Predicción de las razones de un estudiante en su selección de una escuela mediante Regresión Logística

Universidad de Granada
Departamento de Estadística
Granada, España
2022

Predicción de las razones de un estudiante en su selección de una escuela mediante Regresión Logística

Rafael Niño Alfaro

Trabajo para optar el título de Magíster en Estadística Aplicada

Facultad de Ciencias

Director:

Dr. Manuel Escabias Machuca
Dra. Ana María Aguilera del Pino

Universidad de Granada
Departamento de Estadística
Granada, España
2022

Resumen

A partir del análisis de un grupo de variables, se evalúa la incidencia de cada una de estas con el fin de determinar las razones por las cuales un estudiante escoge una institución educativa. Se evalúan las condiciones para la determinación del modelo estudiando la bondad de ajuste de este e interpretando las exponenciales de los parámetros para explicar la relación entre las variables seleccionadas y la de respuesta. Posteriormente se presentan los aspectos teóricos relacionados con la regresión logística multinomial y la formulación del modelo, de acuerdo con la interpretación de exponenciales de parámetros como cocientes de ventajas tanto para las variables cuantitativas como cualitativas, estimación de la máxima verosimilitud, test de razón de verosimilitudes y evaluación de la bondad de ajuste, contrastes de significación de parámetros y análisis de residuos, además del método de selección stepwise, basado en el principio de Akaike, o razón de verosimilitudes condicional, que nos permita determinar el modelo con las variables que mejor sustenten la variable dependiente, implementado con el software de RStudio.

Palabras clave: Modelo de regresión logística multinomial; Variables cuantitativas y cualitativas; Bondad del ajuste; Método de Selección de stepwise; Odds Ratio.

Abstract

From the analysis of a group of variables, the incidence of each of these is evaluated in order to determine the reasons why a student chooses a secondary school. The conditions for determining the model are evaluated by studying the goodness of adjusting this and interpreting the exponentials of the parameters to explain the relationship between the selected variables and the response variable. Subsequently, the theoretical aspects related to multinomial logistic regression and formulation of the model are presented, according to the interpretation of exponential parameters as odds ratios for both quantitative and qualitative variables, estimation of maximum likelihood, likelihood ratio test and evaluation of the goodness of fit, parameter significance tests and residual analysis in addition to the stepwise selection method based on the Akaike principle or conditional likelihood ratio that allows us to determine the model with the variables that best support the dependent variable, this implemented with RStudio software.

Keywords: Multinomial logistic regression model; Quantitative and qualitative variables; Goodness of fit; Stepwise Selection Method; Odds Ratio

Dedicado a:

A la memoria de mi Abuela, Flor María Alfaro Contreras, quien siempre fue mi aliento en los momentos más difíciles y me brindó mucho cariño y alegría, incluso en sus últimos días.

A mi amada Madre Alicia Alfaro, a mi hija hermosa Paula Alejandra Niño, y a mi esposa Carol gil, tres hermosas flores que irradian con su hermosura cada rincón de mi alma.

A Diego Ávila, un hermano del espíritu y de las letras quien me ha enseñado que el verdadero sentido de ser un guerrero está en afrontar cada batalla con entusiasmo y decisión.

Agradecimientos

Quiero dar mis más sinceros agradecimientos, a los profesores Manuel Escabias y Ana Aguilera del Pino, por su apoyo y acertada tutoría durante el desarrollo de este trabajo; lo que me ha permitido profundizar aún más en un área de aplicación tan maravillosa de la estadística.

Deseo dar un reconocimiento especial a Adriana Rincón Gómez, una gran amiga y ex alumna de esta maestría, por su valiosa colaboración y disposición para brindarme su asesoría y consejo en esta investigación.

Finalmente agradezco a mis amigos y familiares, quienes aún en las horas más oscuras y en los momentos más difíciles e inciertos, en aquellos instantes en los que pensé en renunciar, me animaron a seguir adelante y nunca dejaron de creer en mí.

Índice General

Índice de Tablas

Índice de Figuras

Capítulo 1

1. Introducción

1.1 Fundamento del Problema

El modelo de regresión logística multinomial, también conocido como modelo con respuesta politómica, es una generalización del modelo de regresión logístico binomial en el que se desea estimar la probabilidad de que el individuo presente o no un evento especifico, dado un conjunto de variables que explican características particulares de los individuos. En el caso del modelo multinomial, la variable dependiente tiene más de dos alternativas a considerar como posibles respuestas, por lo que la distribución de probabilidadadecuada para modelar este fenómeno es la distribución multinomial. La variable respuesta del modelo de regresión logística multinomial es una variable aleatoria con distribución multinomial, que se puede considerar como el número de éxitos en cada una de las g categorías que se presentan en *n* ensayos independientes.

En este trabajo se evaluó la incidencia de un grupo de variables con el fin de determinar las razones por las que un estudiante escoge una institución educativa. Para lograrlo fue necesario validar la incidencia de estas variables en la determinación del modelo, estudiando la bondad de ajuste de este, e interpretando las exponenciales de los parámetros para explicar la relación entre las variables seleccionadas y la de respuesta.

Usada en muchos campos de la investigación, la regresión logística se ha enfocado en el estudio de aplicaciones sanitarias y sociales por su alta capacidad para la determinación de relaciones entre variables categóricas. Forma parte del conjunto de métodos estadísticos bajo tal denominación y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple.

Para este trabajo se describió la metodología utilizada, los objetivos, los indicadores y cómo fueron medidos. Finalmente se definió el modelo de regresión a implementar, así como

todas las características estadísticas que sustenten su fundamentación, tomando en cuenta las limitaciones.

El análisis de datos ejecutado en este proyecto resulta de utilidad en las investigaciones especializadas en psicología educativa, a partir de la valoración de indicadores claves que facilitan la medición de los parámetros que definen la elección de una institución por parte de la población escolar preuniversitaria.

El objetivo de este trabajo fue determinar los factores que inciden en la selección de una institución educativa por parte de un estudiante, a través del estudio de estas relaciones. Mediante esta investigación se obtuvo una herramienta con el propósito de que las instituciones educativas generen nuevas estrategias para atracción de alumnos. El análisis toma una muestra de estudiantes participantes en dos cursos de dos instituciones educativas diferentes de Portugal durante el año 2017.

En los anexos se incluye una descripción con las sentencias de R utilizadas, así como los resultados. Se describió el proceso de ajuste, inferencia y validación del modelo de regresión empleado, que fue multinomial con variable de respuesta categórica, analizado tanto a través del método parcial como del método marginal.

En el capítulo dos se llevó a cabo un análisis teórico referente a la evaluación del modelo por medio de la selección de las variables con el método de selección stepwise, la interpretación de exponenciales de parámetros como cocientes de ventajas tanto para variables cuantitativas como cualitativas, la estimación de máxima verosimilitud, el test de razón de verosimilitudes para la bondad de ajuste, contrastes de significación de parámetros y análisis de residuos.

Por último, en el capítulo tercero se analizó la aplicación práctica del modelo estudiado a través del software R Studio, a un set de datos obtenido mediante un estudio llevado a cabo en dos escuelas de Portugal en el año 2016. Se tomó un total de 26 variables referentes a diversos parámetros correlacionados con el desempeño académico, las actividades curriculares, el programa diseñado con cada una de sus funciones y líneas de comando empleadas, así como los resultados obtenidos a través del procedimiento de stepwise.

1.1. Objetivos del Trabajo

1.1.1. Objetivo General

Implementar el modelo estadístico que permita establecer el conjunto de variables independientes más adecuadas que permitan predecir las razones por las que un estudiante selecciona una institución educativa, mediante el desarrollo de un modelo de regresión logística multinomial.

1.1.2. Objetivos Específicos:

- Describir la teoría del análisis de regresión logística multinomial y aplicarla a datos reales representados en la encuesta de factores socioeducativos de dos colegios de Portugal en el año 2013.

- Cuantificar e interpretar los efectos de cada variable sobre la probabilidad de seleccionar una institución educativa, a través de las estimaciones obtenidas en el modelo.

- Describir la teoría del análisis de regresión logística multinomial, detallar cómo se aplica esta con el lenguaje de R y finalmente describir una aplicación con datos reales en la que se aplique estos modelos y se realicen los respectivos análisis.

1.2. Antecedentes

La Estadística trasciende el contexto académico y está presente en distintos ámbitos de la actividad humana. Hay una amplia gama de estudios relacionados con el análisis de tendencias a nivel social y educativo. La estadística y la psicología clínica son piezas fundamentales que permiten observaciones y predicciones que contribuyen al estudio científico de los problemas planteados en el ámbito de la educación. Durante los últimos

años se han efectuado algunas investigaciones similares relacionadas con los factores de motivación de los estudiantes enfocados, sobre todo, en la selección de carrerasprofesionales y programas educación superior.

Lizares (2017) trabajo sobre la comparación de modelos de clasificación; en específico, regresión logística y árboles de clasificación para evaluar el *rendimiento académico*. Comparó ambos para la evaluación adecuada del rendimiento académico, a partir de bases de datos (tomando en cuenta factores sociodemográficos y factores académicos) de 3600 registros. Las muestras fueron divididas con un 70% del total para entrenamiento (2520 registros) y el 30% de prueba (1080 registros). El conjunto de entrenamiento (para construir el modelo) estuvo conformado por el 70% de estudiantes universitarios del primer semestre matriculados en el curso de Matemática. A partir de la tabla de clasificación para evaluar la precisión de clasificadores, los resultados indican que la técnica de Árboles de decisión obtuvo el mayor porcentaje de buena clasificación, siendo el mejor modelo bajo la curva ROC, Sensibilidad, Índice de Gini e Índice de Kappa.

Por su parte, Bravo y Vergara (2018) hicieron un estudio relacionado con la predicción de factores que inciden en la elección de una carrera. Fue una investigación de enfoque cuantitativo, corte trasversal y nivel descriptivo, mediante el uso de análisis factorial exploratorio. La muestra fue conformada por 225 estudiantes de último grado (11º) de Bachillerato, pertenecientes a seis (6) colegios de Barrancabermeja (3 colegios públicos y 3 privados), en el Departamento de Norte de Santander, Colombia. Los colegios públicos aportaron a esta muestra un total de 137 estudiantes y los colegios privados un total de 88. Analizando la prevalencia de factores de elección de carrera en totalidad de lamuestra de los estudiantes de bachillerato, los datos mostraron que los intereses personales de los estudiantes son el factor predominante en la elección de carrera, con un 42%; seguido de la intención de generar beneficios a la sociedad con un 32 %; le sigue la posibilidad de optar a un buen salario con un 22 % y finalmente esta la influencia familiar con un 4 %. En las siguientes tablas podemos ver el resumen de la prevalencia según algunos factores de elección, de acuerdo a las conclusiones de este estudio:

Tabla1.

Prevalencia de factores de elección de carrera según la edad

Edad	Intereses personales	Beneficios a la comunidad	Buen salario	Influencia familiar
15-16	17,8 %	12,5%	10,2 %	1,7 %
17-18	20 %	16,4%	13,9%	2,2 %
19-20	4,0 %	3,1%	1,8%	0,0 %

Nota: Adaptado de Bravo y Vergara (2017).

Tabla2.

Prevalencia de factores de elección de carrera según el sexo

Sexo	Intereses personales	Beneficios a la sociedad	Buen salario	Influencia familair	Lo que esta de moda
Femenino	20,9 %	16,9 %	9,8 %	1,8 %	0%
masculino	20,9 %	15,1 %	12,4 %	2,2 %	0%

Nota: Adaptado de Bravo y Vergara (2017)

Tabla3.

Prevalencia de factores de elección de carrera según el tipo de colegio

Tipo de colegio	Intereses personales	Beneficios a la sociedad	Buen salario	Influencia familair	Lo que esta de moda
publico	22,7%	15,6%	17,8 %	1,3 %	0%
privado	19,1%	16,4%	4,4 %	2,7 %	0%

Nota: Adaptado de Bravo y Vergara (2017)

García y Moreno (2012) hicieron un estudio en México relacionado con los factores considerados al seleccionar una universidad. La metodología empleada implicó el diseño de un instrumento de recolección de datos e identificación de los factores que los universitarios tomaron en cuenta antes de elegir una institución en la cual estudiar; recoger *y* analizar la información *y* concluir con base en los resultados. Se basó, esencialmente, en cinco pasos entre los que están la identificación de los atributos, aplicación de encuestas, capturade información, análisis de información y análisis factorial exploratorio en un total de 342 estudiantes. A partir de los resultados del análisis factorial exploratorio realizado a un conjunto de 31 ítems o atributos considerados por 342 alumnos al momento de realizar la selección de una universidad para iniciar sus estudios de licenciatura, se concluye que los tres principales aspectos que éstos tomaron en cuenta se relacionan con factores de índole económico, la calidad o el prestigio

académico de la institución, y aspectos de ingreso/estancia/egreso.

Por otro lado, en el análisis de Van Herpen et al. (2017), se diseñó un estudio basado en predictores tempranos del éxito académico de primer año en la universidad: esfuerzo preuniversitario, autoeficacia preuniversitaria y razones preuniversitarias para asistir a la Universidad, el cual se originó a causa *del* gran número de abandonos en el primer año en la universidad, por lo que se enfocó en investigar los predictores tempranos del éxito académico del primer año. El estudio tomo una muestra *n* de 453 estudiantes de primer año, concentrándose en la transición de la educación secundaria a la educación superior al investigar cómo los factores no cognitivos el esfuerzo preuniversitario y la autoeficacia académica preuniversitaria influyen en la retención del primer año en la universidad.

Los análisis basados en regresión logística multinomial mostraron que el esfuerzo preuniversitario predijo positivamente la retención del primer año, mientras que la autoeficacia académica preuniversitaria no lo hizo. Además, con el análisis factorial exploratorio y el análisis factorial confirmatorio, se identificaron seis motivos preuniversitarios para asistir a la universidad: perspectiva de carrera, desarrollo personal, cumplimiento del entorno social, atractivo de la institución, recomendado por otros, y ubicación. Ninguna de las razones preuniversitarias pareció predecir significativamente la retención en el primer año. Se discuten las implicaciones para la investigación y la práctica.

Otra investigación a cargo de Morales et al. (2018) se basa en un modelo de regresión logística como alternativa para medir la probabilidad de deserción temprana en la universidad de los llanos (Colombia), además de identificar cuáles son las variables de mayor incidencia en la deserción temprana. Para la construcción del modelo se tomó como muestra la información de 574 estudiantes que ingresaron en un periodo inicial y que, un periodo posterior tres años después, eran registrados como vigentes o reportaban como último periodo matriculado cualquiera de los cuatro primeros semestres. Los resultados dan cuenta de que un buen puntaje en la prueba de Estado para ingreso a la educación superior, el ser mujer, no haber reprobado años durante el bachillerato, haber cursado estudios antes y si los padres conviven, disminuye la probabilidad de deserción temprana; así mismo, el haber egresado de un colegio privado aumenta la probabilidad de deserción. Así mismo se ha encontrado que de todas las facultades en las que se puede matricular el estudiante, las que mayor riesgo tienen de deserción temprana son la de Ciencias Básicas e Ingeniería y la Facultad de Ciencias de la salud.

Por último, Ogutu et al. (2017) hace un estudio con el objetivo de examinar la influencia de la autoeficacia en la toma de decisiones de carrera entre estudiantes de la escuela. Los participantes fueron 364 estudiantes de cuarto curso de secundaria en Busia, Condado de Kenia. El género, la edad y el tipo de escuela se utilizaron como variables controladoras de la autoeficacia en la carrera. Mediante la correlación de Spearman, se pudo demostrar que la autoeficacia se correlacionó significativamente con la decisión de carrera de los estudiantes. Cuando se ajustaron las variables en los modelos logísticos multinomiales, la razón de riesgo relativo aumentó o disminuyó, pero el valor p se mantuvo estadísticamente significante. Esto implicó que los factores dentro de la variable de autoeficacia contribuyeron significativamente en la relación entre la autoeficacia y la toma de decisiones profesionales. Sobre la base de los hallazgos, se recomendó que la toma de decisiones profesionales debe mejorarse en las escuelas utilizando estrategias de orientación y asesoramiento profesional.

1.3. Justificación

Para este estudio se consideró la regresión logística como el análisis de regresión más apropiado, teniendo en cuenta la naturaleza de nuestra variable dependiente, del tipo categórica múltiple. Los análisis de regresión permiten describir datos y explicar la relación entre una variable dependiente y una o más variables independientes nominales, ordinales, de intervalo o de nivel de razón. En la actualidad este método es ampliamente utilizado tanto en estudios observacionales como de encuesta y experimentales, muchos de ellos enfocados al ámbito educativo. Tomando en cuenta lo anterior, se pretende demostrar de acuerdo con las variables que componen nuestra data, la viabilidad y el diseño de un modelo que tome comovariable dependiente las razones para escoger una determinada escuela. A diferencia de las investigaciones similares que ya se han discutido previamente, nuestro grupo de datos es másamplio y toma en cuenta una buena cantidad de variables de índole socio económica, que en cierta forma podrán incidir de diversa manera y generar un nivel muy distinto de impacto en la respuesta del modelo.

Capítulo 2

2.1 Modelo de Regresión Logística Multinomial

2.1.1 Formulación del Modelo

El análisis de un modelo de regresión logística binaria, toma una variable dependiente Y, con valores Y=1 (presencia de una característica u otra categoría de la variable) y Y=0 (ausencia de la característica o la otra categoría de la variable), por lo que la ecuación del modelo viene dada por:

$$P(Y = 1|X) = \frac{e^{b_0 + \sum_{s=1}^{n} b_s x_s}}{1 + e^{b_0 + \sum_{s=1}^{n} b_s x_s}}$$

donde $P[Y=1|X]$ representa la probabilidad de que, Y tome el valor 1, en presencia de las covariables X, que denotaremos por p(X).

Si X es un conjunto de n covariables $\{x_1, x_2, \dots, x_n\}$ que forman parte del modelo; b_0 la constante del modelo o término independiente y bi los cocientes de las covariables. Si se aplica la transformación logit de esta ecuación, obtenemos:

$$\ln\left[\frac{p(x)}{1 - p(x)}\right] = b_0 + \sum_{s=1}^{n} b_s x_s$$

Lo cual permite su representación en forma de una función lineal. Ahora bien, si la variable dependiente presenta más de dos categorías, utilizaremos un modelo de regresión logística multinomial que se modela, como se indicó anteriormente, mediante varios logits de manera simultánea, uno para cada una de las restantes categorías respecto a la categoría de referencia que se haya considerado de la variable dependiente.

La regresión logística multinomial es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica) y es una extensión multivariante de la regresión logística binaria clásica. Las variables independientes pueden ser tanto continuas (covariables) como categóricas o mayormente factores (Gonzales, 2015).

Consideremos ahora una variable de respuesta politómica *Y* con más de dos categorías de respuesta que denotaremos por $Y_1, Y_2, \dots, Y_k$. Es posible ajustar un modelo, a partir de la explicación de probabilidad de cada categoría de respuesta en función de un conjunto de covariables $\{x_1, x_2, \dots, x_n\}$ observadas, de la siguiente forma:

$p_j(x) = P[Y = Y_j \,|X = x] = f_j(x) \qquad \forall\, j = 1, \dots, k$ para cada vector x de valores observados de las variables explicativas X.

Cuando se tiene una variable de respuesta binaria, su distribución condicionada a cada combinación de valores observados de las covariables es del tipo Bernouilli. Ahora, si la variable de respuesta es politómica, la distribución de Bernouilli se convierte en una distribución multinomial de parámetros las probabilidades de cada una de las categorías de respuesta. Es decir

$$(\frac{Y}{X} = x) \rightarrow M(1; p_1(x), p_2(x), \dots, p_k(x))$$

Lo que permite que se verifique $\sum_{j=1}^{k} p_j(x) = 1$

Por lo tanto, para la obtención de un modelo lineal, tendremos $\binom{k}{2}$ transformaciones logit, lo que permite que cada par de categorías de la variable respuesta sean comparadas, como se muestra en la siguiente expresión:

$$\ln\left[\frac{\frac{p_i(x)}{(p_i(x)+p_j(x))}}{\frac{p_j(x)}{(p_i(x)+p_j(x))}}\right] = \ln[\frac{p_i(x)}{p_j(x)}]\,, \forall\, j = 1, \dots, k\ (i \neq j)$$

Lo cual representa el logaritmo de la ventaja de respuesta *Yi* frente a *Yj* condicionado a las observaciones de las variables independientes que caen en uno de ambos niveles. Para construir el modelo logit de respuesta multinomial se tendrían que considerar (k − 1) transformaciones logit básicas, definidas con respecto a una categoría de referencia, si se toma la última Y_k como categoría de referencia

Las transformaciones logit generalizadas pueden definirse como:

$$L_j(x) = \ln\left[\frac{p_j(x)}{p_k(x)}\right] \quad \forall\, j = 1, \dots, k-1$$

Donde $L_j(x)$ representa el logaritmo de la ventaja de respuesta *Yj* puesto que las observaciones de las variables independientes caen en alguna de las categorías *Yj* o *Yk*.

Con base en lo anteriormente expuesto, podemos decir que el modelo lineal para cada una

de las transformaciones logit generalizadas, con n variables explicativas, es de la siguiente forma:

$$L_j(x) = \sum_{s=0}^{n} b_{sj} x_s = x' b_j \quad \forall\, j = 1, \dots, k-1$$

Para cada vector de valores observados correspondiente a las variables explicativas $x = (x_1, x_2, \dots, x_n)'$ con $x_0 = 1$ y $b_j = (b_{0j}, b_{1j}, \dots, b_{nj})'$ el vector de parámetros asociado a la categoría Y_j

El modelo, para las probabilidades de respuesta es de la siguiente forma:

$$p_j(x) = \frac{e^{\sum_{s=0}^{n} b_{sj} x_s}}{1 + \sum_{j=1}^{k-1} e^{\sum_{s=0}^{n} b_{sj} x_s}} \quad \forall\, j = 1, \dots, k-1$$

$$p_k(x) = \frac{1}{1 + \sum_{j=1}^{k-1} e^{\sum_{s=0}^{n} b_{sj} x_s}}$$

Una expresión reducida del modelo a partir de ambas expresiones, seria:

$$p_j(x) = \frac{e^{\sum_{s=0}^{n} b_{sj} x_s}}{\sum_{j=1}^{k-1} e^{\sum_{s=0}^{n} b_{sj} x_s}} \quad \forall\, j = 1, \dots, k-1$$

donde $b_{sk} = 0 \quad \forall\, s = 0,1, \dots, n$

2.1.2 Interpretación de parámetros

Este procedimiento depende en gran parte de las variables que hacen parte del modelo que bien pueden ser cualitativas o cuantitativas. A continuación, se describirá cada uno de estos casos:

Una Variable Predictora cuantitativa

Dado un modelo con una única covariable X, por cada valor observado de la forma :

$$L_j(x) = \alpha_j + b_j x \qquad \forall j = 1, \dots, k-1$$

La respuesta de Y_j frente a Y_k al ser incrementado en 1 la covariable independiente X, nos

permite interpretar las exponenciales del conjunto de parámetros b_j en términos de cocientes de ventajas:

$$\theta_j(\Delta X = 1) = \frac{\frac{p_j(x+1)}{p_k(x+1)}}{\frac{p_j(x)}{p_k(x)}} = \frac{e^{(\alpha_j + b_j(x+1))}}{e^{(\alpha_j + b_j)}} = e^{b_j} \qquad \forall j = 1, \dots, k-1$$

Múltiples variables Predictoras Cuantitativas

Para el análisis correspondiente al modelo logit generalizado múltiple, los cocientes de ventajas se definen incrementando una de las variables y controlando fijas las demás.:

$$\theta_s(\Delta X_r = 1 \mid X_s = x_s, s \neq J) = \frac{\frac{P[Y = Y_j | X_r = x_r + 1, X_s = x_s, s \neq r]}{P[Y = Y_k | X_r = x_r + 1, X_s = x_s, s \neq r]}}{\frac{P[Y = Y_j | X_r = x_r, X_s = x_s, s \neq r]}{P[Y = Y_k | X_r = x_r, X_s = x_s, s \neq r]}} = e^{b_{rj}}$$

Luego, el cociente de ventajas de respuesta Yj frente a la última categoría, y Yk cuando aumenta en una unidad la variable Xr y las demás se controlan fijas, está representado por $\theta_j(\Delta X_r = 1 \mid X_s = x_s, s \neq r)$ de acuerdo a las condiciones previamente descritas.

Variables Predictoras Categóricas

Las variables independientes categóricas, son introducidas mediante una transformación para ser interpretadas a través de sus respectivas variables de diseño también denominadas *Dummies.* Esta transformación para la variable C con categorías $C_1, C_2, \dots, C_p$ emplea un método parcial consistente en la asignación de un 1 a la variable asignada a cada categoría y un 0 a las demás, con lo que se obtiene p-1 nuevas variables expresadas como :

$$X_m^C (m = 2, \dots, p)$$

El modelo de regresión logística obtenido, es un para cada uno de los logit generalizados en función de las variables dummie derivadas de la variable C. Así si definiéramos $p_{j/k}$ como la probabilidad de respuesta de la variable dependiente Y_s

$$L_{j/_l} = \ln\left[\frac{p_{j/_l}}{p_{k/_l}}\right] = \beta_{0j} + \sum_{m=2}^{p} \tau_{mj}^{C} X_{lm}^{C} \quad \text{donde } l = 1, \dots, p; j = 1, \dots, k-1$$

Con $\tau_{1j} = 0 \quad \forall\, j$=1, …,$k$−1, $\forall\, l$=2, …,p , podemos obtener un modelo equivalente dado por :

$$L_{j/_l} = \ln\left[\frac{p_{j/_l}}{p_{k/_l}}\right] = \beta_{0j} + \tau_{lj} \quad \text{donde } l = 1, \dots, p; j = 1, \dots, k-1$$

En términos de cocientes de ventajas el modelo puede ser expresado de la siguiente forma:

$$\theta_{j/_{l1}} = \frac{\frac{p_{j/_l}}{p_{k/_l}}}{\frac{p_{j/_1}}{p_{k/_1}}} = \frac{e^{\beta_{0j}+\tau_{lj}}}{e^{\beta_{0j}}} = e^{\tau_{lj}}$$

Lo cual representa el cociente de ventajas de la respuesta Y_j frente a Y_k para la categoría C_j con respecto a C_1 como primera categoría.

2.2. Estimación por Máxima Verosimilitud

Este análisis permite determinar los cocientes del modelo, así como los errores estándar y, por consiguiente, la máxima probabilidad de la variable de estudio, a partir del hecho de que la regresión logística predice probabilidades, en lugar de solo clases, lo que posibilita hacer ajustes mediante el uso de la estimación de máxima verosimilitud. La estimación de los cocientes en el caso de la regresión logística multinomial se lleva a cabo mediante métodos de iteración, concretamente el de Newton -Raphson.

El método de estimación de máxima verosimilitud para el cálculo de los cocientes de nuestro modelo de regresión logística multinomial es como se describe a continuación:

Si se tiene una muestra aleatoria N de tamaño con Q combinaciones diferentes de valores de las variables explicativas X_1, … ,X_n. Cada combinación de valores de las variables explicativas estará dada por $x_q = \left(x_{q0}, x_{q1}, \dots, x_{qn}\right)'$ donde $x_{q0} = 1 \;\; \forall\, q = 1, \dots, Q$. Cada una de estas

combinaciones está asociada con una muestra aleatoria de dq observaciones independientes de la variable de respuesta politómica Y, de entre las cuales identificaremos como *yj/q* al número de observaciones que caen en la categoría de respuesta Yj ∀ j = 1, . . . , k, así que podemos verificar que :

$$\sum_{j=1}^{k} y_{j/q} = d_q \quad \text{y} \quad \sum_{q=1}^{Q} d_q = N$$

Así mismo, los vectores $\left(y_{1/q}, \ldots, y_{k/q}\right)' \quad \forall\, q = 1, \ldots, Q$ siguen una distribución de probabilidad multinomiales independientes $M\left(d_q ; y_{1/q}, \ldots, y_{k/q}\right)$, donde

$$p_{\underset{q}{j}} = P[Y = Y_j \mid X = x_q]$$

Verificando que $\sum_{j=1}^{k} p_{\underset{q}{j}} = 1$

La expresión de la función de verosimilitud de los datos viene dada por:

$$V = \prod_{q=1}^{Q} \left(\frac{d_q!}{\prod_{j=1}^{k} (y_{j/q})!} \prod_{j=1}^{k} p_{j/q}^{\,y_{j/q}} \right)$$

A partir de lo cual, podemos observar que el núcleo de verosimilitud esta dado por la expresión:

$$K = \sum_{q=1}^{Q} \sum_{j=1}^{k} y_{j/q} \ln(p_{j/q})$$

La función de verosimilitud también puede ser expresada con la siguiente función auxiliar:

$$\Lambda = -2\ln(V)$$

A partir de la ecuación del modelo logit generalizado multinomial, y sustituyendo en la expresión anterior, obtenemos la expresión correspondiente núcleo de la log-verosimilitud:

$$K = \sum_{q=1}^{Q} \sum_{j=1}^{k} y_{j/q} \left(\sum_{s=0}^{n} b_{sj}\, x_{qs}\right) - \sum_{q=1}^{Q} \left(\sum_{j=1}^{k} y_{j/q}\right) ln\left(\sum_{j=1}^{k} e^{\sum_{s=0}^{n} b_{sj} x_{qs}}\right)$$
$$= \sum_{q=1}^{Q} \sum_{j=1}^{k} y_{j/q} \left(\sum_{s=0}^{n} b_{sj}\, x_{qs}\right) - \sum_{q=1}^{Q} n_q\, ln\left(\sum_{j=1}^{k} e^{\left(\sum_{s=0}^{n} b_{sj} x_{qs}\right)}\right)$$

Si se deriva con respecto de los parámetros:

$$\frac{\Delta K}{b_{sj}} = \sum_{q=1}^{Q} y_{jq}\, x_{qs} - \sum_{q=1}^{Q} n_q\, x_{qs} \frac{e^{\left(\sum_{s=0}^{n} b_{sj} x_{qs}\right)}}{\sum_{j=1}^{k} e^{\left(\sum_{s=0}^{n} b_{sj} x_{qs}\right)}}$$

Así, obtenemos las ecuaciones de verosimilitud con forma matricial:

$$X'_{((n+1)\times Q)} y_{j(Q\times 1)} = X'_{((n+1)\times Q)} \hat{m}_{j(Q\times 1)}$$

Donde $y_j = \left(y_{j/_1}, y_{j/_2}, \dots, y_{j/_Q}\right)'$ y $\hat{m}_j = \left(\hat{m}_{j/_1}, \hat{m}_{j/_2}, \dots, \hat{m}_{j/_Q}\right)'$ con $\hat{m}_{j/_Q}$ la frecuencia esperada de respuesta *Yj* en la combinación *xq* de valores observados de las variables predictoras, estimada bajo el modelo y que puede definirse como como $\hat{m}_{j/_Q} = d_Q \hat{p}_{j/_Q}$

Los estimadores de máxima verosimilitud son obtenidos mediante la resolución de *k-1* sistemas de *n+1* ecuaciones no lineales, mediante el método iterativo de Newton-Raphson.

Por lo tanto, el estimador de los parámetros $\hat{b}$, que es una matriz de dimensión *(n+1)x(k-1)* formado por las columnas $\hat{b} = (\hat{b}'_1, \hat{b}'_2, \dots, \hat{b}'_{(k-1)})'$ siendo $\hat{b}_j{}'$ el estimador de máxima verosimilitud del vector de parámetros asociado a la categoría de la variable dependiente *Yj*.

Posteriormente, se debe calcular la matriz de covarianzas de $\hat{b}$, que es la inversa de la matriz de información de Fisher. Primero, determinaremos la matriz de covarianzas de cada vector de parámetros $\hat{b}_j$. Para ello hay que calcular las derivadas segundas de K para r≠s:

$$\frac{\Delta^2 K}{\Delta b_{rj} \Delta b_{sj}} = -\sum_{q=1}^{Q} n_q x_{qs} x_{qr} \frac{e^{(\sum_{s=0}^{n} b_{sj} x_{qs})} [\sum_{j=1}^{k} e^{(\sum_{s=0}^{n} b_{sj} x_{qs})} - e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}]}{[\sum_{j=1}^{k} e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}]^2}$$

La expresión de la matriz de covarianzas viene dada por:

$$Cov(\hat{b}_j) = [-E(\frac{}{\Delta b}$$

Para determinar las matrices de covarianzas cruzadas entre cada par de estimadores $\hat{b}_i$ y $\hat{b}_j$ (*i≠j*), se calculan las derivadas segundas de K con *r≠s* y *j≠i:*

$$\frac{\Delta^2 K}{\Delta b_{ri} \Delta b_{sj}} = -\sum_{q=1}^{Q} n_q x_{qs} x_{qr} \frac{-e^{(\sum_{s=0}^{n} b_{sj} x_{qs})} e^{(\sum_{s=0}^{n} b_{si} x_{qs})}}{[\sum_{j=1}^{k} e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}]^2}$$

Dando lugar a la siguiente expresión de la matriz de covarianzas:

$$Cov(\hat{b}_j, \hat{b}_i) = [-E(\frac{\Delta^2 K}{\Delta b_{ri} \Delta b_{sj}})]^{-1} = [X' Diag\left[d_q p_{j/_q} p_{i/_q}\right] X]^{-1}$$

Finalmente, tenemos que la matriz de covarianzas del estimador $\hat{b}$ es:

$$Cov(\hat{b}) = \begin{pmatrix} cov(\hat{b}_1) & cov(\hat{b}_1,\hat{b}_2) & \cdots & cov(\hat{b}_1,\hat{b}_{k-1}) \\ cov(\hat{b}_1,\hat{b}_2) & cov(\hat{b}_2) & \dots & cov(\hat{b}_2,\hat{b}_{k-1}) \\ \dots & \dots & \dots & \dots \\ cov(\hat{b}_1,\hat{b}_{k-1}) & cov(\hat{b}_2,\hat{b}_{k-1}) & \dots & cov(\hat{b}_{k-1}) \end{pmatrix}$$

2.3 Medidas de inferencia en modelos de regresión logística multinomial

La inferencia permite llevar a cabo el análisis de los parámetros del modelo de regresión logística a partir de los datos de la muestra, con el fin de extrapolar los resultados de la misma.

2.3.1 Intervalos de confianza

Para los parámetros del modelo, es posible construir intervalos de confianza a partir de los estimadores de máxima verosimilitud mediante el uso de la distribución normal e intervalos de Odds Ratio mediante transformaciones correspondientes.

2.3.1.1 Intervalos de confianza para los Parámetros y los Odds Ratio

Para cada parámetro de b_{sj} con *j=1, ...,k* construiremos un intervalo de confianza definido como 1-α. Luego, se definirá a N($b_{sj}, \hat{\sigma}^2(\hat{b}_{sj})$) como la distribución asintótica de $\hat{b}_{sj}$, donde $\hat{\sigma}^2(\hat{b}_{sj})$ corresponde al error estándar del estimador asociado con el parámetro b_{sj}

Luego a partir de:

$$P[-Z_{\alpha/2} \leq \frac{\hat{b}_{sj} - b_{sj}}{\hat{\sigma}^2(\hat{b}_{sj})} \leq Z_{\alpha/2}] = 1-\alpha$$

Podremos definir el intervalo de confianza para b_{sj} como

$$IC(b_{sj}) = (\hat{b}_{sj} \pm Z_{\alpha/2}\hat{\sigma}^2(\hat{b}_{sj}))$$

De otra parte, para los Odds ratio, los cocientes de ventajas vienen dados por:

$\theta_j(\Delta X_r=1 X_s=x_s/, s\neq r) = e^{b_{js}}$ ∀ r=1, ...,n; ∀ j=1, ... ,k−1)

A diferencia del caso anterior, para la validación del intervalo de confianza de los odds

ratio se usa el exponencial del intervalo para cada uno de los b_{sj} sobre el nivel $1-\alpha$

$$IC(b_{sj}) = e^{(\hat{b}_{sj} \pm z\alpha_{/2}\hat{\sigma}^2(\hat{b}_{sj}))}$$

Estos intervalos de confianza al 95%, mediante el análisis de los exponenciales de los cocientes, posibilitaran la interpretación del modelo y sus resultados en función de los Odds ratios.

Los Odds Ratio representan el cociente entre de las probabilidades de sucesión que permite establecer una medida de asociación entre dos variables, por lo que es de gran relevancia en modelos de regresión logística que permiten el análisis de estas probabilidades. Según Cerda et al. (2013), son una medida de efecto comúnmente utilizada para comunicar los resultados de investigaciones en áreas de la salud. Matemáticamente corresponde a un cociente entre dos Odds, siendo este una forma alternativa de expresar la posibilidad de ocurrencia de un evento de interés o de presencia de una exposición.

2.4 Pruebas de bondad de ajuste

El análisis de la bondad de ajuste en la regresión logística intenta establecer que tan bien un modelo se ajusta a los datos y por lo general, se aplica después de seleccionar el modelo final. A través de este estudio se valoran los indicadores más relevantes en el análisis del nivel del ajuste, siendo uno de ellos, el doble del logaritmo del estadístico de verosimilitud.

Ahora bien, supongamos que Q representa el número de todas las combinaciones de las variables independientes explicativas que interfieren en el modelo. Si definimos como Y_{qj} el número de observaciones relacionado con la variable dependiente de respuesta, podremos denotar como $\hat{p}_{j/q}$ el estimador de verosimilitud, por lo que las frecuencias esperadas están dadas por:

$$\hat{m}_{j/q} = d_q \hat{p}_{j/q}$$

De esta manera podemos llevar a cabo un análisis de comparación entre las probabilidades observadas y las probabilidades estimadas a través del test global de ajuste con relación al modelo, contrastando la siguiente hipótesis:

$$H_0: p_{j/q} = \frac{e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}}{1+\sum_{j=1}^{k-1} e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}} \quad \forall\, Q = 1, \dots, q;\ \forall j = 1, \dots, k$$

$$H_1: p_{j_{/q}} \neq \frac{e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}}{1+\sum_{j=1}^{k-1} e^{(\sum_{s=0}^{n} b_{sj} x_{qs})}} \quad para\ determinado\ q, j.$$

De esta manera, con la evaluación de H_0 se realizará la comprobación de que los datos permiten un ajuste al modelo de regresión logística multinomial.

2.4.1 Test Chi Cuadrado de Pearson

Se trata de una prueba no paramétrica que permite someter a prueba hipótesis referidas a distribuciones de frecuencias para la evaluación de un modelo de regresión logística multinomial M. Está expresado en la forma:

$$\chi^2(M) = \sum_{q=1}^{Q} \sum_{j=1}^{k} \frac{(y_{j/q} - d_q \hat{p}_{j/q})^2}{d_q \hat{p}_{j/q}}$$

Donde $\hat{p}_{j/q}$ es la estimación por máxima verosimilitud. Puede observarse que sigue una distribución chi cuadrado con grados de libertad establecidos a partir de la diferencia entre el número de parámetros $p_{j/q}$ y Q-(n+1)x(k-1) como el número parámetros independientes, por lo que :

$$\chi^2(M) \rightarrow \chi^2_{Q-(n+1)x(k-1)} \text{ con } d_q \rightarrow \infty$$

Con un nivel de significación para α, si se verifica la condición de:

$$\chi^2(M)_{obs} \geq \chi^2_{Q-(n+1)x(k-1);\alpha}$$

Podremos rechazar la hipótesis nula, o de manera similar, si p_valor $\leq \alpha$, tomamos el p

valor del contraste como la probabilidad acumulada a la derecha de del valor observado :

$$\text{p\_valor} = P[\chi^2(M)_{obs} \geq \chi^2(M)_{obs}]$$

2.4.2 Test de Chi cuadrado Razon de Verosimilitudes: Estadístico de Wilks

El test de chi cuadrado de razón de verosimilitudes (Wiks, 1935), se trata de una prueba alternativa para la medición de contraste de bondad del ajuste para un modelo de regresión logística M, mediante el cual se obtiene como menos dos veces el logaritmo del cociente entre las frecuencias observadas y las frecuencias esperadas, por lo que mientras el estadístico Chi cuadrado de Pearson se basa en las diferencias entre estas frecuencias, la razón de verosimilitud Chi-cuadrado se basa en el cociente entre ellas. La razón de verosimilitud Chi-cuadrado es una alternativa al estadístico Chi-cuadrado cuando el objetivo es contrastar la hipótesis de independencia entre las variables.

Este estadístico es de la forma:

$$G^2(M) = 2\sum_{q=1}^{Q}\sum_{j=1}^{k} y_{j/q} \ln\left(\frac{y_{j/q}}{\hat{m}_{j/q}}\right)$$

Puede observarse que sigue una distribución chi cuadrado con grados de libertad establecidos a partir de la diferencia entre el número de parámetros $\hat{m}_{j/q}$ y b_{sj} bajo el modelo, es decir Q−(n+1)x(k-1) grados de libertad, por lo que :

$$G^2(M) \rightarrow \chi^2_{Q-(n+1)x(k-1)}$$

$$d_q \rightarrow \infty$$

De la misma manera, con un nivel de significación para α, si se verifica la condición de:

$$G^2(M)_{obs} \geq \chi^2_{Q-(n+1)x(k-1);\,\alpha}$$

Podremos rechazar la hipótesis nula.

De manera similar, la hipótesis nula también será rechazada cuando p_valor $\leq \alpha$, si :

p_valor =$P\ [G^2(M) \geq G^2(M)_{obs}] \leq \alpha$

2.5 Valoración de la Calidad de Ajuste

Evaluaremos la calidad del modelo mediante el análisis de algunos indicadores relacionados con la calidad del ajuste, específicamente los cocientes denominados Pseudo R^2. Estos permiten la comparación de modelos de regresión logística multinomial con diferente número de variables predictoras.

Los siguientes son los cocientes utilizados como criterio para nuestro análisis.

2.5.1 Cociente Pseudo-R^2 de McFadden

Los modelos de regresión logística pueden ser ajustados mediante el método de máxima verosimilitud, es decir, las estimaciones de los parámetros son aquellos valores que maximizan la probabilidad de los datos que se han observado. La medida R cuadrado de McFadden se define con:

$$R^2_{McFadden} = 1 - \frac{\ln(L_C)}{\ln(L_0)}$$

$$R^2_{McFadden} = 1 - \frac{\Lambda_f}{\Lambda_0}$$

donde L_c denota el valor de probabilidad (maximizado) del modelo ajustado, y L_0 representa el valor correspondiente, pero para el modelo nulo, el modelo con solo una intersección y sin covariables.

Si el modelo no tiene capacidad predictiva, aunque el valor de probabilidad para el modelo actual será (siempre es) mayor que la probabilidad del modelo nulo, no será mucho mayor. Por lo tanto, la razón de las dos verosimilitudes logarítmicas será cercana a 1, y $R^2_{McFadden}$ será cercana a cero, como debería suponerse.

A continuación, asumiendo que el modelo actual explica toda la variación en el resultado, que denotaremos Y, teniendo en cuenta que el propósito del modelo de regresión logística es dar una predicción para *P (Y = 1)* para cada sujeto, necesitaríamos P (Y = 1) ~ 1 para todos

aquellos casos que tuvieran Y = 1, y P (Y = 1) ~ 0 para aquellos casos que tienen un Y = 0. Si este es el caso, la probabilidad para Y = 1 cuando P (Y = 1) ~ 1 es casi 1, y de manera similar, la probabilidad de ver Y = 0 cuando P (Y = 1) ~ 0 es casi 1. Lo anteriormente expuesto nos da indicios de que el valor de probabilidad para cada observación es cercano a 1. El logaritmo de 1 es 0, por lo que el valor logarítmico de probabilidad log L_c será cercano a 0. Entonces $R^2_{McFadden}$ estará cerca de 1.

2.5.2 Cociente Pseudo-R2 de Cox-Snell.

A diferencia de McFadden, con el Cociente pseudo-R2 de Cox-Snell se utiliza directamente la función de verosimilitud V. Tomando en cuenta el modelo nulo denotaremos el máximo valor de verosimilitud como $V_0 = e^{(-\Lambda 0/2)}$ y así mismo para el modelo ajustado lo definiremos como $V_f = e^{(-\Lambda_f/2)}$, por lo que definiremos el cociente pseudo-R2 de Cox-Snell como:

$$R^2_{CS} = 1 - \left(\frac{V_0}{V_f}\right)^{\frac{2}{N}} = 1 - e^{\left(\frac{\Lambda_f - \Lambda_0}{N}\right)}$$

donde N es el tamaño de la muestra. Es apropiado, entonces, describir esto como un R2 "generalizado" en lugar de un pseudo R^2. Por el contrario, el McFadden R^2 no tiene los mínimos cuadrados ordinarios R^2 como un caso especial.

Esta propiedad de Cox-Snell R2 es buena, especialmente porque la fórmula puede extenderse naturalmente a otros tipos de regresión estimada por máxima verosimilitud.

Sin embargo, el problema del Cox-Snell R^2 es que tiene un límite superior inferior a 1.0. Específicamente, el límite superior es 1 -$V_0^{2/N}$. Esto puede ser mucho menor que 1.0, y depende solo de p.

Este cociente tiene un rango que esta entre $0 \leq R^2_{CS} \leq 1 - V_0^{2/N}$, lo que implica que si se depende totalmente de V_0 sea poco interpretable.

2.5.3 Cociente pseudo R^2 de Nagelkerke.

El Cociente pseudo R^2 de Nagelkerke se puede considerar como un ajuste del cociente de Cox-Snell" para abordar el problema en el que el límite superior de la R al cuadrado de Cox-

Snell no es 1. Para esto dividimos la R de Cox-Snell al cuadrado por su mayor valor posible. La expresión es como sigue:

$$R_N^2 = \frac{R_{CS}^2}{1 - V_0^{\frac{2}{N}}} = \frac{1 - e^{(\frac{\Lambda_f - \Lambda_0}{N})}}{1 - e^{(\frac{-\Lambda_0}{N})}}$$

Obteniéndose un rango de valores entre $0 \leq R_N^2 \leq 1$. Dado que R_{CS}^2 no puede alcanzar un valor de 1, el R^2 de Nagelkerke se desarrolló para tener propiedades más similares al estadístico R^2 utilizado en la regresión ordinaria.

2.6 Medidas de validación de bondad del ajuste

2.6.1 Tasa de clasificaciones correcta

Se trata de un parámetro que permite la verificación con base en los niveles de ajuste del modelo, para establecer la proporción de sujetos clasificados de manera correcta según su categoría observada de la variable dependiente, por lo que este será clasificado de manera correcta si la categoría predicha una vez efectuado el desarrollo del modelo, coincide con la categoría observada.

Este nivel de clasificaciones correctas se calcula mediante el análisis posterior al desarrollo del modelo de regresión logística como un cociente entre el número de individuos clasificados correctamente y el número total de individuos que constituyen la muestra analizada.

Para modelos de regresión logística multinomial, estas clasificaciones se harán con base en aquella categoría donde la probabilidad estimada sea la más alta.

2.7 Contrastes sobre los parámetros del modelo

Esta validación permitirá establecer el nivel de significancia estadística asociado a cada uno de los cocientes de regresión, mediante la evaluación de nulidad para el contraste de un subconjunto de parámetros $b=(b_1, \ldots, b_r)$' en el modelo de regresión analizado.

El siguiente es el contraste de hipótesis:

$H_0: b = 0$

$H_1: b \neq 0$

A continuación, revisaremos dos de los métodos para la validación de los contrastes de parámetros, usados para modelos de regresión logística multinomial.

2.7.1 Contraste de Wald

Este test permite el análisis asociado a la estimación de la máxima verosimilitud $\hat{b}$ del parámetro de interés del modelo comparado con el valor propuesto, asumiendo que la diferencia entre ambos sigue una distribución normal. Se determina esencialmente a partir de los datos asociados a la normalidad asintótica de los estimadores de máxima verosimilitud, bajo la suposición de que la diferencia tipificada entre ambos seguirá aproximadamente una distribución normal. El estimador tiene una media normal b y una matriz de covarianzas $\widehat{cov}(\hat{b})$, por lo que el estadístico de Wald en su forma cuadrática viene dado por la siguiente expresión:

$$\hat{b}'[\widehat{cov}(\hat{b})]^{-1}\hat{b}$$

El cual presenta una distribución cuadrática con r grados de libertad.

Evaluando este estadístico, podemos observar que si el valor observado es mayor o igual que el cuantil de orden 1-α, rechazaremos la hipótesis nula al nivel de significación α, por lo que si

$H_0: b_{sj} = 0$

$H_1: b_{sj} \neq 0$

Entonces el estadístico de Wald estará dado por:

$$W = \frac{\hat{b}_{sj}^2}{\hat{\sigma}^2(\hat{b}_{sj})}$$

Por lo que si $W_{Obs} \geq \chi^2_{1;\alpha}$ con un nivel de confianza de 1- α, rechazaremos la hipótesis

nula, lo que implica que este cociente puede ser tenido en cuenta en el modelo al ser diferente de cero.

Por último, debe de tenerse en cuenta que cuando se analizan modelos con niveles de error estándar significativamente grandes, el estadístico de Wald no brinda fiabilidad en la estimación, al suministrar falsas ausencias de significación, por lo que se sugiere usar otros métodos alternativos como la estimación por máxima verosimilitud.

2.7.2. Contrastes condicionales de razones de verosimilitud

Este método consiste en la eliminación de cada una de las covariables, lo que implica el contraste del modelo resultante al efectuar el anterior procedimiento, frente al modelo completo. La validación del parámetro asociado con la ausencia de significación, permite establecer si el modelo sin la covariable eliminada no empeora con relación al modelo completo. De esta manera se determina el aporte de cada covariable en el modelo por lo que, mediante el principio de parsimonia, es decir, la obtención del modelo más reducido, se eliminan aquellas covariables que no tienen un aporte significativo en el mismo.

Si tenemos el modelo de regresión logística representado como M_G podemos a partir de un subconjunto de parámetros $b = (b_0, \dots, b_r)$ llevar a cabo el contraste que determine la nulidad de los mismos. Supongamos ahora que existe un *Mp* que representa el subconjunto de parámetros nulos., lo que implicaría que este contenido a su vez en el modelo M_G. De esta manera podríamos plantear las siguientes hipótesis:

$H_0: b = 0 \ \ (Verificando\ M_p)$

$H_1: b \neq 0 \ (Si\ se\ asume\ un\ determinado\ M_G)$

Mediante esta evaluación, se contrasta la hipótesis nula de que todos los cocientes toman el valor cero contra hipótesis alternativa de que el modelo que se está considerando actualmente es preciso y difiere significativamente del modelo con parámetros nulos, es decir, brinda un nivel de predicción mejor con respecto a la predicción de la hipótesis nula.

Ahora bien, si cumple que M_p se ha verificado, a través del test de razón de verosimilitudes estará dado por la siguiente expresión:

$$G^2(M_p|M_G) = -2(L_p - L_G) = G^2(M_p) - G^2(M_G)$$

Donde L_p y L_g representan los valores máximos de la log-verosimilitud, suponiendo que los modelos M_p y M_G se han verificado, lo que permite definir el test de razón de verosimilitud como la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada uno de los modelos.

Luego, si tenemos un *r* que representa el número de parámetros que se anulan para H_0 o bien la diferencia entre los grados de libertad de estas distribuciones chi- cuadrado asintóticas $G^2(M_p)\ y\ G^2(M_G)$, entonces para el estadístico $G^2(M_p|M_G)$ se rechazará la hipótesis nula al nivel de significancia α si:

$G^2_{obs}(M_p|M_G) \geq \chi^2_{r;\alpha}$

2.8 Métodos de Selección

Para llevar a cabo la selección del modelo deberemos tener en cuenta la metodología para la introducción de las variables independientes en el análisis. A través del uso de diversos modelos, intentaremos implementar el modelo de regresión logística multinomial a partir del mismo conjunto de variables. El objetivo principal consiste en la formulación de estrategias con el fin de realizar una selección de las variables que mejor explican a la variable de respuesta, es decir, la selección del modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla en términos de cocientes de ventajas.

2.8.1 Criterio de Información de Akaike

El Criterio de información de Akaike (1974) es un método que nos será de gran utilidad a lo largo de este análisis, ya que permite establecer una medida de calidad para el modelo estadístico a partir de un conjunto de datos, a través de la optimización de la bondad del ajuste mediante la selección del menor AIC. Basado, principalmente, en la entropía del proceso, el AIC es un estimador insesgado asintótico, entre un modelo candidato ajustado y el verdadero modelo, ofreciendo una estimación relativa de la información a través de la utilización de este, en la representación del proceso que genera los datos.

El análisis evalúa el ajuste del modelo en los datos de entrenamiento para establecer el resultado deseado mediante el AIC más bajo posible, que indique el mejor equilibrio entre el ajuste del modelo y la generalización. Esto sirve al objetivo final de maximizar el ajuste en datos fuera de la muestra.

$$AIC = -2\,ln(L) + 2k$$

AIC utiliza la estimación de máxima verosimilitud de un modelo (log-verosimilitud) como medida de ajuste.

La probabilidad logarítmica es una medida de la probabilidad de que uno vea sus datos observados, dado un modelo. El modelo con la máxima probabilidad es el que mejor se "ajusta" a los datos. El logaritmo natural de la probabilidad se usa como una conveniencia computacional. Así mismo dentro del método utilizado para ajustar el mejor modelo, se encuentran tres métodos: hacia adelante, hacia atrás, y el modo stepwise. Los procedimientos de cada método son descritos en base a Dueñas (2012):

Hacia adelante:

1. Se inicia con un modelo vacío (sólo la constante).
2. Se ajusta un modelo y se calcula el p-valor del contraste de razón de verosimilitud que resulta de incluir cada variable por separado.
3. Se debe seleccionar el modelo con el p-valor más significativo.
4. Se ajusta de nuevo un modelo con la(s) variable(s) seleccionada(s) y se calcula el p-valor de añadir cada variable no seleccionada anteriormente por separado.
5. Se selecciona el modelo con el más significativo.
6. Se deben repetir los pasos 4 y 5 hasta que no queden variables significativas para incluir.

Hacia atrás

1. Se inicia con un modelo con todas las variables candidatas.
2. Se procede con la eliminación de cada una de las variables, calculando la pérdida de ajuste al eliminar.
3. Se elimina la menos significativa en cada iteracion.
4. Se repite el proceso del paso 2 al 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

Stepwise

1. Este método es una combinación de los procedimientos anteriores explicados.

2. En cada uno de los pasos, se introduce la variable independiente que no se encuentre ya en la ecuación y que tenga la probabilidad para F más pequeña (i.e. hacia adelante).

3. Las variables ya introducidas en la ecuación de regresión pueden ser eliminadas del modelo (i.e. hacia atrás).

4. El método termina cuando ya no hay más variables candidatas a ser incluidas o eliminadas.

Para nuestro caso de interés se utilizó un método backward, que es adecuado cuando hay una cantidad considerable de variables independientes. No obstante, a que se utilizó este método se realizó el análisis con el método completo para analizar las dinámicas de las variables.

2.9 Validación del Modelo: Análisis de Residuos

A través del análisis de los residuos los residuos de Pearson obtendremos una medida relacionada con la evaluación del pronóstico en las observaciones en nuestro modelo, es decir mediante la obtención un indicador de qué tan bien la observación es pronosticada por el modelo mediante los residuos de la devianza a partir de combinación q de las variables explicativas, analizando la categoría j correspondiente a la variable dependiente de respuesta y.

$$r_{\underset{q}{j}} = \frac{y_{\underset{q}{j}} - d_q \hat{p}_{\underset{q}{j}}}{[d_q \hat{p}_{\underset{Q}{s}}]^{\frac{1}{2}}}$$

Capítulo 3

3.1. Estudio de la Aplicación

Basados en los fundamentos teóricos analizados en el capítulo anterior, en éste nos dedicaremos a llevar a la práctica los mismos, mediante la aplicación con datos reales que permitan la implementación de un modelo de regresión logística multinomial.

Para Salinas (2006), los instrumentos de recolección de datos representan cualquier material u objeto que sirva para realizar un análisis y/o observaciones a través de los cuales es posible llevar a cabo un procedimiento altamente eficiente para recopilar información. En cuanto a estas herramientas, por lo general se suelen tener en cuenta dos tipos: aquellas enfocadas en procesos relacionados con investigaciones documentales y/o descriptivas y aquellas usadas en investigaciones experimentales.

Para esto, se emplean los datos de índole socio educativo, correspondientes a una encuesta con datos reales llevada a cabo en dos escuelas de educación básica secundaria en Portugal, en 2017, sobre la cual se recogen diversos datos referentes a aspectos socioeducativos y en la cual se pretende establecer los factores más determinantes por parte de los estudiantes de último curso en la escogencia de una institución. Este estudio consta de 26 variables las cuales serán descritas a continuación:

Escuela: Variable del tipo binario que toma los siguientes valores

- EA= Escuela A
- EB=Escuela B

Sexo: Variable del tipo binario que representa el sexo del estudiante

- F = Femenino
- M =Masculino.

Edad: Variable numérica que representa el rango de edad de los estudiantes de la encuesta y que se encuentra entre los 15 y 22 años.

Ubicación Hogar: Variable binaria que identifica el área de ubicación del hogar del estudiante que bien puede ser:

- U – urbana
- R – rural.

TamaFam: Variable binaria que indica el tamaño de la familia del alumno y que bien puede ser:

- LE3 – menor igual a tres integrantes
- GT3 – mayor a tres integrantes.

Pstatus: Es una variable binaria que indica status de los padres del estudiante

- T – Viviendo juntos
- A – separados

Maedu: Nivel de educación de la madre. Es una variable numérica que puede ser:

- 0 – Ninguna educación
- 1 – educación Primaria (Cuarto grado)
- 2 – Quinto a Noveno grado
- 3 – educación secundaria
- 4 – Universidad.

Paedu: Nivel de educación del padre. Variable numérica que puede ser:

- 0 – Ninguna educación
- 1 – Educación Primaria (Cuarto grado)
- 2 – Quinto a Noveno grado
- 3 – educación secundaria
- 4 – Universidad.

Mtrab: Variable nominal que indica la profesión de la madre del estudiante:

- Profesor
- Sector Salud
- Sector Civil (ejemplo. Administrativo o policía)
- Hogar
- Otros

Patrab: Variable nominal que indica la profesión del padre del estudiante

- Profesor
- Sector Salud
- Sector Civil (ejemplo. Administrativo o policía)
- Hogar
- Otros

AcudienteEstud: Variable nominal que indica quien controla y está pendiente de las actividades del estudiante:

- Madre
- Padre
- Otros

TiempRecorrido: Variable numérica que representa el tiempo en horas que tarda un estudiante en ir de su casa a la escuela.

TiempEstudio: variable numérica que indica el tiempo de estudio de las asignaturas semanal (De 1 a 10 horas).

NumRepro: Esta variable numérica representa el número de veces que el estudiante hareprobado el curso.

SopExtrEdu: Variable binaria que indica si hay soporte extra educacional por parte delestudiante (Si o No)

Famsup: Variable binaria que indica si la familia ayuda en sus tareas y trabajos al estudiante (Si o No).

PagoExtrClas: Variable binaria que indica si el estudiante paga clases extra somo soportea cualquiera de los dos cursos de la muestra (Matemáticas o Idiomas) (Si o No).

ExtraCurricAct: Variable binaria que representa si el estudiante realiza actividades extracurriculares (si o no).

UnivEstudios: Variable binaria mediante la cual se establece si el estudiante espera continuar con sus estudios de educación superior (si o no)

Internet: Es una variable binaria que determina si el estudiante tiene acceso a Internet ensu casa (Si o No).

RelaRomant: Indica si el estudiante tiene una relación romántica (Si o No).

Famrel: Variable numérica que representa el nivel de calidad de las relaciones interfamiliares del estudiante (desde 1 – Muy malas a 5 – Excelentes).

TempLibre: Variable numérica que representa el tiempo libre después de clases desde 1 –Muy Bajo a 5 – Muy Alto.

SalAmigos: Esta variable de tipo numérico representa que tanto sale el estudiante con amigos y abarca desde 1 – muy bajo a 5 – muy alta.

Ausencias: Variable numérica que representa el número de ausencias y que oscila entre 0 y 93.

La variable respuesta en este caso, es categórica nominal y representa el factor que llevo al estudiante, a escoger esta institución para su formación

RazonEscoj : Es la variable de respuesta que determina la razón para escoger una institución educativa y que tiene las siguientes categorías.

- Cercanía a la casa
- Reputación de la institución
- Calidad o nivel del curso
- Otros Factores

A continuación, hemos efectuado los resúmenes de análisis de frecuencias relacionados conalgunas de las variables más relevantes de nuestro estudio. Por ejemplo, para la variable sexo existe una clara prevalencia de mujeres contra hombres en la muestra, como lo indica la figura 1 referente a las frecuencias absolutas, lo que se puede corroborar al analizar las frecuencias relativas en la tabla 4, que nos indican que el 59% de la población son mujeres.

Figura1.

Frecuencia absoluta de la variable Sexo

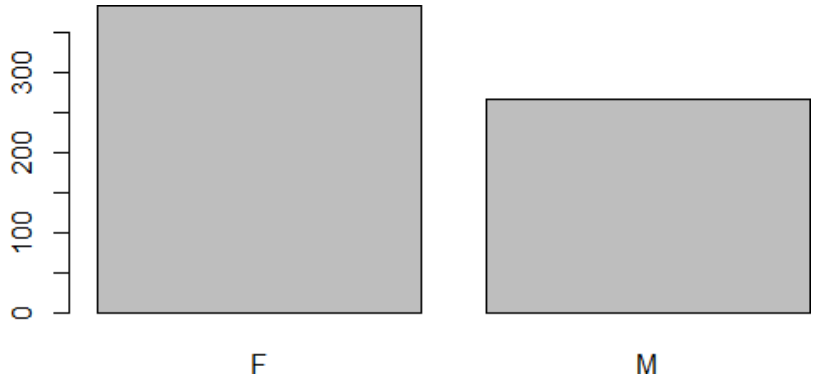


Tabla4.

Frecuencia relativa de la variable sexo

Sexo	
F	M
0.5901387	0.4098613

Otra variable interesante y de gran impacto en este tipo de estudios es la edad. Para este análisis, como puede observarse, las mayores proporciones de estudiantes están en edades que oscilan entre los 16 y los 17 años, con porcentajes de 27,7% y 27,5 respectivamente. Obsérvese el bajo porcentaje de aquellos alumnos con rangos de edad superior, teniendo en cuenta las tendencias estadísticas normales de edad que se manejan a nivel de ingreso a la escuela.

Figura2.

Frecuencia absoluta de la variable edad

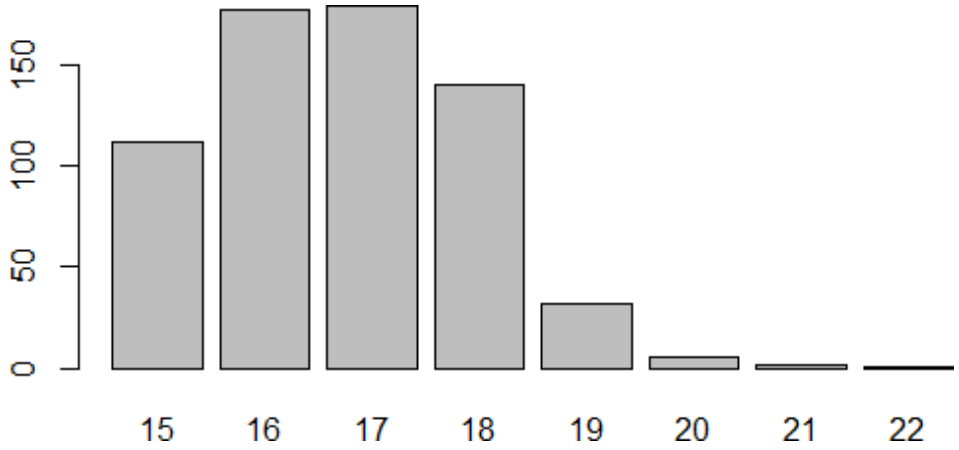


Tabla 5.

Frecuencia relativa de la variable edad.

Edad							
15	16	17	18	19	20	21	22
0.172573190	0.272727273	0.275808937	0.215716487	0.049306626	0.009244992	0.003081664	0.001540832

La variable relacionada con *ubicación hogar*, nos permite ver la distribución de estudiantes de acuerdo a la localización de su vivienda que bien puede ser urbana o rural. Solo un 30,35% de los estudiantes que hacen parte del estudio viven en hogares en zonas rurales.

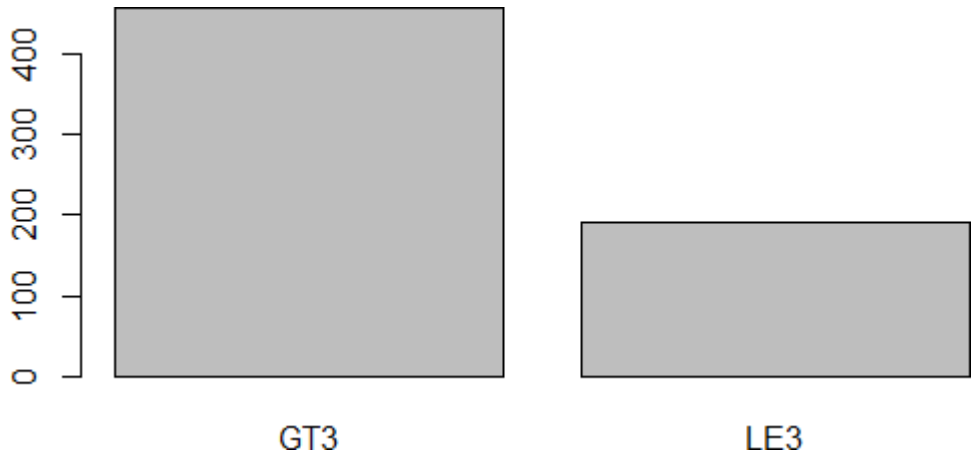
Tabla 6.

Frecuencia relativa de la variable ubicación del hogar

UbicacionHogar	
R	U
0.3035439	0.6964561

Figura 3.

Frecuencia absoluta de la variable tamaño de la familia



El tamaño de la familia es un indicador a través del cual se evalúa el número de integrantes del grupo familiar de cada estudiante. Así, como podemos ver, el 70,45% de los estudiantes viven en núcleos familiares de más de tres personas,mientras el 29,58% restante convive en grupos más pequeños.

Tabla 7.

Frecuencia relativa de la variable composición del núcleo familiar

TamFam	
GT3	LE3
0.7041602	0.2958398

Un factor para evaluar, interesante en el desarrollo del modelo y que no solo podría condicionar el rendimiento académico, sino también las decisiones de cualquier estudiante asociadas a su proceso de formación, es el estatus marital de sus padres. Así podemos ver que en el caso del 87,6% de los estudiantes, sus padres están casados.

Tabla 8.

Frecuencia relativa de la variable Estatus marital de los padres del estudiante

Pstatus	
A	T
0.1232666	0.8767334

Para las variables referentes al nivel educativo de la madre y del padre, observamos por ejemplo el 28,65% de las madres de los alumnos logro terminar un curso entre quinto y noveno grado, mientras que el 26,9% logro completar la educación superior.

Resulta curioso observarque el mayor porcentaje entre los padres, se concentra entre aquellos que solo estudiaron hasta noveno grado con un 32; 2%, mientras que solo el 19% completo sus estudios universitarios.

Tabla 9.

Frecuencia relativa de la variable nivel educativo de la madre

Maedu				
0	1	2	3	4
0.009244992	0.220338983	0.286594761	0.214175655	0.269645609

Tabla 10.

Frecuencia relativa de la variable nivel educativo del padre

Paedu				
0	1	2	3	4
0.01078582	0.26810478	0.32203390	0.20184900	0.19722650

La siguiente tabla, muestra la frecuencia relativa referente a la variable ***TempRecor,*** variable cuantitativa que está relacionada con el tiempo promedio que toma un estudiante en desplazarse hacia su escuela. El mayor porcentaje de estudiantes, en este caso el 56,39% requiere de una hora para llegar a suescuela, en contraste con el 2,4% de la población que necesita de 4 horas en su desplazamiento.

Tabla11.

Frecuencia relativa de la variable tiempo para acudir a la escuela.

TempRecor			
1	2	3	4
0.56394453	0.32819723	0.08320493	0.02465331

Una de las variables analizadas toma en cuenta el sector laboral en el que se desempeñan los padres. Por ejemplo, en este caso, observamos que un porcentaje del 39,75% de las madres se desenvuelve en determinadas profesiones, entre ellas como policía, administradora, vendedora, entre otras. Solo un 7,3% de las madres trabaja con el sector de la salud y un 11% como docente.

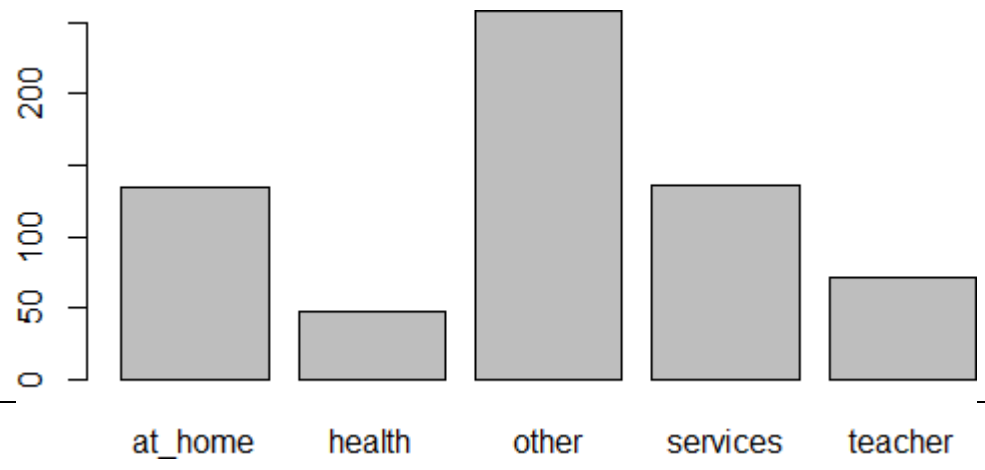
Tabla12.

Frecuencia relativa de la variable profesión de la madre.

MaTrab				
at_home	health	other	services	teacher
0.20801233	0.07395994	0.39753467	0.20955316	0.11093991

Figura 4.

Frecuencia absoluta de la variable profesión de la madre



La variable TempEstudio es cuantitativa y un indicador de las horas destinadas por cada estudiante para la realización de sus actividades académicas. La tabla de frecuencias relativas nos muestra que el 46,9% de los estudiantes destinan dos horas para la preparación de exámenes y otras actividades.

Tabla 13.

Frecuencia relativa de la variable Tiempo dedicado al estudio

TempEstudio			
1	2	3	4
0.32665639	0.46995378	0.14946071	0.05392912

Por último, tenemos los niveles relacionados con el indicador asociado al pago de clases extracurriculares. Solo un 6% de los alumnos ha tomado este tipo de mentorías.

Figura 5.

Frecuencia absoluta de la variable relacionada con pago de clases extracurriculares

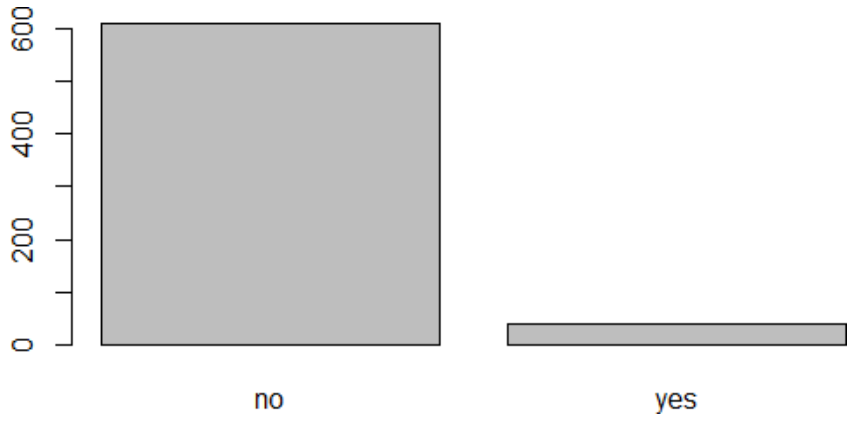


Tabla14.

Frecuencia relativa de la variable pago de clases extracurriculares

PagoExtrClas	
no	yes
0.93990755	0.06009245

3.2 Análisis Preliminares

3.2.1 Análisis Unidimensional

A continuación, se ha implementado un análisis unidimensional con las variables de estudio. Para las variables cualitativas se han calculado porcentajes de las categorías respectivas, mientras que para las variables cuantitativas se validaron estadísticos como la media, desviación típica, mediana, cuartiles y máximos y mínimos.

Tabla 15.

Caracterización de las variables.

VARIABLE	UNIDADES / VALORES QUE TOMA	DESCRIPTIVO
ESCUELA	-Escuela a->	423 (65.18%)
	-escuela b->	226 (34.82%)
SEXO	-femenino->	383 (59.01%)
	-masculino->	266 (40.99%)
UBICACIÓN DEL HOGAR	-rural->	197 (30.35%)
	-urbana->	452 (69.65%)
ESTADO PADRES	-separados->	80 (12.33%)
	-viviendo juntos->	569 (87.67%)
EDUCACIÓN PADRE	-educación pri->	143 (22.03%)
	-Educación secundaria->	139 (21.42%)
	-Educación universitaria->	175 (26.96%)
	-Ninguna educación ->	6 (0.92%)
	-Quinto a noveno->	186 (28.66%)

EDUCACIÓN MADRE	-educación pri->	174 (26.81%)
	-Educación secundaria->	174 (26.81%)
	-Educación universitaria->	131 (20.18%)
	-Ninguna educación ->	128 (19.72%)
	-Quinto a noveno->	7 (1.08%)
TRABAJO PADRE	-En casa->	209 (32.2%)
	-Otro->	135 (20.8%)
	-Profesor->	258 (39.75%)
	-Salud->	48 (7.4%)
	-Servicios->	136 (20.96%)
TRABAJO MADRE	-En casa->	135 (20.8%)
	-Otro->	258 (39.75%)
	-Profesor->	72 (11.09%)
	-Salud->	23 (3.54%)
	-Servicios->	181 (27.89%)
ACUDIENTE DEL ESTUDIANTE	-Madre->	455 (70.11%)
	-Otro->	41 (6.32%)
	-Padre->	153 (23.57%)
SOPORTE EDUCACIÓN	-No->	581 (89.52%)
	-Si->	68 (10.48%)
APOYO FAMILIAR	-No->	251 (38.67%)
	-Si->	398 (61.33%)
RAZÓN DE ELECCIÓN	-Curso ->	285 (43.91%)
	-Hogar->	149 (22.96%)
	-Otros->	72 (11.09%)
	-Reputación->	143 (22.03%)
EDAD	Cantidad de años	Media (DE) : 16 (1.4)
		Min - Max: 15 -22
		Mediana (P25 - P75) : 17 (16;18)
TIEMPO DE RECORRIDO	Cantidad de horas	Media (DE) : 1 (0.5)
		Min - Max: 1 -4
		Mediana (P25 - P75) : 1 (1;2)
TIEMPO DE ESTUDIO	Horas del 1 al 10	Media (DE) : 1 (0.5)
		Min - Max: 1 - 4
		Mediana (P25 - P75) : 2 (1;2)
NÚMERO DE VECES QUE REPROBO UN CURSO	Número de veces	Media (DE) : 0.2219 (3)
		Min - Max: 0 - 0
		Mediana (P25 - P75) : 0 (0;2)
TIEMPO LIBRE	Cantidad de horas	Media (DE) : 3 (1)
		Min - Max: 1 - 5
		Mediana (P25 - P75) : 4 (4;5)

En su mayoría, la muestra estaba compuesta por niñas, quienes correspondieron al 59.01% del total de estudiantes. La escuela más representativa fue la escuela A con un 66.18% del tamaño muestral, el restante correspondía a estudiantes de la escuela B. Alrededor del 70% de los estudiantes estudiaba en escuela urbana y los demás en rural.

El 87.67% de los estudiantes expresó que sus padres vivían juntos y solo el 12.33% reportó que no lo hacían. En cuanto a los niveles educativos de padre y madre, un 26% y un 20% respectivamente indicó que ellos habían terminado estudios de educación media (colegio). Un 21% y 26% respectivamente señaló que no habían terminado secundaria.

En cuanto a los indicadores referentes a las actividades laborales tanto del padre como de la madre, se aprecia que un porcentaje significativo de hombres y mujeres están dedicados al hogar o están desempleados (20%.11% y 23% respectivamente). El segundo empleo más común en las mujeres es el de servicios, el cual corresponde al 27% del total de mujeres. El empleo más común de los hombres fue el de profesor, que correspondió a casi el 40% (39.75%) del total de padres, una cifra bastante considerable que señala que se trata de un empleo muy común en la región.

Alrededor del 70% tienen como acudiente a su madre, situación muy común en los hogares; en general este rol es activamente desempañado por las mismas. Una consideración importante es que un 70% de los estudiantes no tuvo un apoyo significativo por parte de la familia, lo que implica que las actividades extracurriculares quedan en completa responsabilidad individual del estudiante.

En cuanto a la variable dependiente, los estudiantes indicaron que la razón más significativa para elegir la escuela fue el nivel de calidad de contenido del curso (43.91%), seguido por cercanía a la casa,y hogar con una proporción muy similar (22.96% y 22.03% respectivamente).

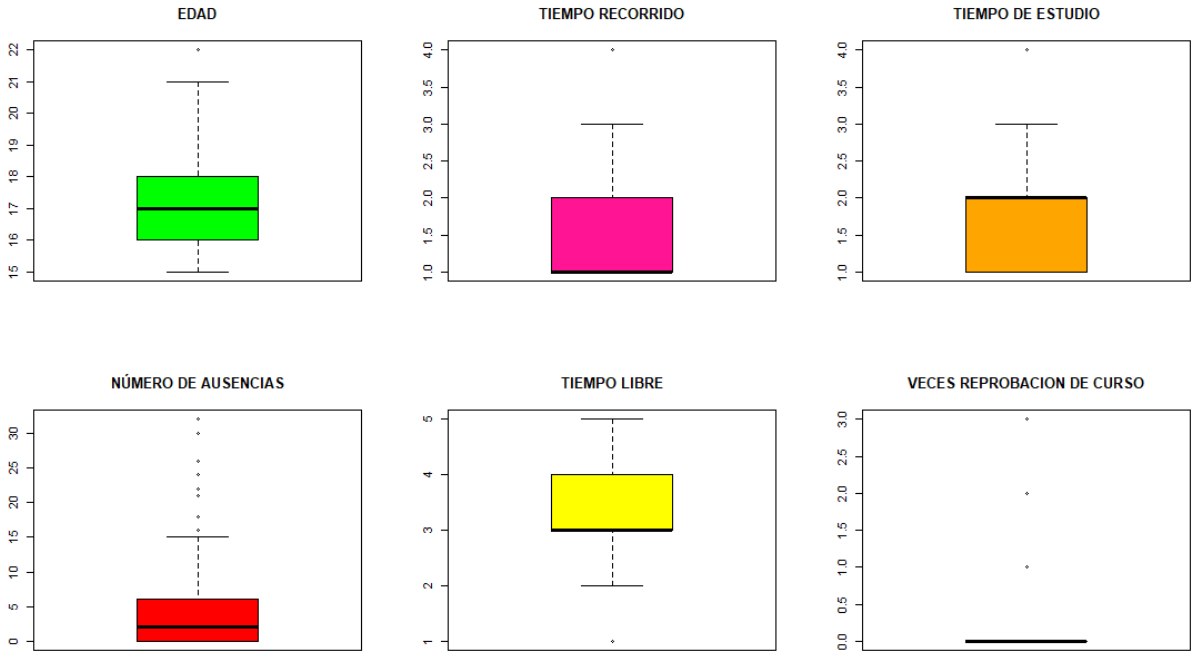
En promedio la edad de los estudiantes varió de 16 a 22 años con una media de 15 años y una desviación de apenas 1.4 años lo que alude a que en cierta forma tenemos datos muy homogéneos o similares. El promedio de las horas de estudio es de solo una hora diaria lo que implica que los estudiantes invierten poco tiempo académico fuera de la escuela.

Por otra parte, los estudiantes invierten media hora, en promedio, para dirigirse a su

centro educativo y consumen 3 horas, en promedio, en actividades de ocio o tiempo libre.

Figura 6.

Diagrama de cajas variables cuantitativas



Nota: la distribución de las variables cuantitativas, conforme a lo evidenciado en la figura, es de poca homogeneidad, a excepción de la variable *edad,* que evidencia mayor simetría.

3.2.2 Análisis Bidimensional

El análisis bidimensional tuvo dos etapas. Primero, un análisis de la variable elección del curso vs variables categóricas. Segundo, una etapa de la variable elección del curso vs variables cuantitativas. Para el caso de los contrastes con variables cuantitativas se realizó una prueba de independencia Chi cuadrado con el fin de establecer si había asociación de cada variable independiente con la variable dependiente de la investigación. Por otra parte, para analizar los contrastes con variables cuantitativas se optó por hacer pruebas de normalidad para establecer el campo de la prueba estadística a realizar (Paramétrico o no paramétrico).

Tabla16.

Análisis Bivariado del conjunto de datos

	Curso	*Hogar*	*Otros*	*Reputación*	*Chi cuadrado*	*P valor*
Escuela						
Escuela A	167	115	27	114	15.37	0.00
Escuela B	118	34	45	29		
	285	**149**	**72**	**143**		
Sexo						
Femenino	176	80	39	88	4.93	0.29
Masculino	109	69	33	55		
UbicacionHogar						
Rural	97	25	30	45	21.59	0.00
Urbano	188	124	42	98		
Tamaño de la familia						
Más de 3	205	102	55	95	5.99	0.20
Menos de tres	80	47	17	48		
Estado de los padres						
Separados	31	22	6	21	16.92	0.00
Viviendo juntos	254	127	66	122		
Educación de la madre						
Ninguna educación	3	1	2	0	11.67	0.02
Educación primaria	76	28	20	19		
Quinto a noveno grado	87	41	18	40		
Educación secundaria	53	40	11	35		
Educación del padre						
Ninguna educación	3	1	1	2	8.85	0.07
Educación primaria	85	35	24	30		
Quinto a noveno grado	94	52	21	42		
Educación secundaria	52	34	10	35		
Universidad	51	27	16	34		

Trabajo de la madre						
En casa	76	23	20	16	11.67	0.02
Salud	14	9	7	18		
Otros	106	70	24	58		
Servicios	55	29	15	37		
Profesor	34	18	6	14		
Trabajo del padre						
En casa	26	8	3	5	8.85	0.07
Salud	6	3	5	9		
Otros	159	93	32	83		
Servicios	79	37	27	38		
Profesor	15	8	5	8		
Acudiente del estudiante						
Padre	63	33	19	38	3.60	0.463
Madre	200	105	52	98		
Otro	22	11	1	7		
Soporte extraeducacional						
No	262	129	64	126	4.30	0.31
Yes	23	20	8	17		
Apoyo familiar						
No	111	54	37	49	2.02	0.59
Yes	174	95	35	94		
Pago de clases extras						
No	268	136	66	140	8.07	0.09
Yes	17	13	6	3		
Actividades extracurriculares						
No	146	88	46	54	4.52	0.34
Si	139	61	26	89		
SupEst						
No	39	12	11	7	3.69	0.45
Si	246	137	61	136		

Internet						
No	82	27	21	21	11.67	0.02
Si	203	122	51	122		
RelaRomant						
No	175	94	42	99	1.98	0.74
Si	110	55	30	44		
FamRela						
1	10	6	2	4	2.57	0.63
2	16	2	4	7		
3	43	31	10	17		
4	137	73	34	73		
5	79	37	22	42		
5	54	20	13	23		

En este caso, el análisis mediante la prueba chi cuadrado tiene cómo hipótesis nula la independencia de las variables. De acuerdo con los resultados de la tabla 16 se aprecia que existe cierta asociación entre la variable escogencia de la institución con variables como trabajo y educación de los padres ($p<0.05$).Por otra parte, las demás variables no evidenciaron asociación significativa con la variable dependiente ($p>0.05$). Es interesante, contrastar estos resultados con los análisis posteriores para la selección de las variables efectuados mediante el procedimiento stepwise.

3.3 Ajuste del modelo logit multinomial para la determinación de los Factores asociados a las razones de escogencia en la selección de una institución de secundaria.

Considerando lo definido en el marco conceptual del presente trabajo, acerca de las bondades y características de los modelos logísticos multinomiales, y luego de determinar que, dado las características de la variable dependiente cualitativa con cuatro categorías de respuesta, dicha técnica se ajusta para el cumplimiento del objetivo de esta investigación. A continuación, se presentan los resultados obtenidos utilizando el programa de libre distribución R y se comentan los hallazgos correspondientes.

La sintaxis básica para los modelos multinomiales en R es: *multinom (fórmula,*

data=datos). El primer término de la fórmula es un factor con más de dos categorías de respuesta y del lado derecho se incluyen tanto los factores (variables cualitativas) como los predictores continuos. Para nuestro estudio, la variable dependiente es la razón de escogencia de instituciones de educación, la cual se introdujo en el modelo como un factor de cuatro categorías donde la primera categoría nivel del curso es tomada como referencia, para el caso, curso que hace referencia al indicador de calidad de curso. Las variables explicativas se definieron previamente (capítulo 3).

Las variables explicativas cualitativas consideradas en el modelo se renombraron adicionando a su nombre original la letra d, indicando que son variables Dummy; para ello, se usó la instrucción *factor* del paquete R. Luego se procedió a identificar las categorías de referencia para lograr una mejor interpretación, a modo de ejemplo se coloca la sintaxis de algunas de ellas.

```
contrasts(Escuelad)
contrasts(UbicacionHogard)
contrasts(Maedud)
contrasts(Paedud)
```

Las variables continuas se han introducido en el modelo con sus valores originales.

3.3.1 Selección del modelo

Para el modelo se seleccionó el método stepwise, del cual se habló en detalle en el capítulo 2 y en el que comenzaremos sólo con la constante.

Habiendo seleccionado el modelo final, de acuerdo con el principio de parsimonia, se contrastaron los parámetros de este mediante el análisis de parámetros como el contraste de Wald, las Odds Ratio de los cocientes, sus intervalos de confianza y los p-valores; lo que posibilita una interpretación óptima del modelo final. El parámetro asociado con la bondad del ajuste global fue medida a través del estadístico de Chi-cuadrado de razón de verosimilitudes y la tasa de clasificaciones correctas. Es importante evaluar la calidad del ajuste, la que fue determinada mediante los cocientes pseudo R-cuadrado de Nagelkerke y McFadden. El último paso consistió en la validación del modelo mediante los residuos de la devianza. El contraste de los modelos iniciales con los nuevos se realiza mediante

contrastes condicionales de razón de verosimilitudes, comparando las devianzas de cada modelo realizándolo en R mediante la función *anova*.

Considerando lo planteado previamente sobre los procedimientos de estimación, se decidió utilizar el procedimiento por pasos sucesivos, es decir, el procedimiento stepwise, para ello, se deben ejecutar los modelos sin variables explicativas (sólo la constante) y otro con todas ellas.

En el anexo incluimos los comandos utilizados en R para ajustar el modelo de manera general, además, junto con el script.

Adicionalmente, se presentan los resultados de la regresión con el modelo vacío (sin variables explicativas) y las interacciones llevadas a cabo por el procedimiento stepwise en el anexo y especificadas allí. Es importante mencionar que, en ese proceso, se han tenido en cuenta los indicadores AIC y el menor de estos es el que se presenta en el último paso, donde quedaron seleccionadas ocho variables cualitativas y ninguna cuantitativa.

El análisis comienza con el análisis del modelo de regresión sin variables explicativas:

```
rlogvacio<-
multinom(formula=RazonEscojd~1,datos,family=binomial(link="logit"))

> summary(rlogvacio,Wald=TRUE)
Call:
multinom(formula = RazonEscojd ~ 1, data = datos)

Coefficients:
           (Intercept)
home        -0.6443565
other       -1.4006575
reputation  -0.6860285

Std. Errors:
           (Intercept)
home         0.1018537
other        0.1344019
reputation   0.1032650

Value/SE (Wald statistics):
           (Intercept)
home         -6.326293
other       -10.421412
reputation   -6.643382

Residual Deviance: 1623.398
AIC: 1629.398
```

Tabla 17.

Indicadores de modelo de regresión con constante.

Variables Independientes	*Variable Dependiente*	*Coeficientes Beta*	*Std. Error*	*Test de Wald*
Contraste	*Cercania a la casa*	-0,6443565	0,1018537	-6,3262930
	Otros Aspectos	-1,4006575	0,1344019	-10,4214120
	Reputacion	-0,6860285	0,1032650	-6,6433820

La siguiente es la instrucción para llevar a cabo la regresión con todas las variables explicativas dummy:

```
rlogcompletodummy<-multinom
(RazonEscojd~Escuelad+Maedud+Paedud+UbicacionHogard+MaTrabd+PaTra
bd+AcudEstudd+TempRecord+TempEstudiod+NumReprod+SopExtraEdud+FamS
upd+PagoExtrClasd+ExtraCurricActd+ExtraCurricActd+SupEstd+Interne
td+RelaRomantd+FamRelad+TempEstudiod,data=datos)
```

Luego de lo anterior, se ejecuta la siguiente instrucción relacionada con el procedimiento stepwise de pasos sucesivos.

```
rlog_stepwise<-step(rlogvacio,
scope=list(lower=rlogvacio,upper=rlogcompletodummy),
                    direction="both")
```

El resultado a través del método stepwise, (detallado en el anexo) nos permitió la selección de ocho variables explicativas de la variable dependiente razón de escogencia, la cual tiene como ya se ha mencionado previamente, cuatro categorías posibles: nivel de calidad del curso, cercanía a la casa, reputación de la institución y otras razones que se agruparon en esta opción. Por lo tanto, las variables explicativas que quedaron en el modelo son:

Escuela: variable del tipo binario que toma los siguientes valores EA= Escuela A o EB= Escuela B.

UbicacionHogar: variable binaria que identifica el área de ubicación del hogar del estudiante que bien puede ser U - urbana o R – rural.

MTrab: variable nominal que indica la profesión de la madre del estudiante: 'Profesor', ‘SectorSalud’, ‘SectorCivil’ (ejemplo. administrativo o policía), 'Hogar' u 'Otros'.

TiempEstudio: variable numérica que indica el tiempo de estudio de las asignaturas

semanal (de 1 a 10 horas).

TiempRecorrido: variable numérica que representa el tiempo en horas que tarda un estudiante en ir de su casa a la escuela.

NumRepro: esta variable numérica representa el número de veces que el estudiante ha reprobado el curso.

PagoExtrClas: variable binaria que indica si el estudiante paga clases extra como soporte a cualquiera de los dos cursos de la muestra (Matemáticas o Idiomas) (Sí o No).

ExtraCurricAct: variable binaria que representa si el estudiante realiza actividades extracurriculares (si o no).

Los cocientes estimados beta no son útiles para interpretar por sí mismos, dado, que son valores en unidades de logaritmos, sin embargo, el signo nos permite identificar cuáles categorías de las variables son menos probables en su aportación para explicar la razón en la elección de una institución. Los cocientes de signos negativos son menos probables y los positivos más. Por ejemplo, la escuela B es menos probable que permita elegir una institución cerca de casa con respecto a la escuela A (-0,599), en ese mismo sentido aclaratorio, se afirma que se da la misma situación en la reputación cuyo valor es de -0,8615.

El error estándar es un valor que se requiere para calcular el estadístico de Wald como se mencionó en la parte conceptual del presente documento. No obstante, se puede observar que dichos valores son relativamente bajos y, por lo tanto, indican que existe poca variabilidad en el estadístico de Wald, factor algo adecuado para el análisis.

Si bien el estadístico de Wald amerita una medida de significancia, también lo es, que la significancia de los cocientes se puede confirmar con el cálculo de los cocientes OR y sus respectivos IC. Antes de ello, serán presentados los gráficos que permiten la identificación de la existencia de la relación entre la razón de escogencia y las variables definidas en el modelo por pasos sucesivos.

3.3.2 Interpretación Grafica Razón de escogencia y las definidas en el Modelo.

A continuación, analizaremos mediante el uso del paquete gráfico *ggplot2*, la

correlación entre las variables independientes con la variable razón de escogencia, la variable dependiente objeto del estudio. La sintaxis usada es la indicada a continuación, donde para este caso coloca sólo para un caso, los demás, se pueden observar en el anexo del Script.

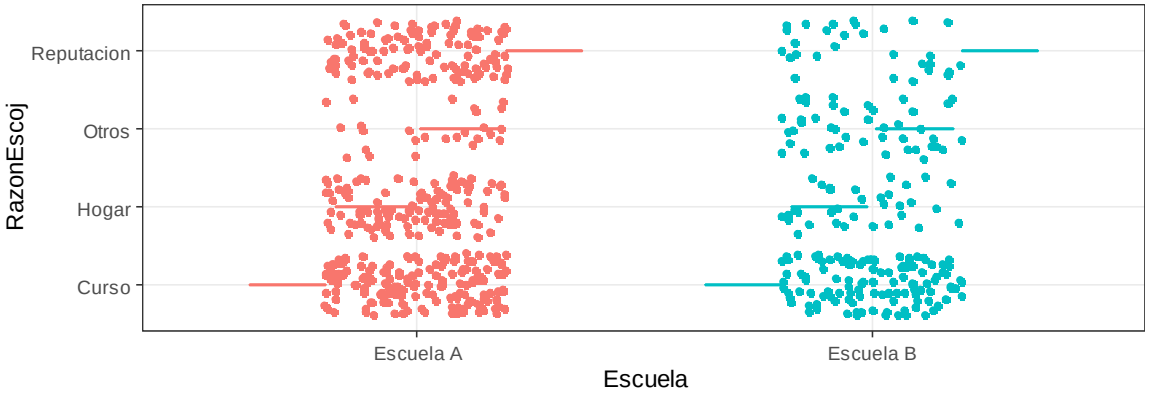
```
ggplot(data = datos, aes(x =Escuela , y = RazonEscoj, color =
Escuela)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.2) +
  theme_bw() +
  theme(legend.position = "null")
```

La figura 7 nos muestra cómo es el comportamiento de las preferencias según la variable independiente correspondiente a la escuela, donde el mayor número de puntos se ubican en la escogencia del curso en ambas escuelas, no obstante, en la escuela A se observa una mayor nube de puntos con relación a la cercanía al hogar y reputación con respecto a la escuela B.

Por el contrario, en la escuela B, se aprecia un nivel de mayor aceptabilidad en la categoría de la variable dependiente "otros factores". Este comportamiento valida la diferencia entre las escuelas en la escogencia de una institución y, por lo tanto, se justifica su entrada al modelo.

Figura 7.

Relación razones de escogencia y escuela.

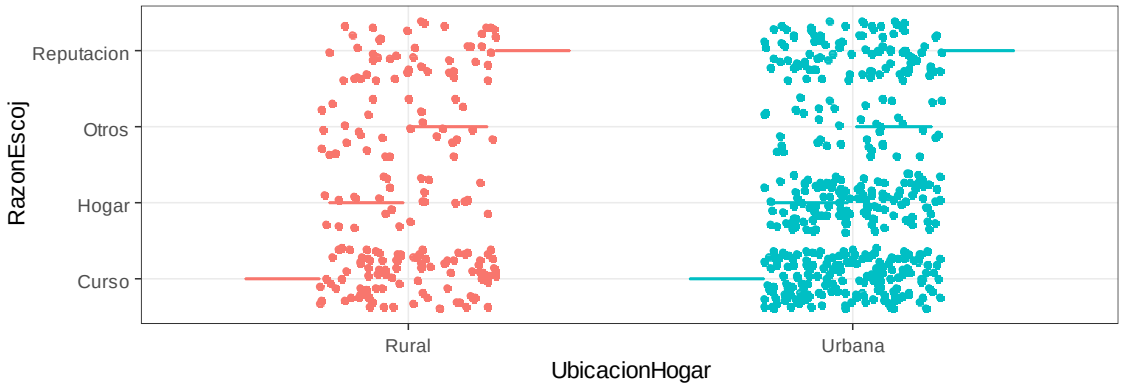


En la figura 8 se marcan diferencias en la distribución de los datos, en relación del análisis con respecto a la variable referente a la ubicación del hogar del estudiante. Con el área urbana, se nota una mayor nube de puntos en las categorías de la variable de respuesta

referentes a calidad de curso, distancia al hogar y reputación, mientras que en la rural se aprecia mayor en la de otros.

Figura 8.

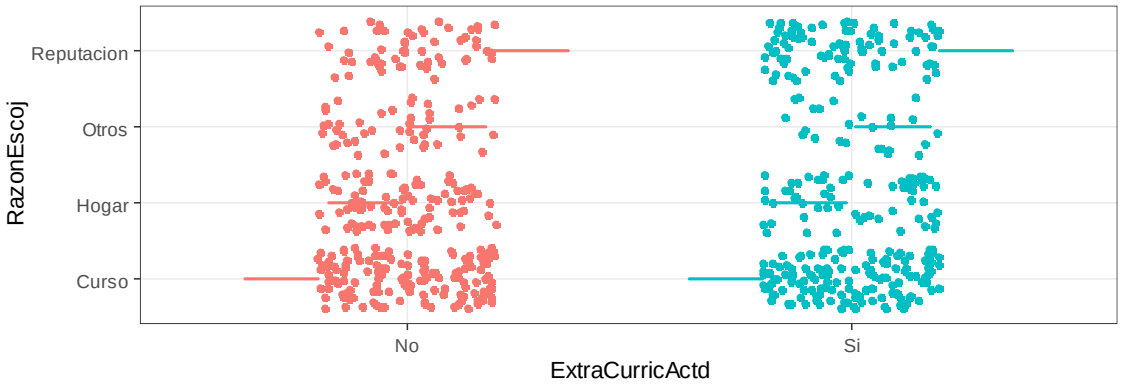
Relación razones de escogencia y ubicación del hogar.



En cuanto a la variable referente a la realización de actividades extracurriculares, la diferencia es menos marcada para opción del nivel del curso, observando que presenta la mayoría de puntos en ambas posibilidades de respuesta. La mayor diferencia se puede observar en la categoría otros aspectos, donde la distribución es más acentuada en las personas que no realizan actividades de formación extracurricular.

Figura 9.

Relación razones de escogencia y actividades extra curriculares

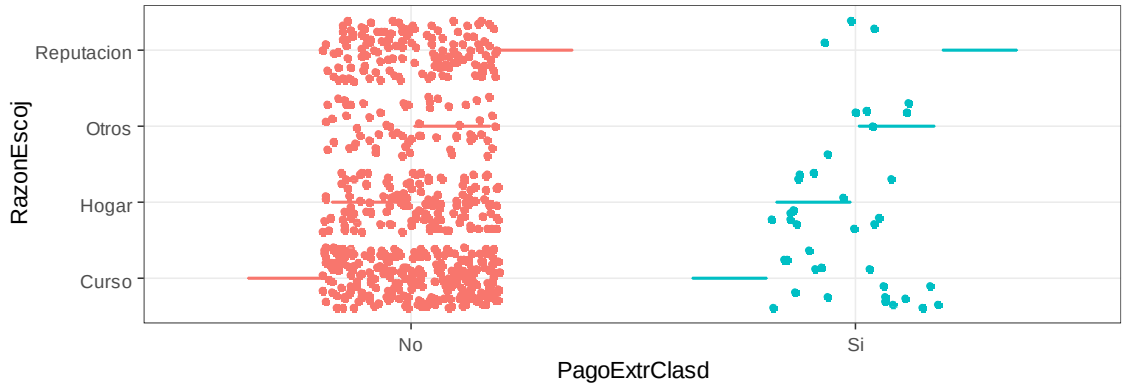


Las diferencias entre los alumnos que realizan pagos de clases extra y aquellos que no, son notorias de acuerdo con lo observado en las distribuciones de la figura. Las nubes

de puntos relacionados con aquellos individuos que no buscan este tipo de refuerzo es bien diferente a los que sí lo hacen, lo que evidencia un mayor nivel de densidad de puntos para este primer caso.

Figura 10.

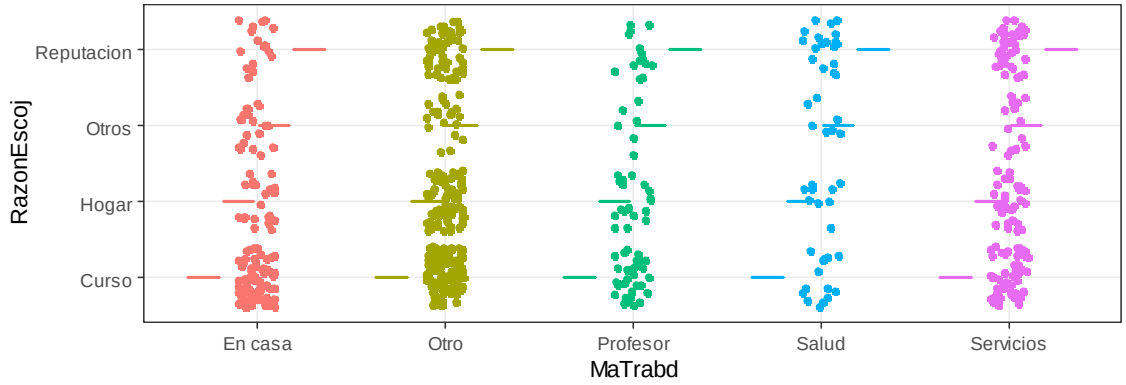
Relación razones de escogencia y pago de clases extra



En cuanto a las ocupaciones de la madre, la aceptación del curso se da mayoritariamente en todas las categorías exceptuando en los individuos cuyas madres tienen profesiones relacionadas con el área de la salud. También puede observarse que la categoría de otras actividades laborales mostró menos aceptación con relaciona a la categoría de la variable referente a otros factores para la elección de un curso.

Figura 11.

Relación razones de escogencia y trabajo de la madre

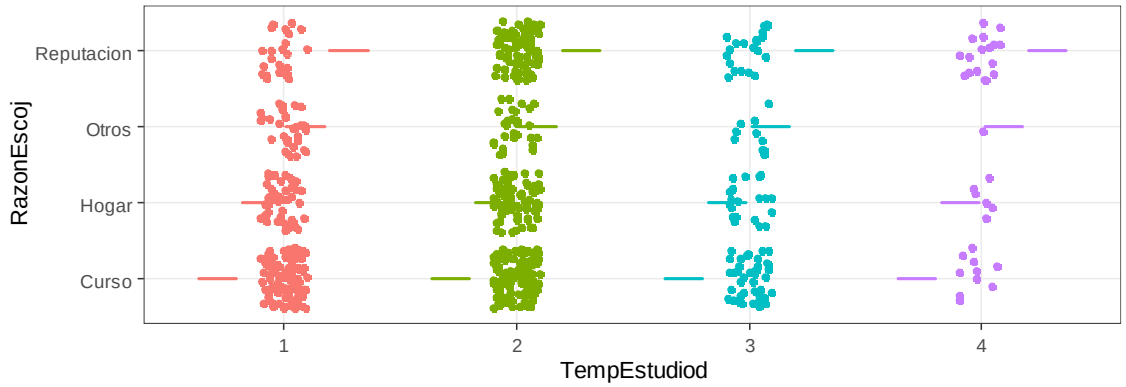


Con base en lo observado en la figura 11, se puede inferir que la mayor densidad de puntos relacionada con el tiempo de estudio de los estudiantes fue de dos horas y la de menos,

de cuatro. Sin embargo, en esta misma categoría de tiempo de estudio correspondiente a dos horas hubo menos aceptación relacionada con la categoría de la variable dependiente “otros factores”. Dentro de los que estudian cuatro horas, la mayoría prefiere escoger su institución por la reputación de esta.

Figura 12.

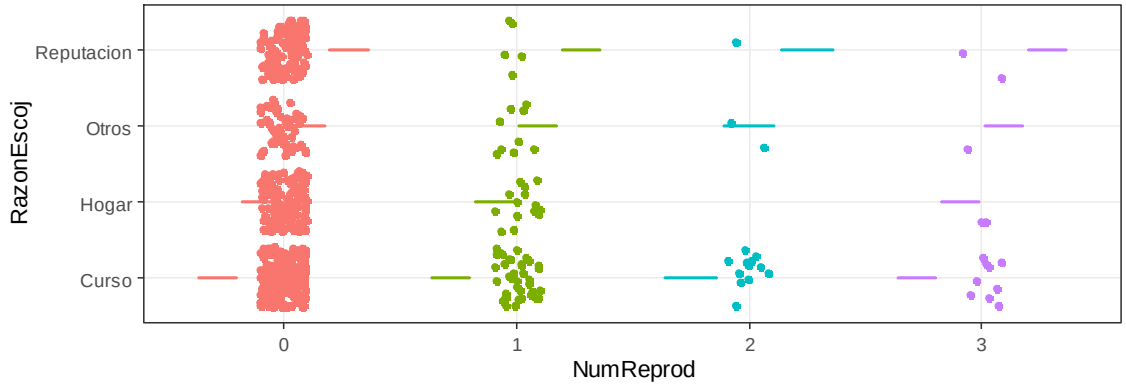
Relación razones de escogencia y tiempo dedicado a estudiar



Es evidente que la gran mayoría de estudiantes no ha repetido materias y por eso la distribución de puntos más densa se encuentra esa categoría. Dentro del grupo de individuos que no ha reprobado cursos, la de menor aceptación es la categoría de otros factores. Si bien en las otras opciones son bajas en cuanto a la densidad de puntos, la opción más mencionada para escoger una institución es la que se refiere a la calidad del curso.

Figura 13.

Relación razones de escogencia y número de veces que se reprobó el curso

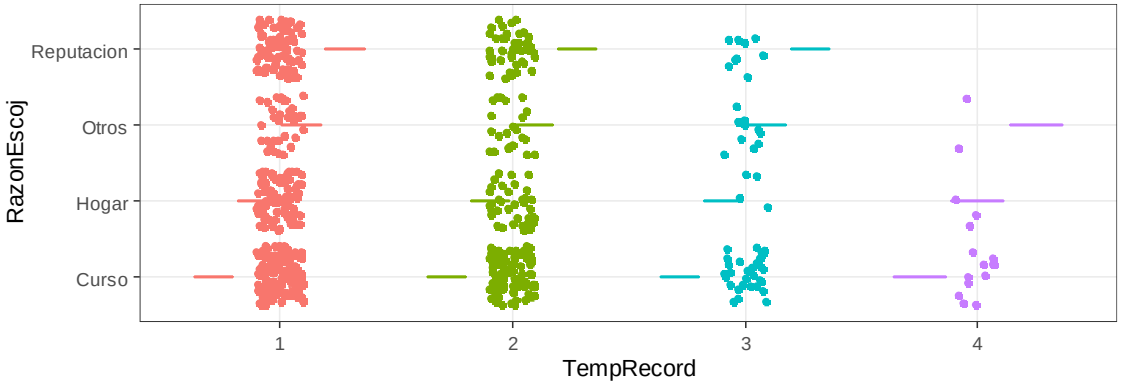


Con respecto a la variable explicativa referente al tiempo de recorrido, la gran

mayoría de encuestados invierten entre una y dos horas para ir de sus casas a la institución. La mayor densidad de puntos se encuentra en las categorías relacionadas con la calidad del curso y la reputación de la institución que lo imparte.

Figura 14.

Relación razones de escogencia y tiempo de recorrido de la casa a la escuela



De lo anterior, también se puede inferir que, en general, con base en el análisis del modelo obtenido, las opciones más reiteradas para escoger una institución educativa son la importancia del curso y la reputación.

A continuación, se indican los parámetros más importantes relacionados con el modelo final obtenido, considerados como elemento esencial de nuestro análisis como los cocientes, sus errores estándar y el estadístico de Wald.

```
> summary(rlog_stepwise, Wald=TRUE)

Call:
multinom(formula = RazonEscojd ~ Escuelad + ExtraCurricActd +
    UbicacionHogard + TempEstudiod + PagoExtrClasd + MaTrabd +
    NumReprod + TempRecord, data = datos)

Coefficients:
           (Intercept) EscueladEB ExtraCurric UbiHogar TEstudiod2 TEstudiod3 TEstudiod4 PagoExtrClas
MaTrabdhealth MaTrabdother
home        -0.9244    -0.5994    -0.3494295   0.51067  0.2009467  -0.07237173  0.1182456 0.5040269
0.2975247    0.6301619
other       -1.3272416  0.9545690 -0.4797079  -0.1399  -0.2210852  -0.08428450 -0.9453975 0.5705283
0.8267888   -0.2755944
reputation  -1.3933518 -0.8615494  0.4644174  -0.6239   0.9074680   0.44720588  1.9658809 -1.377185
1.6201532    0.8307988


           MaTrabdservices MaTrabdteacher NumReprod1 NumReprod2 NumReprod3 TempRecord2 TempRecord3
TempRecord4
home             0.2513626      0.1485063 -0.2780774 -14.639671 -0.8395509 -0.16358781 -1.13427088  -
0.6496610
other            0.2667889     -0.1138832 -0.4261374  -0.346480 -0.9304386 -0.32685396  0.09383476  -
0.3490634
reputation       0.9767575      0.2156322 -1.2636055  -1.913117 -0.7045819  0.07850446 -0.19610611 -
15.1294184


Std. Errors:
           (Intercept) EscueladEB ExtraCurric UbiHogar TEstudiod2 TEstudiod3 TEstudiod4 PagoExtrClas
MaTrabdhealth MaTrabdother
home         0.4272960  0.2619283  0.2175499  0.2864142  0.2446284 0.3341616  0.5715474  0.4097424
0.5139316    0.3016889
other        0.5029221  0.3179567  0.2901525  0.3228650  0.3066366 0.4248910  1.0859322  0.5185661
0.5576610    0.3597407
reputation   0.4669870  0.2846861  0.2298417  0.2724834  0.2814471 0.3677839  0.5053825  0.6680628
0.4940957    0.3473431


            MaTrabdservices MaTrabdteacher NumReprod1   NumReprod2 NumReprod3 TempRecord2 TempRecord3
TempRecord4
home             0.3530132      0.4026938  0.3544586 4.253200e-07  0.8416308   0.2475843   0.5736903
6.889962e-01
other            0.4099095      0.5376703  0.4160963 7.938432e-01  1.0995158   0.3295192   0.4620608
8.355181e-01
reputation       0.3846145      0.4523500  0.5152292 1.064065e+00  0.8276925   0.2579136   0.4513212
4.802319e-07


Value/SE (Wald statistics):


           (Intercept) EscueladEB ExtraCurric UbiHogar TEstudiod2 TEstudiod3 TEstudiod4 PagoExtrClas
MaTrabdhealth MaTrabdother

home         -2.163550  -2.288780  -1.606204   1.7830076  0.8214362  -0.2165770  0.2068869  1.230107
0.5789188    2.0887806
other        -2.639060   3.002198  -1.653296  -0.4335658 -0.7210007  -0.1983673 -0.8705861  1.100204
1.4826011   -0.7660918
reputation   -2.983706  -3.026313   2.020597  -2.2898917  3.2242934   1.2159474  3.8898873 -2.061460
3.2790272    2.3918678


                    MaTrabdservices MaTrabdteacher NumReprod1    NumReprod2 NumReprod3 TempRecord2
TempRecord3   TempRecord4
home             0.7120485      0.3687823 -0.7845132 -3.442037e+07 -0.9975286  -0.6607358  -1.9771486 -
9.429094e-01
other            0.6508482     -0.2118087 -1.0241318 -4.364590e-01 -0.8462257  -0.9919116   0.2030788 -
4.177808e-01
reputation       2.5395757      0.4766933 -2.4525114 -1.797932e+00 -0.8512604   0.3043827  -0.4345156 -
3.150440e+07

Residual Deviance: 1440.737
AIC: 1548.737
```

3.3.3 Contraste condicional de razón de verosimilitud

Con el fin de determinar la significancia del modelo encontrado y determinar la utilidad del mismo en la predicción de la probabilidad de ocurrencia de las categorías recogidas en la variable dependiente, se compara el modelo del paso inicial con el modelo final mediante el contraste condicional de razón de verosimilitud. Para ello, se realiza el contraste utilizando la función *anova*, a fin de comparar las devianzas de estos dos modelos, tal como se indica a continuación:

```
> anova(rlog_stepwise,rlogvacio)

Likelihood ratio tests of Multinomial Models

Response: RazonEscojd
Model 1 sólo la constante.
    Modelo 2 Escuelad + ExtraCurricActd + TempEstudiod + UbicacionHogard +
          PagoExtrClasd + NumReprod + MaTrabd + TempRecord


  Resid. df  Resid. Dev  Test   Df   LR stat.         Pr(Chi)
1      1908   1623.398
2      1857   1440.737   1 vs 2   51 182.6612   1.110223e-16
```

El modelo 2 representa el modelo final en contraste con el modelo 1 sin variables explicativas. Al contrastar los valores correspondientes a las devianzas residuales, observamos que en el modelo 1 es de 1623,398 comparado con un valor menor en el modelo final de 1440,737. Este cambio es estadísticamente significativo con un valor de P para la distribución Chi cuadrado que tiende a cero (1.110223e-16), por lo tanto, el modelo significativo es:

RazonEscojd ~ Escuelad + ExtraCurricActd + TempEstudiod + UbicacionHogard + PagoExtrClasd + NumReprod + MaTrabd + TempRecord)

3.3.4 Intervalos de confianza estimadores Beta.

Otra forma de presentar los cocientes estimados por el modelo, dando cuenta del error de estimación, es presentar los IC que complementan la estimación puntual. En R se obtienen

con la función *confint ()*, que lleva como primer argumento el nombre del modelo del que queremos extraer los intervalos de confianza, y que fueron calculados con una confianza del 95%.

En este caso el objetivo no es identificar si el intervalo contiene la unidad o no, sino detectar entre qué valores se puede encontrar el cociente con una confianza del 95% y observar el ancho del intervalo entre ellos. Se destaca el cociente de TempRecord4 (tiempo de recorrido de la institución a la casa) donde los límites son altos (-15,1294,-15,1294), indicando que quienes estén en dicha categoría no tienen en cuenta la reputación para definir la elección de la institución. Los IC útiles en el análisis se presentan más adelante en la valoración de los Odds Ratio.

Para una mejor comprensión de la interpretación de los resultados es necesario validar las categorías de referencia de cada variable, que está en el modelo final. La categoría base es la que corresponde al cero o el código de la categoría inferior, valga decir: en este caso *Calidad del curso* (identificada como curso) para la variable dependiente, *Escuela A*, *No* para el caso que el estudiante realiza actividades extra curriculares, *rural* para la ubicación del hogar, 1 para el tiempo de estudio, *No* para el pago de clases extra, *en casa* para trabajo de madre, *0* para el número de veces que se reprueba un curso, y *1* para el tiempo en horas que un estudiante tarda en llegar a su casa. Los códigos para las diferentes opciones se muestran a continuación.

RazonEscojd : razones de escoger una institución educativa

	Hogar	Otros	Reputacion
Curso	0	0	0
Hogar	1	0	0
Otros	0	1	0
Reputación	0	0	1

Escuelad: institución donde se hizo el estudio

	[T.Escuela B]
Escuela A	0
Escuela B	1

ExtraCurricActd: actividades extra curriculares

	[T.Si]
No	0
Si	1

UbicacionHogard: ubicación del hogar

	[T.Urbana]
Rural	0
Urbana	1

TempEstudiod: tiempo de estudio

	[T.2]	[T.3]	[T.4]
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

PagoExtrClasd: pago clases extras

	[T.Si]
No	0
Si	1

MaTrabd: profesión de la madre

	[T.Otro]	[T.Profesor]	[T.Salud]	[T.Servicios]
En casa	0	0	0	0
Otro	1	0	0	0
Profesor	0	1	0	0
Salud	0	0	1	0
Servicios	0	0	0	1

NumReprod: veces que un estudiante ha reprobado un curso

	[T.1]	[T.2]	[T.3]
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

TempRecord: tiempo en horas del trayecto de la casa a la institución.

	[T.2]	[T.3]	[T.4]
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Además del valor de las estimaciones de los cocientes de cada variable del modelo, es útil obtener sus correspondientes intervalos de confianza en este caso a un nivel del 95%, por lo que, además, se usó la función *round* para facilitar la lectura de la tabla, la sintaxis en R es la siguiente:

```
confint(object = rlog_stepwise, level = 0.95)
icexp=confint(object = rlog_stepwise, level = 0.95)
round(icexp,4)
```

Tabla 18.

Tabla IC de los cocientes Beta.

Home			Otro			Reputation		
	2.5%	97.5%		2.5%	97.5%		2.5%	97.5%
(Intercept)	-1,7620	-0,0870	(Intercept)	-2,3130	-0,3415	(Intercept)	-2,3086	-0,4781
Escuela EB	-1,1129	-0,0861	Escuela EB	0,3314	1,5778	Escuela EB	-1,4195	-0,3036
ExtraCurricActdyes	-0,7758	0,0770	ExtraCurricActdyes	-1,0484	0,0890	ExtraCurricActdyes	0,0139	0,9149
UbicacionHogardU	-0,0507	1,0720	UbicacionHogardU	-0,7728	0,4928	UbicacionHogardU	-1,1580	-0,0899
TempEstudiod2	-0,2785	0,6804	TempEstudiod2	-0,8221	0,3799	TempEstudiod2	0,3558	1,4591
TempEstudiod3	-0,7273	0,5826	TempEstudiod3	-0,9171	0,7485	TempEstudiod3	-0,2736	1,1680
TempEstudiod4	-1,0020	1,2385	TempEstudiod4	-3,0738	1,1830	TempEstudiod4	0,9753	2,9564
PagoExtrClasdyes	-0,2991	1,3071	PagoExtrClasdyes	-0,4458	1,5869	PagoExtrClasdyes	-2,6866	-0,0678
ManTrabdhealth	-0,7098	1,3048	ManTrabdhealth	-0,2662	1,9198	ManTrabdhealth	0,6517	2,5886
ManTrabdother	0,0389	1,2215	ManTrabdother	-0,9807	0,4295	ManTrabdother	0,1500	1,1516
ManTrabdservices	-0,4405	0,9433	ManTrabdservices	-0,5366	1,0702	ManTrabdservices	0,2229	1,7306
ManTrabdteacher	-0,6408	0,9378	ManTrabdteacher	-1,1677	0,9399	ManTrabdteacher	-0,6710	1,1022
NumReprod1	-0,9728	0,4166	NumReprod1	-1,2417	0,3894	NumReprod1	-2,2734	-0,2538
NumReprod2	-14,6397	-14,6397	NumReprod2	-1,9024	1,2094	NumReprod2	-3,9986	0,1724
NumReprod3	-2,4891	0,8100	NumReprod3	-3,0855	1,2246	NumReprod3	-2,3268	0,9177
TempRecord2	-0,6488	0,3217	TempRecord2	-0,9727	0,3190	TempRecord2	-0,4270	0,5840
TempRecord3	-2,2587	-0,0099	TempRecord3	-0,8118	0,9995	TempRecord3	-1,0807	0,6885
TempRecord4	-2,0001	0,7007	TempRecord4	-1,9866	1,2885	TempRecord4	-15,1294	-15,1294

Nota: Al analizar los signos muestra que es más probable que la selección hecha por individuos de la escuela B aporte en cierta forma a seleccionar una institución con respecto a la calidad del curso que el grupo asociados a las categorías *cercanía a casa* y *reputación de la institución,* en relación con los que prefieren la categoría de la variable respuesta *otros factores,* como ya se había identificado. Esto debido a que tienen signos positivos el IC.

En cuanto la variable explicativa *actividades extracurriculares*, la respuesta de cercanía a casa y otras razones tienen signos alternos, lo que indica que no existe discriminación con respecto al curso. Sin embargo, en la categoría de reputación ambos signos son negativos y, por ende, es menos probable que incida en la elección de la institución si se compara con la preferencia con la calidad del curso.

En ese mismo sentido, la ubicación del hogar en los que consideran la reputación como importante es menos probable que influya en la elección de una institución comparándola con el nivel de calidad curso.

El pago de clases extra marca una mayor diferencia para aquellos miembros de la muestra que aducen tener en cuenta la reputación, pero es menos probable que esta causa influya con la categoría de curso, que es la variable dependiente de referencia.

En las actividades laborales de la madre, se presentan diversas opciones, destacándose

que las mayores diferencias se presentan en la selección de una institución de acuerdo con su reputación donde se encuentran cocientes mayores. En el número de veces que un alumno ha reprobado un curso, sobresale la opción 2 de la categoría asociadas a la selección por cercanía a casa con valores muy altos (-14,6397) indicando que dicha opción es mucho menos probable que influya en la elección de una institución.

En cuanto al tiempo de recorrido entre la casa y la institución, los cocientes son alternos, indicando poca incidencia en la selección de la institución.

3.3.5 Odds ratios (OR) e Intervalos de Confianza (I.C)

Es importante calcular los OR de los exponenciales de los cocientes y sus intervalos de confianza de éstos al 95%, dado que el modelo se interpreta con base a ellos.

3.3.5.1 Odds ratios (OR)

Para el cálculo de las odds ratios de los cocientes del modelo se ha utilizado la función *exp()*. Es útil mencionar que los cocientes mayores de la unidad son aquellas razones con más incidencia a la hora de seleccionar una institución con respecto al curso propiamente dicho, mientras que cuando el cociente del OR es menor del uno es menos probable que influya en la escogencia con respecto a la categoría de referencia “calidad en el curso”.

Para interpretar los odds ratios de cada variable, se asume que el resto de variables explicativas se mantienen fijas. Hay que tener en cuenta que la variable a predecir es razones de escogencia de una institución educativa, donde la categoría de referencia es la calidad del curso.

Considerar otras opciones para la elección de una institución con relación a la calidad de un curso, es 2,59 veces más probable en los casos donde la persona pertenece a la escuela B comparados con la escuela A. El considerar la reputación para la elección de una institución con relación a la calidad del curso, es 1,59 veces más probable en los casos donde la persona sí

lleva a cabo actividades extracurriculares, comparada con la que no.

Considerar la cercanía de la casa para la elección de una institución con relación al interés por un curso, es 1,66 veces más probable en los casos donde la persona viva en el área urbana comparado con aquella que habita en zona rural.

La cercanía de la casa para la elección de una institución con relación al interés por un curso, es 1,66 veces más probable en los casos donde la persona viva en el área urbana comparado con los del área rural.

Tanto los de cercanía a casa como los de otras razones para seleccionar una institución en relación con los de interés por un curso, tiene más posibilidades, es decir, en los de cercanía a casa, es 1,65 veces más probable en los casos donde la persona sí realiza pagos extras, comparados con los que no; y de 1,76 veces más en los de otras razones y que sí pagan extra.

En cuanto al trabajo de la madre, al ser Dummy, se tienen varias variables, todos ellos con cocientes mayores de la unidad, es decir, que tanto en los casos donde la madre se desempeña en salud, servicios, profesora y otros para la elección de una institución con relación al interés por un curso, es más probable que ocurra comparado con las que se quedan en casa.

El considerar la reputación para la elección de una institución con relación al interés por un curso es 7,14 veces más probable en los casos donde la persona dedica mayor tiempo a estudiar (Testudio4), comparado con el grupo de menor dedicación a esta actividad.

En todos los casos, las demás categorías de la variable respuesta *cercanía a casa*, *reputación* y *otros factores* son categorías menos probables para la elección de una institución con relación al interés por un curso, comparado con el grupo asociado a los de menor número de veces que se ha reprobado un curso, esto debido a que los cocientes son menores de la unidad.

En cuanto al tiempo de recorrido entre la casa y la institución, sobresale el hecho de que los que valoran la reputación para la elección de una institución con relación al interés por la calidad de un curso, es menos común en los casos donde el individuo invierte mayor tiempo en recorrer la distancia entre su casa y la institución.

```
> exponente=(exp(coefficients(rlog_stepwise,2)))
> round(exponente,4)
```

```
           (Intercept) EscueladEB ExtraCurric TEstudiod2 TEstudiod3 TEstudiod4 UbiHogardU PagoExtrClas NumReprod1
NumReprod2
home            0.3967     0.5491    0.7051     1.2226    0.9302     1.1255       1.6664    1.6554     0.7572
0.0000
other           0.2652     2.5976    0.6190     0.8016    0.9192     0.3885       0.8694    1.7692     0.6530
0.7072
reputation      0.2482     0.4225    1.5911     2.4780    1.5639     7.1412       0.5358    0.2523     0.2826
0.1476

            NumReprod3 MaTrabdhealth MaTrabdother MaTrabdservices MaTrabdteacher TempRecord2 TempRecord3 TempRecord4
home            0.4319        1.3465       1.8779          1.2858         1.1601      0.8491      0.3217      0.5222
other           0.3944        2.2860       0.7591          1.3058         0.8924      0.7212      1.0984      0.7053
reputation      0.4943        5.0539       2.2952          2.6558         1.2406      1.0817      0.8219      0.0000
```

Con el fin de validar su significancia estadística es necesario calcular los intervalos de confianza (I.C) de los cocientes OR, donde el análisis se implementa a partir de la pertenencia de la unidad a dicho intervalo; de manera tal que si lo contiene, no es significativo. Por el contrario, si el I.C tiene ambos limites por debajo de uno, ese factor es menos probable que aporte a la definición de la selección de la institución educativa con respecto a la calidad del curso. De otra parte, si el I.C tiene valores mayores de la unidad, indica que dicho factor aporta más para la selección de una institución educativa. Es importante, aclarar que en las variables Dummy, con que una sola categoría de la variable sea significativa implica que las demás se deben mantener en el modelo, por eso, algunos cocientes se observan con IC que contienen la

unidad, eso implica que alguna de las otras categorías no lo contienen, incluso en alguna de las otras dos opciones de respuesta de la variable dependiente.

Considerando lo anterior, la escuela en la que se tomó la muestra, influye en la selección de la institución siendo más probable cuando se argumentan otras razones con respecto a la categoría de referencia calidad del curso. Lo anterior ha sido determinado teniendo en cuenta que el cociente en la escuela B es de 2,59 mucho mayor de la unidad, además, las razones asociadas a la cercanía a la casa y la reputación influyen de menor manera con respecto a la calidad del curso.

En cuanto a la variable relacionada al desarrollo de actividades extra curriculares, se tiene que la categoría reputación si influye de mayor manera para seleccionar la institución con respecto a la calidad del curso impartido por esta, pues el cociente es de 1,59, mientras que, en los otros dos casos, estos cocientes son menores de la unidad, (0,70 para cercanía a la casa y 0,61 para otras razones).

Con el mismo razonamiento, se puede afirmar, que el tiempo dedicado a estudiar contribuye a explicar las razones de selección de una institución escolar, al igual que la ubicación del hogar, el pago de clases extra, el número de veces que se ha reprobado un curso, el trabajo al que se decida la madre y el tiempo de recorrido entre casa y escuela. Para mayor facilidad de interpretación, se han calculado los IC de los cocientes, como se indica a continuación.

3.3.5.2 Intervalos de confianza al 95% de los OR.

Para calcular los intervalos de confianza de las odds ratios, se determinaron los cocientes de los estimadores, mediante el comando ***confint()*** y posteriormente los intervalos de los ***odds*** al 95%. Para optimizar el análisis de los resultados, se utilizó la orden ***round()*** como se aprecia a continuación. La tabla permite observar los resultados para las tres categorías de: cercanía a casa, reputación de la institución y otras razones todas ellas comparadas con la categoría base de relevancia del curso. El análisis de los IC de los OR, permite observar la contención de la unidad con el fin de confirmar la significancia estadística del cociente y que tan lejos se encuentra de dicho valor para determinar su influencia en las razones de escogencia de una institución.

En el presente estudio, los cocientes que no contienen el uno, deben de quedar en el modelo dado que pertenecen a una variable Dummy, por lo tanto, con que una de las categorías lo contenga, todas las categorías asociadas a la variable deberían quedar en el modelo. La sintaxis muestra aparte de la instrucción ***confint (),*** el crear un objeto ***(icexp)*** con el fin de lograr el exponente de dichos valores y luego se redondea para la salida presentable.

```
confint(object = rlog_stepwise, level = 0.95)
icexp=confint(object = rlog_stepwise, level = 0.95)
exp(icexp)
round(exp(icexp),4)
```

Tabla 19.

Tabla IC de los Odds Ratio Modelo Final

Home			Otro			Reputation		
	2.5%	97.5%		2.5%	97.5%		2.5%	97.5%
(Intercept)	0,1717	0,9167	(Intercept)	0,0990	0,7107	(Intercept)	0,0994	0,6200
Escuela EB	0,3286	0,9175	Escuela EB	1,3929	4,8471	Escuela EB	0,2418	0,7382
ExtraCurricActdyes	0,4603	1,0800	ExtraCurricActdyes	0,3505	1,0931	ExtraCurricActdyes	1,0140	2,4965
TempEstudiod2	0,7569	1,9747	TempEstudiod2	0,4395	1,4622	TempEstudiod2	1,4274	4,3021
TempEstudiod3	0,4832	1,7906	TempEstudiod3	0,3997	2,1138	TempEstudiod3	0,7606	3,2157
TempEstudiod4	0,3672	3,4503	TempEstudiod4	0,0462	3,2641	TempEstudiod4	2,6521	19,2289
UbicacionHogardU	0,9506	2,9213	UbicacionHogardU	0,4617	1,6369	UbicacionHogardU	0,3141	0,9140
PagoExtrClasdyes	0,7415	3,6955	PagoExtrClasdyes	0,6403	4,8886	PagoExtrClasdyes	0,0681	0,9344
NumReprod1	0,3780	1,5169	NumReprod1	0,2889	1,4761	NumReprod1	0,1030	0,7759
NumReprod2	0,0000	0,0000	NumReprod2	0,1492	3,3516	NumReprod2	0,0183	1,1882
NumReprod3	0,0830	2,2479	NumReprod3	0,0457	3,4027	NumReprod3	0,0976	2,5034
ManTrabdhealth	0,4918	3,6870	ManTrabdhealth	0,7663	6,8195	ManTrabdhealth	1,9189	13,3106
ManTrabdother	1,0396	3,3921	ManTrabdother	0,3751	1,5365	ManTrabdother	1,1619	4,5339
ManTrabdservices	0,6437	2,5683	ManTrabdservices	0,5847	2,9160	ManTrabdservices	1,2497	5,6440
ManTrabdteacher	0,5269	2,5543	ManTrabdteacher	0,3111	2,5598	ManTrabdteacher	0,5112	3,0108
TempRecord2	0,5226	1,3794	TempRecord2	0,3781	1,3757	TempRecord2	0,6525	1,7320
TempRecord3	0,1045	0,9902	TempRecord3	0,4441	2,7168	TempRecord3	0,3394	1,9907
TempRecord4	0,1353	2,0153	TempRecord4	0,1372	3,6274	TempRecord4	0,0000	0,0000

Lo anterior confirma que existe cierto grado de significancia de las variables y su pertinencia en el modelo.

3.3.6 Ajuste global del modelo

Mediante la evaluación del test de máxima verosimilitud, se realiza la comparación entre los modelos saturados (todas las variables) y el modelo encontrado. Para llevar a cabo lo anterior, se utiliza de nuevo la función anova, tendiendo en cuenta sin embargo que no preferiblemente debe existir diferencias entre ellos. Calculando este estadístico y su p-valor se obtiene:

```
> anova(rlog_stepwise,rlogcompletodummy)
Likelihood ratio tests of Multinomial Models

Response: RazonEscojd


Model1

Escuelad + ExtraCurricActd + UbicacionHogard + TempEstudiod + PagoExtrClasd
+ MaTrabd + NumReprod + TempRecord2 Escuelad + Maedud + Paedud +
UbicacionHogard + MaTrabd + PaTrabd + AcudEstudd + TempRecord +
```

```
TempEstudiod + NumReprod + SopExtraEdud + FamSupd + PagoExtrClasd +
ExtraCurricActd + ExtraCurricActd + SupEstd + Internetd + RelaRomantd +
FamRelad + TempEstudiod + SalAmigd
```

```
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1      1857   1440.737
2      1773   1370.913 1 vs 2    84 69.82432 0.8664673
```

Se encontró que el valor de la p de la Chi cuadrado es de 0,8664 y, por lo tanto, se confirma la no diferencia entre los modelos, por lo que, por el principio de parsimonia, el modelo con menos variables es adecuado.

3.3.7 Tasa de clasificaciones correctas

Con base en la proporción de aciertos, se lleva a cabo la cuantificación del parámetro de bondad del ajuste global del modelo. Para la clasificación de cada una de las observaciones sobre la categoría más probable, se ha utilizado la matriz de confusión. Así se ha construido una matriz de clasificación de valores observados-predichos.

Para establecer los componentes de esta matriz, se requiere estimar los predichos y calcular los aciertos para dividirlos sobre el total de la muestra, utilizando la función ***predict()***, explicada previamente, de la siguiente manera:

```
obs<-RazonEscojd
pre<-predict(rlog_stepwise, type="class")
datos$predichos<-rlog_stepwise$fitted.values
View(datos)
table(predict(rlog_stepwise),RazonEscoj)
PAC=(221+39+1+64)/637
```

Tabla 20.

Matriz de confusión para tasa de clasificaciones correctas

	Calidad de Curso	Cercania Hogar	Otros Factores	Reputacion
Calidad de Curso	221	81	56	62
Cercania Hogar	22	39	5	14
Otros Factores	3	0	1	1
Reputacion	34	27	7	64

Nota: Los resultados muestran la distribución de aciertos en cada categoría de la variable

dependiente.

De lo anterior podemos deducir que la proporción de aciertos es del 51,02%, lo que implica que, de los casos analizados, solo este porcentaje logra ser correctamente clasificado. Al coincidir la razón de escogencia con el pronosticado por el modelo, si bien no es lo más deseado, se presenta una tasa de aciertos moderada, indicando que aproximadamente el 49% se puede explicar por otras causas.

3.3.8 Calidad del ajuste del modelo

Para medir la calidad del ajuste del modelo se utiliza como se analizó previamente los cocientes Pseudo-R^2 de Mc-Fadden, de Cox-Snell y de Nagelkerke. No obstante, el de Nagelkerke es una mejora de Cox-Snell y por eso no se requiere su cálculo, en el presente análisis se han calculado el de Mc-Fadden y el más sugerido Nagelkerke. Para ello, y considerando las fórmulas ya presentadas, se utilizan el valor de las devianzas del modelo final y del modelo inicial con sólo la constante, de la siguiente manera:

$$R^2_{McFadden} = 1 - \frac{\Lambda_f}{\Lambda_o} = 0{,}11$$

```
deviance(rlog_stepwise)
deviance(rlogvacio)
dv1=deviance(rlog_stepwise)
dv0=deviance(rlogvacio)
mf=1-(dv1/dv0)
mf
0.11
```

El cociente de Nagelkerke esta dado por:

$$R^2_{CS} = 1 - \left(\frac{V_0}{V_f}\right)^{\frac{2}{N}} = 1 - e^{\left(\frac{\Lambda_f - \Lambda_0}{N}\right)} = 0.2704512$$

```
ng=(1-exp((dv1-dv0)/637))/(1-exp(-dv0/637))
ng

0.2704512
```

El cociente de Mac Fadden no muestra un buen ajuste. Sin embargo, el de Nagelkerke si lo hace, dado que hemos obtenido un valor de 0,2704. Es útil recordar, que se sugiere un buen ajuste, si el valor del cociente es mayor de 0,20, situación que se cumple en este caso.

3.3.9 Validación del modelo

Por último, se realiza la validación del modelo mediante el análisis de los residuos de la devianza, considerando que los residuos que indican una falta de ajuste global son aquellos cuyo valor absoluto son mayores que 4. Estos residuos se calculan mediante la función "residuals" y se realiza un descriptivo de todos los residuos de la siguiente manera:

```
> residuos=residuals(rlog_stepwise)
> #Validación del modelo con los residuos
> #Deben ser menor de cuatro en valor absoluto
> residuos=residuals(rlog_stepwise)
> numSummary(residuos)
                   mean        sd    IQR      0%     25%     50%     75%   100%   n
course     -4.395604e-06 0.4768939 0.9137 -0.8674 -0.4092 -0.2812  0.5045 0.8264 637
home        1.726845e-06 0.4041250 0.2631 -0.5487 -0.2631 -0.1640  0.0000 0.9474 637
other       2.354788e-06 0.3004589 0.0860 -0.5195 -0.1232 -0.0657 -0.0372 0.9742 637
reputation -3.139717e-07 0.3814151 0.2241 -0.7374 -0.2319 -0.1122 -0.0078 0.9755 637
```

Se puede verificar que, entre los máximos y mínimos de los valores anteriores, todos los residuos en valor absoluto son menores de 1 (están en notación exponencial), por lo que no hay ninguna observación que se considere anormal y con falta de ajuste, esto indica que los residuos se ajustan adecuadamente y el modelo con este criterio es adecuado.

Conclusiones

La valoración del modelo de regresión logística propuesto ha permitido un análisis mediante el cual se han realizado diferentes validaciones que en teoría posibilitan la predicción de las razones se escogencia de una escuela en base a una serie de variables independientes. El procedimiento de stepwise, nos condujo a una reducción considerable en el número de variables como predictoras del modelo obteniendo un total de ocho de las 32 iniciales planteadas en el estudio.

En relación a las Odds de las variables que fueron significativas se aprecian situaciones como en las que un estudiante que viva en una ubicación rural tiene 1.304 más probabilidad de elegir la escuela por reputación que por cercanía, de hecho para las demás categorías (nivel y otras razones) la cantidad de veces de probabilidad en relación a la cercanía fue mayor en relación a las Odds de cercanía (>1) ; situación que también ocurrió en la ubicación urbano por lo que se puede inferir que hay mayor probabilidad de que se escoja la escuela por razones diferentes a cercanía sin importar la ubicación donde se encuentre el estudiante.

Otra situación que se aprecia, está relacionada con el hecho de que hubo significancia en los trabajos de los padres y madres que se dedicaban al campo de la educación (profesores) y en ambos casos también se priorizó más factores como la reputación de la institución o la cercanía con el hogar. (Odd >1 para todas las categorías).

Las variables que aportan a explicar las razones de escogencia para la selección de una institución académica fueron: El tipo de escuela, la ubicación geográfica, el realizar actividades extra curriculares, el oficio de las madres, el tiempo de recorrido entre la institución y la casa, el tiempo dedicado a estudiar, el pago de clases extras y el número de veces que un alumno reprueba un curso.

Sin embargo y en base a los resultados obtenidos en el modelo final, donde la Escuela sobre la cual se realizó el estudio hace parte del modelo final obtenido, llama la atención especialmente el hecho de que al elegir una institución educativa es 2,59 veces más probable en los casos donde la persona pertenece a la escuela B comparados con la escuela A. Aunque las dos escuelas hacen parte de la muestra global sobre la cual se realizó el estudio, esto puede

verse condicionado por diversos factores exógenos como el servicio vocacional de la institución y el enfoque misional

Con base en los resultados del análisis de las clasificaciones correctas, la proporción de aciertos establecida representa más bien una tasa moderada que en términos de sustentación del modelo no es muy deseable lo que implica que, solo el cincuenta por ciento de los casos analizados, fue correctamente clasificado.

La evaluación relacionada con la calidad de ajuste del modelo fue realizada mediante los estadísticos de McFadden y Nagelkerke, permitiéndonos obtener un valor indicador de ajuste de nivel aceptable solo con este último.

Los análisis efectuados indican que variables cualitativas como sexo, educación tanto de los padres como de madres, el tamaño de la familia, estatus de los padres, profesión del padre, acudiente pendiente del estudiante, soporte extra educacional, apoyo de la familia en el trabajo estudiantil, salida con amigos, calidad de relaciones interfamiliares, uso de la Internet y relaciones románticas no aportan estadísticamente a explicar la selección de una institución educativa.

También se puede concluir que las razones más mencionadas para elegir una institución educativa fueron la importancia del curso y la reputación.

Anexo: Descripción del Script en R

En esta sección describiremos los aspectos más relevantes para el proceso de implementación computacional del modelo desarrollado en lenguaje R

Función ipack

Tiene la función de instalar y leer diferentes paquetes para evitar su instalación dentro de las órdenes propias de cada procedimiento y para garantizar que se tiene todos los paquetes necesarios. Es útil mencionar que no todos se necesitan en la técnica de la regresión multinomial, pero es preferible tener más librerías activadas que por alguna razón falten.

.

```
#Crear función ipak
#Función ipak: instala y carga múltiples paquetes R.
# verifique si los paquetes están instalados. Instálelos si no lo están, luego cárguelos en la sesión R.
#Creador https://gist.github.com/stevenworthington/3178163


ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}
# usage
packages <- c("effects","InformationValue","haven","car","foreign",
"Rcmdr","nnet",
"reshape2","ggplot2","ggpubr","mvnormtest","MVN","gridExtra","apaTables",
"reshape", "GPArotation", "mvtnorm", "psych", "psychometric","MASS")
ipak(packages)
```

Paquetes instalados.

```
 effects InformationValue         haven           car
        TRUE          TRUE          TRUE          TRUE
     foreign          Rcmdr           nnet      reshape2
        TRUE          TRUE          TRUE          TRUE
     ggplot2        ggpubr     mvnormtest           MVN
        TRUE          TRUE          TRUE          TRUE
   gridExtra     apaTables       reshape   GPArotation
        TRUE          TRUE          TRUE          TRUE
     mvtnorm          psych psychometric          MASS
        TRUE          TRUE          TRUE          TRUE
```

1.Lectura de datos

Se realizo a través de la función **setwd ()** que permite cambiar el directorio de trabajo y pasa como argumento la ruta de la carpeta que se quiere definir como nuevo directorio de trabajo.

```
setwd("d:\\CarpetaRAFAEL-\\Nueva
carpeta\\Maestria_Estadistica_Aplicada\\Trabajo Fin Master 2019-
2020\\TESIS_2021")
datos <- read_sav(""d:\\CarpetaRAFAEL-\\Nueva
carpeta\\Maestria_Estadistica_Aplicada\\Trabajo Fin Master 2019-
2020\\TESIS_2021\\RLM1.sav"")
View(datos)
> #Nombre de variables
> names(datos)
 [1] "Escuela"        "Sexo"
 [3] "Edad"           "UbicacionHogar"
 [5] "TamaFam"        "Pstatus"
 [7] "Maedu"          "Paedu"
 [9] "MaTrab"         "Patrab"
[11] "AcudEstud"      "TempRecor"
[13] "TempEstudio"    "NumRepro"
[15] "SopExtraEdu"    "FamSup"
[17] "PagoExtrClas"   "ExtraCurricAct"
[19] "SupEst"         "Internet"
[21] "RelaRomant"     "FamRela"
[23] "TempLibr"       "SalAmig"
[25] "Ausencias"      " RazonEscoj "
```

2. **Exploración de datos: Identificar valores anormales y errores de digitación. Durante el análisis se detectaron frecuencias muy bajas en la primera categoría de las variables referentes a la educación de la madre y el padre, por lo cual se eliminaron de la base de datos.**

#Estructura del archivo -Tipos de variables

```
str(datos)
```

```
#Función attach para facilitar el manejo de la sintaxis con las variables

attach(datos)

#Función table para definir el conteo de las variables cualitativas
#Se presentan algunas como ejemplo

table(Sexo)
table(RazonEscoj)
table(Maedu)
table(Paedu)
summary(datos)
describe(datos)
```

3. **Adecuación de la base de datos. Recodificación a través del uso de la función within y adecuar para variables Dummy con la función factor.**

```
# Recodificación de Variables Cualitativas

#Sexo
datos <- within(datos, {
  Sexo <- Recode(Sexo, '"M"="Masculino"; "F"="Femenino"', as.factor=TRUE)
})
table(datos$Sexo)

# Escuela
datos <- within(datos, {
  Escuela <- Recode(Escuela, '"EA"="Escuela A"; "EB"="Escuela B"',
as.factor=TRUE)
})
table(datos$Escuela)

# Tamaño familia REVISAR
datos <- within(datos$TamaFam, {Tamafam <- Recode(Tamafam, "GT3"="Menor
igual a tres","LE3"="Mayor a tres", as.factor=TRUE)
})
table(datos$TamaFam)
str(datos)

# Ubicación hogar
datos <- within(datos, {
  UbicacionHogar <- Recode(UbicacionHogar, '"U"="Urbana"; "R"="Rural"',
as.factor=TRUE)
})
table(datos$UbicacionHogar)

# Estado civil

datos <- within(datos, {
  Pstatus <- Recode(Pstatus, '"T"="Viviendo Juntos"; "A"="Separados"',
as.factor=TRUE)
})
table(datos$Pstatus)
```

```
# Nivel de educación madre
datos <- within(datos, {
  Maedu <- Recode(Maedu,
                  '0="Ninguna educación"; 1="Educación primaria"; 2="Quinto a
Noveno"; 3="Educación secundaria"; 4="Educación universitaria"',
                  as.factor=TRUE)
})
table(datos$Maedu)



##Nivel de educación padre

datos <- within(datos, {
  Paedu <- Recode(Paedu,
                  '0="Ninguna educación"; 1="Educación primaria"; 2="Quinto a
Noveno"; 3="Educación secundaria"; 4="Educación universitaria"',
                  as.factor=TRUE)
})
table(datos$Paedu)

##Nivel de educación padre
datos <- within(datos, {
  MaTrab <- Recode(MaTrab,
                   '"at_home"="En casa"; "health"="Salud"; "other"="Otro";
"services"="Servicios"; "teacher"="Profesor"',
                   as.factor=TRUE)
})
table(datos$MaTrab)

##Profesion Padre
datos <- within(datos, {
   Patrab<- Recode(Patrab,
                   '"at_home"="En casa"; "health"="Salud"; "other"="Otro";
"services"="Servicios"; "teacher"="Profesor"',
                   as.factor=TRUE)
})
table(datos$Patrab)

## acudiente

datos <- within(datos, {
  AcudEstud <- Recode(AcudEstud,
                   '"father"="Padre"; "mother"="Madre";
"other"="Otro"',as.factor=TRUE)})
table(datos$AcudEstud)

# Soporte
datos <- within(datos, {
  SopExtraEdu <- Recode(SopExtraEdu,
                        '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$SopExtraEdu)
#Ayuda familia
datos <- within(datos, {
  FamSup <- Recode(FamSup,
                         '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$FamSup)




   #Pago Extraclase
   datos <- within(datos, {
```

```
  PagoExtrClas <- Recode(PagoExtrClas,
                    '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$PagoExtrClas)

#Pago Extracurricu
datos <- within(datos, {
  ExtraCurricAct <- Recode(ExtraCurricAct,
                           '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$ExtraCurricAct)


#Relaciones Romanticas
datos <- within(datos, {
  RelaRomant <- Recode(RelaRomant,
                               '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$RelaRomant)


#Internet

datos <- within(datos, {
  Internet <- Recode(Internet,
                    '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$Internet)

#SUP Supest
datos <- within(datos, {
  SupEst <- Recode(SupEst,
                               '"no"="No"; "yes"="Si"',as.factor=TRUE)})
table(datos$SupEst)

#Razon elección
datos <- within(datos, {
  RazonEscoj <- Recode(RazonEscoj,
                    '"course"="Curso";
"home"="Hogar";"other"="Otros";"reputation"="Reputacion"',as.factor=TRUE)})

nombres=c("Escuela","Sexo","Edad", "Ubicación Hogar", "Tamaño
familiar","Estado civil padres","Educación madre","Educación
padre","Trabajo madre","Trabajo padre","Acudiente","Tiempo libre","Tiempo
de estudio","Num reproducciones","Soporte extraeducacional"," Ayuda
Familiar","Pago extracurricular","Actividades
extracurriculares","SupEst","Internet","Relaciones romáticas","Relaciones
familiares","Tiemopo libre","Salir con amigos","Ausencias","Razón para
escojer")
```

```
#Convertir variables categóricas a Dummy función factor

Escuelad<-factor(Escuela)
str(Escuelad)
is.factor(Escuelad)
Escuelad
UbicacionHogard<-factor(UbicacionHogar)
str(UbicacionHogard)
is.factor(UbicacionHogard)
UbicacionHogard
Maedud<-factor(Maedu)
str(Maedud)
is.factor(Maedud)
Maedud
Paedud<-factor(Paedu)
str(Paedud)
is.factor(Paedud)
```

```
Paedud
TamaFamd<-factor(TamaFam)
str(TamaFamd)
is.factor(TamaFamd)
TamaFamd
Sexod<-factor(Sexo)
str(Sexod)
is.factor(Sexod)
Sexod
Pstatusd<-factor(Pstatus)
str(Pstatusd)
is.factor(Pstatusd)
Pstatusd
Paedud<-factor(Paedu)
str(Paedud)
is.factor(Paedud)
Paedud
MaTrabd<-factor(MaTrab)
str(MaTrabd)
is.factor(MaTrabd)
MaTrab
PaTrabd<-factor(Patrab)
str(PaTrabd)
is.factor(PaTrabd)
PaTrabd
AcudEstudd<-factor(AcudEstud)
str(AcudEstudd)
is.factor(AcudEstudd)
AcudEstudd
TempRecord<-factor(TempRecor)
str(TempRecord)
is.factor(TempRecord)
TempRecord
TempEstudiod<-factor(TempEstudio)
str(TempEstudiod)
is.factor(TempEstudiod)
TempEstudiod
NumReprod<-factor(NumRepro)
str(NumReprod)
is.factor(NumReprod)
NumReprod
SopExtraEdud<-factor(SopExtraEdu)
str(SopExtraEdud)
is.factor(SopExtraEdud)
SopExtraEdud
FamSupd<-factor(FamSup)
str(FamSupd)
is.factor(FamSupd)
FamSupd
PagoExtrClasd<-factor(PagoExtrClas)
str(PagoExtrClasd)
is.factor(PagoExtrClasd)
PagoExtrClasd
ExtraCurricActd<-factor(ExtraCurricAct)
str(ExtraCurricActd)
is.factor(ExtraCurricActd)
ExtraCurricActd
SupEstd<-factor(SupEst)
str(SupEstd)
is.factor(SupEstd)
SupEstd
Internetd<-factor(Internet)
str(Internetd)
```

```
is.factor(Internetd)


Internetd
RelaRomantd<-factor(RelaRomant)
str(RelaRomantd)
is.factor(RelaRomantd)
RelaRomantd
FamRelad<-factor(FamRela)
str(FamRelad)
is.factor(FamRelad)
FamRelad
TempLibrd<-factor(TempLibr)
str(TempLibrd)
is.factor(TempLibrd)
TempLibrd
SalAmigd<-factor(SalAmig)
str(SalAmigd)
is.factor(SalAmigd)
SalAmigd
RazonEscojd<-factor(RazonEscoj)
str(RazonEscojd)
is.factor(RazonEscojd)
RazonEscojd
table(RazonEscojd)
```

```
4. #Identificar las categorías de referencia de las variables Dummy

contrasts(Escuelad)
contrasts(UbicacionHogard)
contrasts(Maedud)
contrasts(Paedud)
contrasts(TamaFamd)
contrasts(Sexod)
contrasts(Pstatusd)
contrasts(Paedud)
contrasts(MaTrab)
contrasts(PaTrabd)
contrasts(AcudEstudd)
contrasts(TempRecord)
contrasts(TempEstudiod)
contrasts(NumReprod)
contrasts(SopExtraEdud)
contrasts(FamSupd)
contrasts(PagoExtrClasd)
contrasts(ExtraCurricActd)
contrasts(SupEstd)
contrasts(Internetd)
contrasts(RelaRomantd)
contrasts(FamRelad)
contrasts(TempLibrd)
contrasts(SalAmigd)
```

```
#Categoría base course - Nivel del curso-

contrasts(RazonEscojd)
```

5. Creación modelo Regresión logística multinomial función multinom()

Para la construcción del modelo multinomial se pueden emplear algunas funciones en R. Sin embargo, en este caso se ha usado la función *multinom ()* del paquete *nnet*. Mediante la

función *multinom ()* también se realiza el ajuste del modelo, y posteriormente se usará la

función *summary ()* para estbalecer los cocientes beta.

```
#Ejecutar el modelo completo con variables Dummy
#Función multinom -Regresión logística multinomial

modelomul<-multinom
(RazonEscojd~Escuelad+Sexod+Maedud+Paedud+Paedud+UbicacionHogard+TamaFamd+P
statusd+MaTrabd+PaTrabd+AcudEstudd+TempRecord+TempEstudiod+NumReprod+SopExt
raEdud+FamSupd+PagoExtrClasd+ExtraCurricActd+SupEstd+Internetd+RelaRomantd+
FamRelad+ TempEstudiod+SalAmigd,data=datos)
summary(modelomul)


#Facilitar la interpretación de los cocientes mediante el exponente
```

La funcion *exp()* en R, como se indica, devuelve el valor exponencial de un número o un vector numérico, dado por los cocientes estimados del modelo evaluado.

```
exp(coefficients(modelomul))

#Regresión logística sin variables explicativas

rlogvacio<-multinom(formula=RazonEscojd~1,datos)
summary(rlogvacio,Wald=TRUE)

#Regresión completa con variables Dummy y estadístico Wald

summary(rlogcompletodummy, Wald=TRUE)
rlogcompletodummy<-
multinom(RazonEscojd~Escuelad+Maedud+Paedud+UbicacionHogard+MaTrabd+PaTrabd
+AcudEstudd+TempRecord+TempEstudiod+NumReprod+SopExtraEdud+FamSupd+PagoExtr
Clasd+ExtraCurricActd+ExtraCurricActd+SupEstd+Internetd+RelaRomantd+FamRela
d+TempEstudiod+SalAmigd,data=datos)
summary(rlogcompletodummy, Wald=TRUE)
```

6. Método aplicado Regresión paso a paso función stepwise

```
#Regresión paso a paso función stepwise

rlog_stepwise<-step(rlogvacio,
                    scope=list(lower=rlogvacio,upper=rlogcompletodummy),
                    direction="both")
#cocientes del modelo, con sus errores estándares y el estadístico de Wald.

summary(rlog_stepwise, Wald=TRUE)

#Verificación de manera gráfica la relación entre las variables independiente
#con la variable dependiente -Razon de escogencia

ggplot(data = datos, aes(x =Escuela , y = RazonEscoj, color = Escuela)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.2) +
  theme_bw() +
```

```
  theme(legend.position = "null")



ggplot(data = datos, aes(x =UbicacionHogar , y = RazonEscoj, color =
UbicacionHogar )) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.2) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = datos, aes(x =ExtraCurricActd , y = RazonEscoj, color =
ExtraCurricActd  )) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.2) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = datos, aes(x =MaTrabd  , y = RazonEscoj, color = MaTrabd  ))
+
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = datos, aes(x =TempEstudiod   , y = RazonEscoj, color =
TempEstudiod   )) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = datos, aes(x =PagoExtrClasd   , y = RazonEscoj, color =
PagoExtrClasd   )) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.2) +
  theme_bw() +
  theme(legend.position = "null")

ggplot(data = datos, aes(x =NumReprod   , y = RazonEscoj, color = NumReprod
)) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")
ggplot(data = datos, aes(x =TempRecord   , y = RazonEscoj, color =
TempRecord   )) +
  geom_boxplot(outlier.shape = NA) +
  geom_jitter(width = 0.1) +
  theme_bw() +
  theme(legend.position = "null")
```

#Además del valor de las estimaciones de los cocientes parciales del modelo conviene obtener sus correspondientes intervalos de confianza

La mejor manera de visualizar los cocientes es a través de intervalos de confianza. Los intervalos de confianza marcan dos puntos dentro de los cuales es esperable que se encuentre el verdadero cociente del modelo, con un determinado nivel de confianza. En general cuanto mayor es la confianza, más amplio el intervalo. Es usual utilizar un intervalo de confianza del 95%. En R

obtenemos los intervalos de confianza de un objeto con la función *confint()*.

```
confint(object = rlog_stepwise, level = 0.95)
icexp=confint(object = rlog_stepwise, level = 0.95)
round(icexp,4)
```

7.1 Estimación de los IC de los OR

```
exp(icexp)
expor=round(exp(icexp),4)
expor
```

#Validación del modelo con la diferencia entre ellos.

```
anova(rlog_stepwise,rlogvacio)
exp(coefficients(rlog_stepwise))
```

#Validación del modelo con la igualdad entre ellos.

A través de la función *anova()*, se obtiene un análisis de varianza (o desviación) para uno o más parámetros del modelo ajustado.

```
anova(rlog_stepwise,rlogcompletodummy)
exp(coefficients(rlog_stepwise))
exponente=(exp(coefficients(rlog_stepwise,2)))
round(exponente,4)
```

Las validaciones efectuadas a continuación se hicieron esencialmente mediante el uso de la función *predict()*, la cual es una función genérica para predicciones a partir de los resultados de las funciones de ajuste de los modelos previamente evaluados.

7.2 Valores predichos

```
predict(rlog_stepwise)
View(datos)
```

7.3 Porcentaje de asertividad

```
obs<-RazonEscojd
pre<-predict(rlog_stepwise, type="class")
datos$predichos<-rlog_stepwise$fitted.values
table(predict(rlog_stepwise),RazonEscoj)
PAC=(221+39+1+64)/637
PAC
```

7.4 Cociente de Mac Faden. Mayores de 0,2 buen ajuste

Estos cocientes son establecidos mediate el uso de la función deviance() la cual permite la obtención del ajuste del modelo asociado a un determinado parámetro,

```
deviance(rlog_stepwise)
deviance(rlogvacio)

dv1=deviance(rlog_stepwise)
dv0=deviance(rlogvacio)
mf=1-(dv1/dv0)
mf
#Cociente de Nagelkerke. Mayores de 0,2 buen ajuste
ng=(1-exp((dv1-dv0)/637))/(1-exp(-dv0/637))
ng
```

7.5 Validación del modelo con los residuos

Lo hacemos mediante el uso de la función *residuals().* Esta es una función genérica que extrae los residuos del modelo de los parámetros obtenidos por las funciones de este.

```
#Deben ser menor de cuatro en valor absoluto
residuos=residuals(rlog_stepwise)
residuos1=round(residuos,4)
numSummary(residuos1)
```

Adicionalmente hemos incluido el anexo con todas las interacciones para el procedimiento Stepwise que permitieron establecer el grupo de variables que hicieron parte del modelo final.

Modelo vacío.

```
> summary(rlogvacio,Wald=TRUE)
Call:
multinom(formula = RazonEscojd ~ 1, data = datos)

Coefficients:
           (Intercept)
Hogar       -0.6443565
Otros       -1.4006575
Reputacion  -0.6860285

Std. Errors:
           (Intercept)
Hogar        0.1018537
Otros        0.1344019
Reputacion   0.1032650

Value/SE (Wald statistics):
           (Intercept)
Hogar        -6.326293
Otros       -10.421412
Reputacion   -6.643382

Residual Deviance: 1623.398
AIC: 1629.398
```

Interacciones por pasos

```
iter  30 value 748.147953
final  value 748.146804
converged
trying - PagoExtrClasd
# weights:  52 (36 variable)
initial  value 883.069508
iter  10 value 753.100363
iter  20 value 743.006731
iter  30 value 741.799512
final  value 741.796769
converged
trying - MaTrabd
# weights:  40 (27 variable)
initial  value 883.069508
iter  10 value 757.606704
iter  20 value 749.272232
iter  30 value 749.152805
final  value 749.152527
converged
trying + Maedud
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 743.531818
iter  20 value 732.894670
iter  30 value 731.689969
iter  40 value 731.664024
final  value 731.663985
converged
trying + Paedud
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 745.107823
iter  20 value 734.927779
iter  30 value 733.615196
iter  40 value 733.592555
final  value 733.592480
converged
trying + PaTrabd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 741.579701
iter  20 value 730.809860
iter  30 value 729.424552
iter  40 value 729.393764
final  value 729.393692
converged
trying + AcudEstudd
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 746.012167
```

```
iter  20 value 735.845043
iter  30 value 734.277088
iter  40 value 734.261943
final  value 734.261924
converged
trying + TempRecord
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 738.566513
iter  20 value 728.899557
iter  30 value 727.720145
iter  40 value 727.607544
iter  50 value 727.587306
iter  60 value 727.585854
iter  60 value 727.585848
iter  60 value 727.585848
final  value 727.585848
converged
trying + NumReprod
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 738.886093
iter  20 value 728.491679
iter  30 value 727.667409
iter  40 value 727.573922
iter  50 value 727.560204
final  value 727.559625
converged
trying + SopExtraEdud
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 746.404468
iter  20 value 735.736818
iter  30 value 734.956082
final  value 734.953635
converged
trying + FamSupd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 745.224421
iter  20 value 735.239981
iter  30 value 734.622599
iter  40 value 734.615024
iter  40 value 734.615017
iter  40 value 734.615015
final  value 734.615015
converged
trying + SupEstd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 746.718483
iter  20 value 737.345157
iter  30 value 736.311051
final  value 736.307862
converged
trying + Internetd
```

```
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 744.881325
iter  20 value 735.195629
iter  30 value 734.582340
iter  40 value 734.577378
iter  40 value 734.577375
iter  40 value 734.577375
final  value 734.577375
converged
trying + RelaRomantd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 744.491485
iter  20 value 736.172610
iter  30 value 735.101913
final  value 735.090999
converged
trying + FamRelad
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 742.619265
iter  20 value 733.843727
iter  30 value 732.698101
iter  40 value 732.620202
final  value 732.619696
converged
trying + SalAmigd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 745.956631
iter  20 value 734.150557
iter  30 value 732.625936
iter  40 value 732.563656
final  value 732.563488
converged
trying + G1
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 761.767380
iter  20 value 738.243650
iter  30 value 736.393453
iter  40 value 736.378451
final  value 736.378437
converged
trying + G3
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 755.872843
iter  20 value 735.354815
iter  30 value 734.139065
iter  40 value 734.127319
final  value 734.127308
converged
                  Df      AIC
+ +NumReprod      48 1551.119
```

```
+ +TempRecord     48 1551.172
<none>            39 1551.400
+ +G3             42 1552.255
- MaTrabd         27 1552.305
+ +Internetd      42 1553.155
+ +FamSupd        42 1553.230
+ +SopExtraEdud   42 1553.907
+ +RelaRomantd    42 1554.182
- G2              36 1554.993
- PagoExtrClasd   36 1555.594
- TempEstudiod    30 1556.294
+ +SupEstd        42 1556.616
+ +G1             42 1556.757
+ +AcudEstudd     45 1558.524
- ExtraCurricActd 36 1559.313
+ +Maedud         48 1559.328
- UbicacionHogard 36 1559.863
+ +PaTrabd        51 1560.787
+ +Paedud         48 1563.185
+ +SalAmigd       51 1567.127
+ +FamRelad       51 1567.239
- Escuelad        36 1572.656
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 738.886093
iter  20 value 728.491679
iter  30 value 727.667409
iter  40 value 727.573922
iter  50 value 727.560204
final  value 727.559625
converged

Step:  AIC=1551.12
RazonEscojd ~ Escuelad + G2 + ExtraCurricActd + UbicacionHogard +
    TempEstudiod + PagoExtrClasd + MaTrabd + NumReprod

trying - Escuelad
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 757.794409
iter  20 value 741.511841
iter  30 value 740.815434
iter  40 value 740.734045
iter  50 value 740.727985
final  value 740.727880
converged
trying - G2
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 741.252931
iter  20 value 730.572151
iter  30 value 730.243787
iter  40 value 730.166675
iter  50 value 730.161258
final  value 730.161175
converged
```

```
trying - ExtraCurricActd
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 748.782854
iter  20 value 734.927785
iter  30 value 734.307095
iter  40 value 734.221311
iter  50 value 734.214551
final  value 734.214325
converged
trying - UbicacionHogard
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 748.206094
iter  20 value 735.872079
iter  30 value 735.039759
iter  40 value 734.940917
iter  50 value 734.931154
final  value 734.930715
converged
trying - TempEstudiod
# weights:  56 (39 variable)
initial  value 883.069508
iter  10 value 750.344334
iter  20 value 739.303340
iter  30 value 738.794099
iter  40 value 738.717096
iter  50 value 738.713818
iter  50 value 738.713814
iter  50 value 738.713814
final  value 738.713814
converged
trying - PagoExtrClasd
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 743.970362
iter  20 value 733.677159
iter  30 value 732.815030
iter  40 value 732.717042
iter  50 value 732.703870
final  value 732.703625
converged
trying - MaTrabd
# weights:  52 (36 variable)
initial  value 883.069508
iter  10 value 752.017400
iter  20 value 740.788005
iter  30 value 740.533978
iter  40 value 740.484001
final  value 740.482939
converged
trying - NumReprod
# weights:  56 (39 variable)
initial  value 883.069508
iter  10 value 747.771531
iter  20 value 737.420154
```

```
iter  30 value 736.701984
final  value 736.699911
converged
trying + Maedud
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 736.689504
iter  20 value 724.209302
iter  30 value 722.913805
iter  40 value 722.830065
iter  50 value 722.798358
iter  60 value 722.796108
final  value 722.796068
converged
trying + Paedud
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 738.161167
iter  20 value 725.980415
iter  30 value 724.635551
iter  40 value 724.542846
iter  50 value 724.496445
iter  60 value 724.494290
final  value 724.494087
converged
trying + PaTrabd
# weights:  84 (60 variable)
initial  value 883.069508
iter  10 value 734.016923
iter  20 value 721.418238
iter  30 value 720.178349
iter  40 value 720.073523
iter  50 value 720.033631
iter  60 value 720.031167
final  value 720.031085
converged
trying + AcudEstudd
# weights:  76 (54 variable)
initial  value 883.069508
iter  10 value 737.974112
iter  20 value 726.412778
iter  30 value 724.958346
iter  40 value 724.864072
iter  50 value 724.835477
iter  60 value 724.833574
final  value 724.833532
converged
trying + TempRecord
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 731.685428
iter  20 value 719.820440
iter  30 value 718.660267
iter  40 value 718.456771
iter  50 value 718.428329
iter  60 value 718.426449
```

```
final  value 718.426379
converged
trying + SopExtraEdud
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 737.670356
iter  20 value 726.901104
iter  30 value 726.115480
iter  40 value 726.025403
iter  50 value 726.009574
iter  60 value 726.008564
final  value 726.008552
converged
trying + FamSupd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 737.792132
iter  20 value 726.871176
iter  30 value 725.656770
iter  40 value 725.528674
iter  50 value 725.505742
iter  60 value 725.503276
final  value 725.503252
converged
trying + SupEstd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 738.311959
iter  20 value 728.504097
iter  30 value 727.479899
iter  40 value 727.383519
iter  50 value 727.370420
iter  60 value 727.369571
iter  60 value 727.369566
iter  60 value 727.369566
final  value 727.369566
converged
trying + Internetd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 737.825041
iter  20 value 726.632246
iter  30 value 725.544617
iter  40 value 725.428359
iter  50 value 725.411782
iter  60 value 725.410689
iter  60 value 725.410683
iter  60 value 725.410683
final  value 725.410683
converged
trying + RelaRomantd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 736.589202
iter  20 value 727.460995
iter  30 value 726.001620
```

```
iter  40 value 725.919825
iter  50 value 725.896571
iter  60 value 725.895388
iter  60 value 725.895384
iter  60 value 725.895384
final  value 725.895384
converged
trying + FamRelad
# weights:  84 (60 variable)
initial  value 883.069508
iter  10 value 735.428221
iter  20 value 725.060276
iter  30 value 723.772516
iter  40 value 723.658195
iter  50 value 723.620155
iter  60 value 723.616179
final  value 723.615766
converged
trying + SalAmigd
# weights:  84 (60 variable)
initial  value 883.069508
iter  10 value 737.047895
iter  20 value 725.229624
iter  30 value 723.583847
iter  40 value 723.457731
iter  50 value 723.395376
iter  60 value 723.390735
final  value 723.390436
converged
trying + G1
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 756.503042
iter  20 value 729.689254
iter  30 value 727.537319
iter  40 value 727.397300
iter  50 value 727.368144
final  value 727.367187
converged
trying + G3
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 749.291791
iter  20 value 726.011011
iter  30 value 724.600850
iter  40 value 724.471210
iter  50 value 724.443041
iter  60 value 724.441761
iter  60 value 724.441756
iter  60 value 724.441756
final  value 724.441756
converged
                  Df      AIC
- G2              45 1550.322
+ +TempRecord     57 1550.853
+ +G3             51 1550.884
```

```
<none>            48 1551.119
- NumReprod       39 1551.400
+ +Internetd      51 1552.821
- MaTrabd         36 1552.966
+ +FamSupd        51 1553.007
+ +RelaRomantd    51 1553.791
+ +SopExtraEdud   51 1554.017
- PagoExtrClasd   45 1555.407
- TempEstudiod    39 1555.428
+ +G1             51 1556.734
+ +SupEstd        51 1556.739
+ +AcudEstudd     54 1557.667
- ExtraCurricActd 45 1558.429
+ +Maedud         57 1559.592
- UbicacionHogard 45 1559.861
+ +PaTrabd        60 1560.062
+ +Paedud         57 1562.988
+ +SalAmigd       60 1566.781
+ +FamRelad       60 1567.232
- Escuelad        45 1571.456
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 741.252931
iter  20 value 730.572151
iter  30 value 730.243787
iter  40 value 730.166675
iter  50 value 730.161258
final  value 730.161175
converged

Step:  AIC=1550.32
RazonEscojd ~ Escuelad + ExtraCurricActd + UbicacionHogard +
    TempEstudiod + PagoExtrClasd + MaTrabd + NumReprod

trying - Escuelad
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 754.071140
iter  20 value 745.116465
iter  30 value 744.898564
iter  40 value 744.834740
iter  50 value 744.827840
final  value 744.827805
converged
trying - ExtraCurricActd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 748.232850
iter  20 value 737.557378
iter  30 value 737.149998
iter  40 value 737.087480
iter  50 value 737.082669
final  value 737.082582
converged
trying - UbicacionHogard
# weights:  60 (42 variable)
```

```
initial  value 883.069508
iter  10 value 742.886391
iter  20 value 737.485486
iter  30 value 737.200686
iter  40 value 737.142759
iter  50 value 737.140336
final  value 737.140319
converged
trying - TempEstudiod
# weights:  52 (36 variable)
initial  value 883.069508
iter  10 value 749.897967
iter  20 value 742.857397
iter  30 value 742.620029
iter  40 value 742.560922
final  value 742.558625
converged
trying - PagoExtrClasd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 747.113974
iter  20 value 735.832753
iter  30 value 735.580428
iter  40 value 735.501486
iter  50 value 735.492556
final  value 735.492507
converged
trying - MaTrabd
# weights:  48 (33 variable)
initial  value 883.069508
iter  10 value 751.662170
iter  20 value 743.929779
iter  30 value 743.753631
iter  40 value 743.719132
final  value 743.718743
converged
trying - NumReprod
# weights:  52 (36 variable)
initial  value 883.069508
iter  10 value 750.248550
iter  20 value 741.684239
iter  30 value 741.496391
final  value 741.496307
converged
trying + Maedud
# weights:  76 (54 variable)
initial  value 883.069508
iter  10 value 739.673482
iter  20 value 726.808507
iter  30 value 725.541916
iter  40 value 725.406588
iter  50 value 725.383294
iter  60 value 725.381295
final  value 725.381270
converged
trying + Paedud
```

```
# weights:  76 (54 variable)
initial  value 883.069508
iter  10 value 739.507075
iter  20 value 728.128555
iter  30 value 727.043488
iter  40 value 726.940809
iter  50 value 726.923271
iter  60 value 726.922030
final  value 726.922019
converged
trying + PaTrabd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 739.976838
iter  20 value 724.033301
iter  30 value 722.908060
iter  40 value 722.831512
iter  50 value 722.804488
iter  60 value 722.802514
final  value 722.802460
converged
trying + AcudEstudd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 740.948321
iter  20 value 728.372957
iter  30 value 727.623844
iter  40 value 727.508515
iter  50 value 727.497050
iter  60 value 727.496358
iter  60 value 727.496355
iter  60 value 727.496355
final  value 727.496355
converged
trying + TempRecord
# weights:  76 (54 variable)
initial  value 883.069508
iter  10 value 734.793642
iter  20 value 721.189872
iter  30 value 720.568457
iter  40 value 720.383598
iter  50 value 720.369546
iter  60 value 720.368517
final  value 720.368498
converged
trying + SopExtraEdud
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 740.180735
iter  20 value 728.988847
iter  30 value 728.596286
iter  40 value 728.513039
iter  50 value 728.506821
final  value 728.506541
converged
trying + FamSupd
```

```
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 742.235546
iter  20 value 728.933422
iter  30 value 728.259645
iter  40 value 728.164746
iter  50 value 728.156160
final  value 728.155889
converged
trying + SupEstd
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 751.956378
iter  20 value 730.498994
iter  30 value 729.834551
iter  40 value 729.733899
iter  50 value 729.724831
final  value 729.724438
converged
trying + Internetd
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 744.485337
iter  20 value 729.144032
iter  30 value 728.119333
iter  40 value 728.030154
iter  50 value 728.016519
final  value 728.015883
converged
trying + RelaRomantd
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 737.890179
iter  20 value 728.716615
iter  30 value 728.191478
iter  40 value 728.094861
iter  50 value 728.084633
final  value 728.084143
converged
trying + FamRelad
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 746.647846
iter  20 value 727.745628
iter  30 value 726.230137
iter  40 value 726.118868
iter  50 value 726.081993
iter  60 value 726.078835
final  value 726.078660
converged
trying + SalAmigd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 737.414823
iter  20 value 727.717334
iter  30 value 726.441677
```

```
iter  40 value 726.325144
iter  50 value 726.292846
iter  60 value 726.290664
final  value 726.290614
converged
trying + G1
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 761.138499
iter  20 value 730.952042
iter  30 value 728.915262
iter  40 value 728.802678
iter  50 value 728.783431
final  value 728.782308
converged
trying + G2
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 738.886093
iter  20 value 728.491679
iter  30 value 727.667409
iter  40 value 727.573922
iter  50 value 727.560204
final  value 727.559625
converged
trying + G3
# weights:  68 (48 variable)
initial  value 883.069508
iter  10 value 736.414568
iter  20 value 727.869200
iter  30 value 727.214304
iter  40 value 727.125312
iter  50 value 727.111876
final  value 727.111060
converged
                  Df      AIC
+ +TempRecord     54 1548.737
+ +G3             48 1550.222
<none>            45 1550.322
+ +G2             48 1551.119
+ +Internetd      48 1552.032
+ +RelaRomantd    48 1552.168
+ +FamSupd        48 1552.312
+ +SopExtraEdud   48 1553.013
- MaTrabd         33 1553.437
+ +G1             48 1553.565
- PagoExtrClasd   42 1554.985
- NumReprod       36 1554.993
+ +SupEstd        48 1555.449
+ +AcudEstudd     51 1556.993
- TempEstudiod    36 1557.117
- ExtraCurricActd 42 1558.165
- UbicacionHogard 42 1558.281
+ +Maedud         54 1558.763
+ +PaTrabd        57 1559.605
+ +Paedud         54 1561.844
```

```
+ +FamRelad       57 1566.157
+ +SalAmigd       57 1566.581
- Escuelad        42 1573.656
# weights:  76 (54 variable)
initial  value 883.069508
iter  10 value 734.793642
iter  20 value 721.189872
iter  30 value 720.568457
iter  40 value 720.383598
iter  50 value 720.369546
iter  60 value 720.368517
final  value 720.368498
converged

Step:  AIC=1548.74
RazonEscojd ~ Escuelad + ExtraCurricActd + UbicacionHogard +
    TempEstudiod + PagoExtrClasd + MaTrabd + NumReprod + TempRecord

trying - Escuelad
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 745.882980
iter  20 value 735.934205
iter  30 value 735.210652
iter  40 value 735.057232
iter  50 value 735.047014
final  value 735.046499
converged
trying - ExtraCurricActd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 736.942575
iter  20 value 727.726112
iter  30 value 727.298377
iter  40 value 727.166803
iter  50 value 727.161566
final  value 727.160962
converged
trying - UbicacionHogard
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 734.580632
iter  20 value 727.295195
iter  30 value 726.818865
iter  40 value 726.668837
iter  50 value 726.656521
final  value 726.655877
converged
trying - TempEstudiod
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 743.474069
iter  20 value 734.556200
iter  30 value 734.044519
iter  40 value 733.961295
iter  50 value 733.957748
```

```
final  value 733.957694
converged
trying - PagoExtrClasd
# weights:  72 (51 variable)
initial  value 883.069508
iter  10 value 739.852802
iter  20 value 726.685533
iter  30 value 726.156459
iter  40 value 725.992941
iter  50 value 725.980540
final  value 725.979836
converged
trying - MaTrabd
# weights:  60 (42 variable)
initial  value 883.069508
iter  10 value 741.038887
iter  20 value 734.190614
iter  30 value 733.670921
iter  40 value 733.556279
iter  50 value 733.550694
final  value 733.550663
converged
trying - NumReprod
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 743.137132
iter  20 value 732.124699
iter  30 value 731.654952
iter  40 value 731.546195
iter  50 value 731.534365
final  value 731.533556
converged
trying - TempRecord
# weights:  64 (45 variable)
initial  value 883.069508
iter  10 value 741.252931
iter  20 value 730.572151
iter  30 value 730.243787
iter  40 value 730.166675
iter  50 value 730.161258
final  value 730.161175
converged
trying + Maedud
# weights:  88 (63 variable)
initial  value 883.069508
iter  10 value 732.405386
iter  20 value 718.074300
iter  30 value 716.173294
iter  40 value 715.914234
iter  50 value 715.873916
iter  60 value 715.870962
final  value 715.870828
converged
trying + Paedud
# weights:  88 (63 variable)
initial  value 883.069508
```

```
iter  10 value 733.818242
iter  20 value 718.638888
iter  30 value 717.385028
iter  40 value 717.156739
iter  50 value 717.117687
iter  60 value 717.113797
final  value 717.113639
converged
trying + PaTrabd
# weights:  92 (66 variable)
initial  value 883.069508
iter  10 value 729.931141
iter  20 value 714.438311
iter  30 value 713.100182
iter  40 value 712.858535
iter  50 value 712.823878
iter  60 value 712.822224
final  value 712.821970
converged
trying + AcudEstudd
# weights:  84 (60 variable)
initial  value 883.069508
iter  10 value 732.819421
iter  20 value 718.673516
iter  30 value 717.868051
iter  40 value 717.656239
iter  50 value 717.635376
iter  60 value 717.634314
final  value 717.634222
converged
trying + SopExtraEdud
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 733.609646
iter  20 value 719.504954
iter  30 value 718.933593
iter  40 value 718.742236
iter  50 value 718.726737
iter  60 value 718.726015
final  value 718.725985
converged
trying + FamSupd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 737.366402
iter  20 value 719.239954
iter  30 value 718.288018
iter  40 value 718.058875
iter  50 value 718.029304
iter  60 value 718.025766
final  value 718.025604
converged
trying + SupEstd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 745.157706
```

```
iter  20 value 721.201111
iter  30 value 720.186322
iter  40 value 719.956309
iter  50 value 719.934477
iter  60 value 719.933230
final  value 719.933182
converged
trying + Internetd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 743.286617
iter  20 value 719.725460
iter  30 value 718.814570
iter  40 value 718.618965
iter  50 value 718.597558
final  value 718.596773
converged
trying + RelaRomantd
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 739.134279
iter  20 value 720.399811
iter  30 value 718.935781
iter  40 value 718.780367
iter  50 value 718.746190
iter  60 value 718.743294
final  value 718.743121
converged
trying + FamRelad
# weights:  92 (66 variable)
initial  value 883.069508
iter  10 value 735.035813
iter  20 value 718.752431
iter  30 value 716.436904
iter  40 value 716.184576
iter  50 value 716.121595
iter  60 value 716.117704
iter  70 value 716.117283
iter  70 value 716.117279
iter  70 value 716.117279
final  value 716.117279
converged
trying + SalAmigd
# weights:  92 (66 variable)
initial  value 883.069508
iter  10 value 735.746082
iter  20 value 718.676096
iter  30 value 716.904676
iter  40 value 716.714741
iter  50 value 716.620242
iter  60 value 716.613058
final  value 716.612773
converged
trying + G1
# weights:  80 (57 variable)
initial  value 883.069508
```

```
iter  10 value 740.724108
iter  20 value 720.789578
iter  30 value 719.544549
iter  40 value 719.332338
iter  50 value 719.294096
iter  60 value 719.292320
final  value 719.292256
converged
trying + G2
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 731.685428
iter  20 value 719.820440
iter  30 value 718.660267
iter  40 value 718.456771
iter  50 value 718.428329
iter  60 value 718.426449
final  value 718.426379
converged
trying + G3
# weights:  80 (57 variable)
initial  value 883.069508
iter  10 value 730.560723
iter  20 value 718.862601
iter  30 value 717.790004
iter  40 value 717.591156
iter  50 value 717.561195
iter  60 value 717.559457
final  value 717.559402
converged
                  Df      AIC
<none>            54 1548.737
+ +G3             57 1549.119
+ +FamSupd        57 1550.051
- TempRecord      45 1550.322
+ +G2             57 1550.853
- MaTrabd         42 1551.101
+ +Internetd      57 1551.194
+ +SopExtraEdud   57 1551.452
+ +RelaRomantd    57 1551.486
+ +G1             57 1552.585
- NumReprod       45 1553.067
+ +SupEstd        57 1553.866
- PagoExtrClasd   51 1553.960
+ +AcudEstudd     60 1555.268
- UbicacionHogard 51 1555.312
- ExtraCurricActd 51 1556.322
+ +PaTrabd        66 1557.644
+ +Maedud         63 1557.742
- TempEstudiod    45 1557.915
+ +Paedud         63 1560.227
+ +FamRelad       66 1564.235
+ +SalAmigd       66 1565.226
- Escuelad        51 1572.093
```

Referencias

Akaige, H. (1974). A new look at the statistical model identification. Doi: 10.1109/TAC.1974.1100705

Bravo, G., y Vergara, M. (2018). Factores que determinan la elección de carrera profesional: en estudiantes de undécimo grado de colegios públicos y privados de Barrancabermeja. *Revista Psicoespacios, 12*(20), 47-58. Doi: 10.25057/21452776.1000

Carrasco, E., Zúñiga, C., y Espinoza, J. (2014). Elección de carrera en estudiantes de nivel socioeconómico bajo de universidades chilenas altamente selectivas. Calidad La Educación, 40, 96–128. https://doi.org/10.4067/S0718-45652014000100004

Cerda, J., Vera, C., y Rada, G. (2013). Odds ratio: aspectos teóricos y prácticos. *Rev Med Chile*, (141), 1329-1335.

García, J., y Moreno, C. (2012). Factores considerados al seleccionar una universidad. Caso Ciudad Juárez, 17(52), 287-305.

Lizares, M (2017). Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico. Tesis para optar al título de licenciada en estadística. Universidad Mayor de San Marcos. Lima, Perú. Disponible en: https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/7122/Lizares_cm.pdf?sequence=3&isAllowed=y

Morales N., Arismendy, C., y Díaz, J. (2020). Análisis de la deserción temprana y sus factores explicativos en la Universidad de los Llanos con datos de ingreso 2015-2. Revista investigaciones Andina. 22(40), 99-121. Doi: 10.33132/01248146.1589

Ogutu, J., Odera, P., y Maragia, S. (2017). Self-Efficacy as a Predictor of Career Decision Making among Secondary School Students in Busia County, Kenya. *Journal of education and practice, 8*(11), 20-29

Salinas, P. (2010). Metodología de la investigación científica. Disponible en http://www.saber.ula.ve/bitstream/handle/123456789/34398/metodologia_investigacion.pdf;jsessionid=FD96511E7A056F24B14EDE98222E834E?sequence=1

Van Herpen, S., Meeuwisse, M., Hofman., Severiens, S., y Arends, L. (2017). Early predictors of first-year academic success at university: pre-university effort, pre-university self-efficacy, and pre-university reasons for attending university. *Educational Research and Evaluation, 23*(1-2), 52-72, Doi: 10.1080/13803611.2017.1301261

Wiks. H. (1935). The Likelihood test of Independence in contingency. *Annals of Mathematical Statistics,* (6), 190.

Bibliografía

Abarca, S. (1995). Psicología de la motivación. San José, C.R.: Editorial Universidad Estatal a Distancia.

Aguilera, A.M. y Escabias, M. (2000). Principal component logistic regression. Proceedings in Computational Statistics 2000.

Aguilera del Pino, A. M. Modelos de Respuesta Discreta. Granada: Copias Coca, Dep. Legal GR-11554-02; 2002

Agresti A. Categorical Data Analysis. Second Edition ed. New York: Wiley; 2002.

Cook, R. D., y Weisberg, S. (1982). Residuals and Influence in Regression. New York. Chapman and Hall.

De, T., Mendoza, L., y Rodríguez Martínez, R. (n.d.). El Efecto de la Orientación Vocacional en la Elección de Carrera, V(13), 10-16. Recuperado de: http://pepsic.bvsalud.org/ pdf/remo/v5n13/v5n13a04.pdf

Dueñas, M. Á. (2012). Modelos de respuesta discreta con datos reales. Granada

Fagerland, M.W., Hosmer, D.W. y Bofin, A.M. (2008). Multinomial goodness-of-fit tests for logistic regression models. Statistics in Medicine, 2008 Sep 20. Vol 27(21), pags 4238 a 4253.

González, V. (2009). Autodeterminación y conducta exploratoria. Elementos esenciales en la competencia para la elección profesional responsable. *Revista Iberoamericana de Educación, 51*(51), 201–220.

Hosmer, D.W. y Lemeshow, S. (1989). Applied Logistic Regression. Wiley.

Kleinbaum, D. G., Klein, M., y Pryor, E. R. (2002). Logistic regression: a self-learning text (2nd ed). New York: Springer.

Fernández, V., y Fernández, R. (2004). Regresión logística multinomial. *Cuadernos de la Sociedad Española de Ciencias Forestales*, *1(18)*, 323-327.

Hernández, P. (2005). La motivación en los estudiantes universitarios. *Revista Electrónica Actualidades Investigativas en Educación, 5*(2), 1-13.

Rivera, S., Larrondo, F., y Ortega, J. (2005). Evaluación de los resultados de un artículo sobre tratamiento. Rev Med Chile 2005; 133: 593-6.

Rodríguez Arjona Ana M, Baas Lara Mario Alberto, Cachon Medina Carlo M (2017). Factores que influyen en los alumnos para la Elección de una carrera.

Martín-Moreno JM, Banegas JR. Sobre la traducción del término inglés odds ratio como oportunidad relativa. *Salud Pública Mex 1997*, (39), 72-4.

Mendoza y Rodríguez. (2008). El efecto de la orientación vocacional en la elección de carrera. Revista Mexicana de Orientación Educativa, 5(13), 10–16.

Montesano, J. C., y Zambrano, E. (2013). Factores que influyen en la elección de una carrera universitaria en la Universidad Católica Andrés Bello. Caracas, Venezuela: Universidad Católica Andrés Bello.

Pineda Barón, L. A. (2015). Factores que afectan la elección de carrera: caso Bogotá. Vniversitas Económica, 15(3), 1–35.

Oztuna, D., Elhan, A.H. & Tuccar, E. (2006). Investigation of four different normality tests in terms of type I. Error rate and power under different distributions. Turkish Journal of Medical Sciences(3), 171-176.

Siegel, S, y Castellan, N. (1995) Estadística no paramétrica aplicada a las ciencias de la conducta. 4ª ed. México: Editorial Trillas,: 151-7

Rodríguez, R. (2004). Ayuda SPSS Chi Cuadrado. Notas Metodológicas.