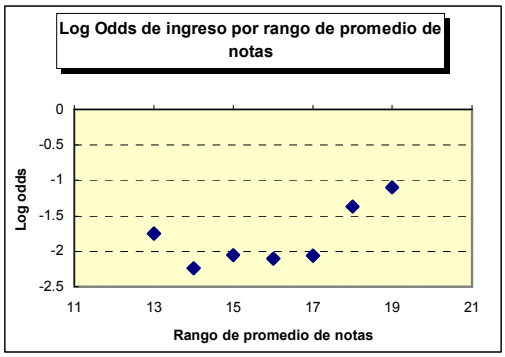
**. Correlation is significant at the 0.01 level (2-tailed).

Del *gráfico (IV.6)* y del *cuadro (IV.1)* podemos concluir que las variables que están fuertemente correlacionados son la *Edad del postulante* y el *Tiempo desde que terminó su educación secundaria*. Las que presentan una correlación considerable son *Índice de pobreza* con el *Gasto per cápita* y un tanto menos las variables *Promedio de notas en el colegio* y la *Edad del postulante*. Todas estas correlaciones son significativas al 1%.

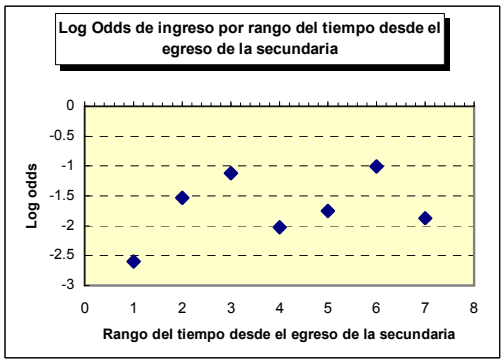
Dado que el Modelo de la Regresión Logística es el Log odds de la variable (Y) ingreso y una combinación lineal de las variables independientes, se intentara mostrar las relaciones lineales existentes entre el Log odds y cada una de las variables independientes con los siguientes gráficos:

Gráfico IV.7
Gráfico de dispersión entre la variable Log Odds y las variables independientes.

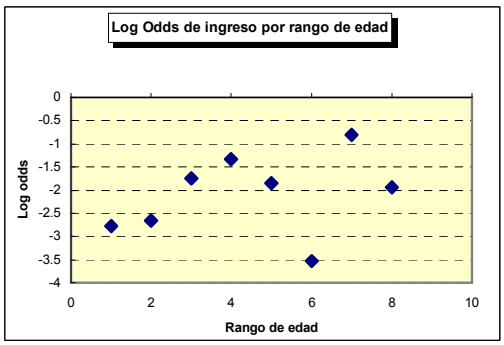
(a)



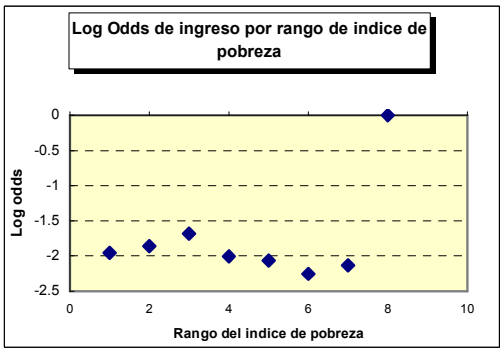
(b)



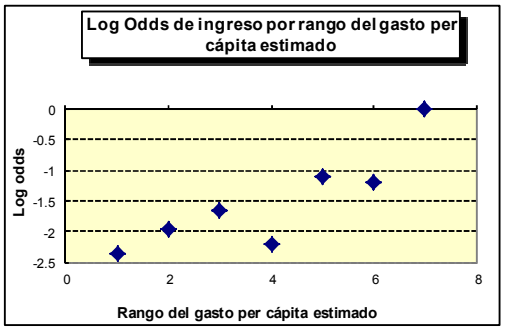
(c)



(d)



(e)



De estos gráficos podemos observar que las variables que tienen una relación aproximadamente lineal con los log odds de la variable ingreso son las siguientes: Promedio de notas, Tiempo desde el egreso de la secundaria y el Gasto per cápita.

4.1.2 Análisis Exploratorio de las Variables Cualitativas

La evaluación de estas variables se hará a través de las tablas de contingencia en la cual se mostrara las medidas de asociación entre la variable dependiente y las variables independientes cualitativas, con las respectivas pruebas Chi-cuadrado. Comentaremos a continuación los resultados de estas variables:

Cuadro IV.2

Tabla de contingencia entre la variable dependiente y la Modalidad de postulación.

			Modalidad de postulación				
			Prueba general	CEPUSM	Primeros puestos	Otras modalidades	Total
¿Ingreso?	No ingreso	Count	1127	204	40	26	1397
		% within ¿Ingreso?	80.7%	14.6%	2.9%	1.9%	100.0%
	Si ingreso	Count	80	81	14	16	191
		% within ¿Ingreso?	41.9%	42.4%	7.3%	8.4%	100.0%
Total		Count	1207	285	54	42	1588
		% within ¿Ingreso?	76.0%	17.9%	3.4%	2.6%	100.0%

Cuadro IV.2.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	142.475[a]	3	.000
Continuity Correction			
Likelihood Ratio	120.459	3	.000
Linear-by-Linear Association	119.092	1	.000
N of Valid Cases	1588		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.05.

- En el *cuadro (IV.2)* vemos que la distribución de los porcentajes difieren por modalidad según si ingresa o no a la universidad. Es decir, de los que no ingresan la mayoría son de la modalidad 'Prueba general' (80.7%), esto en cambio se invierte para los que si ingresan pues el mayor porcentaje se encuentra

en la CEPUSM (42.4%). Esto quiere decir que mayor chance tiene de ingresar los que proviene de la CEPUSM. El "test" Chi-cuadrado (*cuadro IV.2.1*) para ver la dependencia tiene un *p_valor* de 0.000, quiere decir que estas dos variables están relacionados a un nivel de significancia del 5%.

Cuadro IV.3

Tabla de contingencia entre la variable dependiente y el Colegio de procedencia.

			Colegio de Procedencia		
			Estatal	Particular	Total
¿Ingreso?	No ingreso	Count	1057	340	1397
		% within ¿Ingreso?	75.7%	24.3%	100.0%
	Si ingreso	Count	120	71	191
		% within ¿Ingreso?	62.8%	37.2%	100.0%
Total		Count	1177	411	1588
		% within ¿Ingreso?	74.1%	25.9%	100.0%

Cuadro IV.3.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	14.429[b]	1	.000		
Continuity Correction [a]	13.768	1	.000		
Likelihood Ratio	13.503	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	14.420	1	.000		
N of Valid Cases	1588				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 49.43.

- En el *cuadro (IV.3)* vemos que el porcentaje de los que provienen de colegio particular en los postulantes que no ingresan es de 24.3%, este porcentaje crece hasta el 37.2% para los que si ingresan. Estas frecuencias hacen suponer que las variables no son independientes. Según el "test" Chi-cuadrado para la independencia de variables *(cuadro IV.3.1)* vemos que el *p_valor* asociado es de 0.000 y con un nivel de significancia del 5% se confirma la dependencia entre estas variables.

Cuadro IV.4

Tabla de contingencia entre la variable dependiente y el Tipo de preparación pre universitaria.

			Tipo de preparación pre universitaria				
			Solo	Grupo de estudio	Academia	CEPUSM	Total
¿Ingreso?	No ingreso	Count	273	33	871	220	1397
		% within ¿Ingreso?	19.5%	2.4%	62.3%	15.7%	100.0%
	Si ingreso	Count	32	4	68	87	191
		% within ¿Ingreso?	16.8%	2.1%	35.6%	45.5%	100.0%
Total		Count	305	37	939	307	1588
		% within ¿Ingreso?	19.2%	2.3%	59.1%	19.3%	100.0%

Cuadro IV.4.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	98.251[a]	3	.000
Continuity Correction			
Likelihood Ratio	82.932	3	.000
Linear-by-Linear Association	22.721	1	.000
N of Valid Cases	1588		

a. 1 cells (12.5%) have expected count less than 5. The minimum expected count is 4.45.

- En el *cuadro (IV.4)* vemos que la academia (entre los tipos de preparación pre universitaria) es del mayor porcentaje para los postulantes que no ingresaron (62.3%). Esto cambia para los postulantes que si ingresaron, pues en este caso el mayor porcentaje se encuentra en la CEPUSM con un 45.5%. Según el "test" Chi-cuadrado para la independencia de variables *(cuadro IV.4.1)* vemos que el *p_valor* asociado es de 0.000 y con un nivel de significancia de 5% se confirma la dependencia entre estas variables.

Cuadro IV.5

Tabla de contingencia entre la variable dependiente y el Sexo del postulante.

			Sexo		
			Masculino	Femenino	Total
¿Ingreso?	No ingreso	Count	667	730	1397
		% within ¿Ingreso?	47.7%	52.3%	100.0%
	Si ingreso	Count	99	92	191
		% within ¿Ingreso?	51.8%	48.2%	100.0%
Total		Count	766	822	1588
		% within ¿Ingreso?	48.2%	51.8%	100.0%

Cuadro IV.5.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.124[b]	1	.289		
Continuity Correction [a]	.966	1	.326		
Likelihood Ratio	1.123	1	.289		
Fisher's Exact Test				.316	.163
Linear-by-Linear Association	1.124	1	.289		
N of Valid Cases	1588				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 92.13.

- En el *cuadro (IV.5)* vemos que la distribución de los porcentajes por sexo, según los postulantes que ingresaron o no, son parecidas. Esto hace suponer que existe independencia entre estas variables. Según el "test" Chi-cuadrado para la independencia de variables *(cuadro IV.5.1)* vemos que el *p_valor* asociado es de 0.289 y con lo cual afirmamos, con un nivel de significancia del 5%, que existe independencia entre estas variables.

Cuadro IV.6

Tabla de contingencia entre la variable dependiente y la cantidad de veces que postuló anteriormente a la universidad.

			¿Cuántas veces postuló anteriormente a la universidad?					
			Ninguna	Una	Dos	Tres	Cuatro a mas	Total
¿Ingreso?	No ingreso	Count	826	382	126	44	19	1397
		% within ¿Ingreso?	59.1%	27.3%	9.0%	3.1%	1.4%	100.0%
	Si ingreso	Count	50	63	51	21	6	191
		% within ¿Ingreso?	26.2%	33.0%	26.7%	11.0%	3.1%	100.0%
Total		Count	876	445	177	65	25	1588
		% within ¿Ingreso?	55.2%	28.0%	11.1%	4.1%	1.6%	100.0%

Cuadro IV.6.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	110.760[a]	4	.000
Continuity Correction			
Likelihood Ratio	98.818	4	.000
Linear-by-Linear Association	98.931	1	.000
N of Valid Cases	1588		

a. 1 cells (10.0%) have expected count less than 5. The minimum expected count is 3.01.

- En el *cuadro (IV.6)* vemos que la distribución de los porcentajes por la cantidad de veces que postulo anteriormente, según los postulantes que ingresaron o no, son diferentes. Es decir, para los que no ingresaron el mayor porcentaje es de los que nunca postularon anteriormente (59.1%). Esto cambia para los que si ingresaron pues el mayor porcentaje se encuentra para los que ya postularon una vez anteriormente (33.0%). Según el "test" Chi-cuadrado para la independencia de variables *(cuadro IV.6.1)* vemos que el *p_valor* asociado es de 0.000 con lo cual afirmamos, a un nivel de significancia del 5%, que existe dependencia entre estas variables.

Cuadro IV.7

Tabla de contingencia entre la variable dependiente y el Área académica a la que postula.

			Área académica				
			Ciencias basicas e Ingenierías	Ciencias de la salud	Ciencias ecónomico-empresarial	Humanidades	Total
¿Ingreso?	No ingreso	Count	298	446	264	389	1397
		% within ¿Ingreso?	21.3%	31.9%	18.9%	27.8%	100.0%
	Si ingreso	Count	72	31	34	54	191
		% within ¿Ingreso?	37.7%	16.2%	17.8%	28.3%	100.0%
Total		Count	370	477	298	443	1588
		% within ¿Ingreso?	23.3%	30.0%	18.8%	27.9%	100.0%

Cuadro IV.7.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	33.212[a]	3	.000
Continuity Correction			
Likelihood Ratio	33.000	3	.000
Linear-by-Linear Association	3.636	1	.057
N of Valid Cases	1588		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 35.84.

- En el *cuadro (IV.7)* vemos que la distribución de los porcentajes por el tipo de Área académica a la que postula, según los postulantes que ingresaron o no, son diferentes. Es decir, para los que no ingresaron el mayor porcentaje es del área de ciencias de la salud (31.9%). Esto cambia para los que si ingresaron pues el mayor porcentaje se encuentra para los de ciencias básicas e ingenierías (37.7%). Según el "test" Chi-cuadrado para la independencia de variables vemos *(cuadro IV.7.1)* que el *p_valor* asociado es de 0.000 con lo cual afirmamos, a un nivel de significancia del 5%, que existe dependencia entre estas variables.

Cuadro IV.8

Tabla de contingencia entre la variable dependiente y el Nivel de estudio del padre.

			Nivel de estudios del padre						
			Analfabeto/inicial	Primaria	Secundaria/carrera corta	Superior no universitaria	Superior Universitaria	Post grado (Maestría, Doctorado)	Total
¿Ingreso?	No ingreso	Count	6	154	529	288	353	56	1386
		% within ¿Ingreso?	.4%	11.1%	38.2%	20.8%	25.5%	4.0%	100.0%
	Si ingreso	Count	1	14	60	42	62	11	190
		% within ¿Ingreso?	.5%	7.4%	31.6%	22.1%	32.6%	5.8%	100.0%
Total		Count	7	168	589	330	415	67	1576
		% within ¿Ingreso?	.4%	10.7%	37.4%	20.9%	26.3%	4.3%	100.0%

Cuadro IV.8.1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.767[a]	5	.119
Continuity Correction			
Likelihood Ratio	8.785	5	.118
Linear-by-Linear Association	8.253	1	.004
N of Valid Cases	1576		

a. 1 cells (8.3%) have expected count less than 5. The minimum expected count is .84.

- En el *cuadro (IV.8)* vemos que la distribución de los porcentajes por el nivel de estudios del padre, según los postulantes que ingresaron o no, son parecidas. Esto hace suponer que existe independencia entre estas variables. Según el "test" Chi-cuadrado para la independencia de variables *(cuadro IV.8.1)* vemos que el *p_valor* asociado es de 0.119 con lo cual afirmamos, a un nivel de significancia del 5%, que existe independencia entre estas variables.

4.2 AJUSTE DEL MODELO DE REGRESIÓN LOGÍSTICA

Los modelos de regresión, como en el caso lineal, pueden usarse con dos objetivos: 1) *Predictivo*, en el cual el interés del investigador es predecir lo mejor posible la variable dependiente, usando un conjunto de variables independientes. 2) *Estimativo*, en el que el interés se centra en estimar la relación de una o más variables independientes con la variable dependiente.

Debido a estos objetivos es que el análisis de regresión tiene una difícil tarea de establecer un mecanismo general para encontrar el mejor modelo de regresión que son cosas distintas para cada objetivo. En el análisis estimativo que es el caso de esta investigación, pues aquí se trata de encontrar factores determinantes de un 'fenómeno' (si ingresa o no a la universidad), el mejor modelo es el que produce estimaciones mas precisas para el coeficiente de la variable de interés.

Por esto, nos centraremos en los “test” que evalúan los coeficientes y sus significancias ya que estos manifiestan relaciones con la variable independiente.

Para el ajuste inicial del modelo antes tenemos que evaluar si las variables a tomar en cuenta son las apropiadas para el análisis. Estas variables ya mencionadas en la sección (2.2) se eligieron de entre tantas disponibles obtenidas desde las fuentes antes mencionadas. La elección de estas variables es en función de los objetivos de esta monografía, por la consulta de profesionales en sociología y por las conclusiones de diversos estudios e investigaciones hechas sobre los factores influyentes en el rendimiento académico. En consecuencia de todo esto es que se seleccionaron estas variables para el análisis particular de nuestro caso. Ahora, por el análisis exploratorio previo de estas variables concluimos que la variable ‘tiempo desde que termino la secundaria’ no es necesario incluirla para el ajuste del modelo por estar bien relacionada con la ‘edad’. Por último, los otros resultados del análisis exploratorio lo usaremos para confirmar los resultados que se obtendrán al final de este estudio.

Es importante mencionar que el análisis y las salidas que se muestran aquí se efectuaron usando el paquete estadístico SPSS versión 10.0 y con ayuda del programa Excel en algunos casos.

Al revisar las variables que se tomaran en cuenta vemos que existen variables cualitativas con más de dos categorías, estas se tiene que tomar como variables ‘*Dummy*’(o indicadoras). Las variables *‘Numero de veces que postulo anteriormente a la universidad*’ es en principio cuantitativa, pero como solo toma 4 valores la podemos entender como una cualitativa ordinal.

4.2.1 Especificación del Modelo con todas las Variables en Estudio

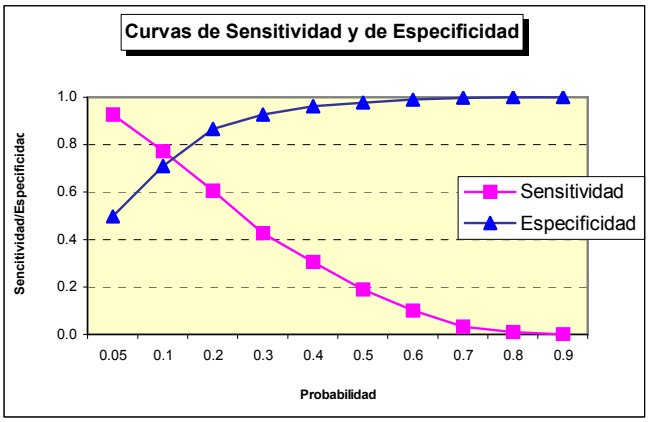
El primero paso será ajustar un modelo saturado, es decir, con todas las variables seleccionadas y declaradas como aceptables después de la revisión exploratoria para tratar de ajustar un buen modelo.

Pero antes de proceder con el ajuste del modelo nos damos cuenta que para efectos de *clasificación*, la cual nos dice la efectividad del modelo, la manera más fácil de 'discriminar' es considerar que si $p>0.5$ entonces la observación pertenece a la clase que uno está interesado. Pero algunas veces esto puede resultar 'injusto' sobre todo si se conoce si una de las clases es menos frecuente que la otra como sucede en este estudio. Un método alternativo es plotear el porcentaje de observaciones que poseen el evento que han sido correctamente clasificados (Sensitividad) versus distintos niveles de probabilidad y el porcentaje de observaciones de la otra clase que han sido correctamente clasificados (Especificidad) versus los mismos niveles de probabilidad anteriormente usados, en la misma gráfica. La probabilidad que se usará para clasificar las observaciones se obtienen interceptando las dos curvas.

Este procedimiento no afecta las estimaciones de los coeficientes, dado que estos nos dice, según los "test" adecuados, si la variable es significativa para explicar la variable de interés. La gráfica de las curvas interceptadas se muestra a continuación:

Gráfico IV.8

Gráfico de curvas de sensitividad y especificidad versus niveles de probabilidad.



Podemos apreciar que el punto de corte es de 0.2 aproximadamente. A partir de ahora tomaremos este punto de corte para todo el análisis posterior.

Ajustaremos un Modelo de Regresión Logística binaria teniendo inicialmente todas las variables mencionadas en la sección (2.2) excepto la variable 'tiempo desde que termino la secundaria' por estar fuertemente correlacionado con la Edad. Esto es el modelo saturado. A continuación se describen los resultados más importantes, los resultados en detalles se presenta en el *ANEXO 2*.

Según estos resultados vemos que el *–2 log likelihood* solo con la constante y con todas las variables tienen los siguientes valores:

Cuadro IV.9

Tabla de valores del -2log likelihood para el modelo solo con la constante y con todas las variables.

Modelo	-2 log likelihood
Solo con la constante	1136.4434
Con todas las variables	883.488

No se ve una gran diferencia de estos valores entre estos modelos. Con la diferencia (que se distribuye como una Chi-cuadrado con 26 grados de libertad) de estos valores podremos contrastar la siguiente hipótesis:

H_0: Todos los coeficientes son iguales a cero.
H_1: Por lo menos un coeficiente es diferente de cero.

La estadística *G* de la *ecuación (III.2.13)* que es la diferencia de la *Desvianza* para cada modelo nos da el siguiente valor G=*1136.4434 - 883.488=252.9554* y el *p-valor* asociado para la prueba es $P\left[\chi^2_{(26)} > 252.955\right] = 0$. A un nivel de significancia del 5%, rechazamos la hipótesis nula y concluimos que al menos un coeficiente es diferente de cero.

La bondad de ajuste del modelo completo se evalúa con la prueba de Hosmer-Lemeshow con la siguiente hipótesis:

H_0: El modelo se ajusta adecuadamente.
H_1: El modelo no se ajusta adecuadamente.

Cuadro IV.10

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.956	8	.652

Podemos decir que el ajuste del modelo es bueno con *p_valor* = 0.6522, pues no hay razones suficientes para rechazar la hipótesis nula a un nivel de significancia del 5%.
El valor del R^2 *de Nagelkerke* es de tan solo 0.290. Además el porcentaje de casos clasificados debidamente por el modelo en una de las dos categorías solo es regular (60.22 % de casos clasificados correctamente por el modelo en cuanto a los casos observados del grupo de los que si ingresan) pero la clasificación global es aceptable

(83.12% de casos correctamente clasificados por el modelo) *(Anexo 2, en la sección BLOQUE 1: MÉTODO ENTER)*.

4.2.2 Ajuste del Modelo Logístico con el Método Forward[10] para la Selección de Variables

En esta sección lograremos encontrar cuál de las variables del modelo saturado realmente explican o están bien relacionados con el fenómeno de interés en este estudio. El método FORWARD utiliza estadísticas las cuales miden el grado de relación que existe entre las variables explicativas y la de interés.

Este método comienza con el modelo ajustado considerando únicamente solo la constante. En los pasos siguientes se va introduciendo variables según su mínimo *p-valor* asociado al estadístico 'Puntuación eficiente de Rao' y eliminando variables según el máximo *p-valor* asociado al estadístico de Wald respecto a los valores críticos que se den en ambos casos.

Describiremos a continuación paso por paso este método *(los valores se muestran para el primero paso y para los pasos siguientes en los cuadros "Variables in the Equation" y "Variables not in the Equation" en las secciones BLOQUE 0: BLOQUE DE INICIO y BLOQUE 1: Método = Forward Stepwise (Wald) del Anexo 3, respectivamente).*

***Paso cero**:* Solo está considerado la constante.

***Primer paso**:* En este paso, de entre todas la variables que no se encuentran en el modelo, la primera candidata a ser seleccionada es la que presenta un mínimo *p_valor* asociado al estadístico Puntuación eficiente de Rao, o al que presenta el máximo valor

[10] Se usó el método Forward Stepwise (Wald) para la selección de variables

de este estadístico. La variable 'Modalidad por la que postula' es la que presenta el mayor valor (138.0766) con un *p_valor* de 0.000 menor a 0.05 con lo que esta será incluida en el modelo. Esta variable como es cualitativa ha generado tres variables indicadoras las cuales serán incluidas todas al mismo tiempo. Se debe recalcar que para esta acción no tiene que observarse las variables indicadoras por separado, pues estas son tratadas en bloque.

Segundo paso: De los que no están en el modelo vemos que el que tiene un mayor valor en el estadístico Puntuación eficiente de Rao (72.521) es la variable "*Veces que postulo anteriormente a la universidad"* con un *p_valor* menor del 0.05 por lo que esta será incluida en el modelo. Ahora tenemos que ver cuál de las variables que se encuentran en el modelo es candidata a ser eliminada según el estadístico de Wald y su *p_valor* asociado. Ninguna será eliminada del modelo.

Tercer paso: De las variables que quedan, el que tiene un alto valor del estadístico Puntuación eficiente de Rao es la variable *"Área académica"* (47.1742) con un *p_valor* asociado de 0.000 menor que 0.05 por lo que será incluida en el modelo. Como en el caso anterior esta variable ha generado cuatro variables indicadoras los cuales serán incluidos todos en bloque. De las variables que se encuentran en el modelo ninguna será eliminada.

Cuarto paso*:* De las variables que no se encuentran en el modelo, el que tiene el mayor valor del estadístico Puntuación eficiente de Rao es la variable *"Colegio de procedencia"* (7.1837) con un *p_valor* de 0.0074 menor de 0.05 por lo que será incluida en el modelo. Vemos aquí también que ninguna que se encuentra en el modelo será

excluida de esta pues el valor del *p_valor* asociado a la estadística de Wald de alguna variable está por debajo de 0.05.

***Quinto paso**:* en este último paso vemos que de las que se encuentran fuera del modelo, la variable *"Promedio de notas en el colegio"* será incluida en el modelo por tener esta un valor mayor en el estadístico Puntuación eficiente de Rao (4.4072), el *p_valor* asociado a esta estadística es de 0.0358 menor de 0.05. De las que se encuentran en el modelo ninguna será eliminada pues los valores de la estadística de Wald y sus respectivos p_valores no permiten esta acción. El último paso de este método se presenta en el *ANEXO 3*.

Las pruebas ómnibus en cada paso se muestra en el siguiente cuadro:

Cuadro IV.11

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Modelo	116.848	3	0.0000
	Bloque	116.848	3	0.0000
	Paso	116.848	3	0.0000
Paso 2	Modelo	185.417	7	0.0000
	Bloque	185.417	7	0.0000
	Paso	68.569	4	0.0000
Paso 3	Modelo	233.413	10	0.0000
	Bloque	233.413	10	0.0000
	Paso	47.996	3	0.0000
Paso 4	Modelo	240.377	11	0.0000
	Bloque	240.377	11	0.0000
	Paso	6.964	1	0.0083
Paso 5	Modelo	244.774	13	0.0000
	Bloque	244.774	13	0.0000
	Paso	4.397	2	0.0294

En donde en cada paso se contrasta las siguientes hipótesis:

Para el 'Modelo' (Modelo con todas las variables consideradas en el paso respectivo) se contrasta la siguiente hipótesis:

H_0: Los coeficientes son iguales a cero.

Para el 'Paso' (cambio que se produce de un paso a otro) se contrasta la siguiente hipótesis:

H_0: La mejora del cambio de un paso a otro no es significativa.

Vemos que los p_valores asociados a la estadística Chi-cuadrado son menores que 0.05 por lo tanto, a un nivel de significancia del 5%, los coeficientes de cada variable en cada paso son diferentes de cero; lo que es lo mismo, la mejora que se produce en cada paso es significativa.

Para contrastar la siguiente hipótesis:

H_0: El modelo se ajusta adecuadamente a los datos en cada paso.

Contra

H_1: El modelo no se ajusta adecuadamente a los datos en cada paso.

Tenemos el *test Goodness-of-fit* de Hosmer-Lemeshow la que se presenta a continuación:

Cuadro IV.12

Test Goodness-of-Fit de Hosmer and Lemeshow			
Pasos	Chi-cuadrado	gl	Sig.
1	0.0000	1	0.9981
2	7.0956	4	0.1309
3	7.2237	7	0.4060
4	10.3982	8	0.2382
5	11.8032	8	0.1602

Con lo que concluimos que los modelos en cada paso si se ajustan adecuadamente.

El valor del R^2 de Nagelkerke es de 0.281 el cual no ha tenido mejora alguna, esto podría deberse a la influencia de los casos discordantes o 'outliers', para confirmar esta

sospecha haremos un estudio de diagnóstico, asumiendo como correcto el modelo ajustado.

En resumen, las variables seleccionadas por este método se presentan en el siguiente cuadro:

Cuadro IV.13

VARIABLES SELECCIONADAS POR EL MÉTODO FORWARD*

N°	Variable
1	Modalidad de ingreso
2	Colegio de procedencia
3	Número de veces que postuló anteriormente a la universidad
4	Área académica
5	Promedio de notas en el colegio

Nota: (*) Forward:Wald

4.3 DIAGNOSTICO DEL MODELO AJUSTADO

En esta sección analizaremos los diferentes tipos de residuos para detectar los casos discordantes u 'outliers' y las medidas de influencia que interfieren en las estimaciones de la variable respuesta y en los parámetros del modelo.

4.3.1 Análisis de los Residuos

Con este fin y para mejorar el modelo propuesto por el método Forward para la selección de variables se identificó los casos discordantes u 'outliers' las cuales influyen significativamente en la estimación de los parámetros. Estos casos son los que tienen residuos estandarizados elevados con respecto a la mayoría de los casos. Un resumen se presenta en el siguiente cuadro:

Cuadro IV.14

OBSERVACIONES DISCORDANTES U 'OUTLIERS' SEGÚN EL TIPO DE RESIDUO

	Residuo			
	Común	Pearson	Pearson Estandarizado	Desvianza
Observaciones discordantes u "outliers"	No se observa casos alejados de los demas	190 478 1016 1035 1194	No se observa casos alejados de los demas	No se observa casos alejados de los demas

Lo podemos visualizar más claramente en los siguientes gráficos:

Gráfico IV.9

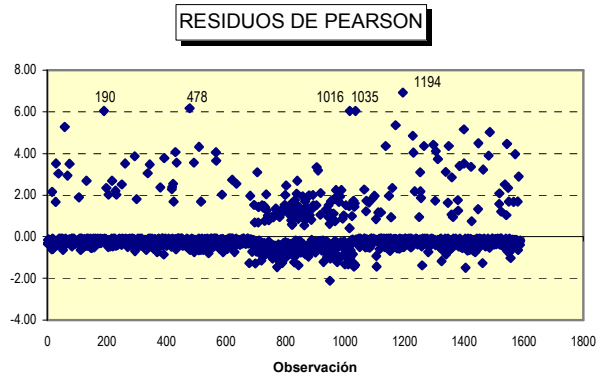
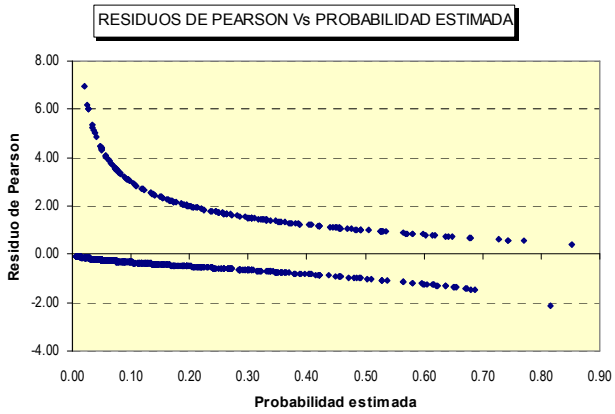


Gráfico IV.10



Vemos que existen algunos casos en los cuales los residuos son elevados, por este motivo solo consideraremos datos influyentes u outliers a los que están más alejados del resto de casos. (Los gráficos de los otros tipos de residuos se presentan en el *ANEXO 4 parte a)*.

Estos casos serán eliminados, luego veremos cómo afecta los estadísticos considerados.

En la sección siguiente evaluaremos las observaciones que son influyentes en los parámetros del modelo.

4.3.2 Análisis de Influencia

En el cuadro *(IV.15)* mencionamos las medidas de influencia y que observaciones son influyentes.

Cuadro IV.15

MEDIDAS DE INFLUENCIA

ESTADISTICA		OBSERVACIONES INFLUYENTES	INFLUYEN EN
Delta Chi-Cuadrado		190, 478, 1016, 1035, 1194.	Las estimaciones de la probabilidad de Ingresar a la UNMSM.
Delta Desvianza		No se aprecia observaciones influyentes	
DFFITS		682, 1030, 1297.	
COOK		628, 1030, 1297, 1305	Las estimaciones de los parámetros del Modelo.
DFBETAS	Modadlidad de Ingreso	No se aprecia observaciones influyentes	Las estimaciones del parámetro que acompaña a cada variable del modelo
	Colegio de procedencia	No se aprecia observaciones influyentes	
	Número de veces que postuló anteriormente a la universidad	28, 250, 682, 699, 797, 827, 866, 935, 945, 1030, 1099, 1106, 1148, 1252, 1297 y 1566	
	Área académica	1297 y 1305	
	Promedio de notas en el colegio	No se aprecia observaciones influyentes	

A continuación graficaremos las medidas de influencia para tener una mayor visualización y detalle de los casos que más influyen en las estimaciones.

4.3.2.1 Gráficos que ayudan a detectar observaciones influyentes en la estimación de la variable respuesta

Gráfico IV.11

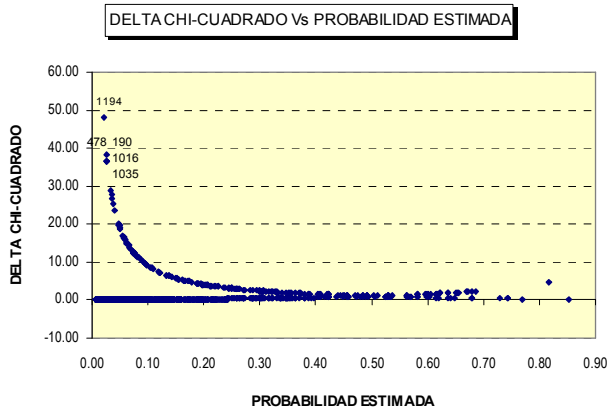
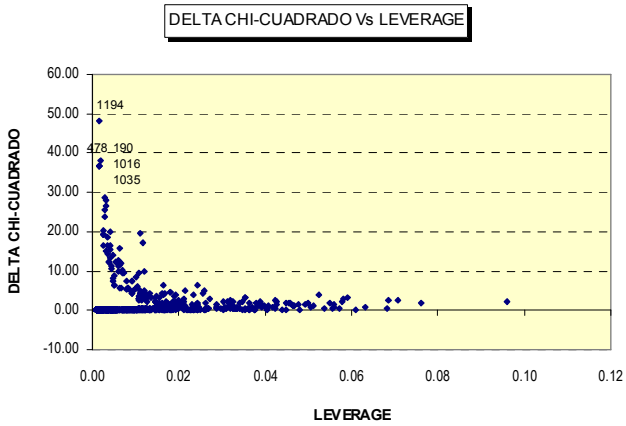


Gráfico IV.12



En estos gráficos del *DELTA CHI-CUADRADO* observamos los siguientes casos que son influyentes en la estimación de la variable respuesta: 190, 478, 1016, 1035 y 1194.

Gráfico IV.13

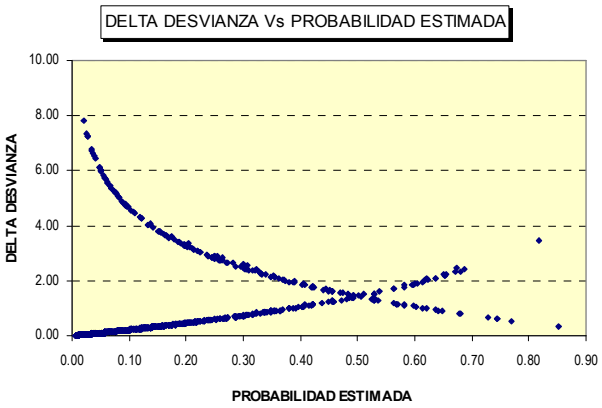
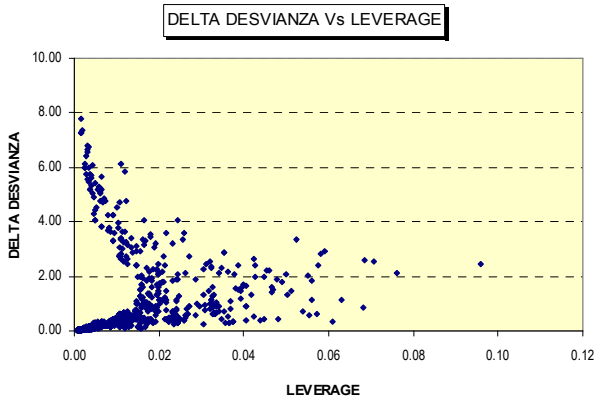


Gráfico IV.14



En estos gráficos del *DELTA DESVIANZA* no es posible detectar casos influyentes en la estimación de la variable respuesta.

Gráfico IV.15

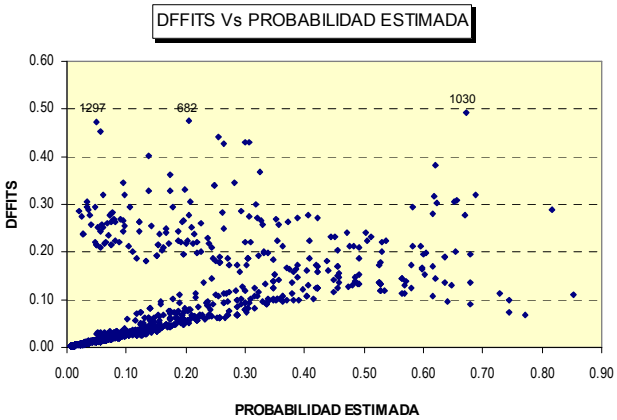
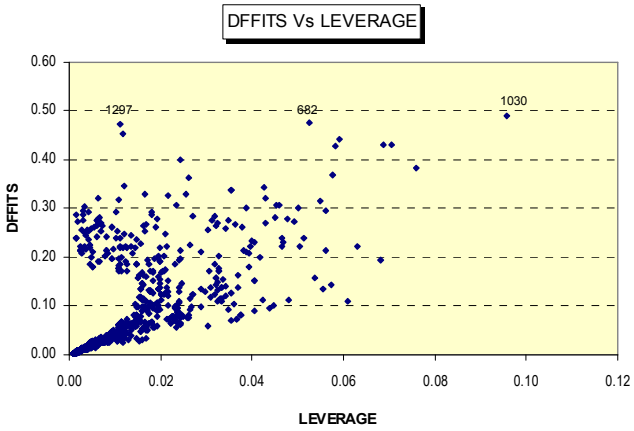


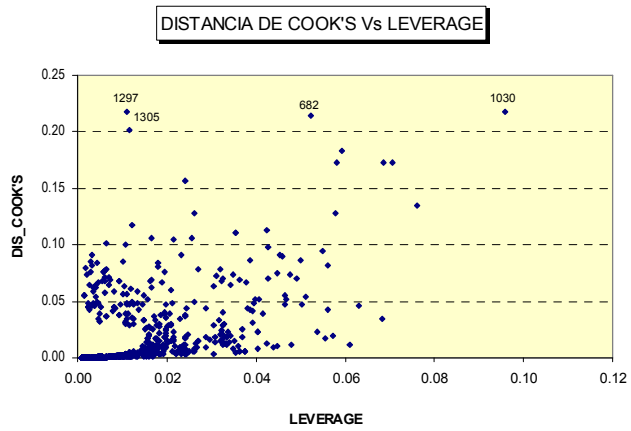
Gráfico IV.16



En estos gráficos de la estadística DFFITS observamos los siguientes casos que son influyentes en la estimación de la variable respuesta: 682, 1030 y 1297.

4.3.2.2 Gráficos que ayudan a detectar observaciones influyentes en la estimación de los parámetros del Modelo de Regresión Logística.

Gráfico IV.17



Podemos observar que los valores de la *DISTANCIA DE COOK'S* no son mayores o iguales a 1 que por teoría deberían de pasar ese valor para ser considerados valores de alta influencia, pero vemos que existen casos en los cuales se alejan del resto (los que están indicados) y que es evidencia suficiente para determinar que son casos influyentes.

En este gráfico observamos los siguientes casos que son influyentes en la estimación de los parámetros: 682, 1030, 1297 y 1305.

4.3.2.3 Gráficos que ayudan a detectar observaciones influyentes en la estimación del parámetro de cada variable.

La estadística que detecta en estos casos las observaciones influyentes para cada variable son los DFBETAS. Mostraremos los gráficos solo de algunas de las variables en los cuales se visualicen observaciones influyentes. Una forma de visualizarlos es identificar cuáles son mayores de 2.0, esto es una forma muy general. Existe un criterio que se adhiere más a cada investigación, esto es, para encontrar una observación influyente se toma como un punto crítico el valor $2/\sqrt{n}$. Este es el valor que tomaremos en cuenta aquí para detectar una observación influyente, este valor es aproximadamente 0.05.

Gráficos para la variable *Modalidad de Ingreso*

Para la variable Modalidad de ingreso (MODALIDA) tenemos que para cada indicadora de esta variable existen muchas observaciones que pasan el 0.05 en los DFBETAS por lo que en esta ocasión no se tomaran dichos casos como influyentes. Tampoco hay observaciones que son mayores que 2.0 en los valores de los DFBETAS. (Estos gráficos se encuentran en el *ANEXO 4 parte b*).

Gráficos para la variable C*olegio de procedencia*

No existen observaciones mayores a 0.05 en los valores del DFBETAS de esta variable, es decir, no existe casos influyentes para la estimación del parámetro de esta variable. (Estos gráficos se encuentran en el *ANEXO 4 parte b*).

Gráficos para la variable *Número de veces que postuló anteriormente a la Universidad*

Gráfico IV.18

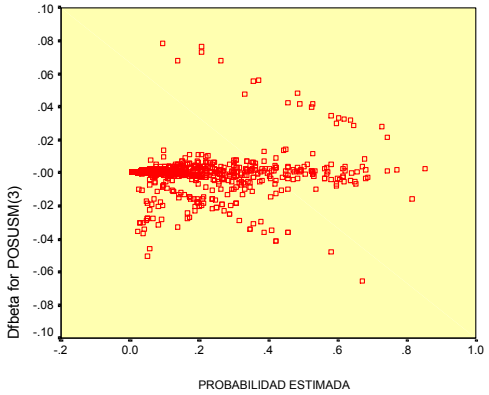
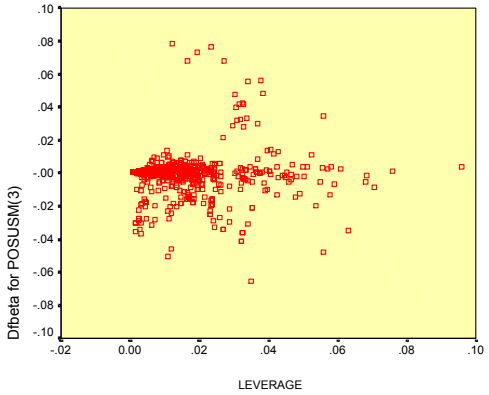


Gráfico IV.19



Para la variable indicadora POSUSM(3) tenemos las siguientes observaciones que pasan el valor de 0.05 en los DFBETAS, estas son: 250, 827, 935, 1099, 1106, 1148, 1252, 1297 y 1566.

Gráfico IV.20

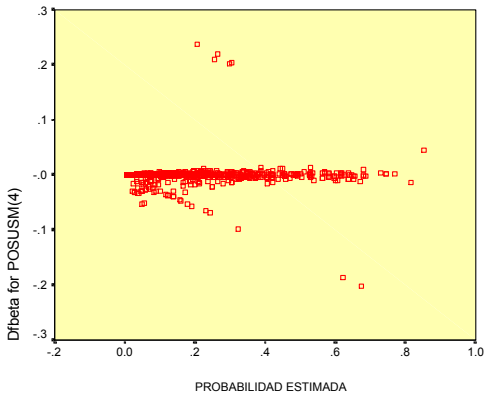
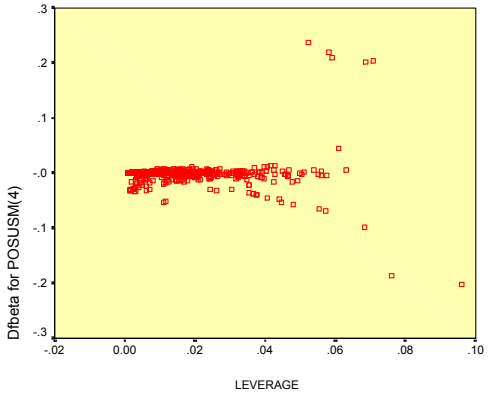


Gráfico IV.21



Para la variable indicadora POSUSM(4) tenemos las siguientes observaciones que pasan el valor de 0.05 en los DFBETAS, estas son: 28, 682, 699, 797, 866, 945, 1030.

Para las variables indicadoras POSUSM(1) y POSUSM(2) no existen observaciones en las cuales sean mayores que el valor critico 0.05 en los DFBETAS. (Estos gráficos se encuentran en el *ANEXO 4 parte b*).

Gráficos para la variable *Área académica*

Gráfico IV.22

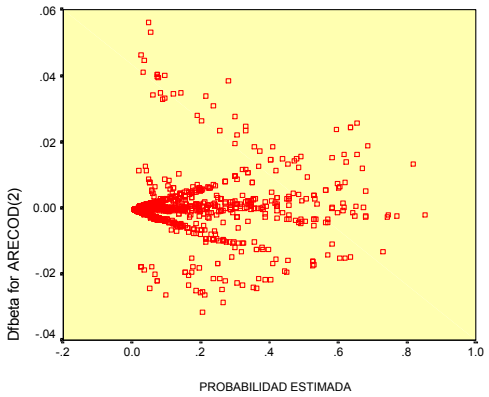
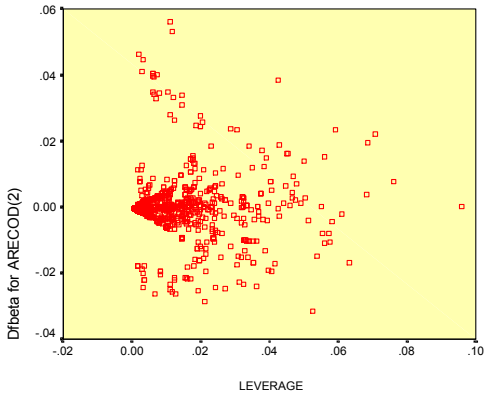


Gráfico IV.23



Para la variable indicadora ARECOD(2) tenemos las siguientes observaciones que pasan el valor de 0.05 en los DFBETAS, estas son: 1297 y 1305.

Para las variables indicadoras ARECOD(1) y ARECOD(3) no existen observaciones en las cuales sean mayores que el valor critico 0.05 en los DFBETAS. (Estos gráficos se encuentran en el *ANEXO 4 parte b*).

Gráficos para la variable *Promedio de notas en el colegio*

No existen observaciones mayores a 0.05 en los valores del DFBETAS de esta variable, es decir, no existe casos influyentes para la estimación del parámetro de esta variable. (Estos gráficos se encuentran en el *ANEXO 4 parte b*).

En resumen, las observaciones que más se repiten en estas medidas de influencia son las siguientes: 682, 1030, 1297 y 1305.

En conclusión, las medidas que tomaremos respecto a las observaciones discordantes u 'outliers' (estos son detectados en los gráficos de los residuos de Pearson) y a las observaciones influyentes en las estimaciones de los parámetros del modelo serán las siguientes: los casos discordantes u 'outliers' serán eliminadas, estas son 5 casos que ya fueron mencionados. Respecto a los casos influyentes, serán eliminadas las que tengan mayor incidencia (las que más se repiten) en las diferentes medidas que detectan los casos influyentes, estas son 4 y también ya fueron mencionadas. Estos casos son diferentes, por lo tanto, la cantidad total de casos a eliminar son 9. (190, 478, 682, 1016, 1030, 1035, 1194, 1297 y 1305)

Una vez eliminados estos casos ajustaremos nuevamente un modelo de regresión logística con el método Forward para la selección de variables.

4.4 AJUSTE DEL MODELO DE REGRESION LOGISTICA SIN CONSIDERAR LOS CASOS DISCORDANTES Y/O INFLUYENTES

Los resultados del último paso del método Forward[11] para la selección de variables se detalla en el *ANEXO 5*. Las variables seleccionadas por este método se presentan en el cuadro siguiente:

[11] Igualmente se utilizó el Método Forward Stepwise (Wald) para la selección de variables

Cuadro IV.16

VARIABLES SELECCIONADAS POR EL MÉTODO FORWARD

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
MODALIDA			77.989	3	0.000	
MODALIDA(1)	-1.790	0.423	17.889	1	0.000	0.167
MODALIDA(2)	-0.197	0.429	0.212	1	0.645	0.821
MODALIDA(3)	0.235	0.550	0.182	1	0.670	1.264
COLPRO	0.542	0.197	7.545	1	0.006	1.720
POSUSM			81.144	4	0.000	
POSUSM(1)	1.059	0.229	21.311	1	0.000	2.883
POSUSM(2)	2.066	0.259	63.535	1	0.000	7.891
POSUSM(3)	2.372	0.353	45.059	1	0.000	10.724
POSUSM(4)	2.100	0.602	12.182	1	0.000	8.163
ARECOD			52.645	3	0.000	
ARECOD(1)	0.776	0.231	11.277	1	0.001	2.173
ARECOD(2)	-1.204	0.281	18.362	1	0.000	0.300
ARECOD(3)	-0.027	0.272	0.010	1	0.922	0.974
P43	0.158	0.070	5.128	1	0.024	1.172
Constant	-4.299	1.125	14.602	1	0.000	0.014

Algunos resultados se muestran en los cuadros siguientes:

Cuadro IV.17

Prueba omnibus para los coeficientes del modelo

	Chi-square	df	Sig.
Step	5.143	1	0.023
Block	262.782	12	0.000
Model	262.782	12	0.000

Cuadro IV.18

-2 Log likelihood	843.532
R^2 de Cox & Snell	0.157
R^2 de Nagelkerke	0.306

Cuadro IV.19

Tabla de clasificación* para INGRESO

Casos		Pronosticado		Porcentaje de clasificación
		No ingreso	Si ingreso	
Observados	No ingreso	1185	174	87.20%
	Si ingreso	69	110	61.45%
			Total	84.20%

Al analizar estos resultados vemos que en algunos aspectos ha mejorado, como la estadística Chi-cuadrado para la prueba de la significancia de los coeficientes cuadro *(IV.17)*, el R^2 de Nagelkerke cuadro *(IV.18)* y la tabla de clasificación cuadro *(IV.19)*, pero estos cambios son mínimos. Un resultado determinante para la decisión de tomar este modelo o el anterior es el resultado sobre el *p_valor* asociado a la estadística Chi-cuadrado del *Goodness-of-fit* test de *Hosmer-Lemeshow* el cual evalúa si el modelo se ajusta adecuadamente. Se presenta en el siguiente cuadro:

Cuadro IV.20

Hosmer and Lemeshow Test		
Chi-square	df	Sig.
15.836	8	0.045

Este *p-valor* es de *0.045* lo que nos dice, a un nivel de significancia del 5%, que el modelo no se ajusta adecuadamente a los datos.

En conclusión, si eliminamos los datos discordantes de la base de datos ganamos en algunas medidas como los porcentajes en la tabla de clasificación, el R^2 de Nagelkerke y en la estadística Chi-cuadrado para la prueba de la significancia de los coeficientes pero perdemos en algo muy importante como el ajuste adecuado del modelo. Por lo tanto nos quedaremos con el modelo ajustado sin la eliminación de los datos discordantes *(Los resultados del método Forward sin la eliminación de los datos discordantes se encuentra en el ANEXO 3)*.

4.5 ANÁLISIS DEL MODELO DE REGRESIÓN LOGÍSTICA CONSIDERANDO TODAS LAS OBSERVACIONES

El resultado del apartado anterior nos dice que el ajuste del modelo sin los datos discordantes u 'outliers' no es adecuado por lo que nos quedamos con el Modelo de Regresión Logística sin la eliminación de estos datos. Como se mencionó en la sección 4.2.2 el método usado para la selección de variables fue el Forward que presenta en el quinto paso de este método las variables significativamente influyentes. A continuación presentaremos estas variables con más detalles:

Cuadro IV.21

VARIABLES SELECCIONADAS POR EL MÉTODO FORWARD

	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
							Lower	Upper
MODALIDA			77.894	3	0.000			
MODALIDA(1)	-1.891	0.400	22.312	1	0.000	0.151	0.069	0.331
MODALIDA(2)	-0.368	0.409	0.812	1	0.368	0.692	0.311	1.541
MODALIDA(3)	-0.001	0.531	0.000	1	0.999	0.999	0.353	2.830
COLPRO	0.458	0.193	5.632	1	0.018	1.581	1.083	2.308
POSUSM			74.604	4	0.000			
POSUSM(1)	0.937	0.220	18.146	1	0.000	2.553	1.659	3.929
POSUSM(2)	1.892	0.250	57.395	1	0.000	6.631	4.065	10.818
POSUSM(3)	2.174	0.344	39.975	1	0.000	8.797	4.483	17.261
POSUSM(4)	1.919	0.558	11.835	1	0.001	6.814	2.283	20.332
ARECOD			46.441	3	0.000			
ARECOD(1)	0.649	0.223	8.436	1	0.004	1.913	1.235	2.965
ARECOD(2)	-1.124	0.267	17.765	1	0.000	0.325	0.193	0.548
ARECOD(3)	-0.134	0.265	0.256	1	0.613	0.874	0.520	1.471
P43	0.142	0.068	4.389	1	0.036	1.153	1.009	1.317
Constant	-3.698	1.082	11.681	1	0.001	0.025		

Entonces el Modelo de Regresión Logística ajustada es el siguiente:

$$p_i = \frac{1}{1+e^{-z}}$$

Donde

z = -3.6982 – 1.8912MODALIDAD(1) – 0.3681MODALIDAD(2) - 0.000MODALIDAD(3) + 0.4581COLPRO + 0.9372POSUSM(1) + 1.8918POSUSM(2) + 2.1744POSUSM(3) + 1.9189POSUSM(4) + 0.6489ARECOD(1) – 1.1240ARECOD(2) – 0.1344ARECOD(3) + 0.1424P43

La descripción de cada una de estas variables se encuentra en el apartado 2.2, cuadro *(II.2)*

Del Modelo Logístico considerado podemos observar que el signo de los coeficientes de algunas variables es positivo, eso significa que la variable aumenta la probabilidad del suceso en estudio, lo que es lo mismo que aumenta la probabilidad de ingresar a la UNMSM de los alumnos postulantes. Se tiene el resumen siguiente:

Cuadro IV.22

VARIABLES CON SIGNO POSITIVO EN LOS COEFICIENTES DE LOS PARÁMETROS		
Variable	B	exp(B)=OR
COLPRO	0.458	1.581
POSUSM(1)	0.937	2.581
POSUSM(2)	1.892	6.631
POSUSM(3)	2.174	8.797
POSUSM(4)	1.919	6.814
ARECOD(1)	0.649	1.913
P43	0.142	1.153

$e^{0.4581}=1.581$ Estima que, Ingresar a la Universidad es 1.581 veces más favorable de ocurrir en alumnos que provienen de colegios privados que estatales.

$e^{0.9372}=2.581$ Estima que, Ingresar a la Universidad es 2.581 veces más factible de ocurrir en alumnos que anteriormente ya postularon una vez, que otros que postularan por primera vez.

$e^{1.892}=6.631$ Estima que, Ingresar a la Universidad es 6.631 veces más factible de ocurrir en alumnos que anteriormente ya postularon dos veces, que otros que postularan por primera vez.

$e^{2.174}=8.797$ Estima que, Ingresar a la Universidad es 8.797 veces más factible de ocurrir en alumnos que anteriormente ya postularon tres veces, que otros que postularan por primera vez.

$e^{1.919}=6.814$ Estima que, Ingresar a la Universidad es 6.814 veces más factible de ocurrir en alumnos que anteriormente ya postularon cuatro veces o más, que otros que postularan por primera vez.

$e^{0.649}=1.913$ Estima que, Ingresar a la Universidad es 1.913 veces más factible de ocurrir en alumnos que postulan al área de Ciencias Básicas, que los que postulan al área de Humanidades.

$e^{0.142}=1.153$ Estima que, Ingresar a la Universidad es 1.153 veces más factible de ocurrir en alumnos que tienen mejor promedio.

Vemos también que existen variables con coeficientes con signo negativo, eso significa que la variable asociada al coeficiente disminuye la probabilidad del suceso en estudio, es decir, disminuye la probabilidad de ingresar a la UNMSM de los alumnos postulantes. Se tiene el resumen siguiente:

Cuadro IV.23

VARIABLES CON SIGNO NEGATIVO EN LOS COEFICIENTES DE LOS PARÁMETROS

Variable	B	exp(B)=OR
MODALIDA(1)	-1.891	0.151
MODALIDA(2)	-0.368	0.692
MODALIDA(3)	-0.001	0.999
ARECOD(2)	-1.124	0.325
ARECOD(3)	-0.134	0.874

En resumen, las variables que no favorecen el "Ingreso a la Universidad" serían: Modalidad de Ingreso (en Prueba general, CEPUSM y Primeros puestos) y Área académica (en Ciencias de la salud y Ciencias Económico-empresariales).

Algo muy importante también para el análisis del modelo es revisar la ***Tabla de Clasificación***. Esta clasificación de los individuos o de los casos se realiza de acuerdo a la probabilidad estimada (punto de corte para la clasificación es de 0.2). Entonces para la clasificación de los individuos, el que tenga una probabilidad estimada mayor de 0.2 será clasificado en el segundo grupo (ingreso) y el que tengo menor de 0.2 será clasificado en el primero (no ingreso). El porcentaje de casos correctamente clasificados será un índice de la efectividad del modelo.

Cuadro IV.24*

Tabla de clasificación* para INGRESO				
Casos		**Pronosticado**		Porcentaje de clasificación
		No ingreso	Si ingreso	
Observados	No ingreso	1179	181	86.69%
	Si ingreso	75	111	59.68%
			Total	83.44%

En el cuadro *(IV.24)* se muestra el resumen de los resultados de clasificación. El porcentaje de casos correctamente clasificados para el grupo de los que 'No ingreso' es igual a 86.69%, mientras que para el grupo de los que 'Si ingresaron' es de 59.68%. En términos generales de 1546 casos, 1290 han sido correctamente clasificados, o lo que es equivalente, el 83.44% ha sido correctamente clasificado. Por lo tanto, este porcentaje nos indica que el modelo, para fines de clasificación de casos, es efectivo.

A continuación presentaremos un ejemplo para ver la calidad de este modelo en cuanto a la estimación de la probabilidad de Ingresar a esta Universidad: *(caso 324)*

Supongamos que un alumno que va a postular tiene las siguientes características: postulará por la modalidad de Prueba general, su colegio de procedencia es Particular, nunca postulo anteriormente a la universidad, postulará al área de Ciencias básicas e ingenierías y su promedio de notas en su educación secundaria fue de 14.3. Tomando en cuenta el modelo hallado y las características mencionadas, la probabilidad de estimada para este caso es de 0.07971, por lo que estimamos que este alumno no ingresara a la universidad. Recordemos que por motivos de clasificación, si una probabilidad estimada pasa el valor de 0.20 será clasificado en el segundo grupo, es decir, en el grupo de los que si ingresarían a la universidad.

* El valor de corte para la clasificación es de 0.2

CAPITULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

El análisis estadístico realizado en el capítulo IV, nos revela las siguientes conclusiones y recomendaciones:

1. Una variable muy importante pudiera ser el estado laboral del postulante, dado que es una característica que de alguna manera complica a cualquier estudiante, pero como la mayoría de los postulantes (95.00% de ellos) no trabajan, esta variable para este estudio no será tomado en cuenta.

2. Para este estudio la cual trata de encontrar cuales son los factores más importantes que influyen en el éxito (ingresar a la universidad), las variables significativamente influyentes a un nivel de significancia del 5% son: Modalidad a la que postula, el colegio de procedencia, la cantidad de veces que ha postulado, área académica a la que postula y el promedio general en la educación secundaria. Todas ellas tienen un *p_valor* asociado a la estadística de *Wald* menores que 0.05. *(Ver cuadro 'Variables in the Equation' del paso 5 en el ANEXO 3).*

3. Las variables que no son influyentes en el examen de admisión son: el tipo de preparación, edad del postulante, sexo, nivel educativo del padre y el gasto per capita. El *p_valor* asociado es mayor que 0.05. *(Ver cuadro 'Variables not in the Equation' del paso 5 en el ANEXO 3).*

4. De entre las variables que influyen en el éxito en el examen de admisión, la más importante es *la cantidad de veces que ha postulado anteriormente*, por tener en sus variables indicadoras los más altos coeficientes en el modelo. La segunda más importante es *modalidad por la que postula* dado que uno de los coeficientes de entre sus variables indicadoras tiene el más alto valor. La tercera variable más importante en el estudio es *área a la que postula* por tener la misma característica del anterior. La cuarta y última variable más importante en el estudio es el *colegio de procedencia* y el *Promedio general de notas en el colegio* respectivamente.

5. La variable *Colegio de procedencia* tiene su coeficientes positivo en el modelo lo que indica que si el postulante proviene de un colegio particular tiene más oportunidad de ingresar a la universidad que el que viene de un colegio estatal *(OR=1.581)*. Aquí vemos indirectamente que la parte económica juega un papel muy importante en el rendimiento de un postulante en el examen de admisión, pues el que venga de un colegio particular y que proporcionalmente son ellos los que más ingresan nos da a entender tal suposición.

6. La variable *cantidad de veces que ha postulado anteriormente* tiene en sus variables indicadoras coeficientes positivo, esto indica que si el postulante ya ha rendido anteriormente un examen de admisión este tendrá más oportunidad que

aquel que está dando su examen por primera vez (el mayor OR de estas variables es de 8.797, esto quiere decir que el que haya postulado anteriormente tres veces tienen más posibilidades de ingresar que cualquier otro). Además podríamos decir que como postuló anteriormente a la universidad el estudiante adquiere mayor 'experiencia' que aquel que nunca ha postulado a la universidad.

7. La variable *Promedio general de notas en el colegio* tiene también un coeficiente positivo en el modelo, esto significa que a mayor nota en el colegio, mayor posibilidad tiene de ingresar a la universidad *(OR=1.153)*. Esto es muy lógico dado que un alumno que ha sobresalido en la época de su educación secundaria tiene más capacidad y rapidez de asimilar los conocimientos necesarios para rendir un examen de admisión y por ende le será menos complicado y le costara menos tiempo en ingresar a la universidad.

5.2 RECOMENDACIONES

A continuación enumeramos algunas recomendaciones necesarias para un estudio más completo de este tipo:

1. Se recomienda obtener información de estas principales variables y de otras que no han sido consideradas en este estudio para poder determinar con más exactitud las principales variables o factores que determina el ingreso a la universidad.

2. Además se recomienda que la información recogida en la ficha de matrícula sea más estructurada ya que presenta serias deficiencias pues se nota muchas inconsistencias en la información.

3. Para solucionar esto, se recomienda disponer de personal capacitada a la hora de recoger esta información, ya que por consulta de algunas personas que trabajaron en el proceso de admisión, no existía responsabilidad a la hora de verificar esta información.

4. Se recomienda también estructurar preguntas acerca de la condición económica del postulante pues, como ya se concluyó, es casi seguro que este factor juega un papel muy importante en la formación del postulante para rendir el examen de admisión. Pues hemos visto también en el análisis exploratorio de datos que la variable '*Gasto per cápita*' tiene una relación lineal con el log odd de la variable que indica el éxito en el examen de admisión.

BIBLIOGRAFÍA

1. DAVID W. HOSMER. Applied Logistic Regression. Second Edition. A Wiley-Interscience Publication. John Wiley & Sons, INC. 2000.

2. LAWRENCE C. HAMILTON. Regression with Graphics. An Second Course in Applied Statistics. Wadsworth Publishing Company Belmont, California. 1992.

3. MANUEL ATO GARCÍA, JUAN JOSÉ LÓPEZ GARCÍA. Análisis Estadístico para Datos Categóricos. Editorial Síntesis S.A. 1996.

4. B. VISAUTA VINACUA (1998). Análisis estadístico con SPSS para Windows. McGraw Hill/Interamericana de España, SAU. 1998.

5. ELENA ABASCAL E ILDEFONSO GRANDE, Métodos Multivariantes para la Investigación Comercial. Editorial Ariel, S.A. Barcelona-España. 1989.

6. MAGDALENA FERRAN ARANAZ. SPSS para Windows Análisis Estadístico. McGraw Hill/Interamericana de España, SAU. 1998.

7. ORDÓÑEZ MERCADO A. Factores Importantes en el Examen de Admisión-UNI. Tecnia, Vol 8 N°1 1998

8. CF. BLOOM, 1972, 1977; ROGERS, 1989; CARRASCO, 1993; GARDNER, 1994

ANEXOS

ANEXO 1

ESTRUCTURA DE LA BASE DE DATOS

ID	MODALIDA	COLPRO	TIPPRE	SEXPOS	EDAPOS	IND_POBR	TIEM_EGR	POSUSM	ARECOD	N_PAD_AG	P43	GASTPER	IND_INGR
N° de caso	Modalidad de postulación	Colegio de procedencia	Tipo de preparación	Sexo	Edad	Indice de pobreza	Tiempo desde que termino su secundaria	N° de veces que postulo anteriormente	Área académica	Nivel de estudio del padre	Promedio de notas en el colegio	Gasto per-cápita estimado	Indica si ingreso o no
1	0	1	3	0	19	12.48	1	1	4	3	14.00	492.33	0
2	0	1	1	1	17	14.60	1	1	4	3	14.00	220.05	0
3	0	0	3	1	19	14.60	2	0	4	2	13.00	305.89	0
4	0	0	3	1	25	4.37	8	0	3	4	13.00	457.96	0
5	0	1	3	1	16	17.81	0	0	2	1	16.00	369.06	0
6	0	0	3	0	17	9.66	1	1	2	4	16.62	699.55	0
7	0	0	3	0	18	14.60	1	0	4	2	15.00	209.56	0
8	3	0	2	1	20	14.60	2	1	2	4	14.00	305.89	0
9	0	0	1	1	18	17.96	0	0	1	4	16.00	186.86	0
10	0	0	1	1	20	1.93	3	0	4	4	14.00	318.54	0
11	0	0	3	0	21	13.64	4	2	2	4	16.00	481.47	0
12	0	0	1	0	20	11.84	4	2	2	4	15.00	238.82	0
13	0	0	1	1	17	12.48	1	0	3	2	16.70	329.27	0
14	0	0	3	0	16	13.64	0	0	1	4	14.00	505.53	0
15	0	0	3	0	24	16.35	6	3	1	1	12.50	257.42	0
16	0	1	3	0	20	1.71	3	2	4	4	12.00	866.08	1
17	0	0	3	1	19	11.78	1	0	2	2	15.00	305.89	0
18	0	0	3	0	17	11.78	2	2	1	2	14.00	484.01	0
19	0	1	3	1	25	17.81	9	4	2	2	16.35	374.30	0
20	0	0	3	0	22	13.64	6	1	4	3	13.00	327.76	0
21	0	0	3	0	16	1.00	0	0	1	2	15.00	288.70	0
22	0	1	3	1	16	8.00	0	0	2	3	16.00	636.63	0
23	0	1	3	0	16	12.48	0	0	4	2	16.00	481.47	0
24	0	0	3	0	21	4.99	4	1	4	2	15.00	480.48	0
25	0	0	3	1	21	13.57	5	4	2	1	14.00	267.83	0
26	0	0	3	1	16	20.56	0	0	1	4	14.50	554.13	0
27	0	0	1	0	17	11.78	0	0	1	2	15.00	352.70	0
28	0	0	3	0	21	14.60	4	4	1	3	14.00	273.20	1
29	0	0	3	1	19	17.96	2	1	4	2	15.00	327.76	1
30	0	0	3	0	19	13.57	3	1	3	3	13.00	251.04	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
1588	0	0	3	1	16	13.74	0	0	4	4	15.00	326.90	0

ANEXO 2

MODELO SATURADO DE LA REGRESIÓN LOGÍSTICA.

Regresión Logística

Case Processing Summary

Unweighted Cases [a]		N	Percent
Selected Cases	Included in Analysis	1546	97.4
	Missing Cases	42	2.6
	Total	1588	100.0
Unselected Cases		0	.0
Total		1588	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No ingreso	0
Si ingreso	1

Categorical Variables Codings

			Parameter coding				
		Frequency	(1)	(2)	(3)	(4)	(5)
Nivel de estudios del padre	Analfabeto/inicial	7	.000	.000	.000	.000	.000
	Primaria	164	1.000	.000	.000	.000	.000
	Secundaria/carrera corta	580	.000	1.000	.000	.000	.000
	Superior no universitaria	324	.000	.000	1.000	.000	.000
	Superior Universitaria	404	.000	.000	.000	1.000	.000
	Post grado (Maestría, Doctorado)	67	.000	.000	.000	.000	1.000
¿Cuántas veces postuló anteriormente a la universidad?	Ninguna	845	.000	.000	.000	.000	
	Una	435	1.000	.000	.000	.000	
	Dos	176	.000	1.000	.000	.000	
	Tres	65	.000	.000	1.000	.000	
	Cuatro a mas	25	.000	.000	.000	1.000	
Tipo de preparación pre universitaria	Solo	292	1.000	.000	.000		
	Grupo de estudio	37	.000	1.000	.000		
	Academia	913	.000	.000	1.000		
	CEPUSM	304	.000	.000	.000		
Modalidad de postulación	Prueba general	1175	1.000	.000	.000		
	CEPUSM	282	.000	1.000	.000		
	Primeros puestos	52	.000	.000	1.000		
	Otras modalidades	37	.000	.000	.000		
Área académica	Ciencias basicas e Ingenierías	367	1.000	.000	.000		
	Ciencias de la salud	457	.000	1.000	.000		
	Ciencias ecónomico-empresarial	290	.000	.000	1.000		
	Humanidades	432	.000	.000	.000		
Rango de edad	Menos de 17 años	244	1.000	.000			
	De 17 a 19 años	964	.000	1.000			
	Mas de 19 años	338	.000	.000			
Rango del gasto per cápita	Menos de 291.72	382	1.000	.000			
	De 291.72 a 543.81	775	.000	1.000			
	Mayor de 543.81	389	.000	.000			

Bloque 0: Bloque de Inicio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.989	.078	647.630	1	.000	.137

Classification Table[a,b]

Observed			Predicted		
			¿Ingreso?		Percentage Correct
			No ingreso	Si ingreso	
Step 0	¿Ingreso?	No ingreso	1360	0	100.0
		Si ingreso	186	0	.0
	Overall Percentage				88.0

a. Constant is included in the model.

b. The cut value is .200

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	MODALIDA	138.077	3	.000
		MODALIDA(1)	134.543	1	.000
		MODALIDA(2)	90.806	1	.000
		MODALIDA(3)	8.551	1	.003
		COLPRO	12.993	1	.000
		TIPPRE	101.385	3	.000
		TIPPRE(1)	.681	1	.409
		TIPPRE(2)	.053	1	.817
		TIPPRE(3)	53.120	1	.000
		SEXPOS	.706	1	.401
		RANG_EDA	18.674	2	.000
		RANG_EDA(1)	9.476	1	.002
		RANG_EDA(2)	.646	1	.422
		IND_POBR	2.919	1	.088
		POSUSM	111.432	4	.000
		POSUSM(1)	2.269	1	.132
		POSUSM(2)	53.891	1	.000
		POSUSM(3)	26.359	1	.000
		POSUSM(4)	3.440	1	.064
		ARECOD	29.385	3	.000
		ARECOD(1)	22.552	1	.000
		ARECOD(2)	16.881	1	.000
		ARECOD(3)	.335	1	.563
		N_PAD_AG	7.769	5	.169
		N_PAD_AG(1)	2.117	1	.146
		N_PAD_AG(2)	2.494	1	.114
		N_PAD_AG(3)	.038	1	.845
		N_PAD_AG(4)	4.111	1	.043
		N_PAD_AG(5)	1.273	1	.259
		P43	7.435	1	.006
		RANG_GAS	7.153	2	.028
		RANG_GAS(1)	4.698	1	.030
		RANG_GAS(2)	.001	1	.970
	Overall Statistics		263.785	26	.000

Bloque 1: Método Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	252.955	26	.000
	Block	252.955	26	.000
	Model	252.955	26	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	883.488	.151	.290

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	5.956	8	.652

Contingency Table for Hosmer and Lemeshow Test

		¿Ingreso? = No ingreso		¿Ingreso? = Si ingreso		
		Observed	Expected	Observed	Expected	Total
Step 1	1	155	153.574	0	1.426	155
	2	154	152.056	1	2.944	155
	3	153	150.814	2	4.186	155
	4	147	149.102	8	5.898	155
	5	146	147.164	9	7.836	155
	6	142	144.293	13	10.707	155
	7	140	139.254	15	15.746	155
	8	130	131.050	25	23.950	155
	9	113	115.594	42	39.406	155
	10	80	77.095	71	73.905	151

Classification Table[a]

Observed			Predicted		
			¿Ingreso?		Percentage Correct
			No ingreso	Si ingreso	
Step 1	¿Ingreso?	No ingreso	1173	187	86.3
		Si ingreso	74	112	60.2
	Overall Percentage				83.1

a. The cut value is .200

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	MODALIDA			34.906	3	.000			
	MODALIDA(1)	-1.825	.435	17.632	1	.000	.161	.069	.378
	MODALIDA(2)	-1.067	.722	2.183	1	.140	.344	.084	1.417
	MODALIDA(3)	-.003	.585	.000	1	.995	.997	.317	3.133
	COLPRO	.352	.204	2.979	1	.084	1.422	.953	2.123
	TIPPRE			1.841	3	.606			
	TIPPRE(1)	-.821	.628	1.711	1	.191	.440	.129	1.506
	TIPPRE(2)	-.531	.826	.414	1	.520	.588	.116	2.967
	TIPPRE(3)	-.740	.593	1.557	1	.212	.477	.149	1.526
	SEXPOS	-.152	.190	.638	1	.424	.859	.592	1.247
	RANG_EDA			.027	2	.987			
	RANG_EDA(1)	-.067	.407	.027	1	.870	.935	.422	2.075
	RANG_EDA(2)	-.023	.243	.009	1	.923	.977	.606	1.574
	IND_POBR	-.008	.019	.156	1	.693	.993	.956	1.030
	POSUSM			57.329	4	.000			
	POSUSM(1)	.910	.242	14.184	1	.000	2.484	1.547	3.989
	POSUSM(2)	1.924	.277	48.310	1	.000	6.849	3.981	11.784
	POSUSM(3)	2.229	.391	32.481	1	.000	9.292	4.317	20.001
	POSUSM(4)	1.968	.602	10.688	1	.001	7.156	2.199	23.282
	ARECOD			44.258	3	.000			
	ARECOD(1)	.610	.232	6.895	1	.009	1.841	1.167	2.903
	ARECOD(2)	-1.169	.273	18.406	1	.000	.311	.182	.530
	ARECOD(3)	-.112	.269	.172	1	.678	.894	.528	1.516
	N_PAD_AG			2.075	5	.839			
	N_PAD_AG(1)	-.992	1.263	.617	1	.432	.371	.031	4.407
	N_PAD_AG(2)	-1.032	1.241	.692	1	.406	.356	.031	4.056
	N_PAD_AG(3)	-.763	1.246	.375	1	.540	.466	.041	5.357
	N_PAD_AG(4)	-.829	1.246	.442	1	.506	.437	.038	5.019
	N_PAD_AG(5)	-.971	1.297	.561	1	.454	.379	.030	4.808
	P43	.152	.071	4.597	1	.032	1.164	1.013	1.338
	RANG_GAS			2.529	2	.282			
	RANG_GAS(1)	-.405	.284	2.025	1	.155	.667	.382	1.165
	RANG_GAS(2)	-.295	.215	1.880	1	.170	.744	.488	1.135
	Constant	-1.796	1.711	1.102	1	.294	.166		

a. Variable(s) entered on step 1: MODALIDA, COLPRO, TIPPRE, SEXPOS, RANG_EDA, IND_POBR, POSUSM, ARECOD, N_PAD_AG, P43, RANG_GAS.

```
               Observed Groups and Predicted Probabilities

       320 •                                                              •
           •  N                                                           •
           •  N                                                           •
F          •  N                                                           •
R      240 •  N                                                           •
E          •  NN                                                          •
Q          •  NN                                                          •
U          • NNN                                                          •
E      160 • NNN                                                          •
N          • NNNN                                                         •
C          • NNNN                                                         •
Y          • NNNNS                                                        •
        80 • NNNNN                                                        •
           • NNNNNNN                                                      •
           • NNNNNNNNNNS                                                  •
           • NNNNNNNNNNNNNNNNNSSSNSS SS          S                        •
Predicted
• • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • • •
• • • • • • • • • • • • •
  Prob:   0            .25            .5             .75             1
  Group:  NNNNNNNNNNNNSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS

          Predicted Probability is of Membership for Si ingreso
          The Cut Value is .20
          Symbols: N - No ingreso
                   S - Si ingreso
          Each Symbol Represents 20 Cases.
```

ANEXO 3

REGRESIÓN LOGÍSTICA CON EL MÉTODO FORWARD PARA LA SELECCIÓN DE VARIABLES

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1546	97.4
	Missing Cases	42	2.6
	Total	1588	100.0
Unselected Cases		0	.0
Total		1588	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No ingreso	0
Si ingreso	1

Bloque 0: Bloque de Inicio

Classification Table[a,b]

Observed			Predicted		
			¿Ingreso?		Percentage Correct
			No ingreso	Si ingreso	
Step 0	¿Ingreso?	No ingreso	1360	0	100.0
		Si ingreso	186	0	.0
	Overall Percentage				88.0

a. Constant is included in the model.

b. The cut value is .200

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.989	.078	647.630	1	.000	.137

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	MODALIDA	138.077	3	.000
		MODALIDA(1)	134.543	1	.000
		MODALIDA(2)	90.806	1	.000
		MODALIDA(3)	8.551	1	.003
		COLPRO	12.993	1	.000
		TIPPRE	101.385	3	.000
		TIPPRE(1)	.681	1	.409
		TIPPRE(2)	.053	1	.817
		TIPPRE(3)	53.120	1	.000
		SEXPOS	.706	1	.401
		RANG_EDA	18.674	2	.000
		RANG_EDA(1)	9.476	1	.002
		RANG_EDA(2)	.646	1	.422
		IND_POBR	2.919	1	.088
		POSUSM	111.432	4	.000
		POSUSM(1)	2.269	1	.132
		POSUSM(2)	53.891	1	.000
		POSUSM(3)	26.359	1	.000
		POSUSM(4)	3.440	1	.064
		ARECOD	29.385	3	.000
		ARECOD(1)	22.552	1	.000
		ARECOD(2)	16.881	1	.000
		ARECOD(3)	.335	1	.563
		N_PAD_AG	7.769	5	.169
		N_PAD_AG(1)	2.117	1	.146
		N_PAD_AG(2)	2.494	1	.114
		N_PAD_AG(3)	.038	1	.845
		N_PAD_AG(4)	4.111	1	.043
		N_PAD_AG(5)	1.273	1	.259
		P43	7.435	1	.006
		RANG_GAS	7.153	2	.028
		RANG_GAS(1)	4.698	1	.030
		RANG_GAS(2)	.001	1	.970
	Overall Statistics		263.785	26	.000

Bloque 1: Método = Forward Stepwise (Wald)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	116.848	3	.000
	Block	116.848	3	.000
	Model	116.848	3	.000
Step 2	Step	68.569	4	.000
	Block	185.417	7	.000
	Model	185.417	7	.000
Step 3	Step	47.996	3	.000
	Block	233.413	10	.000
	Model	233.413	10	.000
Step 4	Step	6.964	1	.008
	Block	240.377	11	.000
	Model	240.377	11	.000
Step 5	Step	4.397	1	.036
	Block	244.774	12	.000
	Model	244.774	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1019.595	.073	.140
2	951.026	.113	.217
3	903.030	.140	.269
4	896.066	.144	.277
5	891.669	.146	.281

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	1	.998
2	7.095	4	.131
3	7.224	7	.406
4	10.398	8	.238
5	11.803	8	.160

Contingency Table for Hosmer and Lemeshow Test

		¿Ingreso? = No ingreso		¿Ingreso? = Si ingreso		Total
		Observed	Expected	Observed	Expected	
Step 1	1	1097	1096.980	78	78.020	1175
	2	39	39.000	13	13.000	52
	3	224	224.000	95	95.000	319
Step 2	1	679	672.759	18	24.241	697
	2	288	289.280	27	25.720	315
	3	100	99.481	15	15.519	115
	4	121	131.045	38	27.955	159
	5	114	114.397	43	42.603	157
	6	58	53.039	45	49.961	103
Step 3	1	193	190.589	0	2.411	193
	2	133	130.268	1	3.732	134
	3	82	83.408	4	2.592	86
	4	203	201.049	5	6.951	208
	5	185	186.420	14	12.580	199
	6	165	165.046	13	12.954	178
	7	142	144.621	25	22.379	167
	8	125	129.258	36	31.742	161
	9	132	129.342	88	90.658	220
Step 4	1	141	139.546	0	1.454	141
	2	154	150.525	0	3.475	154
	3	223	221.591	5	6.409	228
	4	109	108.903	5	5.097	114
	5	164	166.999	13	10.001	177
	6	157	161.061	17	12.939	174
	7	140	137.719	17	19.281	157
	8	121	123.558	31	28.442	152
	9	99	103.919	50	45.081	149
	10	52	46.178	48	53.822	100
Step 5	1	155	153.429	0	1.571	155
	2	149	145.818	0	3.182	149
	3	153	152.581	4	4.419	157
	4	151	149.238	4	5.762	155
	5	145	146.979	10	8.021	155
	6	141	144.411	14	10.589	155
	7	137	138.338	17	15.662	154
	8	136	131.708	20	24.292	156
	9	107	116.374	48	38.626	155
	10	86	81.124	69	73.876	155

Classification Table[a]

			Predicted		
			¿Ingreso?		Percentage Correct
	Observed		No ingreso	Si ingreso	
Step 1	¿Ingreso?	No ingreso	1097	263	80.7
		Si ingreso	78	108	58.1
	Overall Percentage				77.9
Step 2	¿Ingreso?	No ingreso	1216	144	89.4
		Si ingreso	107	79	42.5
	Overall Percentage				83.8
Step 3	¿Ingreso?	No ingreso	1181	179	86.8
		Si ingreso	86	100	53.8
	Overall Percentage				82.9
Step 4	¿Ingreso?	No ingreso	1181	179	86.8
		Si ingreso	81	105	56.5
	Overall Percentage				83.2
Step 5	¿Ingreso?	No ingreso	1179	181	86.7
		Si ingreso	75	111	59.7
	Overall Percentage				83.4

a. The cut value is .200

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MODALIDA			116.789	3	.000			
	MODALIDA(1)	-2.147	.359	35.832	1	.000	.117	.058	.236
	MODALIDA(2)	-.412	.364	1.286	1	.257	.662	.325	1.350
	MODALIDA(3)	-.602	.466	1.667	1	.197	.548	.220	1.366
	Constant	-.496	.339	2.145	1	.143	.609		
Step 2[b]	MODALIDA			86.740	3	.000			
	MODALIDA(1)	-1.826	.381	22.961	1	.000	.161	.076	.340
	MODALIDA(2)	-.375	.387	.939	1	.332	.687	.322	1.467
	MODALIDA(3)	.098	.495	.039	1	.843	1.103	.418	2.912
	POSUSM			65.733	4	.000			
	POSUSM(1)	.903	.212	18.098	1	.000	2.467	1.628	3.741
	POSUSM(2)	1.768	.240	54.394	1	.000	5.860	3.663	9.376
	POSUSM(3)	1.793	.327	30.073	1	.000	6.005	3.164	11.396
	POSUSM(4)	1.547	.522	8.771	1	.003	4.695	1.687	13.067
	Constant	-1.498	.388	14.882	1	.000	.224		
Step 3[c]	MODALIDA			100.978	3	.000			
	MODALIDA(1)	-1.932	.394	24.034	1	.000	.145	.067	.314
	MODALIDA(2)	-.259	.401	.417	1	.518	.772	.352	1.693
	MODALIDA(3)	.275	.509	.290	1	.590	1.316	.485	3.571
	POSUSM			68.348	4	.000			
	POSUSM(1)	.899	.218	16.968	1	.000	2.457	1.602	3.768
	POSUSM(2)	1.797	.246	53.447	1	.000	6.033	3.726	9.768
	POSUSM(3)	1.976	.335	34.758	1	.000	7.215	3.741	13.918
	POSUSM(4)	1.746	.551	10.060	1	.002	5.733	1.949	16.864
	ARECOD			43.511	3	.000			
	ARECOD(1)	.669	.221	9.128	1	.003	1.952	1.265	3.013
	ARECOD(2)	-1.005	.261	14.826	1	.000	.366	.219	.610
	ARECOD(3)	-.188	.264	.508	1	.476	.829	.494	1.389
	Constant	-1.432	.413	12.007	1	.001	.239		
Step 4[d]	MODALIDA			89.698	3	.000			
	MODALIDA(1)	-1.892	.399	22.501	1	.000	.151	.069	.329
	MODALIDA(2)	-.321	.406	.623	1	.430	.726	.327	1.609
	MODALIDA(3)	.270	.513	.277	1	.599	1.310	.479	3.581
	COLPRO	.509	.191	7.113	1	.008	1.663	1.144	2.417
	POSUSM			71.804	4	.000			
	POSUSM(1)	.933	.220	18.056	1	.000	2.542	1.653	3.909
	POSUSM(2)	1.867	.249	56.429	1	.000	6.469	3.975	10.530
	POSUSM(3)	2.038	.338	36.431	1	.000	7.672	3.959	14.867
	POSUSM(4)	1.872	.555	11.369	1	.001	6.504	2.190	19.314
	ARECOD			44.051	3	.000			
	ARECOD(1)	.649	.223	8.488	1	.004	1.914	1.237	2.961
	ARECOD(2)	-1.053	.263	16.016	1	.000	.349	.208	.584
	ARECOD(3)	-.137	.265	.267	1	.606	.872	.519	1.466
	Constant	-1.619	.424	14.547	1	.000	.198		
Step 5[e]	MODALIDA			77.894	3	.000			
	MODALIDA(1)	-1.891	.400	22.312	1	.000	.151	.069	.331
	MODALIDA(2)	-.368	.409	.812	1	.368	.692	.311	1.541
	MODALIDA(3)	-.001	.531	.000	1	.999	.999	.353	2.830
	COLPRO	.458	.193	5.632	1	.018	1.581	1.083	2.308
	POSUSM			74.604	4	.000			
	POSUSM(1)	.937	.220	18.146	1	.000	2.553	1.659	3.929
	POSUSM(2)	1.892	.250	57.395	1	.000	6.631	4.065	10.818
	POSUSM(3)	2.174	.344	39.975	1	.000	8.797	4.483	17.261
	POSUSM(4)	1.919	.558	11.835	1	.001	6.814	2.283	20.332
	ARECOD			46.441	3	.000			
	ARECOD(1)	.649	.223	8.436	1	.004	1.913	1.235	2.965
	ARECOD(2)	-1.124	.267	17.765	1	.000	.325	.193	.548
	ARECOD(3)	-.134	.265	.256	1	.613	.874	.520	1.471
	P43	.142	.068	4.389	1	.036	1.153	1.009	1.317
	Constant	-3.698	1.082	11.681	1	.001	.025		

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

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

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

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

e. Variable(s) entered on step 5: P43.

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	COLPRO	2.527	1	0.112
		TIPPRE	3.497	3	0.321
		TIPPRE(1)	0.367	1	0.544
		TIPPRE(2)	0.020	1	0.889
		TIPPRE(3)	1.572	1	0.210
		SEXPOS	3.593	1	0.058
		RANG_EDA	15.602	2	0.000
		RANG_EDA(1)	7.915	1	0.005
		RANG_EDA(2)	0.430	1	0.512
		IND_POBR	0.996	1	0.318
		POSUSM	72.521	4	0.000
		POSUSM(1)	0.929	1	0.335
		POSUSM(2)	34.234	1	0.000
		POSUSM(3)	12.085	1	0.001
		POSUSM(4)	2.129	1	0.145
		ARECOD	45.172	3	0.000
		ARECOD(1)	33.810	1	0.000
		ARECOD(2)	26.751	1	0.000
		ARECOD(3)	0.396	1	0.529
		N_PAD_AG	3.203	5	0.669
		N_PAD_AG(1)	1.099	1	0.294
		N_PAD_AG(2)	1.180	1	0.277
		N_PAD_AG(3)	0.227	1	0.634
		N_PAD_AG(4)	1.446	1	0.229
		N_PAD_AG(5)	0.132	1	0.717
		P43	0.489	1	0.484
		RANG_GAS	1.475	2	0.478
		RANG_GAS(1)	0.694	1	0.405
		RANG_GAS(2)	0.094	1	0.760
	Overall Statistics		134.348	23	0.000
Step 2	Variables	COLPRO	5.922	1	0.015
		TIPPRE	2.184	3	0.535
		TIPPRE(1)	0.007	1	0.933
		TIPPRE(2)	0.049	1	0.825
		TIPPRE(3)	0.488	1	0.485
		SEXPOS	2.434	1	0.119
		RANG_EDA	0.014	2	0.993
		RANG_EDA(1)	0.012	1	0.913
		RANG_EDA(2)	0.000	1	0.983
		IND_POBR	1.638	1	0.201
		ARECOD	47.174	3	0.000
		ARECOD(1)	33.851	1	0.000
		ARECOD(2)	30.078	1	0.000
		ARECOD(3)	0.224	1	0.636
		N_PAD_AG	4.072	5	0.539
		N_PAD_AG(1)	0.934	1	0.334
		N_PAD_AG(2)	1.871	1	0.171
		N_PAD_AG(3)	0.207	1	0.649
		N_PAD_AG(4)	2.428	1	0.119
		N_PAD_AG(5)	0.043	1	0.835
		P43	2.166	1	0.141
		RANG_GAS	4.558	2	0.102
		RANG_GAS(1)	0.953	1	0.329
		RANG_GAS(2)	1.077	1	0.299
	Overall Statistics		66.158	19	0.000

Variables not in the Equation (continuación)

			Score	df	Sig.
Step 3	Variables	COLPRO	7.184	1	0.007
		TIPPRE	2.740	3	0.433
		TIPPRE(1)	0.129	1	0.719
		TIPPRE(2)	0.185	1	0.667
		TIPPRE(3)	0.171	1	0.679
		SEXPOS	0.083	1	0.773
		RANG_EDA	0.279	2	0.870
		RANG_EDA(1)	0.005	1	0.942
		RANG_EDA(2)	0.167	1	0.683
		IND_POBR	1.689	1	0.194
		N_PAD_AG	4.314	5	0.505
		N_PAD_AG(1)	0.928	1	0.335
		N_PAD_AG(2)	2.120	1	0.145
		N_PAD_AG(3)	0.648	1	0.421
		N_PAD_AG(4)	1.939	1	0.164
		N_PAD_AG(5)	0.005	1	0.946
		P43	5.848	1	0.016
		RANG_GAS	5.442	2	0.066
		RANG_GAS(1)	2.355	1	0.125
		RANG_GAS(2)	0.349	1	0.555
	Overall Statistics		19.880	16	0.226
Step 4	Variables	TIPPRE	2.607	3	0.456
		TIPPRE(1)	0.110	1	0.741
		TIPPRE(2)	0.143	1	0.706
		TIPPRE(3)	0.174	1	0.677
		SEXPOS	0.134	1	0.714
		RANG_EDA	0.017	2	0.991
		RANG_EDA(1)	0.000	1	0.984
		RANG_EDA(2)	0.010	1	0.921
		IND_POBR	1.003	1	0.317
		N_PAD_AG	2.684	5	0.749
		N_PAD_AG(1)	0.386	1	0.535
		N_PAD_AG(2)	1.364	1	0.243
		N_PAD_AG(3)	0.634	1	0.426
		N_PAD_AG(4)	0.838	1	0.360
		N_PAD_AG(5)	0.010	1	0.919
		P43	4.407	1	0.036
		RANG_GAS	3.233	2	0.199
		RANG_GAS(1)	1.155	1	0.282
		RANG_GAS(2)	0.358	1	0.550
	Overall Statistics		12.813	15	0.617
Step 5	Variables	TIPPRE	2.057	3	0.561
		TIPPRE(1)	0.049	1	0.824
		TIPPRE(2)	0.100	1	0.752
		TIPPRE(3)	0.194	1	0.659
		SEXPOS	0.549	1	0.459
		RANG_EDA	0.077	2	0.962
		RANG_EDA(1)	0.003	1	0.956
		RANG_EDA(2)	0.039	1	0.844
		IND_POBR	0.911	1	0.340
		N_PAD_AG	2.787	5	0.733
		N_PAD_AG(1)	0.235	1	0.628
		N_PAD_AG(2)	1.577	1	0.209
		N_PAD_AG(3)	0.600	1	0.439
		N_PAD_AG(4)	0.747	1	0.388
		N_PAD_AG(5)	0.000	1	0.990
		RANG_GAS	3.543	2	0.170
		RANG_GAS(1)	1.346	1	0.246
		RANG_GAS(2)	0.336	1	0.562
	Overall Statistics		8.451	14	0.865

Step number: 5

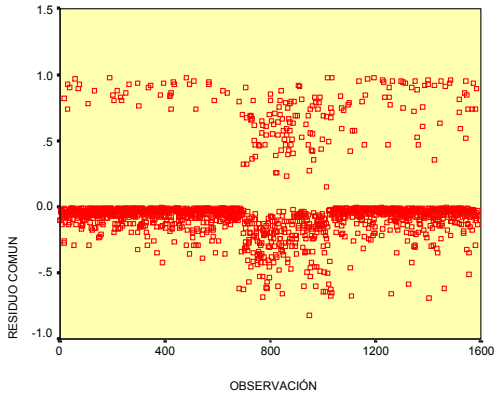
Observed Groups and Predicted Probabilities

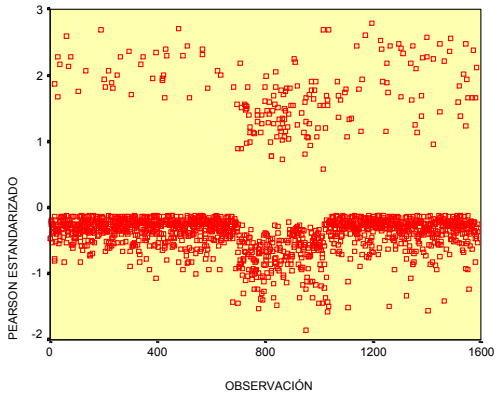
```
       320 • N                                                  •
           • N                                                  •
           • N                                                  •
F          • N                                                  •
R      240 • N                                                  •
E          • N                                                  •
Q          • NS                                                 •
U          •NNN                                                 •
E      160 •NNNS                                                •
N          •NNNN                                                •
C          •NNNN                                                •
Y          •NNNNS                                               •
        80 •NNNNN                                               •
           •NNNNNN S                                            •
           •NNNNNNNNNN N                                        •
           •NNNNNNNNNNNNNNNNNSNNSNS   S     S                   •
Predicted
•••••••••••••••••••••••••••••••••••••••••••••••
•••••••••••••
  Prob:    0           .25          .5           .75           1
  Group:   NNNNNNNNNNNNSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
```

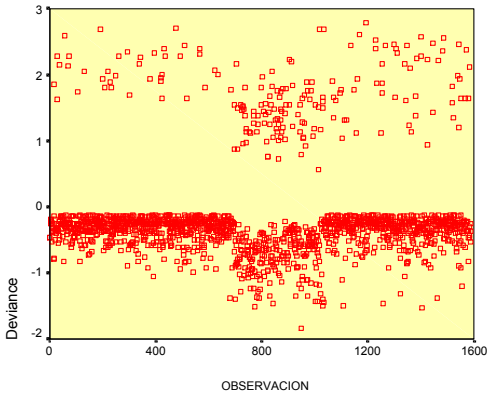
Predicted Probability is of Membership for Si ingreso
The Cut Value is .20
Symbols: N - No ingreso
S - Si ingreso
Each Symbol Represents 20 Cases.

ANEXO 4

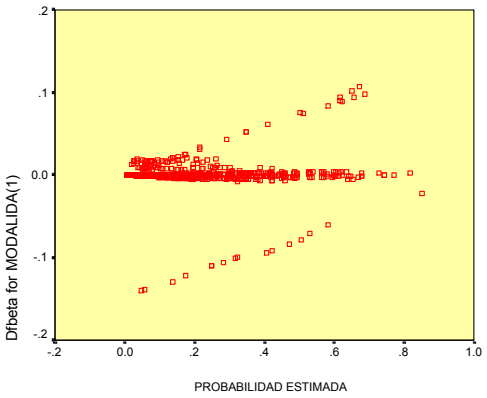
4.a

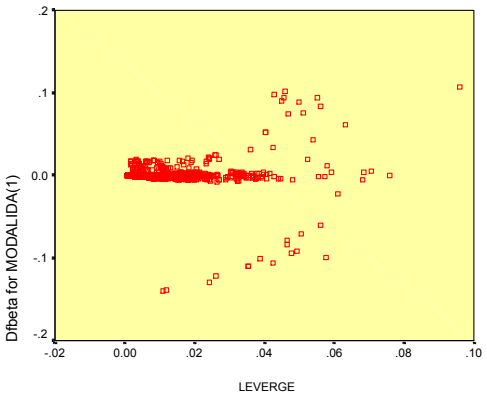

RESIDUO COMUN
1.5
1.0
.5
0.0
-.5
-1.0
0
400
800
1200
1600
OBSERVACIÓN

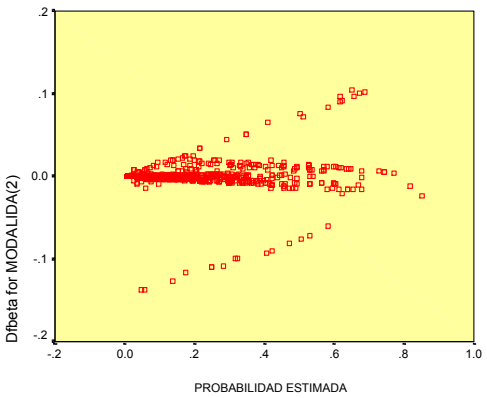

PEARSON ESTANDARIZADO
3
2
1
0
-1
-2
0
400
800
1200
1600
OBSERVACIÓN

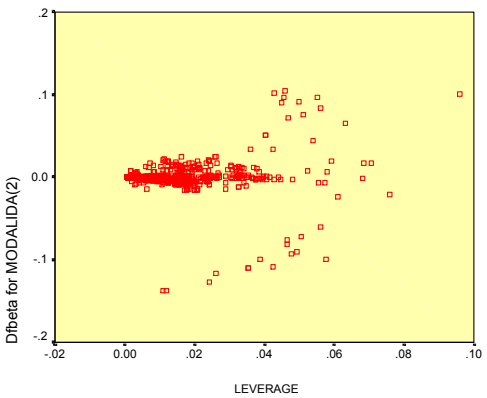

Deviance
3
2
1
0
-1
-2
0
400
800
1200
1600
OBSERVACION

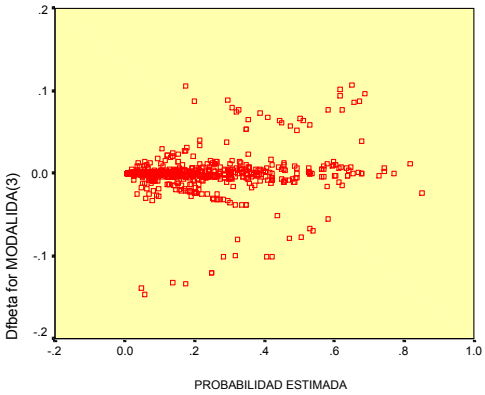
4.b

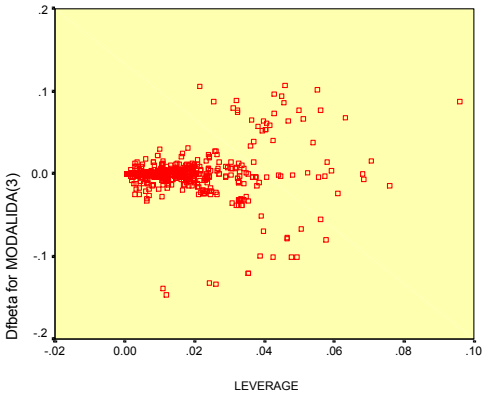

Dfbeta for MODALIDA(1)
.2
.1
0.0
-.1
-.2
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

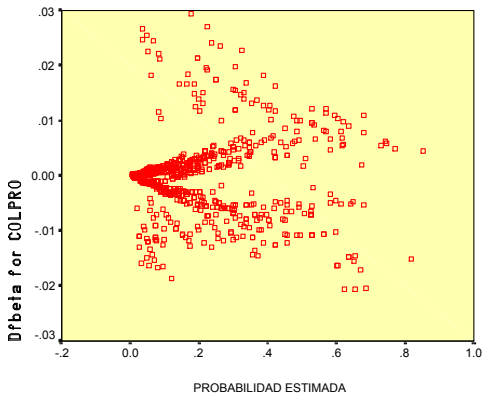

Dfbeta for MODALIDA(1)
.2
.1
0.0
-.1
-.2
-.02
0.00
.02
.04
.06
.08
.10
LEVERGE

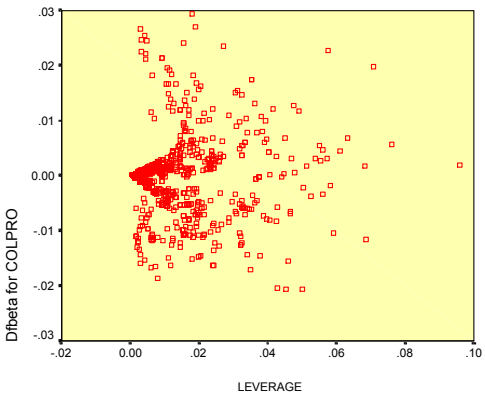

Dfbeta for MODALIDA(2)
.2
.1
0.0
-.1
-.2
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

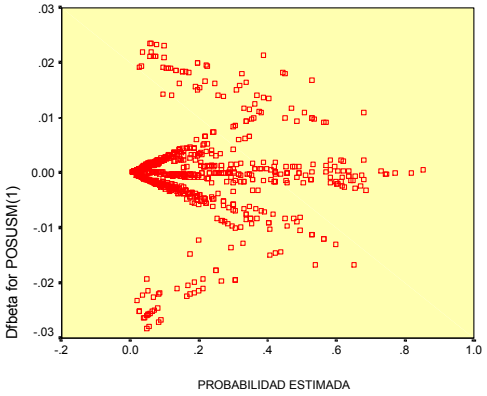

Dfbeta for MODALIDA(2)
.2
.1
0.0
-.1
-.2
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

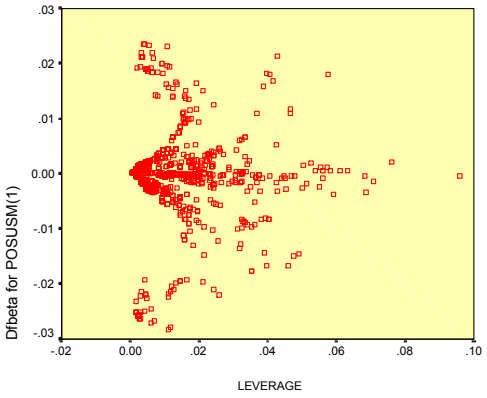

Dfbeta for MODALIDA(3)
.2
.1
0.0
-.1
-.2
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

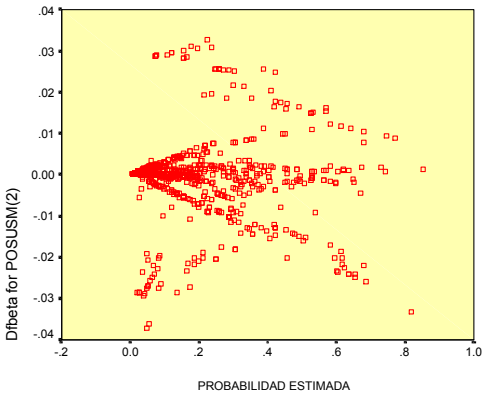

Dfbeta for MODALIDA(3)
.2
.1
0.0
-.1
-.2
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

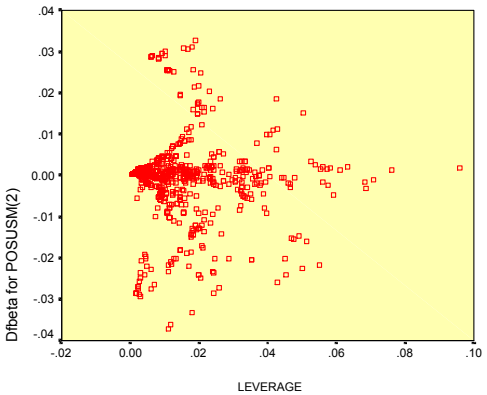

Dfbeta for COLPRO
.03
.02
.01
0.00
-.01
-.02
-.03
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

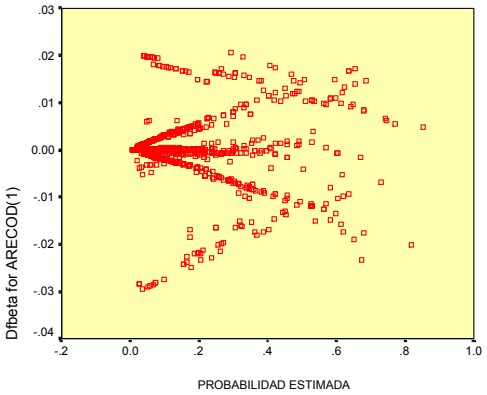

Dfbeta for COLPRO
.03
.02
.01
0.00
-.01
-.02
-.03
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

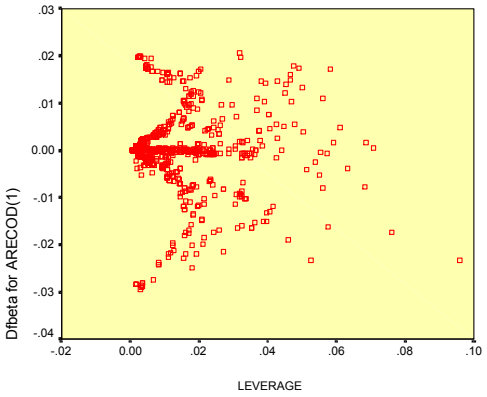

Dfbeta for POSUSM(1)
.03
.02
.01
0.00
-.01
-.02
-.03
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

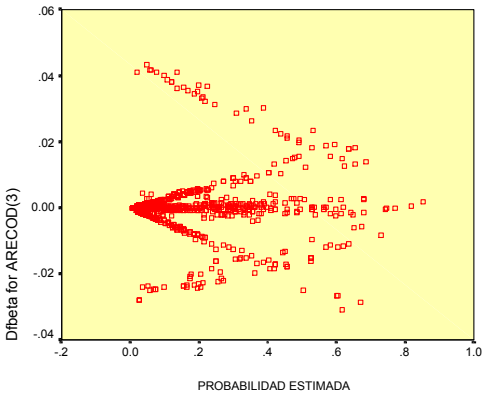

Dfbeta for POSUSM(1)
.03
.02
.01
0.00
-.01
-.02
-.03
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

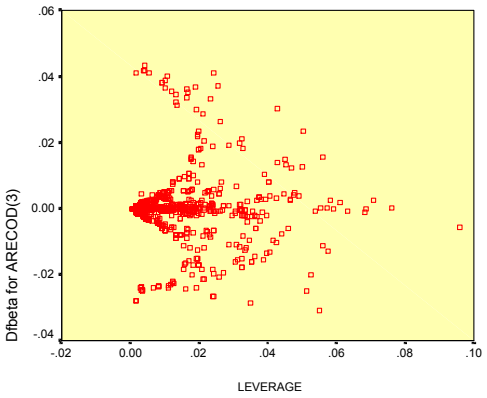

Dfbeta for POSUSM(2)
.04
.03
.02
.01
0.00
-.01
-.02
-.03
-.04
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA

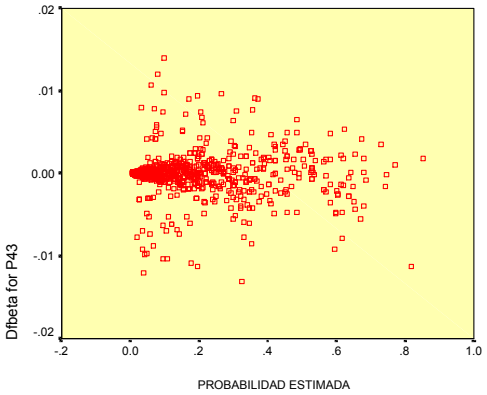

Dfbeta for POSUSM(2)
.04
.03
.02
.01
0.00
-.01
-.02
-.03
-.04
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

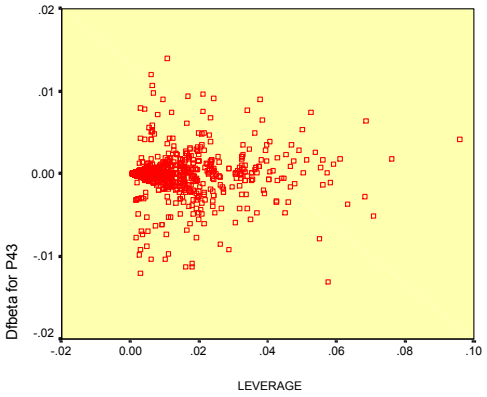

Dfbeta for ARECOD(1)
.03
.02
.01
0.00
-.01
-.02
-.03
-.04
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA


Dfbeta for ARECOD(1)
.03
.02
.01
0.00
-.01
-.02
-.03
-.04
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE


Dfbeta for ARECOD(3)
.06
.04
.02
0.00
-.02
-.04
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA


Dfbeta for ARECOD(3)
.06
.04
.02
0.00
-.02
-.04
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE


.02
.01
0.00
-.01
-.02
Dfbeta for P43
-.2
0.0
.2
.4
.6
.8
1.0
PROBABILIDAD ESTIMADA


.02
.01
0.00
-.01
-.02
Dfbeta for P43
-.02
0.00
.02
.04
.06
.08
.10
LEVERAGE

ANEXO 5

REGRESIÓN LOGÍSTICA CON EL MÉTODO FORWARD PARA LA SELECCIÓN DE VARIABLES (Eliminado los casos discordantes)

Case Processing Summary

Unweighted Cases [a]		N	Percent
Selected Cases	Included in Analysis	1538	97.4
	Missing Cases	41	2.6
	Total	1579	100.0
Unselected Cases		0	.0
Total		1579	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No ingreso	0
Si ingreso	1

Bloque 0: Bloque de Inicio

Classification Table[a,b]

Observed			Predicted		
			¿Ingreso?		Percentage Correct
			No ingreso	Si ingreso	
Step 0	¿Ingreso?	No ingreso	1359	0	100.0
		Si ingreso	179	0	.0
	Overall Percentage				88.4

a. Constant is included in the model.

b. The cut value is .200

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-2.027	.080	649.942	1	.000	.132

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	MODALIDA	140.735	3	.000
		MODALIDA(1)	138.609	1	.000
		MODALIDA(2)	98.012	1	.000
		MODALIDA(3)	9.343	1	.002
		COLPRO	15.580	1	.000
		TIPPRE	105.734	3	.000
		TIPPRE(1)	.933	1	.334
		TIPPRE(2)	.025	1	.874
		TIPPRE(3)	54.548	1	.000
		SEXPOS	.840	1	.359
		RANG_EDA	17.878	2	.000
		RANG_EDA(1)	9.693	1	.002
		RANG_EDA(2)	.399	1	.528
		IND_POBR	3.415	1	.065
		POSUSM	118.748	4	.000
		POSUSM(1)	2.810	1	.094
		POSUSM(2)	58.099	1	.000
		POSUSM(3)	28.194	1	.000
		POSUSM(4)	2.316	1	.128
		ARECOD	33.283	3	.000
		ARECOD(1)	26.181	1	.000
		ARECOD(2)	18.749	1	.000
		ARECOD(3)	.127	1	.722
		N_PAD_AG	8.162	5	.148
		N_PAD_AG(1)	2.379	1	.123
		N_PAD_AG(2)	2.316	1	.128
		N_PAD_AG(3)	.016	1	.900
		N_PAD_AG(4)	4.118	1	.042
		N_PAD_AG(5)	1.556	1	.212
		P43	8.529	1	.003
		RANG_GAS	7.357	2	.025
		RANG_GAS(1)	4.993	1	.025
		RANG_GAS(2)	.001	1	.981
	Overall Statistics		276.959	26	.000

Bloque 1: Método = Forward Stepwise (Wald)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	119.080	3	.000
	Block	119.080	3	.000
	Model	119.080	3	.000
Step 2	Step	74.284	4	.000
	Block	193.364	7	.000
	Model	193.364	7	.000
Step 3	Step	55.105	3	.000
	Block	248.469	10	.000
	Model	248.469	10	.000
Step 4	Step	9.169	1	.002
	Block	257.639	11	.000
	Model	257.639	11	.000
Step 5	Step	5.143	1	.023
	Block	262.782	12	.000
	Model	262.782	12	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	987.233	.075	.145
2	912.949	.118	.230
3	857.844	.149	.291
4	848.675	.154	.301
5	843.532	.157	.306

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	1	.997
2	5.085	4	.279
3	9.255	7	.235
4	17.778	8	.023
5	15.836	8	.045

Contingency Table for Hosmer and Lemeshow Test

		¿Ingreso? = No ingreso		¿Ingreso? = Si ingreso		
		Observed	Expected	Observed	Expected	Total
Step 1	1	1097	1096.972	73	73.028	1170
	2	39	39.000	13	13.000	52
	3	223	223.000	93	93.000	316
Step 2	1	679	672.904	15	21.096	694
	2	288	289.401	26	24.599	314
	3	106	109.107	19	15.893	125
	4	115	120.800	31	25.200	146
	5	114	114.484	43	42.516	157
	6	57	52.305	45	49.695	102
Step 3	1	193	191.178	0	1.822	193
	2	133	130.725	1	3.275	134
	3	82	82.840	3	2.160	85
	4	203	199.511	2	5.489	205
	5	185	187.721	14	11.279	199
	6	166	167.088	13	11.912	179
	7	161	163.972	28	25.028	189
	8	121	126.808	40	34.192	161
	9	115	109.155	78	83.845	193
Step 4	1	141	139.953	0	1.047	141
	2	113	112.002	1	1.998	114
	3	102	99.779	0	2.221	102
	4	162	158.369	0	3.631	162
	5	110	110.534	5	4.466	115
	6	136	141.746	13	7.254	149
	7	141	143.567	12	9.433	153
	8	122	115.592	6	12.408	128
	9	130	134.305	27	22.695	157
	10	202	203.151	115	113.849	317
Step 5	1	155	153.871	0	1.129	155
	2	146	143.547	0	2.453	146
	3	154	150.532	0	3.468	154
	4	149	148.220	4	4.780	153
	5	154	154.716	8	7.284	162
	6	142	148.171	16	9.829	158
	7	139	139.094	15	14.906	154
	8	132	130.571	22	23.429	154
	9	106	115.203	48	38.797	154
	10	82	75.070	66	72.930	148

Classification Table[a]

			Predicted		
			¿Ingreso?		Percentage Correct
	Observed		No ingreso	Si ingreso	
Step 1	¿Ingreso?	No ingreso	1097	262	80.7
		Si ingreso	73	106	59.2
	Overall Percentage				78.2
Step 2	¿Ingreso?	No ingreso	1216	143	89.5
		Si ingreso	100	79	44.1
	Overall Percentage				84.2
Step 3	¿Ingreso?	No ingreso	1181	178	86.9
		Si ingreso	79	100	55.9
	Overall Percentage				83.3
Step 4	¿Ingreso?	No ingreso	1174	185	86.4
		Si ingreso	71	108	60.3
	Overall Percentage				83.4
Step 5	¿Ingreso?	No ingreso	1185	174	87.2
		Si ingreso	69	110	61.5
	Overall Percentage				84.2

a. The cut value is .200

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	MODALIDA			118.151	3	.000			
	MODALIDA(1)	-2.103	.379	30.852	1	.000	.122	.058	.256
	MODALIDA(2)	-.303	.382	.627	1	.428	.739	.349	1.563
	MODALIDA(3)	-.492	.481	1.048	1	.306	.611	.238	1.569
	Constant	-.606	.359	2.853	1	.091	.545		
Step 2[b]	MODALIDA			87.598	3	.000			
	MODALIDA(1)	-1.721	.403	18.227	1	.000	.179	.081	.394
	MODALIDA(2)	-.214	.407	.276	1	.600	.807	.364	1.794
	MODALIDA(3)	.307	.514	.356	1	.551	1.359	.496	3.721
	POSUSM			69.835	4	.000			
	POSUSM(1)	.997	.220	20.613	1	.000	2.711	1.763	4.170
	POSUSM(2)	1.890	.246	58.852	1	.000	6.621	4.085	10.731
	POSUSM(3)	1.911	.332	33.119	1	.000	6.760	3.526	12.959
	POSUSM(4)	1.575	.560	7.925	1	.005	4.832	1.614	14.469
	Constant	-1.742	.414	17.690	1	.000	.175		
Step 3[c]	MODALIDA			103.848	3	.000			
	MODALIDA(1)	-1.859	.415	20.110	1	.000	.156	.069	.351
	MODALIDA(2)	-.098	.419	.054	1	.816	.907	.399	2.064
	MODALIDA(3)	.512	.527	.943	1	.332	1.669	.594	4.690
	POSUSM			73.732	4	.000			
	POSUSM(1)	1.006	.227	19.665	1	.000	2.736	1.754	4.268
	POSUSM(2)	1.945	.254	58.539	1	.000	6.990	4.248	11.503
	POSUSM(3)	2.136	.343	38.814	1	.000	8.464	4.323	16.572
	POSUSM(4)	1.893	.587	10.391	1	.001	6.637	2.100	20.979
	ARECOD			49.172	3	.000			
	ARECOD(1)	.799	.228	12.219	1	.000	2.222	1.420	3.477
	ARECOD(2)	-1.060	.274	14.964	1	.000	.346	.202	.593
	ARECOD(3)	-.094	.270	.121	1	.728	.911	.537	1.544
	Constant	-1.734	.435	15.855	1	.000	.177		
Step 4[d]	MODALIDA			91.338	3	.000			
	MODALIDA(1)	-1.803	.421	18.330	1	.000	.165	.072	.376
	MODALIDA(2)	-.157	.427	.136	1	.712	.854	.370	1.971
	MODALIDA(3)	.522	.532	.961	1	.327	1.685	.594	4.782
	COLPRO	.597	.195	9.361	1	.002	1.817	1.239	2.664
	POSUSM			77.955	4	.000			
	POSUSM(1)	1.052	.229	21.128	1	.000	2.862	1.828	4.482
	POSUSM(2)	2.034	.258	62.304	1	.000	7.644	4.613	12.665
	POSUSM(3)	2.215	.346	40.931	1	.000	9.166	4.650	18.069
	POSUSM(4)	2.043	.595	11.802	1	.001	7.716	2.405	24.753
	ARECOD			49.972	3	.000			
	ARECOD(1)	.778	.230	11.408	1	.001	2.176	1.386	3.417
	ARECOD(2)	-1.120	.277	16.376	1	.000	.326	.190	.561
	ARECOD(3)	-.030	.271	.012	1	.912	.971	.570	1.652
	Constant	-1.973	.449	19.270	1	.000	.139		
Step 5[e]	MODALIDA			77.989	3	.000			
	MODALIDA(1)	-1.790	.423	17.889	1	.000	.167	.073	.383
	MODALIDA(2)	-.197	.429	.212	1	.645	.821	.354	1.904
	MODALIDA(3)	.235	.550	.182	1	.670	1.264	.430	3.714
	COLPRO	.542	.197	7.545	1	.006	1.720	1.168	2.533
	POSUSM			81.144	4	.000			
	POSUSM(1)	1.059	.229	21.311	1	.000	2.883	1.839	4.519
	POSUSM(2)	2.066	.259	63.535	1	.000	7.891	4.748	13.113
	POSUSM(3)	2.372	.353	45.059	1	.000	10.724	5.364	21.438
	POSUSM(4)	2.100	.602	12.182	1	.000	8.163	2.511	26.542
	ARECOD			52.645	3	.000			
	ARECOD(1)	.776	.231	11.277	1	.001	2.173	1.381	3.418
	ARECOD(2)	-1.204	.281	18.362	1	.000	.300	.173	.520
	ARECOD(3)	-.027	.272	.010	1	.922	.974	.571	1.660
	P43	.158	.070	5.128	1	.024	1.172	1.022	1.344
	Constant	-4.299	1.125	14.602	1	.000	.014		

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

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

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

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

e. Variable(s) entered on step 5: P43.

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	COLPRO	3.402	1	0.065
		TIPPRE	2.122	3	0.547
		TIPPRE(1)	0.456	1	0.500
		TIPPRE(2)	0.070	1	0.791
		TIPPRE(3)	1.393	1	0.238
		SEXPOS	4.238	1	0.040
		RANG_EDA	16.136	2	0.000
		RANG_EDA(1)	8.419	1	0.004
		RANG_EDA(2)	0.365	1	0.546
		IND_POBR	1.334	1	0.248
		POSUSM	78.186	4	0.000
		POSUSM(1)	1.163	1	0.281
		POSUSM(2)	37.127	1	0.000
		POSUSM(3)	12.946	1	0.000
		POSUSM(4)	1.579	1	0.209
		ARECOD	50.779	3	0.000
		ARECOD(1)	39.367	1	0.000
		ARECOD(2)	29.175	1	0.000
		ARECOD(3)	0.176	1	0.674
		N_PAD_AG	3.117	5	0.682
		N_PAD_AG(1)	1.146	1	0.284
		N_PAD_AG(2)	1.026	1	0.311
		N_PAD_AG(3)	0.167	1	0.682
		N_PAD_AG(4)	1.305	1	0.253
		N_PAD_AG(5)	0.193	1	0.661
		P43	0.555	1	0.456
		RANG_GAS	1.364	2	0.506
		RANG_GAS(1)	0.746	1	0.388
		RANG_GAS(2)	0.045	1	0.833
	Overall Statistics		147.948	23	0.000
Step 2	Variables	COLPRO	7.635	1	0.006
		TIPPRE	0.940	3	0.816
		TIPPRE(1)	0.018	1	0.892
		TIPPRE(2)	0.132	1	0.716
		TIPPRE(3)	0.351	1	0.553
		SEXPOS	2.957	1	0.086
		RANG_EDA	0.006	2	0.997
		RANG_EDA(1)	0.001	1	0.975
		RANG_EDA(2)	0.006	1	0.940
		IND_POBR	2.026	1	0.155
		ARECOD	54.181	3	0.000
		ARECOD(1)	40.232	1	0.000
		ARECOD(2)	33.722	1	0.000
		ARECOD(3)	0.056	1	0.813
		N_PAD_AG	4.109	5	0.534
		N_PAD_AG(1)	1.058	1	0.304
		N_PAD_AG(2)	1.762	1	0.184
		N_PAD_AG(3)	0.187	1	0.665
		N_PAD_AG(4)	2.337	1	0.126
		N_PAD_AG(5)	0.072	1	0.788
		P43	2.530	1	0.112
		RANG_GAS	4.396	2	0.111
		RANG_GAS(1)	0.901	1	0.342
		RANG_GAS(2)	1.076	1	0.300
	Overall Statistics		74.968	19	0.000

Variables not in the Equation (Continuación)

			Score	df	Sig.
Step 3	Variables	COLPRO	9.488	1	0.002
		TIPPRE	1.461	3	0.691
		TIPPRE(1)	0.133	1	0.715
		TIPPRE(2)	0.352	1	0.553
		TIPPRE(3)	0.072	1	0.788
		SEXPOS	0.149	1	0.699
		RANG_EDA	0.540	2	0.763
		RANG_EDA(1)	0.033	1	0.857
		RANG_EDA(2)	0.268	1	0.604
		IND_POBR	2.191	1	0.139
		N_PAD_AG	4.649	5	0.460
		N_PAD_AG(1)	1.137	1	0.286
		N_PAD_AG(2)	2.093	1	0.148
		N_PAD_AG(3)	0.801	1	0.371
		N_PAD_AG(4)	1.787	1	0.181
		N_PAD_AG(5)	0.010	1	0.921
		P43	6.920	1	0.009
		RANG_GAS	5.277	2	0.071
		RANG_GAS(1)	2.284	1	0.131
		RANG_GAS(2)	0.342	1	0.558
	Overall Statistics		22.070	16	0.141
Step 4	Variables	TIPPRE	1.339	3	0.720
		TIPPRE(1)	0.101	1	0.751
		TIPPRE(2)	0.290	1	0.590
		TIPPRE(3)	0.084	1	0.772
		SEXPOS	0.238	1	0.626
		RANG_EDA	0.087	2	0.958
		RANG_EDA(1)	0.017	1	0.895
		RANG_EDA(2)	0.024	1	0.878
		IND_POBR	1.313	1	0.252
		N_PAD_AG	2.849	5	0.723
		N_PAD_AG(1)	0.449	1	0.503
		N_PAD_AG(2)	1.257	1	0.262
		N_PAD_AG(3)	0.777	1	0.378
		N_PAD_AG(4)	0.627	1	0.428
		N_PAD_AG(5)	0.008	1	0.929
		P43	5.148	1	0.023
		RANG_GAS	2.849	2	0.241
		RANG_GAS(1)	0.952	1	0.329
		RANG_GAS(2)	0.371	1	0.542
	Overall Statistics		12.806	15	0.617
Step 5	Variables	TIPPRE	0.921	3	0.820
		TIPPRE(1)	0.033	1	0.857
		TIPPRE(2)	0.223	1	0.637
		TIPPRE(3)	0.110	1	0.740
		SEXPOS	0.785	1	0.375
		RANG_EDA	0.030	2	0.985
		RANG_EDA(1)	0.003	1	0.956
		RANG_EDA(2)	0.028	1	0.867
		IND_POBR	1.205	1	0.272
		N_PAD_AG	3.054	5	0.692
		N_PAD_AG(1)	0.265	1	0.606
		N_PAD_AG(2)	1.495	1	0.221
		N_PAD_AG(3)	0.725	1	0.394
		N_PAD_AG(4)	0.552	1	0.457
		N_PAD_AG(5)	0.002	1	0.968
		RANG_GAS	3.165	2	0.205
		RANG_GAS(1)	1.150	1	0.284
		RANG_GAS(2)	0.344	1	0.557
	Overall Statistics		7.780	14	0.900