UNIVERSIDAD DE GRANADA

MÁSTER UNIVERSITARIO EN ESTADÍSTICA APLICADA

TRABAJO FIN DE MÁSTER

ANÁLISIS NIVEL DE LOGRO ALCANZADO EN LA NOTA DEL EXAMEN DE GRADO SER BACHILLER – ECUADOR EN FUNCIÓN DE VARIABLES SOCIODEMOGRÁFICAS MEDIANTE LA APLICACIÓN DE MODELOS DE REGRESIÓN LOGÍSTICA MULTINOMIAL

Presentado por:

Tania Paulina Morocho Barrionuevo

Tutores:

Dr. Manuel Escabias Machuca

Dra. Ana María Aguilera del Pino

2020-2021

Agradecimientos

Quiero agradecer a Dios y a mi familia por su apoyo incondicional

A mi hija Evelyn mi motor que impulsa para avanzar,

a mi padre, madre, hermana, por su apoyo y cariño.

A mis tutores el Doctor Manuel Escabias Machuca y

a la Doctora Ana María Aguilera del Pino, por sus

conocimientos y acompañamiento para

ver plasmado este trabajo.

Resumen

El presente trabajo describe el modelo de regresión logística multinomial ordinal y nominal y su aplicación a datos reales con el objetivo de mostrar el uso y utilidad de los modelos para la modelización del nivel de logro alcanzado en la nota del examen de grado Ser Bachiller - Ecuador periodo 2018-2019, en función de las variables sociodemográficas edad, sexo, provincia, tipo de financiamiento de la institución educativa y autoidentificación étnica.

Para el cálculo se utilizó el software libre R ajustando el modelo de respuesta nominal y ordinal, además mediante el método stepwise se halló los mejores predictores para el estudio.

Se concluyó la aplicación del método stepwise estima que las variables independientes utilizadas en el estudio son las adecuadas ya que mediante la selección automática da como predictoras a las mismas variables.

Índice de contenido

Capítulo 1

Introducción

En el Ecuador la educación de calidad es un derecho que todas las personas poseen y que pueden acceder sin excepción alguna, según la Constitución de la Republica.

Los niveles educativos manejados están establecidos por el Sistema Nacional de Educación, mismos que se encuentran divididos en tres niveles: Inicial, Básica y Bachillerato.

- El nivel de Educación Inicial, es el proceso de acompañamiento de niños y niñas para el desarrollo integral de sus capacidades cognitivas, afectivas, psicomotrices, sociales e identitarias, que va desde los 3 hasta los 5 años de edad y se divide en dos subniveles (UNESCO, s.f.):

Inicial 1, que no es escolarizado y comprende a infantes de hasta tres años de edad.

Inicial 2, que comprende a infantes de tres a cinco años de edad.

- El nivel de Educación General Básica está compuesta por diez años de educación obligatoria en los que se refuerzan, amplían y profundizan las capacidades y competencias adquiridas, se introducen disciplinas básicas, para garantizar la diversidad cultural y lingüística, está divida en cuatro subniveles:

Preparatoria, que corresponde al primer grado de Educación General Básica y preferentemente se ofrece a los estudiantes de cinco años de edad.

Básica Elemental, que corresponde al segundo, tercero y cuarto grado de Educación General Básica y preferentemente se ofrece a los estudiantes de 6 a 8 años de edad.

Básica Media, que corresponde a quinto, sexto y séptimo grado de Educación General Básica y preferentemente se ofrece a los estudiantes de 9 a 11 años de edad.

Básica Superior, que corresponde a octavo, noveno y décimo grados de Educación General Básica y preferentemente se ofrece a los estudiantes de 12 a 14 años de edad.

- El nivel de Bachillerato General Unificado comprende tres años de educación obligatoria brindando una formación general y una preparación interdisciplinaria desarrollando en los estudiantes capacidades permanentes de aprendizaje y

competencias ciudadanas, consta de tres cursos primero, segundo, tercero de bachillerato y preferentemente se ofrece a los estudiantes de 15 a 17 años de edad. Las edades estipuladas son sugeridas para la educación en cada nivel, sin embargo, no se niega el acceso de los estudiantes con: necesidades educativas especiales, jóvenes y adultos con escolaridad inconclusa, repetición de un año escolar, entre otros, permitiéndoles acceder a un grado o curso. (Educación, 2015)

Desde el 2010, el Ecuador adoptó la Ley Orgánica de Educación Superior (LOES), la cual permitió la creación de la Secretaría de Educación Superior de Ciencia y Tecnología (SENESCYT) y el Sistema Nacional de Admisión y Nivelación (SNNA) que implementó el Examen Nacional de Educación Superior (ENES) como un instrumento obligatorio para regular el acceso a la Educación Superior, pero en el período 2016-2017, se inició la aplicación de un nuevo examen, denominado Examen Unificado Ser Bachiller, a través del cual los estudiantes obtendrían su título de bachiller y el cupo a una universidad (Guadagni, 2016).

La evaluación educativa es un proceso que integra la elaboración, aplicación y análisis de instrumentos de medición, los cuales deducen las capacidades y destrezas de las personas. El instrumento de medición educativo tiene como función ofrecer información para la toma correcta de decisiones, y numerosas investigaciones consideran que el examen de ingreso a la universidad es una variable imprescindible en el éxito académico del estudiante (Carrión, 2002).

Cuando se emplean instrumentos de gran escala y de alto impacto social, tal como los exámenes de admisión y/o aprobación de la Educación a nivel de Bachillerato, su elaboración debe ajustarse a rigurosos estándares de calidad (Aiken, 1996).

El examen Ser Bachiller es tomado a los estudiantes pertenecientes al tercero de Bachillerato General Unificado (BGU). El procedimiento para la ejecución del examen parte de la autoasignación del estudiante a una sede o una sede preasignada, posteriormente deberán acercase según el cronograma previsto, puesto que hay una fecha asignada para personas con discapacidad y personas privadas de libertad, otra

para la población no escolar y otra para la población escolar y personas ecuatorianas residentes en el extranjero.

La implementación del examen Ser bachiller desde el 2017 fue concebida para evaluar el desarrollo de las aptitudes y destrezas que deben alcanzar los estudiantes al terminal la educación intermedia, y a la vez esta evaluación contribuye al proceso de admisión a la educación superior pública. La información recopilada con este examen sirve para la formulación de mecanismos de mejora para la educación a nivel inicial, básico y bachillerato, enfatizando en acciones de mejora para el acceso a la educación superior.

Siendo así de gran importancia el Examen Unificado Ser bachiller para la población con el fin de continuar con su vida educativa y profesional, se toma como base sustentativa para el estudio la base de datos del Instituto Nacional de Evaluación educativa con su respectivo diccionario de variables denominada "*Base de datos Ser Bachiller Año lectivo 2018-2019"*, tomada a 514852 estudiantes pertenecientes al tercero de Bachillerato, y es oportuno identificar qué factores sociodemográficos (autoidentificación étnica, región natural, tipo de financiamiento, tipo de sexo y edad) influyen en que el estudiante de educación nivel bachillerato obtengan un nivel de logro en el rendimiento académico de tipo (insuficiente, elemental, satisfactorio o excelente).

La autoidentificación étnica es una de las variables para el estudio que vislumbra a personas que se autodefinen con alguna nacional o pueblo de manera libre y voluntaria, sea indígena, mestizo/blanco, afroecuatoriano, montubio u otra étnia.

En relación a las regiones naturales en Ecuador se encuentra definida por criterios de geografía física, principalmente con el clima, la vegetación, la hidrografía de los suelos y otros. La división territorial corresponde a Costa, Sierra, Oriente, Insular, Otro. (INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS, 2018)
Región Costa: Comprende más de la cuarta parte del país, es un área geográfica que se encuentra entre el Océano Pacífico y la Cordillera de los Andes, conformada por las provincias de: Esmeraldas, Manabí, Los Ríos, Santa Elena, Guayas, Santo Domingo de los Tsáchilas y El Oro.

Región Sierra: Se extiende en una doble hilera de montañas y una estrecha meseta conocida como el valle interandino, constituida por: Azuay, Bolívar, Cañar, Carchi, Cotopaxi, Chimborazo, Imbabura, Loja, Pichincha y Tungurahua.
Región Oriente: Al este de los andes, se encuentra cubierta casi en su totalidad con selva, presenta gran cantidad de la flora y la fauna nativa, formada por: Morona Santiago, Napo, Orellana, Pastaza, Sucumbíos y Zamora Chinchipe.
Región Insular: Integrada por el Archipiélago de Colón, comprendida por la Isla Isabela, Santa Cruz y San Cristóbal. (Romero S., 2021)
Región Zona no delimitada: Constituida por cantones que se encuentran en zonas limítrofes o fronterizas como: Loja (Macará, Sozoranga, Espíndola, Puyango, Pindal, Zapotillo, Calvas), Zamora Chinchipe (El Pangui, Yacuambi, Yantzaza, Nangaritza, Palanda, Chinchipe, Paquisha), Pastaza (Pastaza, Arajuno), Orellana (Aguarico) y Morona Santiago (Taisha, Tiwintza, Limón Indanza, San Juan Bosco, Gualaquiza).

El tipo de financiamiento de las instituciones educativas está establecida según su sostenimiento como público, privado y mixto o fiscomisionales.
Públicas: Son instituciones educativas financiadas por el estado.
Mixto o Fiscomisionales: Estas Instituciones Educativas son de carácter religioso o laico, de derecho privado y sin fines de lucro, que garantizan una educación gratuita y de calidad. Los centros cuentan con financiamiento total o parcial del Estado (Gobierno Central a través del MinEduc).
Privado: Constituidas y administradas por personas naturales o jurídicas de derecho privado. La educación en estas Instituciones Educativas puede ser confesional o laica.

El tipo de sexo toma hombre y mujer, y la edad va desde los 16 a los 25 años debido a que la población mayoritaria se encuentra entre este rango de edades.

El nivel de logro alcanzado existen cuatro categorías consideradas dentro de la calificación de la prueba Ser Bachiller: “Excelente” con una calificación de 950 a 1.000 puntos. En esta categoría se ubica el Grupo de Alto Rendimiento, seguido de la calificación “Satisfactoria” con un puntaje de hasta 800. Luego sigue la categoría “Elemental” con un puntaje de hasta 700, el cual es un requisito mínimo para que el

estudiante se pueda graduar. Por último, una calificación "Insuficiente" de hasta 601, aunque sea el mínimo para poder postular a la educación superior, no puede aprobar como requerimiento de examen de grado. (Instituto Nacional de Evaluación Educativa, 2018).

Objetivos y metodología de la investigación

Objetivo General:

Mostrar el uso y utilidad de los modelos de respuesta nominal y de los modelos de respuesta ordinal, para la modelización del nivel de logro alcanzado en la nota del examen de grado Ser Bachiller-Ecuador periodo 2018-2019, en función de las variables sociodemográficas edad, sexo, provincia, tipo de financiamiento de la institución educativa y autoidentificación étnica.

Objetivos Específicos:

- Ajustar un modelo de respuesta nominal para la proyección de nivel de logro alcanzado.
- Ajustar un modelo de respuesta ordinal para estimar el logro alcanzado.
- Seleccionar mediante el método stepwise los mejores predictores tanto para modelo nominal y ordinal
- A través de los resultados establecer la diferencia entre el modelo nominal y ordinal para la modelización de este tipo de datos.

Con la finalidad de cumplir los objetivos propuestos este trabajo se encuentra divido en 4 capítulos.

- Capítulo 1 (Introducción). Descripción introductoria del trabajo, objetivos del mismo y su estructura planteada.

- Capítulo 2 (Antecedentes). Se describe estudios similares tomados como referentes.

- Capítulo 3 (Modelos de Regresión Múltiple). Explicación teórica de los distintos modelos.
- Capítulo 4 (Ajuste del modelo). Descripción del conjunto de datos, utilización de distintos ajustes y conclusiones.

Al final del trabajo se adjunta los respectivos anexos.

Capítulo 2

En este capítulo se muestra de manera resumida estudios similares, publicados en la literatura científica que se encuentran relacionados con los objetivos planteados en este trabajo, los cuales emplean modelos de regresión logística aplicados a base de datos reales.

Antecedentes

En Ecuador, la educación en su trayecto de escolarización y educación media de manera obligatoria en todo el territorio.

Ramiro Efraín Villarruel Meythaler, Karen Irene Tapia Morales, Joselyn Katherine Cárdenas García de la Universidad Central del Ecuador en la revista Economía y Política número 32 del 2020 publican su artículo ***"Determinantes del rendimiento académico de educación media en Ecuador",*** este trabajo se realizó analizando determinantes demográficas, fisiológicas, socioeconómicas, culturales, académicos y psicológicos que inciden en el rendimiento académico de los estudiantes que rindieron el examen Ser bachiller en el periodo lectivo 2016-2017. Para observar el efecto que tienen los mencionados factores en el rendimiento académico, se constituye un modelo econométrico de regresión logística que muestra la magnitud y la relación de las principales variables que determinan el rendimiento académico siendo así que el costo de preparación previo a rendir el examen es (11.42% mayor) y la manifestación, por parte del estudiante de tener una percepción de control interno y responsabilidad personal (10.61% mayor). (Villarruel R., 2020)

Fernando Fajando Bullón, María Maestre Campos, Elena Felipe Castaño, Benito León del Barco, María Isabel Polo del Río, de la Universidad de Extremadura de Mérida – España, en la revista Educación XX1 (Universidad Nacional de Educación a Distancia) del 2017 desarrollaron ***"Análisis del rendimiento Académico de los alumnos de Educación secundaria obligatoria según las variables familiares"*** este trabajo se efectuó

tomando una muestra constituida de 486 alumnos con edades de 12 a 18 años de Enseñanza Secundaria Obligatoria de la ciudad de Cáceres del periodo 2011-20012, se analizaron las variables nivel de estudio y clase ocupacional de los padres, ayuda recibida por parte de algún familiar o persona cercana y autopercepción familiar, como variables determinantes en el rendimiento académico de los alumnos. Se obtuvieron diferencias significativas en el rendimiento académico en función de la formación académica de los padres (F=35.24) y madres (F=38.3), rendimiento académico en función de si se recibe o no ayuda con las tareas (t=2.423) y la percepción que consideran los alumnos que tienen sus familias sobre su valía como estudiantes (F=59.800). Concluye que unas formaciones académicas elevadas de los padres, así como su pertenencia a las clases ocupacionales medias o privilegiadas son predictores de un buen rendimiento académico en sus hijos. (Fajardo Bullón, 2017)

Nancy Dávila, María Dolores García, José María Pérez, Emilio Gómez de la Universidad de Las Palmas de Gran Canaria – España, en la revista de Investigación Educativa 2015, publican su investigación ***"An Asymmetric Logit Model to explain the likelihood of success in academic results",*** se realizó un cuestionario a una población de 569 estudiantes matriculados en matemáticas del grado de Administración y Dirección de empresas (GADE) en el curso académico 2011-2012, con la finalidad de detectar que factores influyen en la probabilidad de éxito en la asignatura. Se aplicó el modelo de regresión logística clásico y un modelo de regresión logística asimétrico Bayesiano. Los resultados concluyen que las variables significativas que podrían determinar el rendimiento académico en términos de la probabilidad relativa de aprobar la asignatura son: la asistencia con regularidad a clases de teoría y prácticas, que el estudiante valore positivamente el material del que dispone para el seguimiento de la asignatura, el tipo de centro en que se cursaron los estudios preuniversitarios y la asistencia a clases de apoyo. (Davila N., 2015)

Sharmin Sharker, Mujibur Rahman, del Departamento de Estadística de la Universidad de Agricultura empresarial y tecnología Dhaka – Bangladesh 2015 realizaron ***" Determinants of academic Performance- A Multinomial Logistic Regression Approach",*** donde se realizó una encuesta a 140 estudiantes graduados de seis

departamentos diferentes y se tomó como factores de estudio el género, universidad de estudio, pre-admisión a escuela secundaria, nivel de asistencia, estado de libertad condicional, tiempo que pasa en el estudio, la educación del padre y participación del estudiante. Se utilizó la regresión logística multinomial para determinan la calificación de un estudiante en el primer semestre de la carrera de pregrado. Los resultados determinaron que la educación del padre, trabajo a tiempo parcial no parece contribuir. Es así que la calificación actual del estudiante depende completamente en la calificación de su primer semestre. (Sharmin S., 2015)

Bereket Tessema, Kidus Meskele Ashine, Wolaita Sodo de la Unicersidad Wolaita Sodo – Entiopía 2012, desarollaron ***"Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students' Academic Achievement: The Case of College of Natural and Computational Science, Wolaita Sodo University, Ethiopia",*** donde se pretendió identificar las principales variables que influyen en el rendimiento académico de los estudiantes, se utilizó el software SPSS versión 22.0 para el análisis teniendo como conclusiones que la variable nivel de educación del padre, el tiempo de estudio organizado y la cantidad de dinero que reciben las familias son variables predictoras significativas. (Tessema B., Meskele K., Sodo W., 2016)

Jorge Hernández Uralde, Alejandro Márquez Jiménez, Joaquina Palomar Lever, investigadores de la ciudad de México en la revista Mexicana de Investigación Educativa volumen 11, número 29 del 2006 publican su artículo ***"Factores asociados con el desempeño académico en el EXANI-1. Zona metropolitana dela ciudad de México 1996-2000",*** esta investigación fue realizada para conocer el grado en que cinco generaciones que han presentado el Examen Nacional de Ingreso a la Educación Media Superior(EXANI-1) en la zona metropolitana de la ciudad de México han variado su desempeño en él, así como la influencia de diversas variables socioeconómicas y educativas; trabajó con datos correspondientes al total de la población evaluada entre 229 y 260 mil individuos utilizando básicamente herramientas de estadística descriptiva y análisis de correlación. Obteniendo como resultados que las variables socioeconómicas, factor cultural, factor de recursos e infraestructura tienen relación con los niveles de desempeño académico (Hernandez J., 2006).

Capítulo 3

En este capítulo se detalla los modelos de regresión logística multinomial que son utilizados tanto con variables dependientes del tipo nominal como del tipo ordinal.

3. Modelos de Respuesta Múltiple

Con una variable aleatoria de respuesta politómica Y, con categorías $Y_1, Y_2, \ldots, Y_S, S > 2$ y un conjunto de variables explicativas $X = (X_1, X_2, \ldots, X_R)'$, los modelos de respuesta múltiple tiene como objetivo explicar la respuesta a partir de las variables explicativas (Agresti, 2002).

Si se dispone de:

- Una muestra aleatoria de tamaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, X_2, \ldots, X_R, x_q = (x_{q1}, \ldots, x_{qR})' \; \forall q = 1, \ldots, Q$;
- Una muestra aleatoria de n_q observaciones independientes de las variables de respuesta politómica Y para cada x_q, $(y_{q1}, \ldots, y_{qS})' \forall q = 1, \ldots, Q$ con y_{qS} el número de observaciones que caen en la categoría de respuesta $Y_s \; \forall s = 1, \ldots, S$.

De tal manera que los vectores tienen distribuciones de probabilidad multinomiales independientes,

$$(y_{q1}, \ldots, y_{qS})' \rightarrow M(n_q; p_1, \ldots, p_S)$$

Verificando que:

- $\sum_{s=1}^{S} p_s = 1$, donde $p_s = (Y = Y_s | X = x_q)$
- $\sum_{s=1}^{S} y_s = n_q$ y $\sum_{q=1}^{Q} n_q = N$

Según si la respuesta tiene carácter ordinal o nominal, se pueden formular distintos modelos de respuesta múltiple.

3.1 Formulación de los modelos

3.1.1 Modelos logit de Respuesta Nominal

Al tener una variable de respuesta nominal el modelo logit se formula en base a las *transformaciones logit generalizadas* definidas con respecto a una categoría de referencia:

$$L_s(x) = ln\left[\frac{p_s(x)}{p_S(x)}\right], \quad \forall t, s = 1, \dots, S-1,$$

Se ha tomado como categoría de referencia la última aunque en la aplicación los programas R suelen tomar la primera; para cualquier transformación logit para un par de categorías se puede obtener a partir de sus transformaciones logit generalizadas asociadas en la forma (Agresti, 2002):

$$ln\left[\frac{p_t(x)}{p_s(x)}\right] = L_t(x) - L_s(x) \quad \forall t, s$$

Dadas las R variables explicativas $X_1, \dots, X_R$ y las Q combinaciones diferentes de valores de las variables explicativas $x_q = (x_{q1}, \dots\dots, x_{qR})' \; \forall q = 1, \dots, Q$; el modelo es de la forma (Agresti, 2002):

$$L_s(x_q) = \sum_{r=0}^{R} \beta_{rs}\, x_{qr} = x'_q \beta_s \qquad \forall s = 1, \dots, S-1, \qquad q = 1, \dots, Q$$

Con $x_{q0} = 1$, y $\beta_s = (\beta_{0s}, \beta_{1s}, \dots, \beta_{Rs})'$ el vector de parámetros asociados a la categoría Y_s.

Equivalentemente, el modelo en términos de las probabilidades de respuesta:

$$p_s(x_q) = \frac{exp(\sum_{r=0}^{R} \beta_{rs}\, x_{qr})}{1+\sum_{s=1}^{S} exp(\sum_{r=0}^{R} \beta_{rs}\, x_{qr})} \qquad \forall s = 1, \dots, S \qquad (1)$$

Definiendo $\beta_{rs} = 0 \;\; \forall s = 1, \dots, S-1$ (Agresti, 2002).

La exponencial de los parámetros β_s asociados a cada categoría de respuesta se interpreta como el cociente de ventajas de respuesta Y_s frente a la última Y_S cuando se

incrementa en una unidad la variable asociada al parámetro y el resto permanece inalteradas (Agresti, 2002).

Cuando alguna de las variables explicativas es categórica y se introduce en el modelo mediante sus variables del diseño asociadas (Agresti, 2002), la exponencial del parámetro asociado a una variable de diseño se interpreta como el cociente de ventajas de respuesta Y_s frente a la última Y_S para la categoría asociada al parámetro respecto de la primera (tomando codificación parcial con la primera como categoría de referencia).

3.1.2 Modelo logit de Respuesta Ordinal

En el caso de una variable de respuesta cualitativa ordinal se definen transformaciones logit que tienen en cuenta el orden entre las distintas categorías de respuesta. Se los conoce como modelos logit acumulativos y las transformaciones logit acumulativas se definen como (Agresti, 2002):

$$L_s(x_q) = \ln\left(\frac{P[Y \leq Y_s | X = x_q]}{1 - P[Y \leq Y_s | X = x_q]}\right) = ln\left(\frac{F(x_q)}{1 - F(x_q)}\right), \qquad \forall s = 1, \dots, S-1$$

Siendo F la función de distribución de la variable respuesta.

Existen diferentes formulaciones para un modelo de respuesta ordinal aunque el que aquí se explicará es el de efectos homogéneos o ventajas proporcionales (Agresti, 2002).

Dadas R variables explicativas $X_1, X_2, \dots, X_R$ y Q observaciones diferentes x_q el modelo logit acumulativo de efectos homogéneos es de la forma:

$$L_s(x) = \alpha_s + \sum_{r=0}^{R} \beta_r \, x_{qr} \; = x'_q \beta_s$$

Siendo $x_q = (x_{q1}, \dots\dots, x_{qR})'$ el vector de valores observados de las variables explicativas con $\beta = (\beta_0, \beta_1, \dots, \beta_R)'$ el vector de parámetros (Agresti, 2002).

En este modelo las exponenciales de los parámetros asociados a las variables explicativas se interpretan como el cociente de ventajas de categoría inferior a Y_s cuando se incrementa en una unidad la variable asociada al parámetro y las demás se controlan fijas.

Además de los modelos logit acumulativos de efectos homogéneos o ventajas proporcionales existen otras alternativas como los modelos logit para categorías adyacentes, o los modelos logit para categorías adyacentes de ventajas paralelas que no se explicaran aquí por no ser usados en la aplicación con datos reales y que se pueden ver en (Agresti, 2002).

3.2 Estimación por máxima verosimilitud

Independientemente de si se trata del modelo de respuesta nominal u ordinal, la estimación de parámetros se realiza mediante la estimación por máxima verosimilitud. Dada una muestra aleatoria de tamaño N con Q combinaciones diferentes de valores de las variables explicativas $X_1, \dots, X_R$, $x_q = (x_{q0}, x_{q1}, \dots, x_{qR})'$ con $x_{q0} = 1 \; \forall q = 1, \dots, Q$; y una muestra aleatoria de n_q observaciones independientes de la variable de respuesta politómica Y,$(y_{q1}, \dots, y_{qS})' \forall q = 1, \dots, Q$, la función de verosimilitud de los datos es entonces (Agresti, 2002):

$$\prod_{q=1}^{Q}\left(\frac{n_q!}{\prod_{s=1}^{S}(y_{qs})!}\prod_{s=1}^{S}P_{qs}^{y_{qs}}\right)$$

El núcleo de la log-verosimilitud esta dado:

$$K = \sum_{q=1}^{Q}\sum_{s=1}^{S} y_{qs}\ln(p_s)$$

Derivando con respecto a los parámetros en cada modelo se tienen las ecuaciones de verosimilitud cuya resolución por el método de Newton-Raphson proporciona las estimaciones de los parámetros (Jobson, 1991).

3.3 Significación de parámetros

La matriz de covarianzas de los estimadores de los parámetros de los distintos modelos aquí abordados $\hat{\beta}$ es la inversa de la matriz de información de Fisher dada por las derivadas parciales de segundo orden del núcleo de la verosimilitud. (Agresti, 2002)

$$Cov(\hat{\beta}) = \left[-E\left(\frac{\partial^2 K}{\partial\beta_j \partial\beta_k}\right)\right]^{-1}$$

$$Cov(\hat{\beta}) = \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1, \hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_h) \\ Cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_h) \\ \cdots & \cdots & \cdots & \cdots \\ Cov(\hat{\beta}_1, \hat{\beta}_h) & Cov(\hat{\beta}_2, \hat{\beta}_h) & \cdots & Var(\hat{\beta}_h) \end{pmatrix}$$

Las raíces cuadradas de los elementos de la diagonal de esta matriz son los errores estándar (ASE) de los estimadores de los parámetros del modelo.

Como estimadores de máxima verosimilitud tienen distribución normal asintótica (Jobson, 1991).

$$\hat{\beta} \xrightarrow[N\to\infty]{d} N\left(\beta, Cov(\hat{\beta})\right)$$

Todos los programas que ajustan modelos de respuesta ordinal proporcionan la estimación de los parámetros y de sus errores estándar de estimación, con lo que es fácil a partir de la distribución normal asintótica, estudiar la significación de parámetros (Hosmer, 2000).

Donde el test de significación de parámetros está dado por:

$$H_0: \beta_j = 0 \quad vs \quad H_1: \beta_j \neq 0$$

Y su estadístico de contraste:

$$W = \frac{\beta_j}{S.E(\beta_j)}$$

3.4 Bondad de ajuste: Test chi-cuadrado de razón de verosimilitudes

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste de un modelo de regresión logística multinomial M se obtiene como menos dos

veces el logaritmo del cociente entre el supremo de la verosimilitud bajo la hipótesis nula y el supremo de la verosimilitud en la población, es decir (Jobson, 1991):

$$G^2(M) = 2\left[\sum_{q=1}^{Q}\sum_{s=1}^{S} y_{qs}\, ln\left(\frac{y_{qs}}{p_{qs}}\right)\right]$$

Este estadístico tienen distribución asintótica chi-cuadrado con grados de libertad mismos que son obtenidos mediante la diferencia entre la dimensión del espacio paramétrico y la dimensión de este espacio bajo la hipótesis nula. Para un modelo de regresión logística multinomial los grados de libertad es la diferencia entre el número de parámetros p_{qs} y el número de parámetros β_{rs} bajo el modelo, es decir $Q - (N-1) * (S-1)$ grados de libertad (Hosmer, 2000).

$$G^2(M) \underset{d_q}{\overset{d}{\rightarrow}} X^2_{Q-(N+1)*(S-1)}, si\ d_q \rightarrow \infty$$

Así que se rechaza la hipótesis nula con un nivel de significación α cuando $G^2(M)_{Obs} \geq X^2_{Q-(N+1)*(S-1);\alpha}$. O equivalentemente cuando $\text{p} - \text{valor} = \text{p}[G^2(M) \geq G^2(M)_{Obs}] \leq \alpha$.

Al estadístico de Wilks, $G^2(M)$ se lo llama devianza. La teoría asintótica se aplica de forma más natural a modelos logísticos con variables explicativas categóricas (Jobson, 1991).

El contraste se puede expresar como:

$$H_0: p(x_q) = \frac{e^{(\sum_{r=0}^{R}\beta_{rs}x_r)}}{1 + \sum_{s=2}^{S} e^{(\sum_{r=0}^{R}\beta_{rs}x_r)}}$$

$$H_1: p(x_q) \neq \frac{e^{(\sum_{r=0}^{R}\beta_{rs}x_r)}}{1 + \sum_{s=2}^{S} e^{(\sum_{r=0}^{R}\beta_{rs}x_r)}}$$

3.5 Medidas globales de bondad de ajuste

Para cuantificar la bondad del ajuste global del modelo se dispone de medidas como la *Tasa de clasificaciones correctas* (CCR).

3.5.1 Tasa de clasificaciones correctas

Para cuantificar la bondad del ajuste global del modelo se dispone también de otra medida como es la tasa de clasificaciones correctas. Es decir a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación observados-predichos y se utiliza el porcentaje de clasificaciones correctas como una medida de la calidad de predicción, del mismo modo que se hace en el análisis discriminante (Pando Fernández V., 2004). Se define como la proporción de individuos clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N. Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta Y_s coincide con su valor estimado por el modelo.

3.6 Significación de variables. Contrastes condicionales de razón de verosimilitudes

Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las variables frente al modelo completo. La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (es decir, da igual su presencia o su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo (Hosmer, 2000, pág. 31).

Al suponer que tenemos un modelo de regresión logística multinomial M_G que se ajusta bien y se desea contrastar si un subconjunto de parámetros, $\beta = (\beta_1, \dots, \beta_r)$ son nulos. Sea M_P el modelo con ese subconjunto de parámetros ceros.
Así que M_P esta anidado en el modelo general M_G. Así que planteamos el contraste:

$$H_0: \beta = 0 \quad (M_P \; se \; verifica)$$

$$H_1: \beta \neq 0 \quad (asumiendo \; cierto \; M_G)$$

Si asumimos que M_G se verifica, el estadístico del test de razón de verosimilitudes para contrastar si M_P se verifica es: $G^2(M_P|M_G) = -2(L_P - L_G) = G^2(M_P) - G^2(M_G)$, siendo

L_P y L_G los máximos de la log-verosimilitud bajo la suposición de que se verifican los modelos saturados, M_P y M_G, respectivamente. Es decir, el test de razón de verosimilud para contrastar dos modelos anidados es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo. Al rechazar la hipótesis nula se concluye que al menos una de las variables independientes contribuye significativamente al modelo (Jobson, 1991).

El estadístico $G^2(M_P|M_G)$ tienen distribución chi-cuadrado con grados de libertad la diferencia entre los grados de libertad de las distribuciones chi-cuadrado asintóticas de $G^2(M_P)$ y $G^2(M_G)$, es decir el número de parámetros que se anulan para H_0, r (Hosmer, 2000).

Así que se rechaza la hipótesis nula al nivel de significación α cuando $G^2_{obs}(M_P|M_G) \geq X^2_r; \alpha$.

3.7 Selección de modelos logit - Stepwise

La selección Stepwise es el método que combina los modelos adelante y atrás, el cual se puede empezar por el modelo vacío o el completo, pero en cada paso se examinan las variables incluidas, si deben salir y las no seleccionadas, si deben ingresar. El método stepwise, está basado en contrastes condicionales de razón de verosimilitudes (Hosmer, 2000).

Partiendo de un modelo vacío, sólo con la constante, el método consiste en partir de este modelo inicial, de modo que en cada paso se ajustarán todos aquellos modelos que resulten de incluir cada una de las variables explicativas que no están en el modelo seleccionado en el paso anterior. Llevándose así acabo contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis alternativa el modelo resultante de la inclusión de cada variable. De modo que se seleccionarán las variables para las que el constaste sea significativo, y se incluirá en modelo aquella variable asociada al mínimo $p - valor$ de entre todos los menores o iguales que $\alpha 1$. La inserción de variables mediante este método continúa hasta que ninguno de estos contrastes condicionales sea significativo (Jobson, 1991).

Seguidamente se considera en cada paso la posibilidad de eliminar alguno de los parámetros del modelo seleccionado en el método hacia atrás. Pero no se puede eliminar en un paso la variable que acaba de entrar en el paso anterior, por lo que se fijará para la eliminación de variables un nivel de significación $\alpha 2$ mayor que $\alpha 1$ (Jobson, 1991).

De la misma manera, para la eliminación de variables se realizarán contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el modelo que resulta de la eliminación de cada una de las variables y en la hipótesis alternativa el modelo seleccionado en el paso anterior. Así, las variables candidatas a eliminar serán aquellas cuyo $p - valor$ sea mayor que $\alpha 2$ y se eliminara la variable asociada con el mayor $p - valor$ de estos. La eliminación de variables continúa hasta que todos estos contrastes condicionales resulten significativos (Agresti, 2002).

El procedimiento *stepwise* continuará hasta llegar a un paso en el que ninguno de los contrastes condicionales de introducción de variables sea significativos y todos los de eliminación de variables sean significativos (Agresti, 2002).

3.8 Modelos con variables explicativas categóricas: variables del diseño

Asociadas a una variable cualitativa A con categorías denotadas por $A_i (i = 1, \dots, I)$, se define un total de $(I - 1)$ variables del diseño o variables ficticias. La razón para definir una variable ficticia menos que el número de categorías es que la matriz de diseño resultante del modelo de regresión logística múltiple correspondiente sea invertible (tenga columnas linealmente independientes) (Agresti, 2002). A continuación se presentan distintas formas de codificación de las variables del diseño.

3.8.1 Método parcial

Recibe también el nombre de codificación respecto a un grupo de referencia. Consiste en elegir una categoría de referencia, de modo que todas las variables del diseño asignan el valor 0 a dicha categoría de referencia. Asociada a cada una de las restantes categorías se define una variable de diseño que toma el valor de 1 para su categoría asociada y 0 para todas las demás (Agresti, 2002).

De acuerdo con lo expuesto anteriormente, la *m-ésima* variable de diseño va asociada con la categoría A_m, y se define en la siguiente forma:

$$X_{im}^{A} = X_{m}^{A}(A = A_i) = \begin{cases} 1 \; i = m \\ 0 \; i \neq m \end{cases} \quad \forall \, m = 2, \dots, I; i = 1, \dots, I.$$

3.8.2 Método marginal

Se llama también codificación mediante desviación respecto a la media. De nuevo se toma una categoría de referencia a la que todas las variables del diseño asignan el mismo valor que en este caso es -1.

Por lo tanto, la *m-ésima* variable del diseño está asociada con la categoría A_m y se define como:

$$X_{im}^{A} = X_{m}^{A}(A = A_i) = \begin{cases} 1 & i = m \\ -1 & i = 1 \\ 0 & i \neq m, 1 \end{cases} \quad \forall \, m = 2, \dots, I; i = 1, \dots, I.$$

Cabe mencionar que el estudio realizado se ejecutó por el método parcial tomando la primera categoría como referencia.

Capítulo 4

Análisis nivel de logro alcanzado en la nota del examen de grado ser bachiller en función de variables sociodemográficas. Modelos de regresión multinomial

El trabajo pretende conocer si está relacionado el nivel de logro alcanzado por los sustentantes en el examen Ser bachiller con las variables sociodemográficas, ya que conociendo este aspectos se genera mejores soluciones a las condiciones en las que se encuentren los sustentantes.

En este capítulo se desarrolla el ajuste del modelo de Regresión de respuesta ordinal y nominal respectivamente con variables explicativas cuantitativas y cualitativas utilizando el software libre R, comparando sus resultados en el nivel de logro alcanzado en la prueba Ser Bachiller en función de las variables sociodemográficas en el Ecuador en el periodo lectivo 2018-2019,

4.1 Preparación de los datos de estudio

En el Ecuador la educación está reglamentada por el Ministerio de educación y la prueba "Ser Bachiller" está en base a estándares aprobados por la misma institución para medir el desempeño de los estudiantes con la finalidad de mejorar la calidad de la educación y también ayuda a establecer si los estudiantes de tercero de Bachillerato General Unificado BGU pueden postularse a una carrera universitaria.

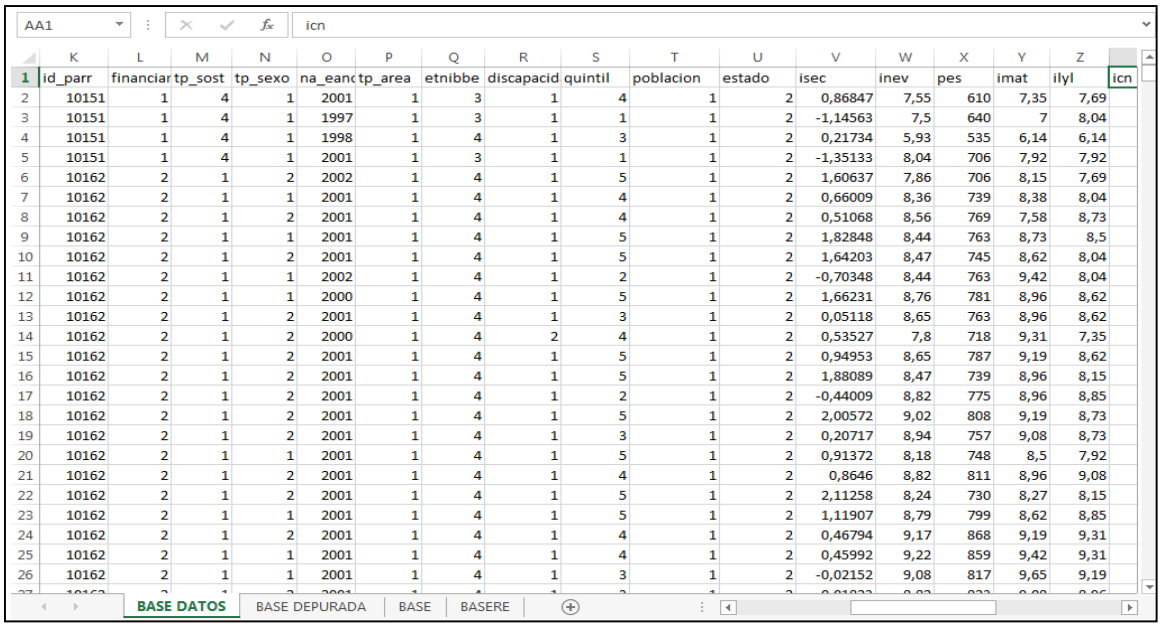
Para conocer el nivel de logro alcanzado, se analizará distintas variables sociodemográficas (edad, sexo, autoidentificación étnica, región natural y tipo de financiamiento de la institución educativa). Con las variables planteadas se desea explicar aquellos factores que mayormente influyan en el estudio.

Los datos a utilizar corresponden a la base de datos Ser Bachiller del Instituto Nacional de Evaluación Educativa del Ecuador correspondiente al año lectivo 2018-2019 proporcionado por su página web (http://evaluaciones.evaluacion.gob.ec/BI/bases-de-datos-ser-bachiller/).

Los archivos se encuentran divididos en diccionario de variables especificando (instituciones educativas, distrito, circuito, provincia, cantón, parroquia) y la base de datos (sin etiqueta inicial de fila y el resto de filas contiene información registrada de cada sujeto perteneciente a la muestra). En columnas se encuentra la información de cada una de las variables recogidas mismas que están codificadas.

Inicialmente la muestra se encuentra conformada de 514852 personas y 34 variables.

Los datos se encuentran en un fichero *.xlsx* de Excel.



AA1 | icn

	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
1	id_parr	financiar	tp_sost	tp_sexo	na_eanc	tp_area	etnibbe	discapacid	quintil	poblacion	estado	isec	inev	pes	imat	ilyl	icn
2	10151	1	4	1	2001	1	3	1	4	1	2	0,86847	7,55	610	7,35	7,69	
3	10151	1	4	1	1997	1	3	1	1	1	2	-1,14563	7,5	640	7	8,04	
4	10151	1	4	1	1998	1	4	1	3	1	2	0,21734	5,93	535	6,14	6,14	
5	10151	1	4	1	2001	1	3	1	1	1	2	-1,35133	8,04	706	7,92	7,92	
6	10162	2	1	2	2002	1	4	1	5	1	2	1,60637	7,86	706	8,15	7,69	
7	10162	2	1	1	2001	1	4	1	4	1	2	0,66009	8,36	739	8,38	8,04	
8	10162	2	1	2	2001	1	4	1	4	1	2	0,51068	8,56	769	7,58	8,73	
9	10162	2	1	1	2001	1	4	1	5	1	2	1,82848	8,44	763	8,73	8,5	
10	10162	2	1	2	2001	1	4	1	5	1	2	1,64203	8,47	745	8,62	8,04	
11	10162	2	1	1	2002	1	4	1	2	1	2	-0,70348	8,44	763	9,42	8,04	
12	10162	2	1	1	2000	1	4	1	5	1	2	1,66231	8,76	781	8,96	8,62	
13	10162	2	1	2	2001	1	4	1	3	1	2	0,05118	8,65	763	8,96	8,62	
14	10162	2	1	2	2000	1	4	2	4	1	2	0,53527	7,8	718	9,31	7,35	
15	10162	2	1	2	2001	1	4	1	5	1	2	0,94953	8,65	787	9,19	8,62	
16	10162	2	1	2	2001	1	4	1	5	1	2	1,88089	8,47	739	8,96	8,15	
17	10162	2	1	2	2001	1	4	1	2	1	2	-0,44009	8,82	775	8,96	8,85	
18	10162	2	1	2	2001	1	4	1	5	1	2	2,00572	9,02	808	9,19	8,73	
19	10162	2	1	2	2001	1	4	1	3	1	2	0,20717	8,94	757	9,08	8,73	
20	10162	2	1	1	2001	1	4	1	5	1	2	0,91372	8,18	748	8,5	7,92	
21	10162	2	1	2	2001	1	4	1	4	1	2	0,8646	8,82	811	8,96	9,08	
22	10162	2	1	2	2001	1	4	1	5	1	2	2,11258	8,24	730	8,27	8,15	
23	10162	2	1	1	2001	1	4	1	5	1	2	1,11907	8,79	799	8,62	8,85	
24	10162	2	1	2	2001	1	4	1	4	1	2	0,46794	9,17	868	9,19	9,31	
25	10162	2	1	1	2001	1	4	1	4	1	2	0,45992	9,22	859	9,42	9,31	
26	10162	2	1	1	2001	1	4	1	3	1	2	-0,02152	9,08	817	9,65	9,19	

BASE DATOS | BASE DEPURADA | BASE | BASERE

Figura 4.1: Base de datos Ser Bachiller 2018-2019

Se procede a filtrar la base considerando las variables para el estudio y eliminando las respuestas vacías, con la base establecida se procede hacer el cambio de nominación tomando su valor numérico por su nombre respectivo teniendo así un total de 275498 individuos y 6 variables (Ver Figura 4.1).

El estudio se realizó con base en datos agrupados mismos que constan con 6000 filas con 7 columnas, tomando la frecuencia de cada una de las variables (Ver Figura 4.2).

	edad	tp_sexo	financiamiento	etnibbe	nm_regi	nl_inev	Freq
1	16	hombre	publico	mestizo/blanco	costa	insuficiente	3
2	17	hombre	publico	mestizo/blanco	costa	insuficiente	655
3	18	hombre	publico	mestizo/blanco	costa	insuficiente	6152
4	19	hombre	publico	mestizo/blanco	costa	insuficiente	4150
5	20	hombre	publico	mestizo/blanco	costa	insuficiente	2253
6	21	hombre	publico	mestizo/blanco	costa	insuficiente	910
7	22	hombre	publico	mestizo/blanco	costa	insuficiente	577
8	23	hombre	publico	mestizo/blanco	costa	insuficiente	404
9	24	hombre	publico	mestizo/blanco	costa	insuficiente	270
10	25	hombre	publico	mestizo/blanco	costa	insuficiente	238
11	16	mujer	publico	mestizo/blanco	costa	insuficiente	15
12	17	mujer	publico	mestizo/blanco	costa	insuficiente	737
13	18	mujer	publico	mestizo/blanco	costa	insuficiente	6219
14	19	mujer	publico	mestizo/blanco	costa	insuficiente	3855
15	20	mujer	publico	mestizo/blanco	costa	insuficiente	1665
16	21	mujer	publico	mestizo/blanco	costa	insuficiente	834

Showing 1 to 17 of 6,000 entries, 7 total columns

Figura 4.2: Datos agrupados Ser Bachiller 2018-2019

Donde se tiene:

Variable dependiente nivel de logro alcanzado siendo: 0=insuficiente, 1=elemental, 2=satisfactorio y 3=excelente.

Variables Independientes que ofrecen información sociodemográfica del individuo: edad, sexo, autoidentificación étnica, provincia y tipo de financiamiento de las instituciones educativas detallamos a continuación:

EDAD: De 16 - 25 años

SEXO: Se compone en dos categorías:

- 1: Mujer
- 2: Hombre

AUTOIDENTIFICACIÓN ÉTNICA: Se compone en cinco categorías:

- 1: Afroecuatoriano
- 2: Montubio

- 3: Indígena
- 4: Mestizo/ Blanco
- 5: Otro

REGIONES NATURALES: Se compone de cinco categorías:

- 1: Costa
- 2: Sierra
- 3: Oriente
- 4: Insular
- 90: Otras (Zona No Delimitada)

TIPO DE FINANCIAMIENTO DE LA INSTITUCIÓN EDUCATIVA: Se compone de tres categorías

- 1: Público (Fiscal y Municipal)
- 2: Privado (Particular)
- 3: Mixto (Fiscomisional)

NIVEL DE LOGRO ALCANZADO POR EL SUSTENTANTE: Se compone de cuatro categorías

- 0: Insuficiente
- 1: Elemental
- 2: Satisfactorio
- 3: Excelente

Dentro de la base tenemos variables categóricas las regiones naturales, el tipo de financiamiento, la autoidentificación étnica, el sexo y el nivel de logro alcanzado; una vez definidas todas las variables, el fichero se convierte en *.csv*.

La herramienta que se va a utilizar para los análisis es RStudio. El fichero en formato .csv puede leerse en R cargando la **librería readr** (Ver Anexo).

4.2 Análisis Unidimensional

Realizamos un análisis descriptivo de las variables para extraer información contenida dentro del conjunto de observaciones.

- ***Regiones Naturales***

Regiones Naturales	
Sierra	116519
Costa	142585
Oriente	15576
Insular	403
Otras	414
Total	275497

Tabla 1: Tamaño de la población por regiones

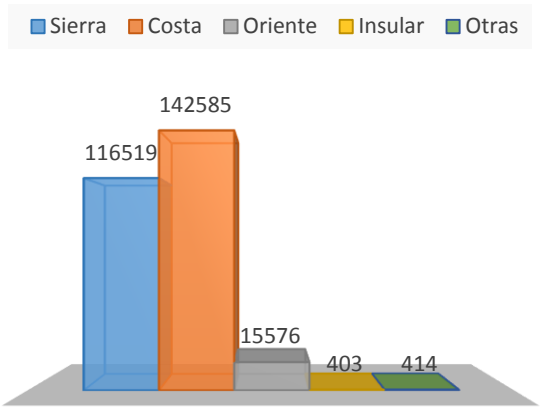


Figura 4.3: Distribución de Regiones

Podemos observar que la mayor cantidad de la población se encuentra en la región costa con 142585 sujetos correspondiéndole un 51.76%. Siendo la muestra total 275497 individuos correspondiéndole a la Sierra un 42.3%, Oriente un 5.65%, Otras y región que cada una tiene un 0.15%.

- ***Tipo de financiamiento de la entidad educativa***

Financiamiento entidad educativa	
Público	200008
Privado	52961
Mixto	22528
Total	275497

Tabla 2: Tamaño de la población por

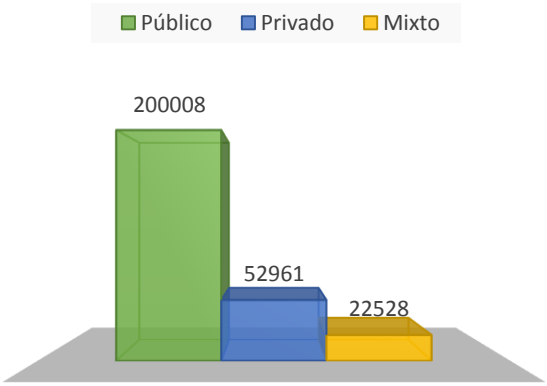


Figura 4.4: Distribución de Tipo de Financiamiento

Según el tipo de financiamiento podemos ver que los individuos de la población proviene del financiamiento público pues se tiene 200008 individuos, correspondiéndole al financiamiento mixto un 8.18%, al privado un 19.22% y al público 72.60%.

- ***Tipo de Sexo***

Tipo de Sexo	
Hombre	137118
Mujer	138379
Total	275497

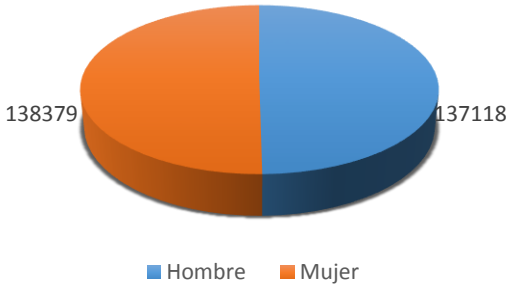


Tabla 3: Tamaño de la población por Tipo de sexo

Figura 4.5: Distribución de Tipo de sexo (Mujer, Hombre)

Podemos observar de la población total hay 138379 mujeres que corresponde al 50.23% y 137118 hombres que corresponde al 49.77%.

- ***Autoidentificación Étnica***

Autoidentificación Étnica	
Afroecuatoriano	11775
Indígena	15139
Montubio	11019
Mestizo/Blanco	236468
Otro	1096
Total	275497

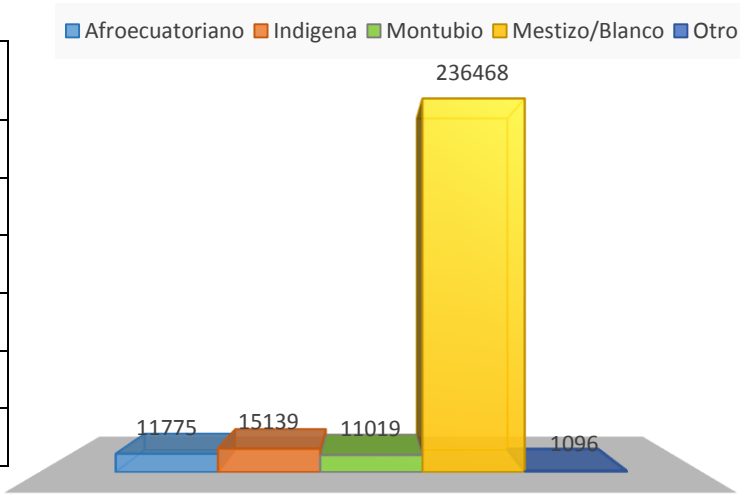


Tabla 4: Tamaño de la población por Autoidentificación étnica

Figura 4.6: Distribución de Autoidentificación étnica

Se puede observar que el tipo de autoidentificación étnica mayoritaria en este estudio es mestizo/blanco con 236468 individuos, mientras que para afroecuatoriano 4.27%, indígena le corresponde 5.50%, mestizo/blanco un 85.83%, montubio 4% y otro 0.40%.

- ***Nivel de logro alcanzado***

Nivel de logro alcanzado	
Insuficiente	56698
Elemental	116480
Satisfactorio	95603
Excelente	6716
Total	275497

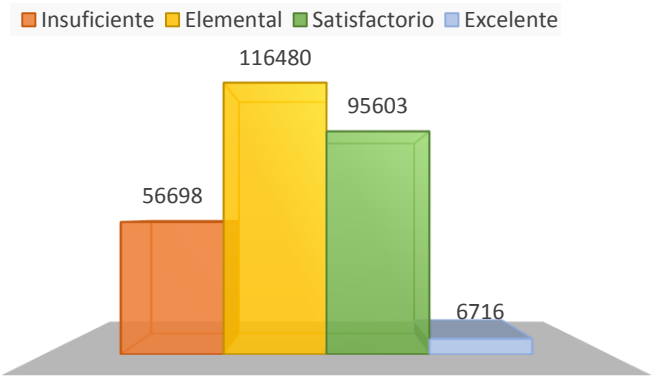


Tabla 5: Tamaño de la población por Logro alcanzado **Figura 4.7: Distribución del nivel de logro alcanzado**

El nivel de logro alcanzado mayormente alcanzado es el nivel elemental con 116480 que le corresponde el 42.28%, mientras al nivel insuficiente el 20.58%, al satisfactorio el 34.70% y al nivel excelente el 2.44%.

- ***Edad***

Edad	
16	287
17	26746
18	143610
19	54687
20	22520
21	10537
22	6593
23	4445
24	3231
25	2841
Total	275497

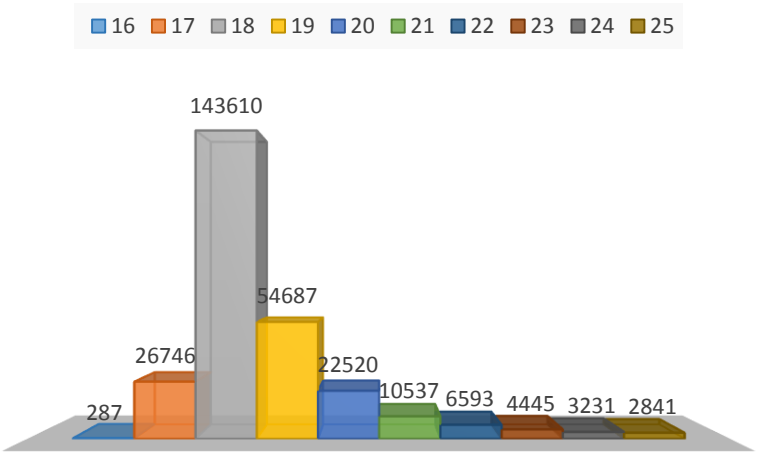


Tabla 6: Tamaño de la población por edad **Figura 4.8: Distribución de edad**

La edad mayoritaria de los individuos de la población corresponde a los 18 años y la minoritaria es 16 años con un 0.10%.

4.3 Análisis Bidimensional

La paradoja de Simpson "Cambio en el sentido de una asociación entre dos variables (numéricas o cualitativas) cuando se controla el efecto de una tercera variable".

A pesar de que sabemos que los análisis bidimensionales de variables (relación entre la variable respuesta y cada una de las variables explicativas) pueden ir en sentido contrario de lo que ocurra después con el ajuste de modelo de respuesta nominal o de respuesta ordinal, en esta sección se desea observar como la respuesta se relaciona con cada una de las posibles variables explicativas, aunque el análisis bidimensional puede ir en contra de lo que ocurra en el ajuste de los modelos.

El análisis bivariante permite descubrir relaciones existentes entre la variable dependiente y las variables independientes. En este caso como todas las variables independientes son categóricas, podremos realizar un contraste de chi-cuadrado que nos permitirá establecer si, existe o no relación entre ambas variables.

- ***Nivel de Logro alcanzado – Regiones Naturales***

Logro alcanzado - Regiones Nacionales					
	Costa	**Sierra**	**Otras**	**Oriente**	**Insular**
Insuficiente	42821	9973	102	3747	55
Elemental	60367	48038	203	7672	200
Satisfactorio	37285	54026	105	4044	143
Excelente	2112	4482	4	113	5

Tabla 7: Distribución de la población nivel de logro alcanzado – regiones naturales

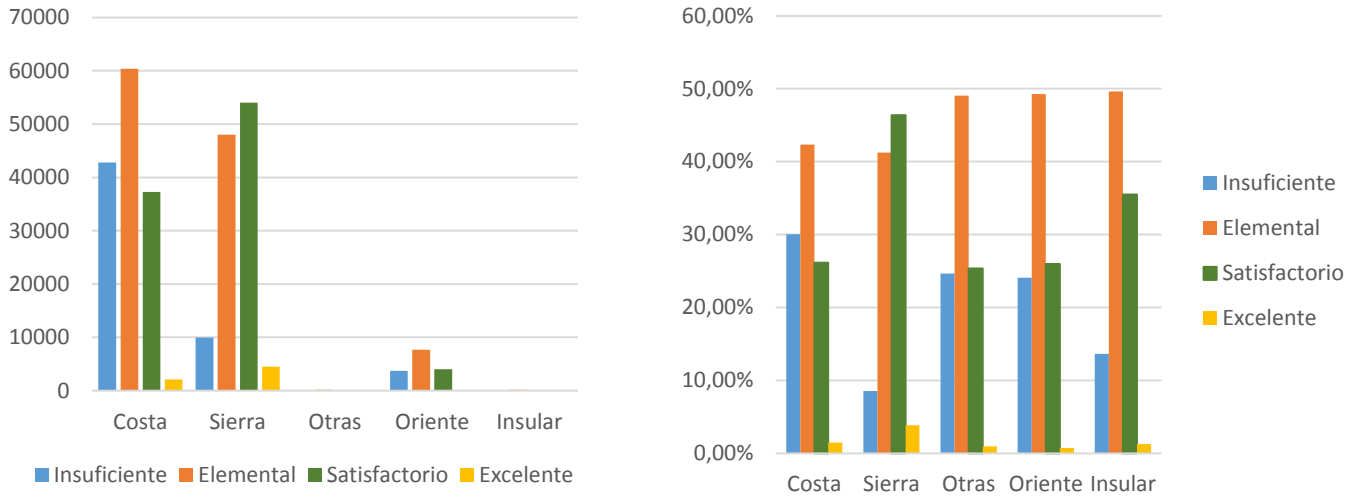


Figura 4.9: Frecuencias nivel de logro alcanzado – regiones naturales

En la tabla de frecuencias (Tabla 7) podemos ver la relación entre la variable dependiente **nivel de logro alcanzado** y la variable independiente **regiones naturales**; encontramos que el nivel de logro alcanzado en la región Costa un 30.03% está en la categoría insuficiente, 26.15% satisfactorio, 1.48% excelente y 42.34% elemental misma donde se encuentra la mayor cantidad de postulantes, en la región Sierra se tiene 8.46% es insuficiente, 41.23% elemental, 3.85% excelente y 46.37% satisfactorio, en la región Otras se tiene 24.64% es insuficiente, 25.36% satisfactorio, 0.97% excelente y 49.03% elemental, en la región Oriente se tiene 24.06% es insuficiente, 25.96% satisfactorio, 0.73% excelente y 49.26% elemental y en la región Insular se tiene 13.65% es insuficiente, 35.48% satisfactorio, 1.24% excelente y 49.63% elemental, dichos porcentajes son tomados por columna (Ver Figura 4.9).

Prueba Chi-Cuadrado

```
chisq.test(t_rg)


  Pearson's Chi-squared test


data:  t_rg
X-squared = 24299, df = 12, p-value < 2.2e-16
```

La prueba muestra la asociación entre las variables categóricas **nivel de logro alcanzado** (insuficiente, elemental, satisfactorio, excelente) y **regiones naturales** (costa, sierra, oriente, insular, otras) es altamente significativa, el p-valor obtenido es menor a 0.05 por lo que existe una relación de dependencia y se rechaza la hipótesis nula.

De la variable independiente regiones naturales fue seleccionada como categoría de referencia ***Costa***, misma que será utilizada posteriormente en el estudio aplicación de modelos de regresión logística.

- ***Nivel de Logro alcanzado – Tipo de financiamiento***

Logro alcanzado - Tipo de financiamiento			
	Público	**Privado**	**Mixto**
Insuficiente	47860	5069	3769
Elemental	90587	17479	8414
Satisfactorio	58203	27898	9502
Excelente	3358	2515	843

Tabla 8: Distribución de la población nivel de logro alcanzado – Tipo de financiamiento

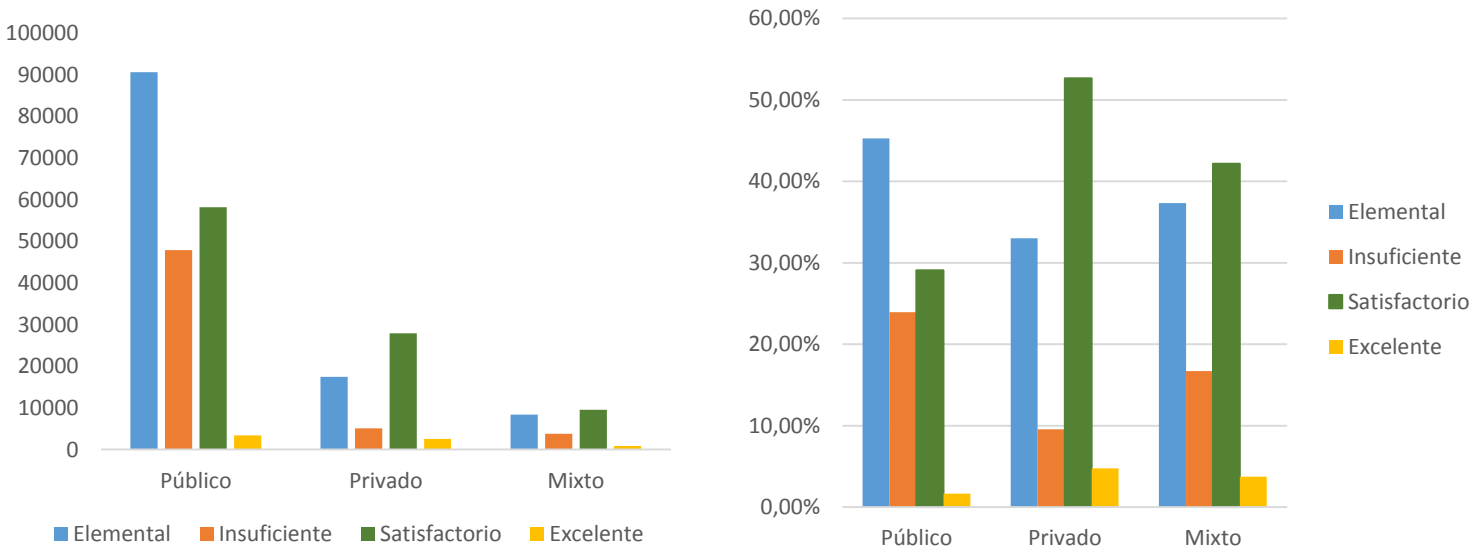


Figura 4.10: Frecuencias de nivel de logro alcanzado – tipo de financiamiento

La relación entre la variable independiente tipo de financiamiento y el nivel de logro alcanzado se encuentra: en el financiamiento Público con 23.93% insuficiente, 29.10% satisfactorio, 1.68% excelente y 45.29% elemental; en el financiamiento Privado 9.27% insuficiente, 33% elemental, 4.75% excelente y 52.68%; en el financiamiento Mixto 16.73% insuficiente, 37.35% elemental, 4.75% excelente y 42.18% satisfactorio. (Ver figura 4.10).

Prueba Chi-Cuadrado

```
chisq.test(t_f)


  Pearson's Chi-squared test


data:  t_f
X-squared = 14900, df = 6, p-value < 2.2e-16
```

En la prueba de chi-cuadrado se observa una relación de dependencia entre las variables categóricas **nivel de logro alcanzado** (insuficiente, elemental, satisfactorio, excelente) y **tipo de financiamiento** (público, privado, mixto), cuyo p-valor obtenido es menor a 0.05 por lo que existe una asociación entre variables.

La categoría de referencia tomada de la variable independiente tipo de financiamiento es ***Público*** que será utilizada posteriormente en la aplicación de modelos de regresión logística.

- ***Nivel de Logro alcanzado – Tipo de sexo***

Logro alcanzado - Tipo de sexo		
	Mujer	**Hombre**
Insuficiente	27232	29466
Elemental	56840	59640
Satisfactorio	50564	45039
Excelente	3743	2973

Tabla 9: Distribución de la población nivel de logro alcanzado – Tipo de sexo

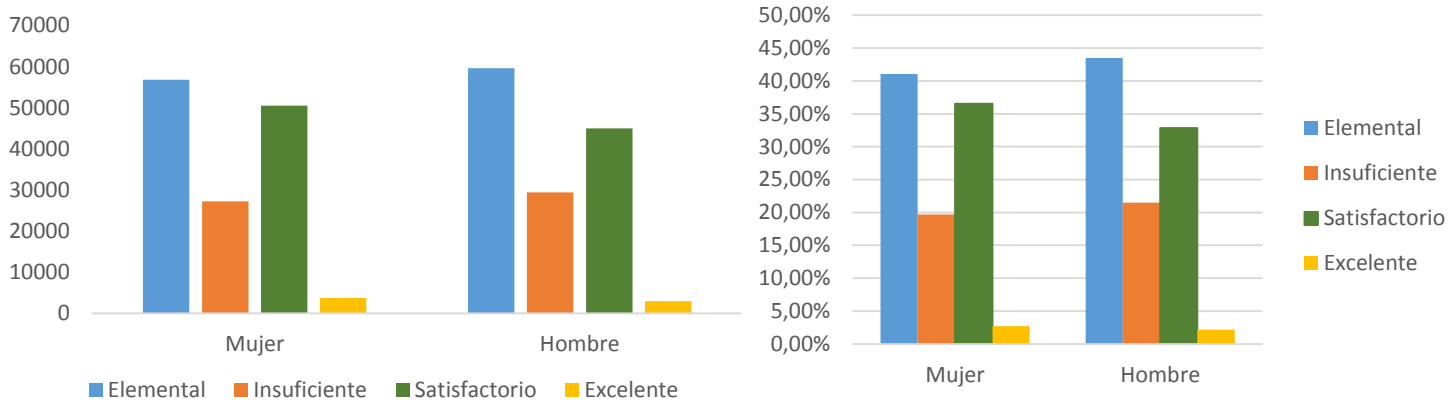


Figura 4.11: Frecuencias de nivel de logro alcanzado – tipo de sexo

La variable tipo de sexo con relación a la variable nivel de logro alcanzado se encontró que en Mujer el 19.68% es insuficiente, 41.08% elemental, 36.54% satisfactorio y 2.70% excelente; en Hombre el 21.49% insuficiente, 43.50% elemental, 32.85 satisfactorio y2.17% excelente. (Ver figura 4.11).

Prueba Chi-Cuadrado

```
chisq.test(t_s)


  Pearson's Chi-squared test


data:  t_s
X-squared = 557.15, df = 3, p-value < 2.2e-16
```

Realizando la prueba del chi-cuadrado entre las variables categóricas **nivel de logro alcanzado** (insuficiente, elemental, satisfactorio, excelente) y **tipo de sexo** (hombre, mujer) muestra que es altamente significativa y el p-valor obtenido es menor a 0.05 por lo que existe una relación de dependencia y se rechaza la hipótesis nula. A pesar que el test resulte significativo no se aprecian mayores diferencias entre la variable tipo de sexo, esto puede deberse al gran número de individuos de la población.

La categoría de referencia tomada de la variable independiente tipo de sexo es ***Mujer*** que será utilizada posteriormente en la aplicación de modelos de regresión logística.

- ***Logro alcanzado – Autoidentificación Étnica***

Logro alcanzado - Autoidentificación Étnica					
	Mestizo/Blanco	**Indígena**	**Afroecuatoriano**	**Otro**	**Montubio**
Insuficiente	45384	3472	4070	309	3463
Elemental	99137	7093	5250	479	4521
Satisfactorio	85615	4434	2369	299	2886
Excelente	6332	140	86	9	149

Tabla 10: Distribución de la población nivel de logro alcanzado – Autoidentificación étnica

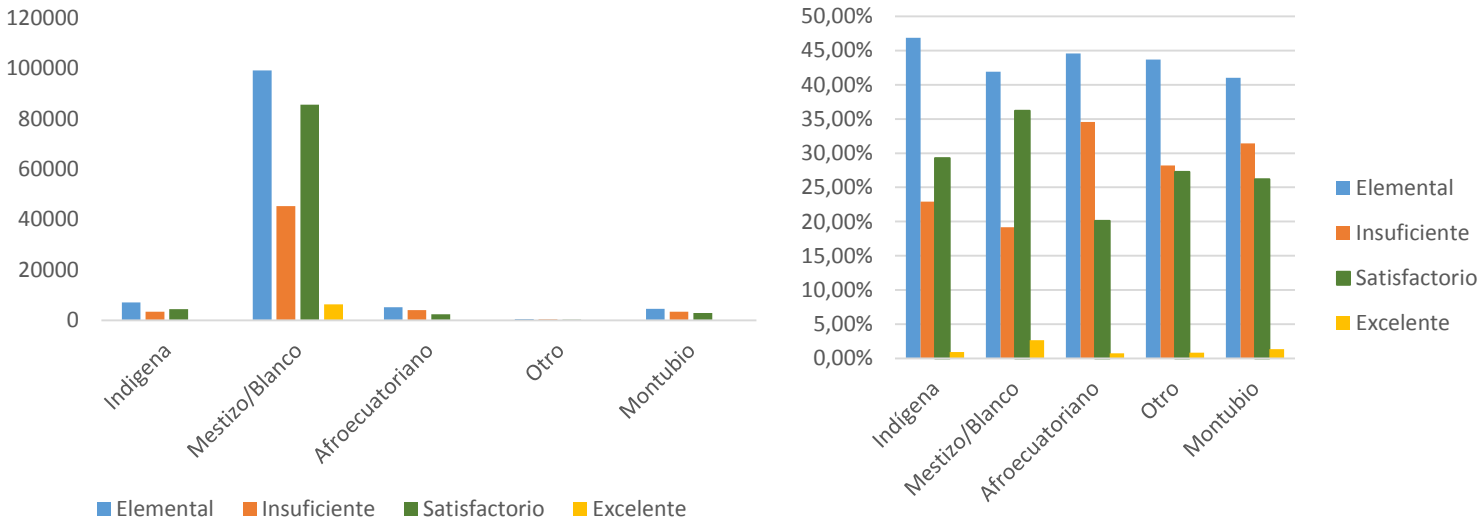


Figura 4.12: Frecuencias de autoidentificación étnica -nivel de logro alcanzado

La relación entre la variable independiente autoidentificación étnica y el nivel de logro alcanzado se encuentra: En autoidentificación étnica Mestizo/ Blanco 19.19% insuficiente, 36.21% satisfactorio, 2.68% satisfactorio y 41.92% elemental; en autoidentificación étnica Indígena el 22.93% insuficiente, 29.29% satisfactorio, 0.92% excelente y 46.85% elemental; en autoidentificación Afroamericano 34.56% insuficiente, 20.12% satisfactorio, 0.73 % excelente y 44.59% elemental; en autoidentificación étnica Otro 28.19% insuficiente, 26.19% satisfactorio, 0.82% excelente y 43.70% elemental; en autoidentificación étnica Montubio 31.43%

insuficiente, 26.19% satisfactorio, 1.35% excelente y 41.03% elemental(Ver figura 4.12).

Prueba Chi-Cuadrado

```
chisq.test(t_et)


  Pearson's Chi-squared test


data:  t_et
          X-squared = 3798, df = 12, p-value < 2.2e-16
```

Con la prueba del chi-cuadrado entre las variables categóricas **nivel de logro alcanzado** (insuficiente, elemental, satisfactorio, excelente) y **autoidentificación étnica** (indígena, mestizo/blanco, montubio, Afroecuatoriano, otro) se observa que son altamente significativa, el p-valor obtenido es menor a 0.05 por lo que existe una relación de dependencia y se rechaza la hipótesis nula.

La categoría de referencia tomada de la variable independiente autoidentificación étnica es ***Mestizo/Blanco*** que será utilizada posteriormente en la aplicación de modelos de regresión logística.

- ***Logro alcanzado – Edad***

Logro alcanzado - Edad										
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**
Insuficiente	47	2699	20450	14046	7858	4134	2801	1979	1422	1262
Elemental	113	9778	57224	25732	10785	4960	3024	2027	1516	1321
Satisfactorio	111	13039	61104	14397	3796	1418	760	433	291	254
Excelente	16	1230	4832	512	81	25	8	6	2	4

Tabla 11: Distribución de la población nivel de logro alcanzado – edad

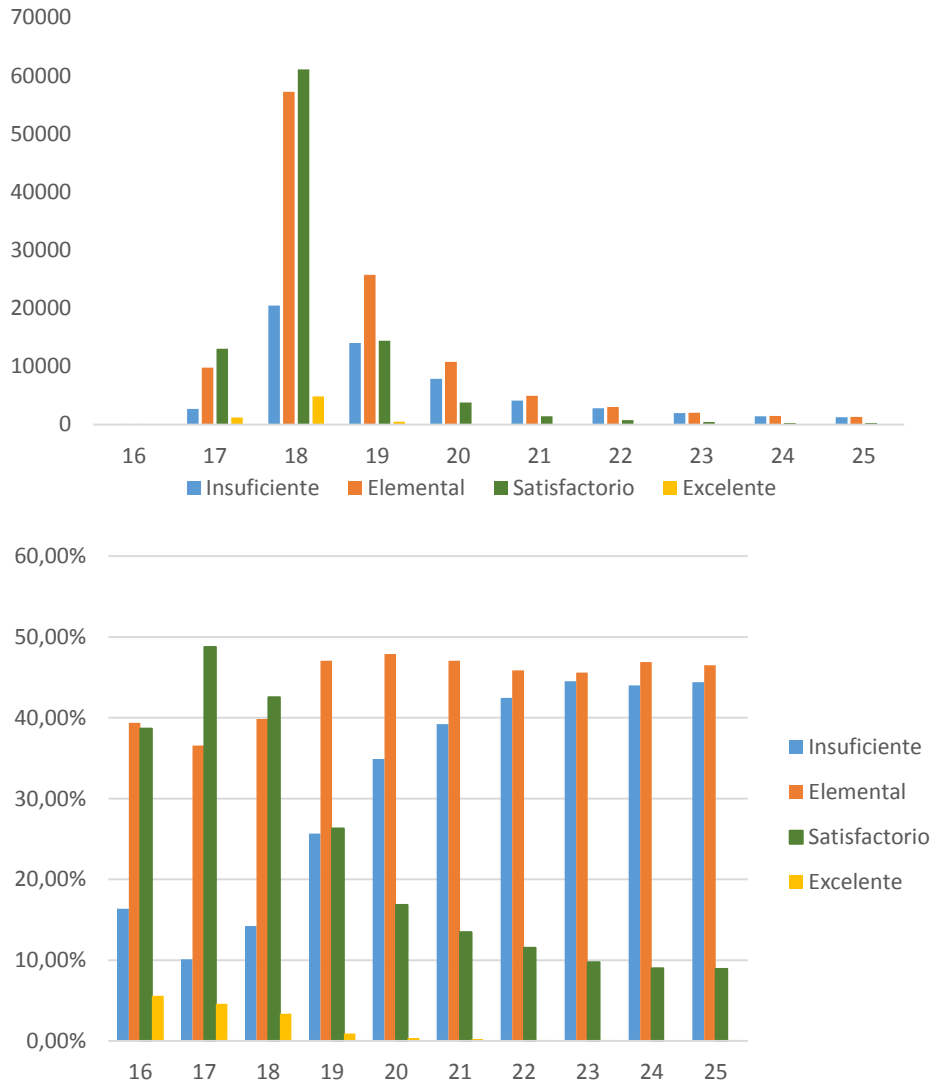


Figura 4.13: Frecuencias de nivel de logro alcanzado - Edad

En la tabla de frecuencias (Tabla 11) podemos ver la relación entre la variable dependiente **nivel de logro alcanzado** y la variable independiente **edad**; encontramos que el nivel de logro alcanzado en 16 años es un 16.38% insuficiente, 39.37% elemental, 38.68% satisfactorio y 5.57% excelente; en 17 años el 10.09% es insuficiente, 36.56% elemental, 48.75% satisfactorio y 4.60% excelente; en 18 años el 14.24% es insuficiente, 39.85% elemental, 42.55% satisfactorio y 3.36% excelente; en 19 años el 25.68% es insuficiente, 47.05% elemental, 26.33% satisfactorio y 0.94% excelente, en 20 años el 34.89% es insuficiente, 47.89% elemental, 16.86% satisfactorio y 0.36% excelente; en 21 años el 39.23% es insuficiente, 47.07% elemental, 13.46% satisfactorio

y 0.24% excelente; en 22 años el 42.48% es insuficiente, 45.87% elemental, 11.53% satisfactorio y 0.12% excelente; en 23 años el 44.52% es insuficiente, 45.60% elemental, 9.74% satisfactorio y 0.13% excelente; en 24 años el 44.01% es insuficiente, 46.92% elemental, 9.01% satisfactorio y 0.06% excelente; en 25 años el 44.42% es insuficiente, 46.50% elemental, 8.94% satisfactorio y 0.14% excelente (Ver Figura 4.9).

Prueba Chi-Cuadrado

```
chisq.test(t_e)


  Pearson's Chi-squared test


 data:  t_e
          X-squared = 28475, df = 27, p-value < 2.2e-16
```

Realizando la prueba del chi-cuadrado entre las variables categóricas **nivel de logro alcanzado** (insuficiente, elemental, satisfactorio, excelente) y **edad** (16, 17, 18, 19, 20, 21, 22, 23, 24, 25) se observa que es altamente significativa, el p-valor obtenido es menor a 0.05 por lo que existe una relación de dependencia.

4.4 Ajuste de Regresión de respuesta nominal

Para realizar el ajuste se tomo encuenta las categorias de referencia con las cuales se va ha trabajar, siendo así: Para región natural, igual a Costa; para autoidentificación étnica mestizo/blanco, tipo de financiamiento público, tipo de sexo hombre y nivel de logro alcanzado incuficiente.

Se modeliza el nivel de logro alcanzado en función de la edad (cuantitativa), tipo de financiamiento, región natural, tipo de sexo y autoidentificación étnica (las cuatro cualitativas). En este caso la variable respuesta Y (nivel de logro alcanzado) toma cuatro valores que denotaremos por Y_0 =elemental (categoría de referencia), Y_1 =insuficiente, Y_2 =satisfactorio, Y_3 =excelente (S=4).

El ajuste se realiza utilizando el software libre *R* con la librería *nnet* y la función *multinom ()*, que toma las transformaciones logit generalizadas con la categoría de referencia

$$L_s(x) = ln\left[\frac{P_s(x)}{P_0(x)}\right], \forall_s = 1,2,3.$$

Donde:

$L_s(x)$, Logaritmo de la ventaja de respuesta Y_s frente a la respuesta Y_1.

```
library(nnet)

Ajuste.Multinom.Cuanli<-

multinom(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data
= libro1)
```

4.4.1 Ajuste del modelo

Coefficients:	elemental		satisfactorio		excelente	
nm_regisierra	$\hat{\tau}_{s1}$	-1,344	$\hat{\tau}_{12}$	0,665	$\hat{\tau}_{13}$	1,034
nm_regiotras	$\hat{\tau}_{s2}$	-0,453	$\hat{\tau}_{22}$	0,148	$\hat{\tau}_{23}$	-0,028
nm_regioriente	$\hat{\tau}_{s3}$	-0,611	$\hat{\tau}_{32}$	-0,075	$\hat{\tau}_{33}$	-0,812
nm_regiinsular	$\hat{\tau}_{s4}$	-0,992	$\hat{\tau}_{42}$	0,111	$\hat{\tau}_{34}$	-0,410
financiamientoprivado	$\hat{\tau}_{s1}$	-0,660	$\hat{\tau}_{12}$	0,955	$\hat{\tau}_{13}$	1,400
financiamientomixto	$\hat{\tau}_{s2}$	-0,159	$\hat{\tau}_{22}$	0,667	$\hat{\tau}_{23}$	1,168
tp_sexomujer	$\hat{\tau}_{s1}$	-0,042	$\hat{\tau}_{12}$	0,147	$\hat{\tau}_{13}$	0,245
edad	$\hat{\beta}_{11}$	0,224	$\hat{\beta}_{21}$	-0,407	$\hat{\beta}_{31}$	-0,833
etnibbeindigena	$\hat{\tau}_{s1}$	0,511	$\hat{\tau}_{12}$	-0,291	$\hat{\tau}_{13}$	-1,054
etnibbeafroecuatoriano	$\hat{\tau}_{s2}$	0,281	$\hat{\tau}_{22}$	-0,451	$\hat{\tau}_{23}$	-1,049
etnibbeotro	$\hat{\tau}_{s3}$	0,248	$\hat{\tau}_{32}$	-0,245	$\hat{\tau}_{33}$	-1,052

etnibbemontubio	$\hat{\tau}_{s4}$	0,147	$\hat{\tau}_{42}$	0,040	$\hat{\tau}_{34}$	-0,092
Intercepts:	**insuficiente**		**satisfactorio**		**excelente**	
	$\hat{\beta}_{10}$	-4,515	$\hat{\beta}_{20}$	6,693	$\hat{\beta}_{30}$	11.311
Residual Deviance:	563519.1					
AIC:	563597.1					

Tabla 12: Estimación de parámetros del modelo nominal

La modelización de las transformaciones *logit* son de la forma:

$L_s(x) = -4{,}515 + 6{,}693 + 11{,}311 + 0{,}224X_1 - 0{,}407X_1 - 0{,}833X_1 - 1{,}344X_{21} + 0{,}665X_{21} + 1{,}034X_{21} - 0{,}453X_{22} + 0{,}148X_{22} - 0{,}028X_{22} - 0{,}611X_{23} - 0{,}075X_{23} - 0{,}812X_{23} - 0{,}992X_{24} + 0{,}111X_{24} - 0{,}410X_{24} - 0{,}660X_{31} + 0{,}955X_{31} + 1{,}400X_{31} - 0{,}159X_{32} + 0{,}667X_{32} + 1.168X_{32} - 0{,}042X_{41} + 0{,}147X_{41} + 0{,}245X_{41} + 0{,}511X_{51} - 0{,}291X_{51} - 1{,}054X_{51} + 0{,}281X_{52} - 0{,}451X_{52} - 1{,}049X_{52} + 0{,}248X_{53} - 0{,}245X_{53} - 1{,}052X_{53} + 0{,}147X_{54} + 0{,}040X_{54} - 0{,}092X_{54}$

Errores Estándar:

Coefficients:	insuficiente	satisfactorio	excelente
nm_regisierra	0,013	0,009	0,02
nm_regiotras	0,123	0,122	0,507
nm_regioriente	0,023	0,022	0,099
nm_regiinsular	0,154	0,113	0,456
financiamientoprivado	0,017	0,011	0,028
financiamientomixto	0,021	0,016	0,041
tp_sexomujer	0,010	0,009	0,025
edad	0,003	0,004	0,017
etnibbeindigena	0,024	0,021	0,087
etnibbeafroecuatoriano	0,022	0,026	0,110
etnibbeotro	0,075	0,077	0,338
etnibbemontubio	0,023	0,025	0,086

Intercepts:	insuficiente	satisfactorio	excelente
	0,061	0,079	0,320

Tabla 13: Errores estándar del modelo nominal

4.4.2 Significación de parámetros

La significación de parámetros nos permite contrastar las hipótesis de tal manera que:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

El resultado del ajuste muestra los parámetros estimados y sus errores estándar mediante el Test de Wald.

Los parámetros significativos para el estudio son: Edad, en Regiones Naturales (región sierra, región oriente), Autoidentificación étnica, Sexo, Tipo de financiamiento.

Los parámetros no significativos son: región otras y región insular, ya que al ser su p-valor mayor a 0.05 se acepta la hipótesis nula asumiendo que los dos parámetros deberían ser cero, por tal motivo no se extraen del modelo puesto que tienen parámetros no nulos.

Interpretaciones

Influencia de Regiones Naturales

- La $exp(\hat{\tau}_{s1}) = 0{,}260$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en la región sierra en lugar de la región costa permaneciendo constantes (edad, autoidentificación étnica, financiamiento, sexo). Analogamente se interpreta para el resto de categorías de nivel de logro.
- $exp(\hat{\tau}_{s1}) = 0{,}542$, es el cambio multiplicativo que se produce en la ventaja de nivel de logro insuficiente frente a elemental, cuando se pasa de región costa a oriente

permaneciendo constantes (edad, autoidentificación étnica, financiamiento, sexo). Analogamente se interpreta para el resto de categorías de nivel de logro.

Influencia de Tipo De Financiamiento

- La $exp(\hat{\tau}_{s1}) = 0{,}516$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en el tipo de financiamiento privado en lugar de público permaneciendo constantes (edad, autoidentificación étnica, región natural, sexo). Analogamente se interpreta para el resto de categorías de nivel de logro.
- La $exp(\hat{\tau}_{s2}) = 0{,}852$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en el tipo de financiamiento mixto en lugar de público permaneciendo constantes (edad, autoidentificación étnica, región natural, sexo). Analogamente se interpreta para el resto de categorías de nivel de logro.

Influencia de Tipo De Sexo

- La $exp(\hat{\tau}_{s1}) = 0{,}958$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en el tipo de sexo mujer en lugar de hombre permaneciendo constantes (edad, autoidentificación étnica, región natural, tipo de financiamiento). Analogamente se interpreta para el resto de categorías de nivel de logro.

Influencia de Edad

- La $exp(\hat{\beta}_{11}) = 1{,}251$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo aumenta a su edad en un año permaneciendo constante (autoidentificación étnica, región natural, tipo de financiamiento). Analogamente se interpreta para el resto de categorías de nivel de logro.

Influencia de Autoidentificación Étnica

- La $exp(\hat{\tau}_{s1}) = 1{,}666$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en autoidentificación étnica indigena

en lugar de mestizo/blanco permaneciendo constantes (edad, sexo, región natural, tipo de financiamiento). Analogamente se interpreta para el resto de categorías de nivel de logro.

- La $exp(\hat{\tau}_{s2}) = 1{,}325$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en autoidentificación étnica afroecuatoriano en lugar de mestizo/blanco permaneciendo constantes (edad, sexo, región natural, tipo de financiamiento). Analogamente se interpreta para el resto de categorías de nivel de logro.
- La $exp(\hat{\tau}_{s4}) = 1{,}159$, es el cociente de ventaja de nivel de logro insuficiente frente a elemental cuando un individuo se encuentra en autoidentificación étnica montubio en lugar de mestizo/blanco permaneciendo constantes (edad, sexo, región natural, tipo de financiamiento). Analogamente se interpreta para el resto de categorías de nivel de logro.

4.4.3 Significación de variables

Mediante el Test condicional de razón de verosimilitudes permite la comparación de modelos para estudiar la significación de variables.

$$H_0: p(x_q) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$H_1: p(x_q) = \frac{e^{\beta_0 + \beta_1 x_q}}{1 + e^{\beta_0 + \beta_1 x_q}}$$

- ***Autoidentificación étnica***

	Model	Resid. df	Resid. Dev
1	nm_regi + financiamiento + tp_sexo + edad	826464	565276.8
2	nm_regi + financiamiento + tp_sexo + edad + etnibbe	826452	563519.1

	Test	Df	LR stat.	Pr(Chi)
1				
2	1 vs 2	12	1757.772	0

El valor experimental del test es 1757.772 que para una Chi-cuadrado con 12 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable autoidentificación étnica debe estar en el modelo.

- ***Edad***

```
                                                Model    Resid. df   Resid. Dev
1        nm_regi + financiamiento + tp_sexo + etnibbe    826455      588072.2
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    826452      563519.1
    Test    Df LR stat.  Pr(Chi)
1
2   1 vs 2  3  24553.13   0
```

El valor experimental del test es 24553.13 que para una Chi-cuadrado con 3 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable edad debe estar en el modelo.

- ***Tipo de sexo***

```
                                                Model    Resid. df   Resid. Dev
1           nm_regi + financiamiento + edad + etnibbe    826455      563907.7
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    826452      563519.1
    Test    Df LR stat.   Pr(Chi)
1
2   1 vs 2  3  388.6435    0
```

El valor experimental del test es 388.6435 que para una Chi-cuadrado con 3 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable tipo de sexo debe estar en el modelo.

- ***Tipo de financiamiento***

```
                                                 Model    Resid. df   Resid. Dev
1                    nm_regi + tp_sexo + edad + etnibbe    826458      577766.0
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    826452      563519.1
    Test    Df  LR stat.   Pr(Chi)
1
2   1 vs 2  6   14246.91     0
```

El valor experimental del test es 14246.91 que para una Chi-cuadrado con 6 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable tipo de financiamiento debe estar en el modelo.

- ***Regiones Naturales***

```
                                                 Model    Resid. df   Resid. Dev
1           financiamiento + tp_sexo + edad + etnibbe    826464      588043.7
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    826452      563519.1
    Test    Df LR stat.   Pr(Chi)
1
2   1 vs 2  12  24524.66    0
```

El valor experimental del test es 24524.66 que para una Chi-cuadrado con 12 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable regiones naturales debe estar en el modelo.

Se determina que todas las variables son significativas.

4.4.4 Predicción

El modelo permite predecir cada valor de la variable explicativa nivel de logro alcanzado (elemental, insuficiente, satisfactorio, excelente) con mayor probabilidad de éxito. En la Tabla 14 podemos observar las predicciones y que las probabilidades por filas suman la unidad.

	insuficiente	elemental	satisfactorio	excelente
1	0,1181	0,4582	0,4092	0,0144
2	0,3494	0,5530	0,0969	0,0006
3	0,5646	0,2851	0,1486	0,0014
4	0,1670	0,5179	0,3078	0,0070

Tabla 14: Predicción de varios modelos

1. Para cuya persona con características región natural sierra, financiamiento público, tipo de sexo mujer, edad 18 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 12%, elemental de 46%, satisfactorio 41% y excelente 1%.
2. Para una persona con características región natural sierra, financiamiento público, tipo de sexo mujer, edad 22 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 35%, elemental de 55%, satisfactorio 9% y excelente 1%.
3. Para una persona con características región natural sierra, financiamiento público, tipo de sexo mujer, edad 21 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 56%, elemental de 29%, satisfactorio 15% y excelente 1%.
4. Para cuya persona con características región natural sierra, financiamiento público, tipo de sexo mujer, edad 19 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 17%, elemental de 52%, satisfactorio 31% y excelente 1%.

4.4.5 Tabla de clasificación

El modelo logit se puede usar como test de diagnóstico para clasificar las categorías de la variable respuesta en función de los valores de la variable explicativa, en la tabla las filas contiene el estado real en la muestra y las columnas la categoría predicha por el modelo.

	insuficiente	elemental	satisfactorio	excelente
insuficiente	9163	41681	5854	0
elemental	6237	73911	36332	0
satisfactorio	1467	35860	58276	0
excelente	31	1282	5403	0

Tabla 15: Tabla de Clasificación del modelo nominal

La tabla de clasificación establece cuantas veces se repite la predicción siendo así:

- Entre los individuos de logro alcanzado insuficiente, el modelo acierta en un 16,16%.
- Entre los individuos de logro alcanzado elemental, el modelo acierta en un 63,45%.
- Entre los individuos de logro alcanzado satisfactorio, el modelo acierta en un 20.39%.
- Entre los individuos de logro alcanzado excelente, el modelo acierta en un 0%.
- Entre todos los casos, el modelo acierta en un 51,31%.

4.4.6 Bondad del Ajuste

Al ser datos agrupados nuestra data.frame la bondad del ajuste se lo realiza mediante:

```
Ajuste.Multinom.Cuali$deviance
[1] 563519.1
pchisq(Ajuste.Multinom.Cuali$deviance,826452,lower.tail= F)
[1] 1
```

4.5 Estimación del modelo utilizando Ajuste de Regresión de respuesta ordinal

Para realizar el ajuste de regresión ordinal logit de respuesta ordinal con variables de diseño cualitativas y cuantitativas para datos agrupados con la finalidad de explicar el nivel de logro a partir de todas las variables, sin embargo hay que tener encuenta cuales

son las categorias de referencia con las cuales se va ha trabajar, siendo así: Para región natural, igual a Costa; para autoidentificación étnica mestizo/blanco, tipo de financiamiento público, tipo de sexo mujer y nivel de logro alcanzado elemental.

Para el desarrollo se utilizo el paquete MASS con la función *polr* del sofware libre R, que ajusta un modelo de regresión logística a una respuesta factorial ordenada.

```
Ajuste.Ordinal.Logro<-
polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data =
base
```

4.5.1 Ajuste del modelo

Coefficients:	Value	Std. Error	t value	P-value
nm_regisierra	1,182	0,0081	145,61	0,0000
nm_regiotras	0,374	0,0939	3,982	6,841e-05
nm_regioriente	0,302	0,0168	17,944	5,366e-72
nm_regiinsular	0,580	0,0933	6,223	4,892e-10
financiamientoprivado	1,105	0,0098	112,538	0,0000
financiamientomixto	0,644	0,0139	46,122	0,0000
tp_sexomujer	0,122	0,0073	16,614	7,383e-89
edad	-0,382	0,0026	-144,110	0,0000
etnibbeindigena	-0,469	0,0165	-28,338	1,18e-176
etnibbeafroecuatoriano	-0,506	0,0183	-27,622	6,09e-168
etnibbeotro	-0,363	0,0576	-6,296	3,054e-10
etnibbemontubio	-0,095	0,0191	-4,998	5,801e-07
Intercepts:	**Value**	**Std, Error**	**t value**	**P-value**
insuficiente|elemental	-7,941	0,0508	-156,222	0,0000
elemental|satisfactorio	-5,702	0,0496	-115,049	0,0000
satisfactorio|excelente	-2,228	0,0503	-44,288	0,0000

Residual Deviance:	5648176,48
AIC:	568206,48

Tabla 16: Estimación de parámetros del modelo

Los valores estimados puntuales dependientes e independientes $\hat{\beta}$ son:

$$(\hat{\tau}_{21}) = 1{,}182 \qquad (\hat{\tau}_{41}) = 0{,}122$$
$$(\hat{\tau}_{22}) = 0{,}374 \qquad (\hat{\beta}_{10}) = -0{,}382$$
$$(\hat{\tau}_{23}) = 0{,}302 \qquad (\hat{\tau}_{51}) = -0{,}469$$
$$(\hat{\tau}_{24}) = 0{,}580 \qquad (\hat{\tau}_{52}) = -0{,}506$$
$$(\hat{\tau}_{31}) = 1{,}105 \qquad (\hat{\tau}_{53}) = -0{,}363$$
$$(\hat{\tau}_{32}) = 0{,}644 \qquad (\hat{\tau}_{54}) = -0{,}095$$

La modelización de las transformaciones *logit* son de la forma:

$$L_s(x) = -0{,}382 + 1{,}182X_{21} + 0{,}374X_{22} + 0{,}302X_{23} + 0{,}580X_{24} + 1{,}105X_{31} + 0{,}644X_{32} + 0{,}122X_{41} - 0{,}469X_{51} - 0{,}506X_{52} - 0{,}363X_{53} - 0{,}095X_{54}$$

Utilizando β y las interacciones tenemos las siguientes ecuaciones:

$$logit[\,P(Y \leq 1)] = -7{,}941 - L_s(x)$$

$$logit[\,P(Y \leq 2)] = -5{,}702 - L_s(x)$$

$$logit[\,P(Y \leq 3)] = -2{,}228 - L_s(x)$$

4.5.2 Significación de parámetros

La significación de parámetros nos permite contrastar las hipótesis de tal manera que:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

El resultado del ajuste muestra los parámetros estimados y sus errores estándar mediante el Test de Wald.

Los parámetros significativos para el estudio son: En Regiones Naturales (región sierra, región otras, región oriente, región insular), Tipo de financiamiento (financiamiento privado, financiamiento mixto), Tipo de Sexo (sexo hombre), Edad, Autoidentificación étnica (etnia indígena, etnia ecuatoriana, etnia otro, etnia montubio).

Interpretaciones

Influencia de Regiones Naturales

- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0,306, si es región sierra frente a si es región costa, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.688, si es región otras frente a región costa, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.739, si es región oriente frente a si es región costa, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.559, si es región insular frente a región costa, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.

Influencia de Tipo De Financiamiento

- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.331, si es tipo de financiamiento privado frente a financiamiento público, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.524, si es tipo de financiamiento es mixto frente a financiamiento público, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.

Influencia de Tipo De Sexo

- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 0.836, si es si es mujer frente a si es hombre, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.

Influencia de Edad

- La ventaja de tener un resultado por debajo de insuficiente, es multiplicar 1,466 cuando se aumenta un año su edad, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.

Influencia de Autoidentificación Étnica

- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 1,599, si la autoidentificación étnica es indígena frente a autoidentificación étnica mestizo/blanco, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 1,659, si la autoidentificación étnica es afroecuatoriano frente a autoidentificación étnica mestizo/blanco, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 1,437, si la autoidentificación étnica es otro frente a autoidentificación étnica mestizo/blanco, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.
- La ventaja de tener un resultado por debajo de insuficiente, se multiplica por 1,100, si la autoidentificación étnica es montubio frente a autoidentificación étnica mestizo/blanco, de la misma manera se procede si se desea conocer la ventaja por debajo de elemental, satisfactorio y/o excelente.

4.5.3 Significación de variables

Mediante el Test condicional de razón de verosimilitudes permite la comparación de modelos para estudiar la significación de parámetros.

$$H_0: p(x_q) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

$$H_1: p(x_q) = \frac{e^{\beta_0 + \beta_1 x_q}}{1 + e^{\beta_0 + \beta_1 x_q}}$$

- ***Autoidentificación étnica***

	Model	Resid. df	Resid. Dev
1	nm_regi + financiamiento + tp_sexo + edad	275486	569729.8
2	nm_regi + financiamiento + tp_sexo + edad + etnibbe	275482	568176.5

	Test	Df	LR stat.	Pr(Chi)
1				
2	1 vs 2	4	1553.347	0

El valor experimental del test es 1553.347 que para una Chi-cuadrado con 4 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable autoidentificación étnica debe estar en el modelo.

- ***Edad***

	Model	Resid. df	Resid. Dev
1	nm_regi + financiamiento + tp_sexo + etnibbe	275483	590832.1
2	nm_regi + financiamiento + tp_sexo + edad + etnibbe	275482	568176.5

	Test	Df	LR stat.	Pr(Chi)
1				
2	1 vs 2	1	22655.58	0

El valor experimental del test es 22655.58 que para una Chi-cuadrado con 1 grado de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable edad debe estar en el modelo.

- ***Tipo de sexo***

```
                                                Model    Resid. df   Resid. Dev
1           nm_regi + financiamiento + edad + etnibbe    275483      568576.1
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    275482      568176.5
    Test    Df LR stat.   Pr(Chi)
1
2   1 vs 2  1  399.6563      0
```

El valor experimental del test es 399.6563 que para una Chi-cuadrado con 1 grado de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable tipo de sexo debe estar en el modelo.

- ***Tipo de financiamiento***

```
                                                Model    Resid. df   Resid. Dev
1                 nm_regi + tp_sexo + edad + etnibbe    275484      582284.6
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    275482      568176.5
    Test    Df  LR stat.   Pr(Chi)
1
2   1 vs 2  2   14108.08     0
```

El valor experimental del test es 14108.08 que para una Chi-cuadrado con 2 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable tipo de financiamiento debe estar en el modelo.

- ***Regiones Naturales***

```
                                                Model    Resid. df   Resid. Dev
1           financiamiento + tp_sexo + edad + etnibbe    275486      590769.3
2 nm_regi + financiamiento + tp_sexo + edad + etnibbe    275482      568176.5
    Test    Df LR stat.   Pr(Chi)
1
2   1 vs 2  4  22592.83    0
```

El valor experimental del test es 22592.83 que para una Chi-cuadrado con 4 grados de libertad arroja un p-valor de 0, que al 5% de significación, indica que la variable regiones naturales debe estar en el modelo.

4.5.4 Predicción

En la Tabla 17 podemos observar las predicciones de varios modelos, tomando las categorías de referencia

	insuficiente	elemental	satisfactorio	excelente
1	0,1285	0,4520	0,3975	0,0219
2	0,4052	0,4594	0,1304	0,0048
3	0,3173	0,4961	0,1794	0,0070
4	0,1778	0,4920	0,3150	0,0150

Tabla 17: Predicción de varios modelos

1. Para cuya persona con características región natural costa, financiamiento público, tipo de sexo mujer, edad 18 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 13%, elemental de 45%, satisfactorio 40% y excelente 2%.
2. Para una persona con características región natural costa, financiamiento público, tipo de sexo mujer, edad 22 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 40%, elemental de 46%, satisfactorio 13% y excelente 1%.
3. Para una persona con características región natural costa, financiamiento público, tipo de sexo mujer, edad 21 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 32%, elemental de 50%, satisfactorio 17% y excelente 1%.
4. Para cuya persona con características región natural costa, financiamiento público, tipo de sexo mujer, edad 19 años, autoidentificación étnica indígena tiene la probabilidad de obtener el nivel de logro alcanzado insuficiente 18%, elemental de 49%, satisfactorio 32% y excelente 1%.

4.5.5 Tabla de clasificación

El modelo logit se puede usar como test de diagnóstico para clasificar las categorías de la variable respuesta en función de los valores de la variable explicativa mediante la tabla de clasificación, una tabla que en las filas contiene el estado real en la muestra y en las columnas la categoría predicha por el modelo.

	insuficiente	elemental	satisfactorio	excelente
insuficiente	9626	41752	5320	0
elemental	7469	74608	34403	0
satisfactorio	1638	38798	55167	0
excelente	31	1519	5166	0

Tabla 17: Tabla de Clasificación del modelo

- Entre los individuos de nivel de logro alcanzado insuficiente, el modelo acierta en un 16,98%.
- Entre los individuos de nivel de logro alcanzado elemental, el modelo acierta en un 64,05%.
- Entre los individuos de nivel de logro alcanzado satisfactorio, el modelo acierta en un 18,97%.
- Entre los individuos de nivel de logro alcanzado excelente, el modelo acierta en un 0%.
- Entre todos los casos, el modelo acierta en un 50,59%.

4.5.6 Bondad del Ajuste

Al ser datos agrupados nuestra data.frame la bondad del ajuste se lo realiza mediante:

```
Ajuste.Ordinal.Logro$deviance
[1] 568176,5
pchisq(Ajuste.Ordinal.Logro$deviance,275482,lower.tail= F)
[1] 1
```

Que indica que el modelo de respuesta ordinal es adecuado.

4.6 Método de selección de variable – Ajuste STEPWISE nominal

Nuestro modelo, la selección parte del modelo que tiene cinco parámetros independientes y se pretende utilizar el procedimiento stepwise en ambas direcciones para encontrar los mejores predictores, dando como resultado las frecuencias en base a una variable independiente:

```
Ajuste.Ordinal.0<-polr(nl_inev~1,data=libro1)
Ajuste.Multinom.Step<-step(Ajuste.Ordinal.0,scope=list(lower=
nl_inev~1,upper=nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibb
e),direction="both")
```

```
Start:  AIC=632066.7
 nl_inev ~ 1
trying + edad
# weights:  44 (30 variable)
initial  value 8317.766167
final  value 8317.766167
converged
trying + nm_regi
# weights:  24 (15 variable)
initial  value 8317.766167
final  value 8317.766167
converged
trying + financiamiento
# weights:  16 (9 variable)
initial  value 8317.766167
final  value 8317.766167
converged
```

```
trying + tp_sexo
# weights:  12 (6 variable)
initial  value 8317.766167
final  value 8317.766167
converged
trying + etnibbe
# weights:  24 (15 variable)
initial  value 8317.766167
final  value 8317.766167
converged
```

```
    Df      AIC
+ +tp_sexo          6 16647.53
+ +financiamiento   9 16653.53
+ +nm_regi         15 16665.53
+ +etnibbe         15 16665.53
+ +edad            30 16695.53
```

Se observa que el modelo stepwise selecciona todas las variables (edad, tipo de financiamiento, tipo de sexo, región natural y étnia) como predictores del nivel de logro alcanzado.

4.7 Método de selección de variable – Ajuste STEPWISE ordinal

Para realizar una selección stepwise para regresión logística con R se utiliza la función *step()*:

```
Ajuste.Ordinal.0<-polr(nl_inev~1,data=base1)
Ajuste.Ordinal.Step<-step(Ajuste.Ordinal.0,scope=list(lower=
nl_inev~1,upper=nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibb
e),direction="both")
```

```
Start:  AIC=632066.7
nl_inev ~ 1

                 Df    AIC
+ edad            1 607908
+ nm_regi         4 608705
+ financiamiento  2 617313
+ etnibbe         4 628516
+ tp_sexo         1 631570
```

Nuestro modelo, mediante la selección stepwise consta de todas las variables edad, autoidentificación étnica, regiones naturales, tipo de financiamiento y tipo de sexo.

Conclusiones:

- En este trabajo se aplicaron los modelos de respuesta nominal y respuesta ordinal, para la modelización del nivel de logro alcanzado en la nota de examen de grado Ser Bachiller periodo 2018-2019, en función de las variables sociodemográficas edad, sexo, provincia, tipo de financiamiento de la institución educativa y autoidentificación étnica.
- Mediante la aplicación del ajuste de respuesta nominal para la proyección del nivel de logro alcanzado se analizó la incidencia de las variables sociodemográficas en la probabilidad de que un estudiante de educación media obtenga un rendimiento académico insuficiente, elemental, satisfactorio o excelente en el examen Ser Bachiller en el periodo 2018-2019. Con la construcción del modelo de regresión logística y la bondad del ajuste se validó a través del método de máxima verosimilitud, donde al mostrar significancia estadística no se desestimó ninguna variable, aceptándose la hipótesis alternativa.
- Aplicando el ajuste de respuesta ordinal para la estimación del logro alcanzado se validó la bondad del ajuste bajo el método de máxima verosimilitud y por significancia estadística no se desestimó ninguna variable.
- Se comprobó que para el modelo de ajuste nominal y ordinal las variables utilizadas son las adecuadas, llegando a esta conclusión mediante la aplicación del método stepwise ya que la selección automática de las variables independientes como predictoras son las mismas.
- Se observa que en el modelo ordinal hay 15 parámetros como variables estimadas, mientras que para el modelo nominal hay 39 parámetros como variables por categoría estimadas, siendo así la diferencia entre el modelo nominal y el modelo ordinal.

Bibliografía

Agresti, A. (2002). *Categorical Data Analysis.* New York: Willey.

Aiken, L. (1996). *Tests psicológicos y evaluación.* México: Prentice Hall Hispanoamérica.

Andersen, E. (1990). *The Statistical Analysis of Categorical Data.* New York: Springer-Verlag.

Carrión, E. (2002). Validación de características al ingreso como predictores del rendimiento académico en la carrera de medicina. *Educación Media Superior*, 6.

Davila N., G. M. (2015). An Asymmetric Logit Model to explain the likelihood of sucess in academic results. *Investigación Educativa*, 27-45.

Educación, M. d. (05 de 01 de 2015). *Ministerio de Educación*. Obtenido de https://educacion.gob.ec/wp-content/uploads/downloads/2017/02/Reglamento-General-a-la-Ley-OrgAnica-de-Educacion-Intercultural.pdf

Fajardo Bullón, F. M. (2017). ANÁLISIS DEL RENDIMIENTO ACADÉMICO DE LOS ALUMNOS DE EDUCACIÓN. *Educación XX1*, 209-232. Obtenido de https://www.redalyc.org/pdf/706

Guadagni, A. (Febrero de 2016). *Ingreso a la universidad en Ecuador, Cuba y Argentina*. Obtenido de Red Latinoamericana de Cooperación Universitaria: http://www.rlcu.org.ar/recursos/E_0000046_004_cea_numero_44.pdf

Hernandez J., M. A. (2006). Factores asociados con el desempeño académico en el EXANI-1. Zona Metropolitana de la Ciudad de México 1996-200. *Revista Mexicana de Investigación Educativa*, 547-581.

Hosmer, D. &. (2000). *Applied Logistic Regresion.* USA: Wiley & Sons.

INSTITUTO NACIONAL DE ESTADÍSTICA Y CENSOS. (11 de 09 de 2018). *INEC.* Obtenido de https://anda.inec.gob.ec/anda/index.php/catalog/266/datafile/F1/V4118

Instituto Nacional de Evaluación Educativa. (08 de 02 de 2018). *Ser Bachiller: Ficha técnica y conceptual*. Obtenido de http://evaluaciones.evaluacion.gob.ec/BI/ser-bachiller/

Jobson, J. (1991). *Categorical & multivariate methods.* USA: Springer verlag.

Pando Fernández V., S. M. (2004). *Regresión logística multinomial.* Esp. Cien.

Romero S. (16 de 03 de 2021). *Ecología Verde*. Obtenido de https://www.ecologiaverde.com/cuales-son-las-regiones-naturales-del-ecuador-3269.html

Sharmin S., M. R. (2015). Determinants of Academic Performance-A Multinomial Logistic Regression Approach. *International Journal of Scientific & Engineering Research*, 1212-1216.

Tessema B., M. k. (2016). Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students' Academic . *Journal of Education and Practice* , 3-7.

Tessema B., Meskele K., Sodo W. (2016). Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students' Academic. *Journal of Education and Practice*, 3-7.

UNESCO. (s.f.). *SITEAL*. Obtenido de https://siteal.iiep.unesco.org/sites/default/files/sit_informe_pdfs/dpe_ecuador-_25_09_19.pdf

Villarruel R., T. K. (03 de 07 de 2020). *Revista Economía y Política.* Obtenido de Universidad de Cuenca: https://www.redalyc.org/jatsRepo/5711/571163421008/html/index.html#redalyc_571163421008_ref15

Anexos

Anexo I.- Sintaxis en R Análisis Descriptivo

```
#CARGAR DATOS #

library(readr)
base <- read_delim("Libro1.csv", ";", escape_double = FALSE,
                   col_types = cols(nm_regi = col_factor(levels = c()),
                                    financiamiento = col_factor(levels = c()),
                                    tp_sexo = col_factor(levels = c()),
                                    etnibbe = col_factor(levels = c()),
                                    nl_inev = col_factor(levels = c()),
                   trim_ws = TRUE)
head(base)

contrasts(base$nm_regi)
contrasts(base$financiamiento)
contrasts(base$tp_sexo)
contrasts(base$etnibbe)
contrasts(base$nl_inev)
#Tomando el primer modelo como base segun el glosario de terminos#
base$nm_regi<-relevel(base$nm_regi,ref = "costa")
contrasts(base$nm_regi)
base$etnibbe<-relevel(base$etnibbe,ref = "mestizo/blanco")
contrasts(base$etnibbe)
base$tp_sexo<-relevel(base$tp_sexo,ref = "hombre")
contrasts(base$tp_sexo)
base$nl_inev<-relevel(base$nl_inev,ref = "insuficiente")
contrasts(base$nl_inev)

#ANALISIS DESCRIPTIVO#

#V_REGIONES
base$nm_regi <- factor(base$nm_regi)
levels(base$nm_regi)
table(base$nm_regi)
#porcentaje
t(prop.table(table(base$nm_regi)))
#grafica
rg<-barplot(table(base$nm_regi),las=2, xlab = "FRECUENCIA", horiz= TRUE,col =
rainbow(5),cex.names = 0.9,cex.axis = 0.7,main = "REGIONES")
```

```
#V_FINANCIAMIENTO
base$financiamiento<- factor(base$financiamiento)
levels(base$financiamiento)
table(base$financiamiento)
#porcentaje
t(prop.table(table(base$financiamiento)))
#grafica
f<-barplot(table(base$financiamiento), las=2, xlab = "FRECUENCIA", horiz= TRUE,col =
rainbow(3),cex.names = 0.9,cex.axis = 0.7,main = "FINANCIAMIENTO")

#V_SEXO
base$tp_sexo<- factor(base$tp_sexo)
levels(base$tp_sexo)
t<-table(base$tp_sexo);t
#porcentaje
t(prop.table(table(base$tp_sexo)))
#grafica
pie(table(base$tp_sexo),labels=t, clockwise =TRUE,col = rainbow(2),main = "SEXO")
legend("topright",c("Mujer","Hombre"),cex = 0.5,fill = rainbow(2))

#V_ETNIA
base$etnibbe<- factor(base$etnibbe)
levels(base$etnibbe)
table(base$etnibbe)
#porcentaje
t(prop.table(table(base$etnibbe)))
#grafica
barplot(table(base$etnibbe), ylab = "FRECUENCIA", col = rainbow(5),cex.names = 0.9,cex.axis
= 0.7,main = "AUTOIDENTIFICACION ETNICA")

#V_NIVEL DE LOGRO ALCANZADO
base$nl_inev<- factor(base$nl_inev)
levels(base$nl_inev)
table(base$nl_inev)
#porcentaje
t(prop.table(table(base$nl_inev)))
#grafica
barplot(table(base$nl_inev),ylab = "FRECUENCIA",col = rainbow(4),cex.names = 0.9,cex.axis =
0.7,main = "NIVEL DE LOGRO ALCANZADO")

#V_EDAD
base$edad
levels(base$edad)
table(base$edad)
summary(base$edad)
#porcentaje
t(prop.table(table(base$edad)))
```

```
#gráfica
barplot(table(base$edad), ylab = "FRECUENCIA", col = rainbow(5),cex.names = 0.9,cex.axis = 0.7,main = "EDAD")
```

Anexo II.- Análisis Bivariante

```
#LOGRO ALCANZADO - REGIONES

table(base$nl_inev,base$nm_regi)

t_rg<-table(base$nl_inev,base$nm_regi)

#Porcentajes

prop.table(t_rg)

prop.table(t_rg)*100

#Grafica

colores<-c("red","yellow","green","blue")

barplot(t_rg,las=2,xlab="FRECUENCIA DE LOGRO", horiz= TRUE,col = colores,cex.names = 0.9,cex.axis = 0.5,main = " LOGRO ALCANZADO - REGIONES")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.5,fill = colores)

#TEST

chisq.test(t_rg)

#LOGRO ALCANZADO - FINANCIAMIENTO

table(base$nl_inev,base$financiamiento)

t_f<-table(base$nl_inev,base$financiamiento);t_f

#Porcentajes

prop.table(t_f)

prop.table(t_f)*100

#Grafica

colores<-c("pink","yellow","blue","orange")

barplot(t_f,las=2,xlab="FRECUENCIA DE LOGRO", ylab="FINANCIAMIENTO", horiz= TRUE,col = colores,cex.names = 0.8,cex.axis = 0.7,main = " LOGRO ALCANZADO - FINANCIAMIENTO")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.5,fill = colores)
```

```
#TEST

chisq.test(t_f)

#LOGRO ALCANZADO - SEXO

table(base$nl_inev, base$tp_sexo)

t_s<-table(base$nl_inev, base$tp_sexo)

#Porcentajes

prop.table(t_s)

prop.table(t_s)*100

#Grafica

colores<-c("purple","yellow","blue","pink")

barplot(t_s,las=2, horiz= TRUE,col = colores,cex.names = 0.9,cex.axis = 0.6,main = " LOGRO ALCANZADO - SEXO")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.5,fill = colores)

#Mujer

porcentaje<-c(27232,56840,50564,3743)

etiqueta<-paste(porcentaje,sep="")

colores1<-c("purple","yellow","blue","pink")

pie(porcentaje,labels=etiqueta,clockwise = TRUE,col = colores1, main="LOGRO ALCANZADO - MUJERES")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.6,fill = colores1)

#HOMBRE

porcentaje<-c(29466,59640,45039,2973)

etiqueta<-paste(porcentaje,sep="")

colores2<-c("purple","yellow","blue","pink")

pie(porcentaje,labels=etiqueta,clockwise = TRUE,col = colores2, main="LOGRO ALCANZADO - HOMBRES")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.6,fill = colores2)

#TEST

chisq.test(t_s)
```

```
#LOGRO ALCANZADO - ETNIA

table(base$nl_inev,base$etnibbe)

t_et<-table(base$nl_inev,base$etnibbe);t_et

#Porcentajes

prop.table(t_et)

prop.table(t_et)*100

#Grafica

colores<-c("green","purple","blue","orange")

barplot(t_et,las=2,xlab="FRECUENCIA DE LOGRO", horiz= TRUE,col = colores,cex.names = 0.48,cex.axis = 0.6,main = " LOGRO ALCANZADO - ETNIA")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.6,fill = colores)

#TEST

chisq.test(t_et)

#LOGRO ALCANZADO - EDAD

t_e<-table(base$nl_inev,base$edad)

#Gráfica

colores<-c("pink","purple","blue","orange")

barplot(t_e,las=2,xlab="FRECUENCIA DE LOGRO", ylab="EDAD", horiz= TRUE,col = colores,cex.names = 0.8,cex.axis = 0.7,main = " LOGRO ALCANZADO - EDAD")

legend("topright",c("Insuficiente","Elemental","Satisfactorio","Excelente"),cex = 0.6,fill = colores)

#TEST

chisq.test(t_e)
```

Anexo III.- Ajuste de respuesta ordinal

```
#Utilizaremos la librería MASS y la función polr permite el ajuste del modelo de respuesta ordinal.

#Determinamos el logro alcanzado a partir de las variables regiones, financiamiento, sexo, edad y etnia, utilizamos la siguiente función R.

library(MASS)
```

```
Ajuste.Ordinal.Logro<-polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data =
base)

Ajuste.Ordinal.Logro

##Significación de parámetros

summary(Ajuste.Ordinal.Logro)

summary(Ajuste.Ordinal.Logro)$coefficients

exp(summary(Ajuste.Ordinal.Logro)$coefficients)

## p-value

2*pnorm(abs(summary(Ajuste.Ordinal.Logro)$coefficients[,"t value"]),lower.tail=F)

Z<-fitted.values(Ajuste.Ordinal.Logro);Z

##Significación de parámetros ANOVA

anova(polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data=base),polr(nl_inev
~nm_regi+financiamiento+tp_sexo+edad,data = base))

anova(polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data=base),polr(nl_inev
~nm_regi+financiamiento+tp_sexo+etnibbe,data = base))

anova(polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data=base),polr(nl_inev
~nm_regi+financiamiento+edad+etnibbe,data = base))

anova(polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data=base),polr(nl_inev
~nm_regi+tp_sexo+edad+etnibbe,data = base))

anova(polr(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data=base),polr(nl_inev
~financiamiento+tp_sexo+edad+etnibbe,data = base))

##PREDICCIÓN

head(predict(Ajuste.Ordinal.Logro,type = "probs")[1:10])

head(predict(Ajuste.Ordinal.Logro,type = "class"))

##TABLA DE CLASIFICACION

table(base1$nl_inev,predict(Ajuste.Ordinal.Logro,type="class"))

#LA BONDAD DEL AJUSTE

Ajuste.Ordinal.Logro$deviance

pchisq(Ajuste.Ordinal.Logro$deviance,275482,lower.tail= F)

#METODO DE SELECCION DE VARIALBLES AJUSTE STEPWISE

Ajuste.Ordinal.0<-polr(nl_inev~1,data=base)
```

```
Ajuste.Ordinal.Step<-step(Ajuste.Ordinal.0,
scope=list(lower=nl_inev~1,upper=nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe)
,
                direction="both")

#Significacion de parametros

summary(Ajuste.Ordinal.Step)

2*pnorm(abs(summary(Ajuste.Ordinal.Step)$coefficients[,"t value"]),lower.tail=F)
```

##PREDICCIÓN

```
head(predict(Ajuste.Ordinal.Step,type = "probs")[1:10,])
```

#LA BONDAD DEL AJUSTE

```
Ajuste.Ordinal.Step$deviance

pchisq(Ajuste.Ordinal.Step$deviance,5397,lower.tail= F)
```

Anexo IV.- Ajuste de respuesta nominal

```
#Utilizaremos el data.frame

libro1<-
data.frame(xtabs(~edad+tp_sexo+financiamiento+etnibbe+nm_regi+nl_inev,data=base))

library(nnet)

Ajuste.Multinom.Cuanli<-
multinom(nl_inev~nm_regi+financiamiento+tp_sexo+edad+etnibbe,data = libro1)

summary(Ajuste.Multinom.Cuanli)

exp(summary(Ajuste.Multinom.Cuanli)$coefficients)
```

#SIGNIFICACION DE PARAMETROS ANOVA

```
anova(multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe,

        data=libro1),multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo,
data=libro1))

anova(multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe,

        data=libro1),multinom(nl_inev~edad+nm_regi+financiamiento+etnibbe,
data=libro1))

anova(multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe,

        data=libro1),multinom(nl_inev~edad+nm_regi+tp_sexo+etnibbe, data=libro1))

anova(multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe,
```

```
        data=libro1),multinom(nl_inev~edad+financiamiento+tp_sexo+etnibbe, data=libro1))
anova(multinom(nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe,
        data=libro1),multinom(nl_inev~nm_regi+financiamiento+tp_sexo+etnibbe,
data=libro1))
#OBTENCION DE LOS P-VALORES CON LA PROBABILIDADES DE LA DISTRIBUCION NORMAL
2*pnorm(abs(summary(Ajuste.Multinom.Cuanli)$coefficients/
        summary(Ajuste.Multinom.Cuanli)$standard.errors),lower.tail=F)
#PREDICCION
predict(Ajuste.Multinom.Cuanli,type = "probs")[1:10,]
predict(Ajuste.Multinom.Cuanli,type = "class")[1:10]
#tabla de clasificacion
table(libro1$nl_inev, predict(Ajuste.Multinom.Cuanli,type="class"))
#BONDAD DEL AJUSTE
Ajuste.Multinom.Cuanli$deviance
pchisq(Ajuste.Multinom.Cuanli$deviance,826452,lower.tail= F)
## METODO DE SELECCION DE VARIALBLES AJUSTE STEPWISE
Ajuste.Multinom.0<-multinom(nl_inev~1,data=libro1)
Ajuste.Multinom.Step<-step(Ajuste.Multinom.0,
scope=list(lower=nl_inev~1,upper=nl_inev~edad+nm_regi+financiamiento+tp_sexo+etnibbe)
,
                direction="both")
summary(Ajuste.Multinom.Step)
exp(summary(Ajuste.Multinom.Step)$coefficients)
#PREDICCION
head(predict(Ajuste.Multinom.Step,type = "probs")[1:10,])
head(predict(Ajuste.Multinom.Step,type = "class")[1:10])
#LA BONDAD DEL AJUSTE
Ajuste.Multinom.Step$deviance
pchisq(Ajuste.Multinom.Step$deviance,5397,lower.tail= F)
```