ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

DETERMINANTES SOCIO-ECONÓMICOS Y DEMOGRÁFICOS DEL MÁXIMO NIVEL DE INSTRUCCIÓN ALCANZADO POR LA POBLACIÓN ECUATORIANA MEDIANTE UN MODELO DE REGRESIÓN ORDINAL.

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

PROYECTO DE INVESTIGACIÓN

PATRICIA JHOANA GUERRERO TUGÁ
patytuga@hotmail.com
patricia.guerrero@epn.edu.ec

Director: NELSON ALEJANDRO ARAUJO GRIJALVA, MS.C.
alejandro.araujo@epn.edu.ec

QUITO, FEBRERO 2019

DECLARACIÓN

Yo PATRICIA JHOANA GUERRERO TUGÁ, declaro bajo juramento que el trabajo aquí escrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración cedo mis derechos de propiedad intelectual, correspondientes a este trabajo, a la Escuela Politécnica Nacional, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normatividad institucional vigente.

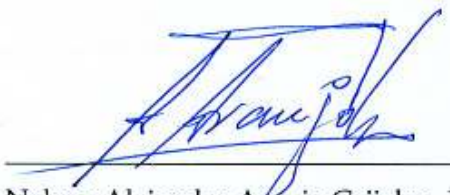
Patricia Jhoana Guerrero Tugá

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por:

PATRICIA JHOANA GUERRERO TUGÁ

bajo mi supervisión.



Nelson Alejandro Araujo Grijalva, Ms.c.
Director del Proyecto

AGRADECIMIENTOS

Gracias a Dios por mi familia y todas las bendiciones que he recibido y las que me brinda día a día al estar siempre conmigo.

Gracias a mi madre Inés Tugá, que con su ejemplo me enseña a siempre seguir, jamás rendirme y creer en mi, que con su amor y apoyo incondicional me ayuda a ser mejor persona. En mi vida sin ella nada sería posible, pues mis sueños los cumplo siempre en tu nombre, mi "mamita linda" siempre voy a ser tu "negrita", la "niña de tus ojos" como tú lo eres para mi. Ni las palabras ni la vida me va a alcanzar para agradecerte por ser esa mujer fuerte, sabia y amorosa con la que Dios me ha bendecido como madre. Gracias por ser mi mamá.

Gracias a mis hermanos y mis hermanas en especial a Esther Guerrero, por su confianza y compañerismo, desde que llegué a estudiar a Quito y de hecho desde siempre, me ha dado su amor con palabras de aliento para seguir en los momentos más difíciles. Mi hermana querida gracias por ser mi ejemplo a seguir, estoy muy orgullosa de tenerte como hermana.

Gracias a mi tío Reinaldo Tugá, que ha llegado a ser como el padre que nunca tuve y un gran amigo, compartimos la pasión por las matemáticas, los libros y deportes, gracias por inculcar en mi la pasión por el conocimiento.

Gracias a todas las personas que forman la Escuela Politécnica Nacional, porque hacen que sea posible estudiar en una excelente universidad y cada día trabajan por mejorarla.

Gracias al Ms.c. Nelson Alejandro Araujo Grijalva por ser el director de este proyecto de investigación que con sus consejos y sugerencias ha hecho esto posible.

Gracias a mi gran amigo Alexito, por su amistad, por la convivencia tan agradable que tuvimos y por todo lo compartido a lo largo de esta trayectoria, loquito, sabes que el cariño que te tengo es de ñaños.

Gracias a mi mejor amiga Natalia, a quien conozco desde la escuela y la vida nos volvio a juntar en la universidad, por su amistad y todo lo que hemos compartido juntas.

Gracias a mi amigos/as y compañeros/as del inicio de la carrera Elizabeth, Gissela, Jessica , Alex, Cristhian y Jhon por los momentos compartidos y su amistad sincera, en especial a Cris que me ha enseñado, aconsejado y compartido sus conocimientos.

Gracias a mis amigos/as y compañeros/as Karen y Jairo que llegue conocer a lo largo de mis estudios en la carrera, por su compañerismo y apoyo sincero.

Gracias a mis amigos/as y compañeros/as de la universidad Gaby, Leonardo y Santiago por todos los momentos y aventuras juntos.

Gracias a mis amigos/as y compañeros/as de la universidad que me acompañaron en las actividades recreativas como los deportes a Sebas, Cris, Jaime, Sebitas, Jeny, Pauli, Eve, Karen, Mabe, Tefa y demás, por todos los juegos compartidos.

Gracias a mis amigos/as y compañeros/as Carlitos, Miguel y Jonathan que al volver a estudiar tuve la maravillosa oportunidad de conocerlos.

En especial a mi mejor amigo Carlos Rivera (Carlitos [36]) a quien quiero como a un hermano, por su amistad sincera y siempre estar a mi lado, por todas las aventuras, locuras, consejos, risas, llantos, etc. Mi Carlitos, gracias amigo mío, sabes que a cualquier lugar que la vida te lleve, mi corazón y mi buena vibra estarán contigo.

Finalmente, gracias a todas las personas que han formado parte de esta etapa de mi vida incluyendo a las que no mencioné, amigos, compañeros, profesores, conocidos, etc. Estoy realmente agradecida por absolutamente cada momento que he vivido para llegar aquí, he disfrutado el proceso, pues todo ha sido extraordinario gracias a ustedes.

DEDICATORIA

A mi madre por su apoyo y amor incondicional.

A mis hermanas y hermanos, por todo los momentos que hemos compartido.

A mi tío por su cariño y amistad.

"No mires al mundo como es, sino como debería de ser"

Paty Tugá

Índice general

Resumen

El presente proyecto de investigación tiene como objetivo construir un modelo analítico de regresión logística ordinal que permita estimar el máximo nivel de educación que alcanzará un determinado individuo de la población ecuatoriana pudiendo ser Ninguno, Primario, Secundario o Superior, es decir la variable dependiente a predecir será una variable cualitativa con 4 posibles categorías, mediante la utilización de variables explicativas relacionadas con nivel socio-económico y variables demográficas. Se determinan los factores sociales, económicos y demográficos más influyentes para obtener un buen nivel de educación en el Ecuador.

Para obtener el modelo se va a usar la información pública del último Censo de Población y Vivienda (2010) descargada de la página web oficial del INEC.

Palabras clave: INEC, niveles de educación en el Ecuador, modelo logit

Abstract

The objective of this research project is to construct an analytical model of ordinary logistic regression that allows estimating the maximum level of education that will allow the dependent variable to be identified to be predicted. It will be a qualitative variable with 4 possible categories, through the use of explanatory variables related to socioeconomic level and demographic variables. The most influential social, economic and demographic factors are determined to obtain a good level of education in Ecuador.

To obtain the model, public information from "Censo de Población y Vivienda (2010) " downloaded from the official INEC website can be used.

Keywords: INEC, levels of education in Ecuador, logit model

Capítulo 1

Introducción

1.1. La Educación en el Ecuador

La educación en el Ecuador a través de los años ha pasado por varios cambios, hay amplia literartura sobre lo que es la educación y la importancia de la misma, siempre me he preguntado que es lo que más influye para que una persona tenga una buena educación haciendo referencia al nivel de educación, entonces algunas de las preguntas que surgen son: ¿cuáles son los factores que influye para que una persona alcance un cierto nivel de educación?, ¿de qué depende?, ¿es posible cuantificar eso?, entre otras. De esta forma y con el fin de responder a estás preguntas, se ha realizado este proyecto de investigación.

En este capítulo se realizará una breve reseña histórica de la educación en el Ecuador, la importancia y de como abordaremos este estudio para el presente proyecto de investigación.

1.1.1. Breve reseña histórica de la educación ecuatoriana

Desde los primeros años de la fundación de la república del Ecuador (1830), el Estado promueve y fomenta la educación pública. Al inicio de la época republicana se organizó el sistema educativo mediante la Dirección General de Estudios y la primera ley de Instrucción Pública.

Con la presidencia de Vicente Rocafuerte (1835-1839) se va configurando una política educativa de alcance nacional, pues se crea la Dirección General de Instrucción e Inspección y se emite el Decreto reglamentario de Instrucción Pública de 1836, con lo que buscaba regular ocho colegios (uno femenino) y 290 escuelas (30 femeninas)

que sumaba una población estudiantil de algo más de 13.000 estudiantes. Esta regulación no regía para los estudiantes universitarios que no eran más de ochenta (Ministerio de Educación, s/f, p. 1). En aquel periodo, la cobertura de este número de alumnos primarios y secundarios es muy bajo si se estima que hacia 1840 el país tenía una población de 617.382 habitantes (Saenz y Palacios, 1983).

Por iniciativa legal del presidente García Moreno (1861-1965/1869-1875), en el año "1871 se dispuso que la educación primaria fuera gratuita y obligatoria" (Freire, 2015: 4), se exige a que los directores de las instituciones educativas profesen la religión católica y se da el derecho a un establecimiento por cada comunidad que albergue a 500 niños. "Para entonces, el número de escolares era de alrededor de 32.000 y el Estado invertía el 11 % de su presupuesto en instrucción pública" (Ministerio de Educación, s/f, p. 1).

Durante el régimen "progresista" de Plácido Caamaño (1883-1888) se crea el Ministerio de Instrucción Pública en el año 1884. La política educativa de este periodo dio a la educación primaria "un empuje inédito tanto en número de escuelas como de estudiantes: Ecuador llegó a ser el país de América Latina con mayor cantidad de escolares en proporción al número de habitantes" (Freire, 2015, p. 4). Además, "se editaron diversos libros de texto por autores ecuatorianos, algunos de los cuales se usaron en varios países de habla española" (Ibíd., p. 5).

Con la revolución liberal de 1895 liderada por Eloy Alfaro, la educación pública recibe un gran impulso. Se institucionalizó una política que propugnaba una enseñanza laica y su financiamiento estatal exclusivo (Ministerio de Educación, s/f, p. 2). En las décadas anteriores se ejecutaba en las instituciones fiscales y religiosas una educación acorde a los lineamientos doctrinarios de la iglesia católica.

El liberalismo alfarista con la política de separación Iglesia-Estado, sentó las bases para el incremento sostenido de la educación pública. "En 1911, por ejemplo, 1.197 (77 %) escuelas eran fiscales respecto de un total de 1.551. Tres décadas después, en 1941, el porcentaje había subido ligeramente (79 %), aunque ahora respecto de un total mayor de 3.114 escuelas, de las cuales 2.470 eran fiscales" (Terán Najas y Soasti, 2006, p. 44).

Esta consolidación de la educación fiscal, también se puede constatar con los datos de la población escolar que corresponden al comienzo de los años cuarenta del siglo pasado: de 243 781 estudiantes, el 73 % pertenecías a escuelas públicas, el 10 % a municipales y el 17 % a particulares. No tenemos información sobre la enseñanza secundaria, solo sabemos que se educaban 10.000 colegiales en planteles

oficiales (Terán Najas y Soasti, 2006). Pese al gran avance en la cobertura educativa, estas cifras siguen siendo pobres si consideramos que la población del Ecuador en 1940 era 2'586.000 personas (Saenz y Palacios, 1983).

La educación confesional consiguió alcanzar de nuevo una presencia relevante con el presidente Velasco Ibarra desde 1946 (año en que autoriza la fundación de Pontificia Universidad Católica del Ecuador), gracias a un modus vivendi del año 1937 que restableció las relaciones entre el Estado ecuatoriano y la Iglesia. No obstante, la educación laica crecía de modo sostenido, pues la matriculación en la escuela primaria alcanzaba en 1947 el 86 % de los niños habilitados para estudiar, aunque el abandono escolar era muy alto. El Estado en este año financiaba a la educación pública con el 17,5 % del presupuesto nacional (Luna Tamayo, 2014).

A los inicios de los años de 1950 se impone una nueva era, la desarrollista, orientada por la CEPAL (Comisión Económica para América Latina, de las Naciones Unidas) que proponía el intervencionismo del Estado para alcanzar el desarrollo económico y social mediante políticas gubernamentales planificadas. El ciclo desarrollista tiene su fin en los años de 1980 (Ossenbach, 1999). El periodo desarrollista se deja sentir en las políticas educativas de los diversos gobiernos.

El país intentó ejecutar el modelo económico cepalino que perseguía un proceso de sustitución de importaciones mediante la industrialización. Esto requería la ampliación de la educación pública con la finalidad de formar trabadores calificados que puedan laborar en las empresas manufactureras. En esta perspectiva, el Estado invirtió un porcentaje cada vez mayor del presupuesto nacional en educación (ver cuadro 1.1)[1].

Año	**1940**	**1950**	**1960**	**1970**	**1975**	**1979**
Porcentaje (%)	15,64	19,77	15,41	21,26	22,61	25,21

Tabla 1.1: Evolución del porcentaje del presupuesto nacional dedicacdo a la educación 1940-1979

En el cuadro precedente se observa un aumento consistente del porcentaje del Presupuesto Nacional dedicado a la educación gracias a la bonanza bananera y al boom petrolero, exceptuando parte del decenio de 1960 por la disminución de las exportaciones del banano. El efecto positivo de esta política educativa fue un incremento de las tasas brutas de escolaridad: La educación primaria pasa de 56,7 % en

[1]Fuente: Ossenbach (1999)

1950 a 93,7 % en 1975, en tanto que la educación media en este mismo periodo sube de un insignificante 4,4 % a un 27,5 % (Ídem).

La baja tasa de escolaridad en la educación media se explica, en gran medida, porque la mayoría de niños de las zonas rurales no podían acceder a este nivel de estudios, al cumplir por ley una primaria de cuatro años, pues se les exigía una educación primaria de seis años. Tal discrimen se superó en 1964 con el primer plan decenal de educación 1964 – 1974, que aumento a seis años de estudio la educación primaria (Luna Tamayo, 2014) en todo el territorio nacional. Cabe destacar que se amplió la obligatoriedad y gratuidad de la educación hasta la enseñanza básica, completando nueve años escolares de estudio (Ossenbach, 1999).

En tanto que la educación superior afrontó un proceso de masificación estudiantil. "Si en 1950 hubo 4.122 matriculados a nivel universitario, en 1970 alcanzó el número de 39.861" (Luna Tamayo, 2014, p. 28). La presencia de este fenómeno se debe, principalmente, a la eliminación de las pruebas de ingreso decretada en 1969 y a la creación de siete nuevas universidades públicas, que se adhirieron a las cinco fundadas antes de 1944 (Ossenbach, 1999).

El modelo desarrollista en el Ecuador no se consolidó ni pudo materializar sus propuestas. Llegado los años de 1980, cuando los recursos petroleros disminuyen y la crisis de la deuda externa se pone de manifiesto, la corriente cepalina cede su influencia a una concepción económica neoliberal. De hecho "ni murió el modelo desarrollista y el neoliberal nació incompleto" afirma Luna Tamayo, (2014, p. 122). En este contexto, el Estado se debilita y su accionar es limitada. La política educativa restringe el gasto e inversión. Así, en 1980 la inversión en educación con relación al PIB fue de 5,3 %, en 1991 se contrajo al 2,8 %; su participación en el Presupuesto General del Estado disminuyó desde 34 % en 1980 hasta 17 % en 1990 (Luna Tamayo, 2014). A finales de la última década del siglo XX, la crisis económica y social se profundiza. Como consecuencia, en el año 2000 el gasto en educación con relación al PIB fue apenas el 1,7 % y su participación en el Presupuesto General del Estado representaba el exiguo 8,5 % (Briones Rugel et. al., 2011).

A pesar de esta situación financiera difícil del sector educativo, teniendo como base lo conseguido en años precedentes, no dejó de mejorar. Entre 19882 y 1999 la matrícula de educación primaria se elevó de 68,6 % a 90,3 %, la secundaria de 29,5 % a 51,4 % y la universitaria de 7,4 % a 14,9 % (Luna Tamayo, 2014). Si hubiese contado con mayores recursos la educación, quizá hubiese progresado más, ya que la matrícula de la educación secundaria y universitaria, está lejos de alcanzar el cien

por ciento; y talvez hubiese logrado mejores estándares de calidad.

En los primeros años del nuevo milenio, la inversión educativa empezó a recuperarse lentamente, en 2001 alcanzó el 2,3 % del PIB, hasta llegar al 3 % en el 2007 (Ídem). En el sistema educativo del país no se realizaron mayores cambios. Será un nuevo boom petrolero el que permitirá elevar, por lo menos cuantitativamente, los indicadores de la educación.

1.1.2. Acceso de la población ecuatoriana a la educación 2007-2017

En este epígrafe se realiza un somero estudio sobre acceso de la población ecuatoriana a la educación desde el año 2007 hasta el 2017, es decir, la cobertura estudiantil del sistema educativo, periodo que coincide con la presidencia de la república del Economista Rafael Correa, gobierno que se benefició de los ingentes recursos generados por la segunda bonanza petrolera; esto permitió el retorno del protagonismo del Estado en la sociedad.

Una de sus primeras gestiones del entonces presidente Correa fue convocar a una asamblea constituyente para que elaborara una nueva constitución. La Constitución Política del Estado (2008), aún vigente, señala en el artículo 26 que la "educación es un derecho de las personas a lo largo de su vida y un deber ineludible e inexcusable del Estado", es obligatoria hasta el bachillerato o su equivalente y en el artículo 28 que es "gratuita hasta el tercer nivel de educación superior".

En el régimen del presidente Correa, uno de los años más representativo del presupuesto educativo en relación al Presupuesto General del Estado es en el año 2009 con 13,8 %; si bien este porcentaje nos es el más alto desde 1950, si es uno de los más importantes en dólares corrientes ($ 2 071 158 501,33) (Briones Rugel et. al., 2011). En tanto que, el mayor presupuesto de la educación en relación al PIB lo alcanzó en 2014 con el 5 % (Banco Central del Ecuador, 2014-2015), un punto menos de lo exigido por la Política 8 del Plan Decenal de Educación 2006-2015 y tomado como compromiso por la Constitución en la disposición transitoria Democtava.

Como consecuencia de este mayor financiamiento a la educación se mejoró el acceso a ella. La tasa neta de asistencia a la Educación General Básica (EGB) pasó de 91,18 % en 2006 a 96,23 % en 2016. (Correa Delgado, 2017).[15] En la tabla 1.2 [2] se pude constatar el crecimiento de la tasa neta de matrícula en la población indígena y afroecuatoriana.

[2]FUENTE: Correa Delgado, 2017, p. 150 - ELABORACIÓN: Propia

Año	Indígena	Afroecuatoriano	Mestizo	Blanco
2006	89,2	88,3	91,9	93,6
2016	96,2	95,6	96,4	93,7

Tabla 1.2: Tasa neta de matrícula por etnia 2006 y 2016

En lo que respecta al bachillerato, en la tabla 1.3[3] se muestra el progreso alcanzado en la tasa neta de asistencia correspondientes a los años representativos de 2006 y 2011 a nivel nacional, considerando el género y la etnia.

Etnia y género	Año 2006	Año 2011
Nacional	47,9	62,1
Hombre	46,0	61,7
Mujer	50,0	62,5
Indígena	24,2	46,4
Afroecuatoriano	44,1	57,6
Mestizo	50,0	65,6
Blanco	49,3	64,6

Tabla 1.3: Tasa neta de asistencia al bachillerato por etnia y género 2006 y 2011

Por otra parte, la gestión y política educativa del gobierno del Economista Rafael Correa tuvo honda repercusión en las Instituciones de Educación Superior (IES). En 2012 el país tenía 71 universidades y escuelas politécnicas. Estas y sus extensiones fueron evaluadas y categorizadas; se cerraron 17 de estas instituciones, de las mismas 13 eran privadas (Correa Delgado, 2017).

La inversión en educación superior fue importante; si en 2006 la inversión en educación superior fue de 0,7 % del PIB (335 millones USD), en 2014 se elevó a 2,1 %, que representan 2.160 millones de dólares americanos (SENESCYT, 2015), "una cifra inédita, siendo la segunda más alta del mundo para el año 2014 según el análisis de la Unesco" (Correa Delgado, 2017, p.154). No cabe duda que el monto de la inversión es notable, sin embargo se lo canalizó a la fundación de nuevas IEP públicas, hubiese sido mejor que se las fortalezca a las existentes.

Tal magnitud de recursos no tuvo un impacto determinante en la matrícula de las IES como se esperaba. En la tabla 1.4 [4] ayuda darse cuenta que el periodo 2006-2014, el incremento de la matrícula neta y bruta no es más alto que el ciclo económico

[3]FUENTE: Luna Tamayo, 2014, p. 253 ELABORACIÓN: Propia
[4]FUENTE: SENESCYT (2015) ELABORACIÓN: Propia

crítico 1999-2006; más aún, en este periodo la tasa bruta de matrícula crece el doble que los ocho años siguientes.

Año	Tasa bruta de matrícula	Tasa neta de matrícula
1999	22,61	15,11
2006	28,76	18,93
2014	31,86	21,23

Tabla 1.4: Tasa neta y bruta de matrícula en las IES 1999-2006-2014

En alguna medida, se logró una relativa democratización del acceso a la educación superior, ya que "ha permitido aumentar la participación en la matrícula bruta del 40 % más pobre: entre 2006 y 2014 subió en un 101 %. En el 2014, la probabilidad de estar en el sistema de educación Superior fue cuatro veces mayor que en 2006 para personas con padres sin ningún nivel educativo".[15]

1.2. Importancia y análisis de la educación

La educación es uno de los factores más influyentes en el progreso de sociedades, debido a que, una educación de calidad permite alcanzar mejores niveles de bienestar social y de crecimiento económico, impulsando la ciencia, tecnología e innovación.

La educación siempre ha sido importante para el desarrollo, pero ha adquirido mayor relevancia en el mundo que actualmente vive profundas transformaciones, motivadas en gran parte tanto del vertiginoso avance de la ciencia y sus aplicaciones, como del acelerado desarrollo de medios y tecnologías de la información.

En economías modernas, el conocimiento se ha convertido en uno de los factores claves para su desarrollo; las sociedades que han alcanzado mejores niveles económicos y sociales son aquellas que han conseguido cimentar y optimizar los sistemas educativos y de investigación científica y tecnológica, prestando mayor énfasis a la escolarización y la educación superior, debido a la estrecha relación existente en entre ellos, por ejemplo, estudios de la Organización para la Cooperación y el Desarrollo Económicos (OCDE) prueban que un año adicional de escolaridad incrementa el PIB per cápita de un país entre 4 % y 7 % [5].

[5]OCDE, Perspectivas económicas para América Latina, 2009.

A lo largo de la historia existen varias bibliografías acerca el desarrollo de la educación, entre ellos la teoría del capital humano, la cual argumenta que debido a la correlación entre el nivel educativo del trabajador y el salario percibido en el mercado de trabajo, como también la evidencia que invertir más en educación le da el trabajador mayores ingresos durante toda su vida que si no hubiese alcanzado dicho nivel educativo.[38]

Lester Thurow [46] plantea una discusión sobre la dimensión macroeconómica de la teoría del capital humano, al analizar la evolución de la educación y de la economía en Estados Unidos. En los años de 1950 a 1970 hubo mejor igualdad en la educación que en la renta, incrementando el número promedio de años de estudios y pese a eso, los pobres de 1969 al compararlos eran más pobres que los de 1949. Además, en ese periodo la proporción de trabajadores con estudios universitarios aumentó del 3 % al 6 % anual que la productividad y la economía nacional en cambio no pasó del 3 %. Por lo tanto, Thurow, encuentra que la relación entre la educación, productividad e ingresos dede ser diferente y mucho más compleja, de lo que supone la teoría del capital humano, sobre todo cuando hay períodos de recesión con desempleo. [38]

El "shooling model" de Mincer (1974) [30] enfatiza que la experiencia en el empleo genera niveles distintos de ingresos, como así también diferentes pendientes en el perfil edad-ingresos de las ocupaciones. Sin embargo, luego de algún período de la vida activa la productividad es decreciente debido incluso a factores biológicos.

En cualquier período de la vida activa, esta teoría considera que los ingresos se componen del rendimiento de la dotación de capital humano que posee la persona y de su tasa de inversión. Es así como al comienzo de la vida activa el individuo consagra una gran parte del tiempo a su educación. Dicha tendencia según Becker (1964) [3] se debería a que en general el perfil edad - ingresos tiene un rápido crecimiento durante la primera década de vida activa, luego un menor crecimiento y un nivel constante en la tercera o cuarta década.[24]

En el informe de la OCDE, *Escuelas y calidad de la enseñanza(1991)*, se expone sobre la teoría del capital humano para hacer hincapié en mejorar la calidad de la educación que se requiere por las nuevas exigencias de la demanda laboral como la flexibilidad, versatilidad y trabajo en equipo.[34]

Actualmente existe consenso con que el conocimiento y el capital humano (la teoría del capital humano) son factores determinantes para el desarrolo económico, además, se determina que la educación incrementa la productividad de los indivi-

duos y que también es usada por los empleadores para seleccionarlos al momento de emplearlos. Aunque en muchas ocasiones, estos últimos asignan una mayor importancia a las actitudes y habilidades asociadas indirectamente con la obtención de un título educativo que a los conocimientos propios impartidos y desarrollados por esa formación. [24]

Independientemente de los diferentes puntos de vista respecto a la teoría del capital humano, existe un acuerdo amplio para interpretar la relación entre la economía y la educación, es así que se quiere analizar la educación en el Ecuador y por supuesto teniendo claro que el capital humano no se forja solo en elsistema educativo, ya que también ahora gracias a los diferente cambios tecnológicos y sociales surgen nuevas formas y vías informales de formación, incluida la experencia laboral y otras causas. [38]

Adicionalmente en un país como Ecuador, en vías de desarrollo, necesita enfocarse en la generación de estrategias que contribuyan al fortalecimiento de su sistema educativo, en particular el acceso y culminación satisfactoria de estudios superiores, ya que según el último Censo de Población y Vivienda, ejecutado por el Instituto Nacional de Estadísticas y Censos (INEC), en el artículo "El Censo informa: Educación" se establece que el 8,74 % de la población mayor o igual a 17 años (Ver Tabla 1.5) que ha obtenido un título(universitario o no universitario) en una Institución de Educación Superior.

Además, la distribución de ecuatorianos con título registrado en la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT) de acuerdo a su auto identificación, permite identificar que la tasa más pequeña de personas con título registrado corresponde a las etnias: Indígena, Afroecuatoriano y Montuvio, con un 2,11 %, 3,68 % y 2,75 % respectivamente, en este caso, se refleja la existencia de desigualdad de acuerdo a la etnia generando ciertos indicios de desigualdad, en el acceso y culminación de estudios superiores. Considerando esta desigualdad se plantea analizar los aspectos sociales, culturales, y económicos que podrían influir en el estudio.

Auto identificación según su cultura y costumbres	Población con título registrado	Población Ecuatoriana mayor o igual a 17 años	Porcentaje de la Población Ecuatoriana	Tasa de población con título registrado
Indígena	12.533	592.663	6,32 %	2,11 %
Afroecuatoriano	24.238	658.782	7,03 %	3,68 %
Montuvio/a	19.976	726,115	7,74 %	2,75 %
Mestizo/a	688.324	6.777.294	72,27 %	10,16 %
Blanco/a	71.354	584.884	6,24 %	12,2 %
Otro/a	3.344	37.886	0,40 %	8,83 %
Total	**819.769**	**9.377.604**	**100 %**	**8.74 %**

Tabla 1.5: Población ecuatoriana con título registrado por auto identificación

Figura 1.1: Población ecuatoriana con título registrado por auto identificación

Es importante también recalcar una de las problemáticas actuales que atraviesa la sociedad ecuatoriana, debido a que en los últimos años se ha evidenciado un importante crecimiento en la demanda de cupos universitarios, que las actuales instituciones de educación superior no han sido capaces de cubrir, ocasionando que un elevado número de bachilleres no consigan obtener una plaza para continuar sus estudios universitarios. Tal como como muestra la tabla 1.6, en donde se presentan el número de cupos ofertados y el número de postulantes en el periodo comprendido entre el primer semestre 2016 y el segundo semestre 2017, de acuerdo a los datos

proporcionados por el Área de la Gestión de Información (SENESCYT) solicitados mediante petición formal para fines académicos presentes aquí. [12]

Periodo	Cupos ofertados	Postulantes	Brecha	Porcentaje Brecha
1er. Semestre 2016	87.981	136.260	48.279	35,4 %
2do. Semestre 2016	100.931	163.408	62.477	38,2 %
1er. Semestre 2017	91.173	172.495	81.322	47,1 %
2do. Semestre 2017	111.753	190.376	78.623	41,3 %

Tabla 1.6: Cupos ofertados vs número de postulantes

Figura 1.2: Cupos ofertados vs número de postulantes

De acuerdo a la Figura 1.2, los cupos ofertados no son suficientes para cubrir la demanda de postulantes, provocando una brecha creciente e importante entre la oferta y la demanda de cupos universitarios en los periodos analizados. En el 2016, en promedio 55.378 (36,8 %) y en el 2017, en promedio 79.973 (44,2 %) cupos de los postulantes no podrían ser cubiertos.

Considerando que en Ecuador no existen suficientes estudios técnicos enfocados a solucionar la problemática descrita, surge la necesidad de disponer de una herramienta analítica, matemática o estadística, que permita estimar el máximo nivel de instrucción que alcanzará un determinado estudiante, mediante la utilización de información relacionada con el nivel socio-económico, aspectos demográficos y nivel educativo alcanzado, la misma que representaría un aporte técnico importante en el proceso de acceso a instituciones de educación, en especial el acceso a instituciones de educación superior, tema que será abordado en el presente estudio, con el fin de

generar estrategias para atenuar problemas ocasionados por oferta insuficiente de cupos universitarios e incrementar el acceso universitario de poblaciones minoritarias, ya sea por su etnia o su nivel socio-económico.

Entonces se puede construir una herramienta analítica para estimar el máximo nivel de educación alcanzado por un determinado estudiante, pudiendo ser Ninguno, Primaria, Secundaria o Superior, constituye un aporte técnico importante a la gestión de procedimientos relacionados con el acceso a los distintos niveles de educación, debido a que un pronóstico adecuado del nivel de educación máximo que se alcanzará, dependiendo de las características sociales, económicas, demográficas, etc. del estudiante, permite generar estrategias cuyo propósito principal sea el de mitigar los principales problemas sociales ocasionados por el reducido acceso a los distintos niveles de educación.

El modelo construido tendrá la capacidad de estimar el número de estudiantes más propensos a alcanzar cierto nivel de instrucción, en el presente estudio se dará mayor énfasis al nivel de educación Superior, permitiendo al Ente Regulador diseñar políticas de gestión en el procedimiento de acceso a entidades de educación superior, en particular en la asignación de cupos universitarios, el cual es uno de los problemas más importantes que atraviesa actualmente el país, en lo que se refiere al sistema educativo.

Adicionalmente la identificación adecuada de los factores sociales, económicos y demográficos más influyentes en el nivel de educación que alcanza un estudiante, permitirá conocer los pilares fundamentales en los cuales los entes reguladores deberían enfocarse con el fin de incrementar las tasas de acceso a la educación de poblaciones minoritarias de diferente etnia, status económicos y lugar de procedencia.

Los modelos de regresión logística multinomial son ampliamente utilizados en la estimación o predicción de una variable dependiente cualitativa con más de 2 categorías, conocidas como variables politómicas, las cuales pueden ser de tipo nominal u ordinal.

Cuando se ordenan las categorías de respuesta, se puede ejecutar un modelo de regresión multinomial. La desventaja es que está descartando información sobre el objetivo. Un modelo de regresión logística ordinal conserva esa información, pero es un poco más complicado[6].

En el libro de David G. Kleinbaum, "Logistic Regression" en el capítulo 13, deja

[6]Logistic Regression Models for Multinomial and Ordinal Variables. The Analysis Factor.

claro que la regresión logística ordinal, a diferencia de la regresión politómica, tiene en cuenta cualquier orden inherente de los niveles o variable de resultado, haciendo así un uso más completo de la información ordinal. Así en el presente estudio se tiene variables ordinales, ya que tienen un orden natural entre los niveles, por lo tanto se utilizará un modelo de regresión logística ordinal para estimar la variable dependiente que describe el máximo nivel de educación alcanzado que presenta las siguientes categorías: Ninguno, Primaria, Secundaria, Superior.

Entonces se va a estudiar el modelo de regresión logística ordinal para obtener un mejor análisis y una mejor solución para el problema propuesto. El modelo de regresión logística ordinal se ajusta perfectamente a los requerimientos técnicos deseados, debido a que, aparte de estimar la propensión que presenta un individuo de alcanzar cada nivel de educación, permite identificar y medir el nivel de la influencia de los principales factores sociales, económicos y demográficos relacionados con el acceso a educación de cualquier nivel.

En la actualidad se conocen varias metodologías posibles para abordar problemas de estimación de variables dependientes ordinales con más de 2 categorías posibles, entre las más utilizadas e importantes tenemos las técnicas paramétricas, tales como los modelos de regresión logística ordinal y las técnicas no paramétricas como los árboles de decisión (sin la necesidad de asumir distribuciones a priori). Dependiendo de la naturaleza del problema y la aplicación que tendrá el modelo final obtenido, es posible elegir una de entre las distintas metodologías posibles.

Las ventajas de la utilización de la técnica de regresión logística ordinal son principalmente: medir la influencia de cada una de las variables explicativas incluidas, simplicidad en la interpretación del modelo estimado, facilidad de implementación del modelo para futuras ejecuciones, y finalmente niveles de predicción bastante acertados.

Diferentes estudios históricos, en diversas áreas de la educación, han permitido conocer las bondades de la utilización de la regresión logística ordinal, por ejemplo Liu y Koirala (2012) utilizan esta técnica para estimar diferentes niveles de competencia matemática. Estudios similares como el de Chen, C. y Hughes, J. (2004) utilizan un modelo de regresión logística ordinal para establecer una relación entre el nivel de satisfacción de estudiantes (relacionada a la experiencia universitaria en general) y variables explicativas tanto demográficas como factores referentes al entorno de aprendizaje.

El modelo de regresión logística ordinal puede aplicarse en diferentes campos del análisis, entre ellos: educación, medicina, marketing, siniestralidad, etc. Podemos citar varios estudios adicionales: en medicina, Silva y Otros (2008) emplean el modelo de regresión logística ordinal para estimar el nivel de calidad de vida de pacientes con esquizofrenia; en nutrición, Sumonkanti y Rajwanur (2011) utilizan un modelo de regresión logística ordinal para identificar los factores determinantes de la malnutrición infantil con datos de encuestas demográficas y de salud; en predicción de siniestralidad, Quispe (2016) utiliza un modelo de regresión logística ordinal para predecir el nivel de gravedad en accidentes de tránsito.

En comparación con los modelos de regresión logística multinomial, los modelos ordinales son los adecuados cuando existe una jerarquía en los niveles de la variable dependiente, y respecto a la regresión logística binaria que genera perdida de información al definir dos categorías posibles en la variable dependiente, los modelos ordinales no registran esta limitante (Silva y otros, 2008).

En el presente trabajo el objetivo es construir un modelo analítico de regresión logística ordinal que permita estimar el máximo nivel de educación que alcanzará un determinado individuo, pudiendo ser Ninguno, Primario, Secundario o Superior identificando los factores determinantes socio-económicos y demográficos más influyentes para el máximo nivel de educación alcanzado.

La información necesaria para la construcción del modelo se obtendrá de fuentes públicas, entre las principales podemos enumerar: Censo de Población y Vivienda 2010, boletines estadísticos del INEC, SENECYT, etc. Para la construcción del modelo analítico se empleará como herramienta fundamental el lenguaje de programación estadística R (R Core Team, 2018), que actualmente es muy utilizada en la solución de problemas de este tipo y también dado que la base se encuentra en el programa estadístico SPSS, se realizará algunos análisis con este programa mencionado, en el apéndice B se encuentra la información de estos paquetes estadísticos.

Capítulo 2

Aspectos metodológicos

En el presente capítulo se va desarrollar los modelos de regresión binaria abordando el modelo logit, para así comprender mejor el modelo de regresión logística ordinal el cual es ampliamente utilizado cuando se tiene una variable dependiente cualitativa con más de 2 categorías de tipo ordinal, como se ha expuesto en el capítulo 1 y también se explicará ésto en la siguiente sección 2.2 del mismo. Además, se realiza detalladamente la formulación del modelo econométrico, estudiando las suposiciones, estimaciones e interpretación del modelo.

2.1. Modelos de regresión de respuesta cualitativa

Se realizará una breve revisión de los modelos de regresión binarios, ya que el modelo de regresión logística binario se puede modificar para aplicar a una variable dependiente de tipo ordinal, definiendo las probabilidades de ocurrencia de modo diferente. Esta revisión esta basada en un resumen del libro de Damodar N. Gujarati y Dawn C. Porter [21].

En los modelos en donde Y (variable dependiente) es cualitativa, el objetivo es encontrar la probabilidad de que un acontecimiento suceda.

Comenzaremos el estudio de los modelos con respuesta cualitativa, en primer lugar, el modelo de regresión con respuesta binaria. Hay algunos métodos para crear un modelo de probabilidad para una variable de respuesta binaria como:

1. El modelo lineal de probabilidad (MLP)
2. El modelo logit

2.1.1. Modelos binarios

1. **Modelo lineal de probabilidad (MLP)**

   Para establecer las ideas, considere el siguiente modelo simple:

   $$Y_i = \beta_1 + \beta_2 X_i + \mu_i \tag{2.1}$$

   donde X_i el ingreso familiar, y $Y_i = 1$ si la familia tiene casa propia y 0 si la familia no tiene casa propia.

   El modelo 2.1 parece un modelo de regresión lineal cualquiera, pero debido a que la variable dependiente es binaria, o dicótoma, se le denomina modelo lineal de probabilidad (MLP). Esto es porque la expectativa condicional de Y_i dado X_i, $E(Y_i|X_i)$ puede interpretarse como la probabilidad condicional de que el suceso tenga lugar dado X_i; es decir, $Pr(Y_i = 1|X_i)$. Así, en el ejemplo, $E(Y_i|X_i)$ da la probabilidad de que una familia tenga casa propia y perciba ingresos por una cierta cantidad X_i.

   La esperanza condicional del modelo 2.1 es más, se interpreta como la probabilidad condicional de Y_i, es decir: $E(Y_i|X_i) = \beta_1 + \beta_2 X_i = P_i$; con el supuesto de que $E(\mu_i) = 0$ y donde P_i es la probabilidad de que ocurra el evento con $Y_i = 1$ y $(1 - P_i)$ de que no ocurra con $Y_i = 0$. Es decir, Y_i sigue la distribución de probabilidades de Bernoulli.

   El MLP tiene algunos problemas, como:

   a) la no normalidad de los u_i

   b) la heteroscedasticidad de u_i

   c) la posibilidad de que Y_i se encuentre fuera del rango [0,1]

   d) los valores generalmente bajos de R^2

   Sin embargo estos problemas se pueden resolver, en el texto utilizado se explica detalladamente. [21]

   El problema fundamental con el MLP es que lógicamente no es un modelo muy atractivo porque supone que $P_i = E(Y = 1|X)$ aumenta linealmente con X, es decir, el efecto marginal o incremental de X permanece constante todo el tiempo. En verdad se esperaría que P_i estuviera relacionado en forma no lineal con X_i: con ingresos muy bajos, una familia no será propietaria de una casa, pero en un nivel de ingresos lo bastante altos, por ejemplo, X^*, es

muy probable que sí tenga casa propia. Cualquier incremento en el ingreso más allá de X^* tendrá un efecto pequeño sobre la probabilidad de tener casa propia. Así, en ambos extremos de la distribución de ingresos, la probabilidad de ser dueño de una casa prácticamente no se verá afectada por un pequeño incremento en X.

Por consiguiente, lo que necesitamos es un modelo (probabilístico) que tenga estas dos características:

1) a medida que aumente Xi, $Pi = E(Y = 1|X)$ también aumente pero nunca se salga del intervalo [0;1], y

2) la relación entre Pi y Xi sea no lineal, es decir, "uno se acerca a cero con tasas cada vez más lentas a medida que se reduce Xi, y se acerca a uno con tasas cada vez más lentas a medida que Xi se hace muy grande".[1]

En términos geométricos, el modelo que deseamos tendría la forma de S, o sigmoidea, en este modelo la probabilidad se encuentra entre 0 y 1, y que éste varía en forma no lineal con X.

Además la curva en forma de S, o sigmoidea, se va a parecer mucho a la función de distribución acumulativa de una variable aleatoria (FDA). Por consiguiente,se puede utilizar fácilmente la FDA en regresiones de modelos en los cuales la variable de respuesta es dicótoma, para adquirir valores 0-1. La pregunta práctica ahora es, ¿cuál FDA?, aunque todas las FDA tienen forma de S, para cada variable aleatoria hay una FDA única. Por razones tanto históricas como prácticas, las FDA que suelen seleccionarse para representar los modelos de respuesta 0-1 son: 1) la logística y 2) la normal; la primera da lugar al modelo logit, y la última, al modelo probit (o normit).

2. **El modelo logit**

Se considera ahora la siguiente representación:

$$P_i = \frac{1}{1 + e^{-Z_i}} = \frac{e^Z}{1 + e^Z} \tag{2.2}$$

donde $Z_i = \beta_1 + \beta_2 X_i$

La ecuación 2.2 representa lo que se conoce como función de distribución logística (acumulativa).

Ahora Z_i se encuentra dentro de un rango de $-\infty$ a $+\infty$, P_i se encuentra dentro de un rango de 0 a 1, y que P_i no está linealmente relacionado con Zi (es decir, con X_i), lo que satisface los dos requisitos considerados antes. Pero parece que

al satisfacer estos requisitos creamos un problema de estimación, porque P_i es no lineal no sólo en X sino también en las β, como se ve a partir de 2.2 pero puede linealizarse, lo cual se demuestra de la siguiente manera:

Si Pi es la probabilidad de tener casa propia, como en la ecuación 2.2, entonces $(1 - Pi)$ es la probabilidad de no tener casa propia, así:

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \tag{2.3}$$

Ahora la razón de las probabilidades en favor de tener una casa propia es $\frac{P_i}{1-P_i}$, así se tiene:

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{Z_i}}{1 + e^{-Z_i}} = e^{Z_i} \tag{2.4}$$

Finalmente, tomando el logaritmo natural de la ecuación 2.4, se tiene que:

$$L_i = ln\left(\frac{P_i}{1 - P_i}\right) = Z_i \tag{2.5}$$

$$= \beta_1 + \beta_2 X_i \tag{2.6}$$

L se llama logit, y de aquí el nombre del modelo logit.

2.2. Modelos ordinales

En el modelo anteriormente expuesto logit, el interés residía en construir un modelo para una variable de respuesta del tipo sí o no. Sin embargo la variable dependiente puede tener más de dos resultados, y con mucha frecuencia son ordinales por naturaleza; es decir, no pueden expresarse en una escala de intervalo. Suele suceder que en las investigaciones del tipo de encuesta las respuestas se den en una escala de tipo Likert, por ejemplo, “totalmente de acuerdo”, “algo de acuerdo”, o “desacuerdo”. O las respuestas de una encuesta sobre educación quizá sean “menor a la educación media superior”, “educación media superior”, “licenciatura”, “posgrado”.

Dado que se quiere estudiar este tipo de respuetas, las cuales son codificadas, como por ejemplo: 0 (menor a la educación media superior), 1 (educación media superior), 2 (licenciatura) y 3 (posgrado), por lo tanto estas son escalas ordinales, pues hay un orden claro y natural entre las categorías, pero no se puede decir que 2

(licenciatura) es dos veces 1 (educación media superior), o que 3 (posgrado) es tres veces 1 (educación media superior), pues no tendría sentido.[21]

En el presente proyecto de investigación se estudiará el modelo de regresión logística ordinal "donde el logit es el logaritmo de la razón entre las probabilidades acumuladas de dos segmentos de la escala, divididos arbitrariamente por determinados puntos de corte", se sustenta en el supuesto de que la variable dependiente Y ordinal representa la medida en bruto de una variable de intervalo o de razón subyacente, aunque esta suposición no se necesita para justificar su aplicación (McCullagh y Nelder, 1989).[28] Se supone que el efecto de las variables predictoras es el mismo en cada categoría de la variable dependiente Y. Es lo que se denomina supuesto de paralelismo entre las líneas de regresión.

El modelo de regresión logística cuando la variable dependiente es ordinal fue llamado originalmente modelo logit acumulativo (Walker y Duncan, 1967) [50] y posteriormente McCullagh (1980) [27] lo denominó modelo de razones proporcionales. Se considera la probabilidad de dicho suceso y de los demás sucesos que se encuentran antes o después del mismo, en el orden que se esté analizando las categorías previamente ordenadas para el análisis.

A pesar de que los valores de las categorías se ordenan en rangos, la distancia real entre las categorías es desconocida. Este modelo se usa para estimar las *Odds* de estar por encima o debajo de un determinado nivel de la variable de dependiente Y.

2.2.1. Formulación del modelo de regresión logit ordinal

Sea Y una variable de respuesta categórica, con $y_1, ..., y_g$ categorías a partir de variables explicativas $X = (X_1, ..., X_m)$, con $g \geq 3$ y siendo una regresión ordinal, se obtiene el siguiente modelo:

$$f(\gamma_j(X)) = f(P(Y \leq y_j|X)) \tag{2.7}$$

donde $j = 1, ..., g-1$, f es la función de enlace $(Logit, Log-Log, probit)$ y además notar que $\gamma_j(X) = P(Y \leq y_j|X)$.

Ahora las funciones de enlace pueden ser:

Logit

$$f(Y) = Logit(P(Y)) = log\left(\frac{P(Y)}{1-P(Y)}\right) = logOdds(Y) \tag{2.8}$$

Así al cociente entre ambos se le denomina Odds, es decir se compara la probabilidad de ocurrencia de un evento con la probabilidad de que no ocurra.

Log-Log

$$f(Y) = Log[-Log(1 - P(Y))] \tag{2.9}$$

Probit

$$f(Y) = \Phi^{-1}(P(Y)), \Phi(x) = P(Z \sim N(0,1) \leq x) \tag{2.10}$$

La función de enlace *Logit* es adecuado cuando la respuesta esta uniformemente representada,así en adelante se considera el enlace *Logit*.

Ahora se presenta como se van a dedifinir las probabilidades sobre el modelo:

$$P(Y \leq y_j|X) = \frac{1}{1 + e^{-(\alpha_j - \beta^T X)}}$$

$$\begin{aligned} 1 - P(Y \leq y_j|X) &= 1 - \frac{1}{1 + e^{-(\alpha_j - \beta^T X)}} \\ &= \frac{e^{-(\alpha_j - \beta^T X)}}{1 + e^{-(\alpha_j - \beta^T X)}} \\ &= P(Y > y_j|X) \end{aligned}$$

Al analizar las probabilidades *odds* se tiene lo siguiente:

$$Odds = \frac{P(Y \leq y_j|X)}{1 - P(Y \leq y_j|X)} = \frac{P(Y \leq y_j|X)}{P(Y > y_j|X)} = \frac{\frac{1}{1+e^{-(\alpha_j - \beta^T X)}}}{\frac{e^{-(\alpha_j - \beta^T X)}}{1+e^{-(\alpha_j - \beta^T X)}}}$$

$$Odds = e^{(\alpha_j - \beta^T X)}$$

Así la función de enlace logit, es la siguiente:

$$f(Y) = Logit(P(Y)) = log\left(\frac{P(Y)}{1 - P(Y)}\right) = logOdds(Y) = \alpha_j - \beta^T X \tag{2.11}$$

Finalmente, se formula el modelo logit como sigue:

Sean Y la variable dependiente ordinal, con $g \geq 3$ categorías tales que $y_1, ..., y_g$, X es el vector de variables explicativas $(X_1, X_2, ..., Xm)$ con $k = 1, ..., m$ se tiene el siguiente modelo lineal:

$$f(\gamma_j(X)) = f(P(Y \leq y_j|X)) = \alpha_j - \beta^T X \tag{2.12}$$

donde: $j = 1, ..., g-1$, $P(Y \leq y_j|X)$ es la probabilidad acumulada del evento $(Y \leq j)$, f es conocida como la función de enlace, la cual para el presente proyecto la función de enlace es: $logit$, $\alpha_j - \beta^T X$ que es el predictor lineal en el cual α_j son los parámetros de los interceptos desconocidos, que satisfacen la condición de $\alpha_1 \leq \alpha_2 \leq ... \leq_{g-1}$ y $\beta = (\beta_1, ..., \beta_m)^T$ también son parámetros a estimar.

Además notar que β no depende de las y_j categorías ordinales, de forma que el modelo asume que la relación entre X y Y, es independiente de las y_j categorías ordinales, y es por esto el modelo también se denomina de razones proporcionales porque el log de la $OddsRatio$ es idéntico a lo largo de los intervalos entre los puntos de corte de las y_j categorías.

El modelo proporciona las probabilidades:

Acumuladas:

$$P(Y \leq y_j|X) = [1 + e^{-(\alpha_j - \beta^T X)}]^{-1}$$

Absolutas:

$$P(Y = y_j|X) = P(Y \leq y_j|X) - P(Y \leq y_{j-1}|X)$$

Así si existen j niveles de una variable dependiente de tipo ordinal, el modelo llevará a cabo $j-1$ predicciones, cada una estimando las probabilidades acumuladas en un determinado nivel o por debajo de la categoría g de la variable dependiente Y.

Dado que se tiene los odds, donde los $Odds = \frac{P(Y)}{1-P(Y)}$, es decir los odds son una razón de probabilidad, indicando que la razón entre las dos probabilidades de ocurrencia/no ocurrencia de un evento es grande o pequeña, con la cual se puede analizar la razón entre Odds, a esto se le llama $OddRatio(OR)$, lo cual permite comparar el pronóstico bajo dos condiciones distintas.

Una vez que el modelo de probabilidades proporcionales se ajusta y los parámetros se estiman, el proceso para calcular la razón de probabilidades es el mismo que en la regresión logística estándar.

Primero consideraremos el caso especial donde la exposición es la única variable independiente y se codifica 1 y 0.

Sea X_k una variable independiente tal que: $X_k = 1$ o $X_k = 0$, se tiene:

$$Odds(Y \leq y_j) = \frac{P(Y \leq y_j|X_k)}{P(Y > y_j|X_k)} = \exp\left(\alpha_j - \beta_k X_k\right)$$

$$OR = \frac{\frac{P(Y \leq y_j|X_k=1)}{P(Y>y_j|X_K=1)}}{\frac{P(Y \leq y_j|X_k=0)}{P(Y>y_j|X_K=0)}} = \frac{\exp\left(\alpha_j - \beta_k(1)\right)}{\exp\left(\alpha_j - \beta_k(0)\right)} = \frac{exp(\alpha_j - \beta_k)}{\exp\left(\alpha_j\right)} = e^{(\beta_k)}$$

Y de esto se puede analizar mejor la interpretación de los estimadores la cual se expone más adelante.

Finalmente, al expandir el modelo para agregar más variables independientes es sencillo, como se demostró anteriormente para llegar al modelo, ahora con la última idea desarrollada, el modelo se puede resumir como sigue:

$$P(Y \leq y_j|X) = \frac{1}{1 + \exp\left[-\left(\alpha_j - \sum_{k=1}^{m} \beta_k X_k\right)\right]}$$

donde: $j = 1, 2, .., g-1$ con g el número de categorías.

Notar que: $P(Y \leq g|X) = 1$

Las *odds* para el resultado menor o igual que el nivel y_j son entonces e de la cantidad α_j más la suma de las X_k variables independientes por su respectivo coeficiente β_k.

$$Odds = \frac{P(Y \leq y_j|X)}{P(Y > y_j|X)} = \exp\left(\alpha_j - \sum_{k=1}^{m} \beta_k X_k\right)$$

Los Odds Ratio (OR) se calcula de la forma usual, con la exponencial del coeficiente β_k, si X_k es decodificada con 0 o 1.Al igual que en la regresión logística estándar, el uso de múltiples variables independientes permite la estimación de una razón de probabilidades para una variable que controla los efectos de las otras covariables en el modelo.

$$OR = \exp\left(\beta_k\right), \quad si \quad X_k \quad es \quad codificada \quad (0,1)$$

2.2.2. Suposiciones y estimaciones

Se tienen las siguientes suposiciones sobre el modelo:

1. Residuos

   No se supone normalidad, homocedasticidad ni incorrelación de los residuos.

2. Datos

   El modelo de regresión ordinal supone una condición sobre los datos a modelar: odds proporcionales o líneas paralelas

   El modelo de probabilidades proporcionales es una suposición importante. Según este modelo, la odds ratio que evalúa el efecto de una variable de exposición para cualquiera de estas comparaciones será la misma independientemente de dónde se realice el punto de corte.Esta suposicón será más estudiada en el capitulo 4.

   Además, con lo desarrollado del modelo se deduce que:

$$\widehat{OR}_{\Delta X_k=c}(Y \leq y_j) = e^{\beta_k c}$$

   para todo $j = 1, ..., g-1$

Para la estimación y significación de los parámetros del modelo, se tiene lo siguiente:

Los parámetros del modelo se pueden estimar por máxima verosimilitud, maximizando numéricamente la función de verosimilitud:

$$L(\alpha, \beta|Y, X) = \prod_{i=1}^{n}\prod_{j=2}^{g-1}\left[\frac{1}{1+e^{-(\alpha_1+\beta' X_i)}}\right]^{\delta_i 1}\left[\frac{1}{1+e^{-(\alpha_j+\beta' X_i)}} - \frac{1}{1+e^{-(\alpha_{j-1}+\beta' X_i)}}\right]^{\delta_i j} \tag{2.13}$$

donde:

$$\delta_i j = \begin{cases} 1 & \text{si el i-ésimo individuo muestra} \quad Y = y_j \\ 0 & \text{caso contrario} \end{cases}$$

Se realizará la prueba de Wald para resolver el contraste de hipótesis:

$H_0 : \beta_k = 0$ contra $H_1 : \beta_k \neq 0$; según el estadístico:

$\frac{\widehat{\beta_k}}{\sqrt{\widehat{I_{kk}^{-1}}}} \quad \underset{\sim}{H_0} \quad N(0,1)$

o equivalentemente,

$\frac{\widehat{\beta_k^2}}{\widehat{I_{kk}^{-1}}} \quad \underset{\sim}{H_0} \quad \chi_1^2$

donde I es la matriz de información de fisher.(Ver una explicación más detallada en el apéndice B)

Análogamente, el test de razón de verosimilitudes permite contrastar:

$H_0 : \beta_{k_1} = ... = \beta_{k_s} = 0$ contra $H_1 : \exists r : \beta_{k_r} \neq 0$

2.2.3. Interpretación

Ahora para interpretar los parámetros, se debe considerar que:

$$log\left(\frac{P(Y \leq y_j|X)}{1 - P(Y \leq y_j|X)}\right) = log(Odds(Y)) = \alpha_j - \beta^T X$$

con $j = 1, ..., g-1$ donde:

- α_j

  $$\widehat{logOdds}(Y \leq y_j|X = 0) = \alpha_j \iff e^{\alpha_j} = \widehat{Odds}(Y \leq y_j|X = 0)$$

  Es decir, el intercepto de α_j es el log de $Y \leq y_j$ donde todas las variables independientes X_k son igual a cero.

  Esto es similar a la interpretación de la intercepción para otros modelos logísticos, excepto que, con el modelo de probabilidades proporcionales, estamos modelando las probabilidades de registro de varias desigualdades. Esto produce varias intercepciones, cada una de las cuales corresponde a las probabilidades de registro de una desigualdad diferente (dependiendo del valor de j).[20]

- β_k:

  Si X_k es un factor:

  $$\widehat{logOR}_{X_k}(Y \leq y_j) = \beta_k \iff e^{\beta_k} = \widehat{OR}_{X_k}(Y \leq y_j)$$

  Ahora, para los Odds Ratio, si el valor de OR es menor que uno, lo cual sucede cuando el coeficiente de la variable regresora es negativo, indica que, si las otras variables

explicativas permanecen constantes, los cambios en la variable explicativa analizada incrementan la probabilidad de obtener categorías de mayor valor en la variable objeto del estudio. En cambio, si los valores de Odds Ratio son mayores que uno, esto demuestra que las variaciones en la variable independiente disminuye el riesgo de obtener categorías de mayor valor de la variable objetivo.

Como ya se ha explicado los $Odds = \frac{P(Y)}{1-P(Y)}$, son una razón de probabilidad, indicando que la razón entre las dos probabilidades de ocurrencia/no ocurrencia de un evento es grande o pequeña, una forma más sencilla para comprender mejor de como interpretar los *Odds* es con un ejemplo numérico, así por ejemplo, una $Odds = 3,5$ significa que la probabilidad de las respuestas en la categoría y_j o por debajo de la categoría y_j es $3,5$ veces más probable que la probabilidad de respuestas por encima de dicha categoría.

Capítulo 3

Análisis de datos y construcción del modelo

A partir de la información pública del último Censo de Población y Vivienda (2010) descargada de la página web oficial del INEC, que contiene variables relacionadas con factores sociales, económicos y demográficos, como por ejemplo: discapacidad, edad, provincia o región, etnia, lengua, etc., se construirá un modelo de regresión logística ordinal que permita pronosticar el máximo nivel de educación alcanzado.

3.1. Depuración de la base de datos

La población que será utilizada en la construcción del modelo analítico corresponde a los 14'483.499 habitantes del Ecuador de acuerdo con el último Censo de Población y Vivienda, cuya distribución por nivel de instrucción alcanzado se resume en la Tabla 3.1. Partiendo de la población total se procede a descartar ciertos grupos ya sea por no formar parte del universo de estudio, reducida representatividad o inconsistencias de la información. Las exclusiones que se realizan se enumeran a continuación:

- Se descartan los siguientes grupos: Centro de Alfabetización/(EBA), Pre escolar, Ciclo Postbachillerato, Postgrado, No declara, los cuales conjuntamente concentran a 855.254 sujetos (5,9 % del total).
- Se descarta el grupo de sujetos que no se encontraban en edad de estudiar, es decir menores a cinco años de edad, que son 1'462.277 (10,1 % del total).

De esta manera la población inicial que se utilizará corresponde a 12'165.968 de habitantes luego de las exclusiones realizadas.

Nivel de instrucción más alto alcanzado	Total	Porcentaje
Ninguno	654.682	4,5 %
Centro de Alfabetización/(EBA)	96.411	0,7 %
Pre escolar	140.801	1,0 %
Primario	5.803.415	40,1 %
Secundario	3.954.373	27,3 %
Ciclo Postbachillerato	140.045	1,0 %
Superior	1.753.498	12,1 %
Postgrado	140.459	1,0 %
No declara	337.538	2,3 %
No en edad de estudiar	1.462.277	10,1 %
Total	**14.483.499**	**100,0 %**

Tabla 3.1: Población ecuatoriana por nivel de instrucción alcanzado

Para el proceso de análisis de consistencia de los datos se aplican técnicas de depuración de datos faltantes y determinación de datos atípicos descritas en el trabajo de Castro (2008).[4]

A continuación se eliminan los datos perdidos por el sistema o nulos utilizando el paquete estadístico R, además los datos se ingresan al programa estadístico SPSS, con lo cual finalmente se tiene 9'221.116 millones de registros.

Estadísticos

Nivel_Educativo

N	Válido	9221116
	Perdidos	0

Nivel_Educativo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Ninguno	434394	4,7	4,7	4,7
	Primaria	4332119	47,0	47,0	51,7
	Secundaria	3013534	32,7	32,7	84,4
	Superior	1441069	15,6	15,6	100,0
	Total	9221116	100,0	100,0	

Figura 3.1: Población ecuatoriana según su nivel educativo

Para realizar el modelo se ha dividido la base, en dos partes con una división 80 % y 20 %, obteniendo dos bases de datos con las mismas caraterísticas, esta es una partición apropiada de acuerdo al libro citado de N.Siddiqi [40]. Además con esto se va a realizar las respectivas matrices de confusión, las cuales se exponen en el capítulo 4.

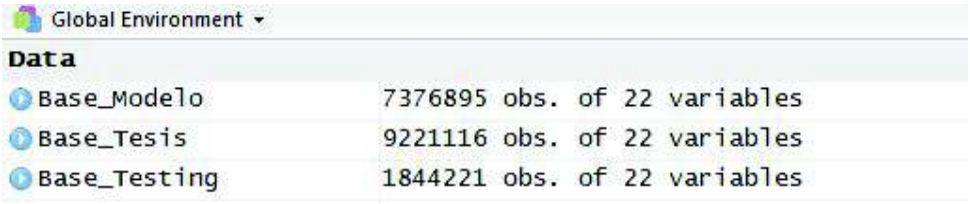


Figura 3.2: Bases de datos

En la figura 3.2 los datos de la partición obtenida del 80 % de los datos totales, se la nombra como base modelo con un total de 7'376.895 millones de registros y al 20 % como base testing con un total de 1'844.221 millones de registros.

3.2. Filtrado y contrucción de variables

La variable dependiente Y será una variable cualitativa ordinal con 4 categorías que describirán el máximo nivel de educación alcanzado por el individuo, la cual se define a continuación:

$$Y = \begin{cases} 0 & \text{Si el nivel de educación máximo alcanzado fue Ninguno} \\ 1 & \text{Si el nivel de educación máximo alcanzado fue Primaria} \\ 2 & \text{Si el nivel de educación máximo alcanzado fue Secundaria} \\ 3 & \text{Si el nivel de educación máximo alcanzado fue Superior} \end{cases}$$

Es importante recalcar que para el presente estudio se descartan títulos de cuarto nivel, PHD, y títulos obtenidos en el extranjero por no formar parte del alcance del análisis.

En la filtración de las variables se emplearán ciertos estadísticos (Arnold y Emerson, 2011) que permiten medir la discriminación generada por una variable explicativa sobre los grupos de la variable dependiente, entre ellos el estadístico de Kolmogorov-Smirnov, KS (Massey, 1951), que se aplica a variables cuantitativas, sin embargo en el análisis de las

variables independientes se tienen variables cualitativas que al realizar la prueba de Chi-cuadrado dichas variables son significativas.

Para la construcción de variables se debe notar que partiendo de las variables originales se procederán a construir nuevas variables que generen una mayor discriminación en los grupos de la variable dependiente, así las variables han sido recodificadas, las cuales se comprueban que las variables tienen una asociación estadísticamente significativa ya que se rechaza H_0 porque el valor p es menor o igual al nivel de significancia, así se rechaza la hipótesis nula y se concluye que hay una asociación estadísticamente significativa para la variable de estudio (se puede ver esto en el apéndice A y el apéndice B).

Por lo tanto, finalmente se tienen las variables para el modelo: Region, Area, Sexo, GrupoEdadRec, TieneSeguroP, TieneDiscapacidad, LeerEscribir, InternetU6M, ComputadorU6M, ComoTrabajaRec, EstadoConyugalRec, Idioma, Etnia. En el apéndice A, se presenta la tabla A.1 con la información de cada variable.

A continuación se describe cada una de las variables cualitativas:

- Region : es una variable nominal que representa 4 regiones naturales del Ecuador que son la Costa (entre el océano Pacífico y la cordillera), la Sierra (la zona andina) y el Oriente o la Amazonía (al este de la cordillera). Además de esto, el país cuenta con la región insular (las islas Galápagos). Así, se tienen las siguientes categorías:
  - Costa
  - Sierra
  - Oriente
  - Insular

Regiones del Ecuador

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Costa	4399687	47,7	47,7	47,7
	Sierra	4357948	47,3	47,3	95,0
	Oriente	447131	4,8	4,8	99,8
	Insular	16350	,2	,2	100,0
	Total	9221116	100,0	100,0	

Figura 3.3: Tabla de frecuencias de las regiones

Figura 3.4: Diagrama de las regiones

En las figuras 3.3 y 3.4 se puede apreciar que donde hay mayor población son en las regiones de la Costa y la Sierra con 47,7 % y 47,3 % respectivamente.

- Area : es una variable nominal que representa las áreas del Ecuador.

  Las áreas urbanas son capitales provinciales y cabeceras cantonales o municipios según la división político administrativa (DPA) vigente en el país, sin tomar en cuenta su tamaño.

  Las áreas rurales incluyen las cabeceras parroquiales, otros centros poblados, las periferias de los núcleos urbanos y la población dispersa.

  Estas definiciones son las oficiales, por ello, todos los resultados de los Censos de población publicados por el INEC asumen esta definición según el Sistema Integrado de Indicadores Sociales del Ecuador (SIISE).[8]

  A continuación se definen las siguientes categorías:

  - Área urbana
  - Área rural

Área urbana o rural

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Area urbana	5914413	64,1	64,1	64,1
	Area rural	3306703	35,9	35,9	100,0
	Total	9221116	100,0	100,0	

Figura 3.5: Tabla de frecuencias del área

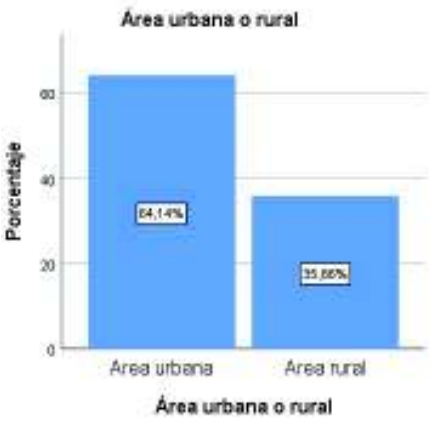


Figura 3.6: Diagrama del área

En las figuras 3.5 y 3.6 se observa que hay mayor población en el área urbana con el 64,1 % que en el área rural con tan solo 35,9 %.

- Sexo : es una variable nominal que representa el sexo del individuo donde la definición con la que trabaja la OMS (Organización Mundial de la Salud) dice que " 'Sexo' se refiere a las características biológicas y fisiológicas que definen a hombres y mujeres, son categorías sexuales"[1]. [11]

  Por lo cual, las categorías de dicha variable quedan como sigue:

  - Hombre
  - Mujer

Cúal es el Sexo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Hombre	4514834	49,0	49,0	49,0
	Mujer	4706282	51,0	51,0	100,0
	Total	9221116	100,0	100,0	

Figura 3.7: Tabla de frecuencias del sexo

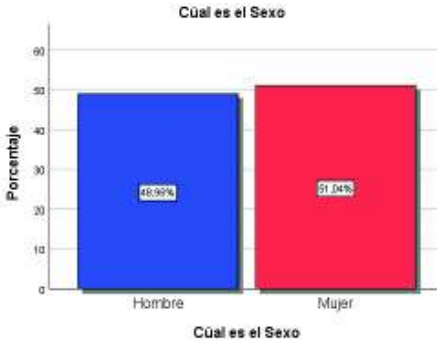


Figura 3.8: Diagrama del sexo

En las figuras 3.7 y 3.8 se refleja que a pesar de que hay más mujeres que hombres, la diferencia es poca ya que hay 49,0 % de hombres y 51,0 % de muejres.

- GrupoEdadRec : es una variable ordinal que representa la división de los grupos de edad que se ha considerado de la población a estudiar, los cuales se definen como sigue:

  - Edad entre 5 y 12
  - Edad entre 13 y 18
  - Edad entre 19 y 25
  - Edad entre 26 y 40
  - Edad mayor a 41

[1] «Gender, equity, human rights». World Health Organization (en inglés británico)

Edad cumplida

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Edad entre 5 y 12	1773731	19,2	19,2	19,2
	Edad entre 13 y 18	1291544	14,0	14,0	33,2
	Edad entre 19 y 25	1313412	14,2	14,2	47,5
	Edad entre 26 y 40	2234890	24,2	24,2	71,7
	Edad mayor a 41	2607539	28,3	28,3	100,0
	Total	9221116	100,0	100,0	

Figura 3.9: Tabla de frecuencias de la edad

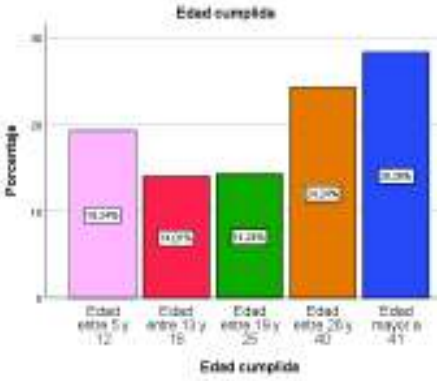


Figura 3.10: Diagrama de la edad

En las figuras 3.9 y 3.10 se muestra los porcentajes de los diferentes grupos de edad, en donde se puede ver que el grupo de edad entre 13 y 18 con el grupo de edad entre 19 y 25, tienen aproximadamente 14 % cada uno.

- TieneSeguroP : es una variable nominal que representa si el individuo posee seguro privado y no caso contrario, con lo cual las categorías de esta variable son:
  - Si
  - No

Tiene seguro de salud privado

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Si	936060	10,2	10,2	10,2
	No	8285056	89,8	89,8	100,0
	Total	9221116	100,0	100,0	

Figura 3.11: Tabla de frecuencias del seguro privado

Figura 3.12: Diagrama del seguro privado

En las figuras 3.11 y 3.12 se puede ver que la mayoría de la población que es el 89,8 % no tiene seguro de salud privado.

- TieneDiscapacidad : es una variable nominal que representa si el individuo tiene discapacidad permanente por más de un año, donde la discapacidad puede ser intelectual, fisico-motora, visual, auditiva y mental.

Por lo tanto, la variable tiene la siguientes categorías:

- Si
- No

Tiene discapacidad permanente por más de un año

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Si	562963	6,1	6,1	6,1
	No	8658153	93,9	93,9	100,0
	Total	9221116	100,0	100,0	

Figura 3.13: Tabla de frecuencias de la discapacidad

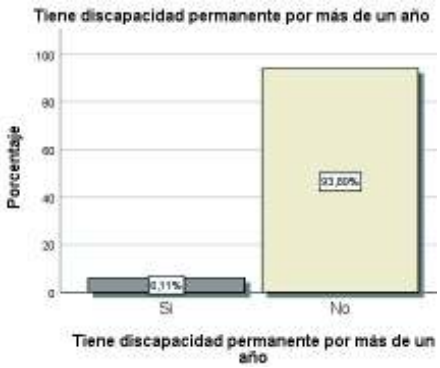


Figura 3.14: Diagrama de la discapacidad

En las figuras 3.13 y 3.14 muestran que el 93,9 % que es la mayoría de la población no tienen discapacidad permanente por más de un año.

- LeerEscribir : es una variable nominal que representa si el individuo sabe leer y escribir, y no caso contrario. Así, se tienen las siguientes categorías:
  - Si
  - No

Sabe leer y escribir

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Si	8646221	93,8	93,8	93,8
	No	574895	6,2	6,2	100,0
	Total	9221116	100,0	100,0	

Figura 3.15: Tabla de frecuencias de saber leer y escribir

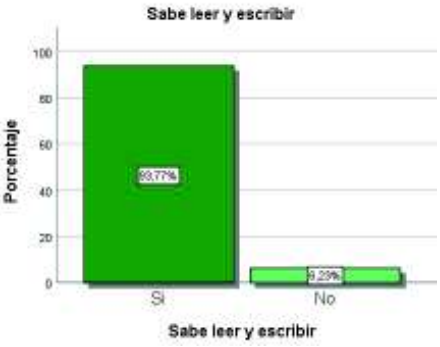


Figura 3.16: Diagrama de saber leer y escribir

En las figuras 3.15 y 3.16 se tiene que el 93,8 % de la población sabe leer y escribir.

- InternetU6M : es una variable nominal que representa si el individuo en los ultimos 6 meses ha utilizado el internet, y no caso contrario. Así, se tiene las siguientes categorías:

  - Si
  - No

En los últimos 6 meses ha utilizado Internet

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Si	2600613	28,2	28,2	28,2
	No	6620503	71,8	71,8	100,0
	Total	9221116	100,0	100,0	

Figura 3.17: Tabla de frecuencias del internet

Figura 3.18: Diagrama del internet

En las figuras 3.17 y 3.18 se puede ver que el 71,8 % de la población no ha usado el internet en los últimos 6 meses, mientras que el resto si.

- ComputadorU6M : es una variable nominal que representa si el individuo en los ultimos 6 meses ha utilizado una computadora y no caso contrario. Así, se tiene las siguientes categorías:

  - Si
  - No

En los útlimos 6 meses ha utilizado Computadora

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Si	3255356	35,3	35,3	35,3
	No	5965760	64,7	64,7	100,0
	Total	9221116	100,0	100,0	

Figura 3.19: Tabla de frecuencias de la computadora

Figura 3.20: Diagrama de la computadora

En las figuras 3.19 y 3.20 se puede ver que el 64,7 % de la población no ha usado la computadora en los últimos 6 meses, mientras que el resto si.

- ComoTrabajaRec : es una variable nominal que representa el cargo o trabajo que ha realizado, dividiendo a la variable en las siguientes categorías:

  - Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales
  - Empleado/a u obrero/a privado
  - Jornalero/a o peón
  - Patron/a
  - Socio/a
  - Cuenta propia
  - Trabajador/a no remunerado
  - Empleado/a doméstico/a
  - No trabaja

En el lugar indicado trabaja o trabajó como

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales	460828	5,0	5,0	5,0
	Empleado/a u obrero/a privado	1503754	16,3	16,3	21,3
	Jornalero/a o peón	565436	6,1	6,1	27,4
	Patron/a	136437	1,5	1,5	28,9
	Socio/a	49230	,5	,5	29,5
	Cuenta propia	1247833	13,5	13,5	43,0
	Trabajador/a no remunerado	63437	,7	,7	43,7
	Empleado/a doméstico/a	168610	1,8	1,8	45,5
	No trabaja	5025551	54,5	54,5	100,0
	Total	9221116	100,0	100,0	

Figura 3.21: Tabla de frecuencias del trabajo

En las figuras 3.21 y 3.22 se observa que los trabajos que son independientes del estado como Empleado/a u obrero/a privado y cuenta propia con 16,3 % y 13,5 % respectivamente, tienen mayor población que los de la categoría Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales con tan solo el 5 % de la población.

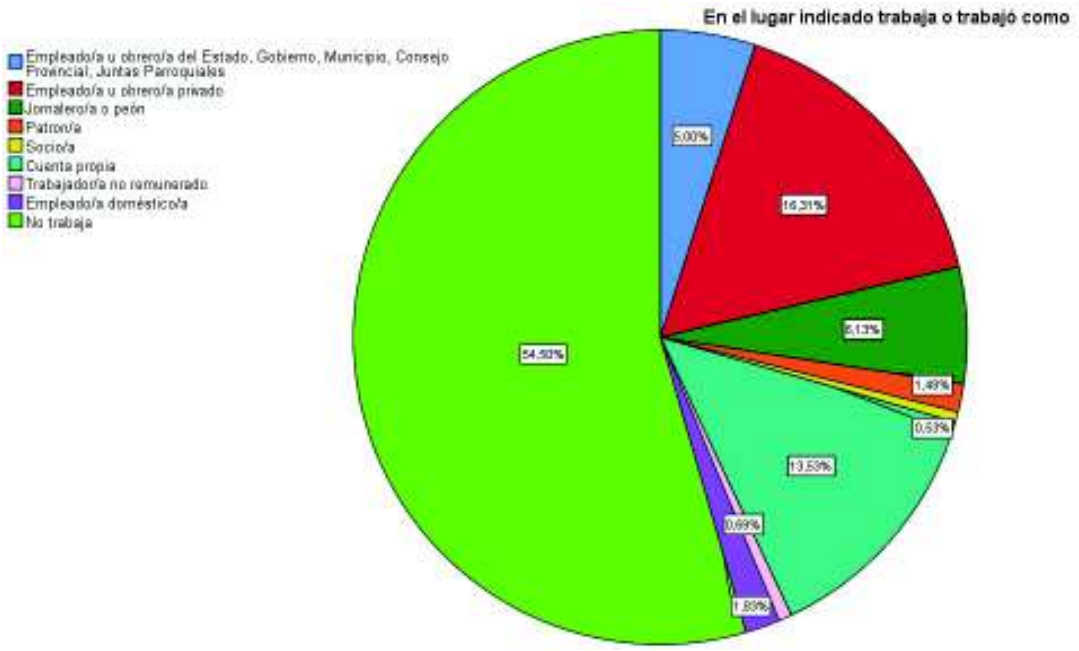


Figura 3.22: Diagrama del trabajo

- EstadoConyugalRec : es una variable nominal que representa el estado civil de cada individuo, donde el estado civil es la condición de una persona según el Registro Civil del Ecuador en función de si tiene o no pareja y su situación legal respecto a esto.

  El estado civil es un dato que se encuentra en la parte frontal de la cédula de identidad del Ecuador, las distinciones del estado civil de una persona pueden ser variables de un estado a otro.[14]

  Los estados civiles disponibles son las siguientes categorías:

  - Casado/a
  - Unido/a
  - Separado/a
  - Divorciado/a
  - Viudo/a
  - Soltero/a
  - No aplica(no tiene ningún estado civil)

Estado Conyugal

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Casado/a	2513532	27,3	27,3	27,3
	Unido/a	1530862	16,6	16,6	43,9
	Separado/a	358452	3,9	3,9	47,7
	Divorciado/a	138835	1,5	1,5	49,3
	Viudo/a	263159	2,9	2,9	52,1
	Soltero/a	2846755	30,9	30,9	83,0
	No aplica	1569521	17,0	17,0	100,0
	Total	9221116	100,0	100,0	

Figura 3.23: Tabla de frecuencias del estado conyugal

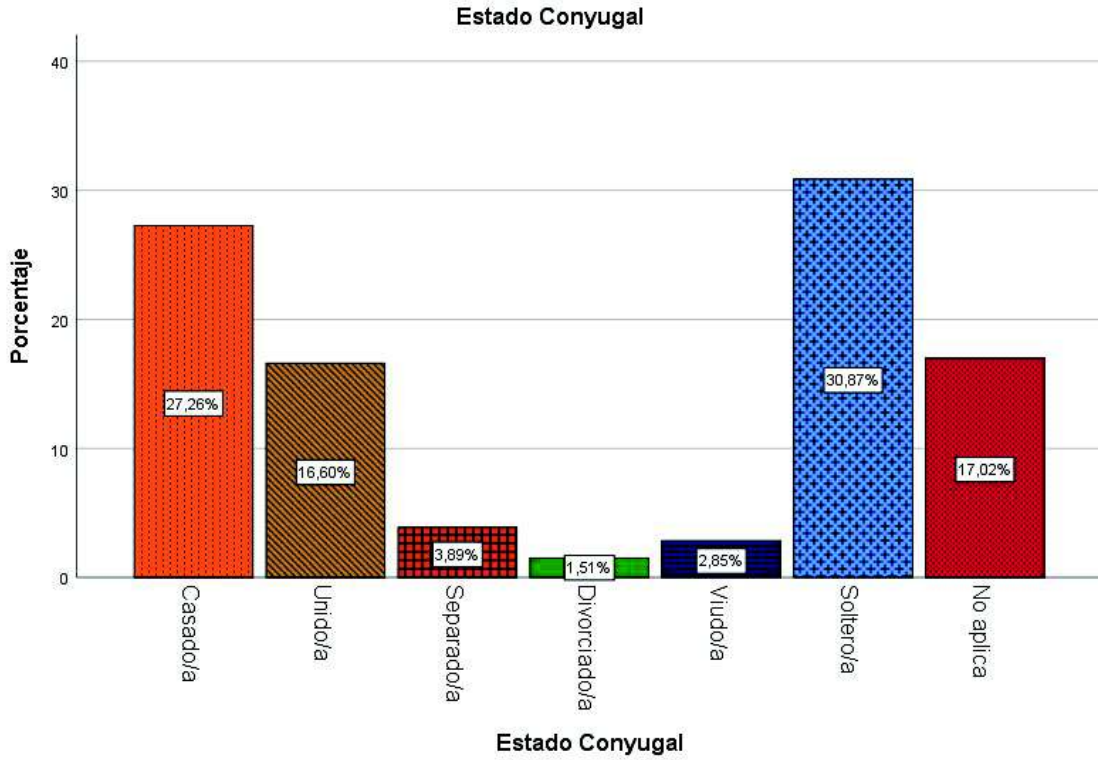


Figura 3.24: Diagrama del estado conyugal

En las figuras 3.23 y 3.24 se puede ver que la mayoría de la población tienen estado civil casada/o con 27,3 %, y soltero/a con 30.9 %.

Se observa también que apenas el 1,5 % de la población se encuentran divorciados/as, lo cual es el grupo con menos individuos.

- Idioma : es una variable nominal que representa el idioma que habla una persona "el idioma es la lengua propia de un pueblo o nación o de varios pueblos y naciones"[2].

  Se puede definir como sigue :"el idioma es un sistema de comunicación lingüístico, que puede ser tanto oral como escrito, y se caracteriza por regirse según una serie de convenciones y normas gramaticales que garantizan la comunicación entre las personas. De allí que idioma y lengua sean términos sinónimos". [41]

  Entonces cada individuo puede hablar uno o más idiomas, así las categorías de esta variable se definen como sigue:

  - Indígena
  - Castellano Español
  - Extranjero
  - Indígena/Castellano Español
  - Indígena/Extranjero
  - Castellano Español/Extranjero
  - Indígena/Castellano Español/Extranjero
  - No habla

  En las figuras 3.25 y 3.26 se puede constatar como era de esperarse que la mayoría de la población hable el idioma castellano español con el 92,77 %, luego el 2,17 % el idioma indígena, sin embargo hay 2,42 % de personas que hablan ambos idiomas.

  Además, hay un porcentaje bajo de 1,77 % de personas que hablan castellano español y además algún idioma extranjero.

  Las personas que hablan el idioma indígena y algún idioma extranjero son apenas el 0.01 % de la población.

  Las personas que hablan solamente un idioma extranjero son apenas el 0.61 % de la población.

  Finalmente, las personas que hablan los tres idiomas, es decir en la categoría Indígena/Castellano Español/Extranjero, es solamente el 0,03 % de la población.

[2] "Idioma". En: Significados.com. Recuperado de `https://www.significados.com/idioma/`

Idioma que habla recodificado

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Indígena	200159	2,2	2,2	2,2
	Castellano Español	8554095	92,8	92,8	94,9
	Extranjero	55897	,6	,6	95,5
	Indígena/Castellano Español	222823	2,4	2,4	98,0
	Indígena/Extranjero	732	,0	,0	98,0
	Castellano Español/Extranjero	162901	1,8	1,8	99,7
	Indígena/Castellano Español/Extranjero	2910	,0	,0	99,8
	No habla	21599	,2	,2	100,0
	Total	9221116	100,0	100,0	

Figura 3.25: Tabla de frecuencias del idioma

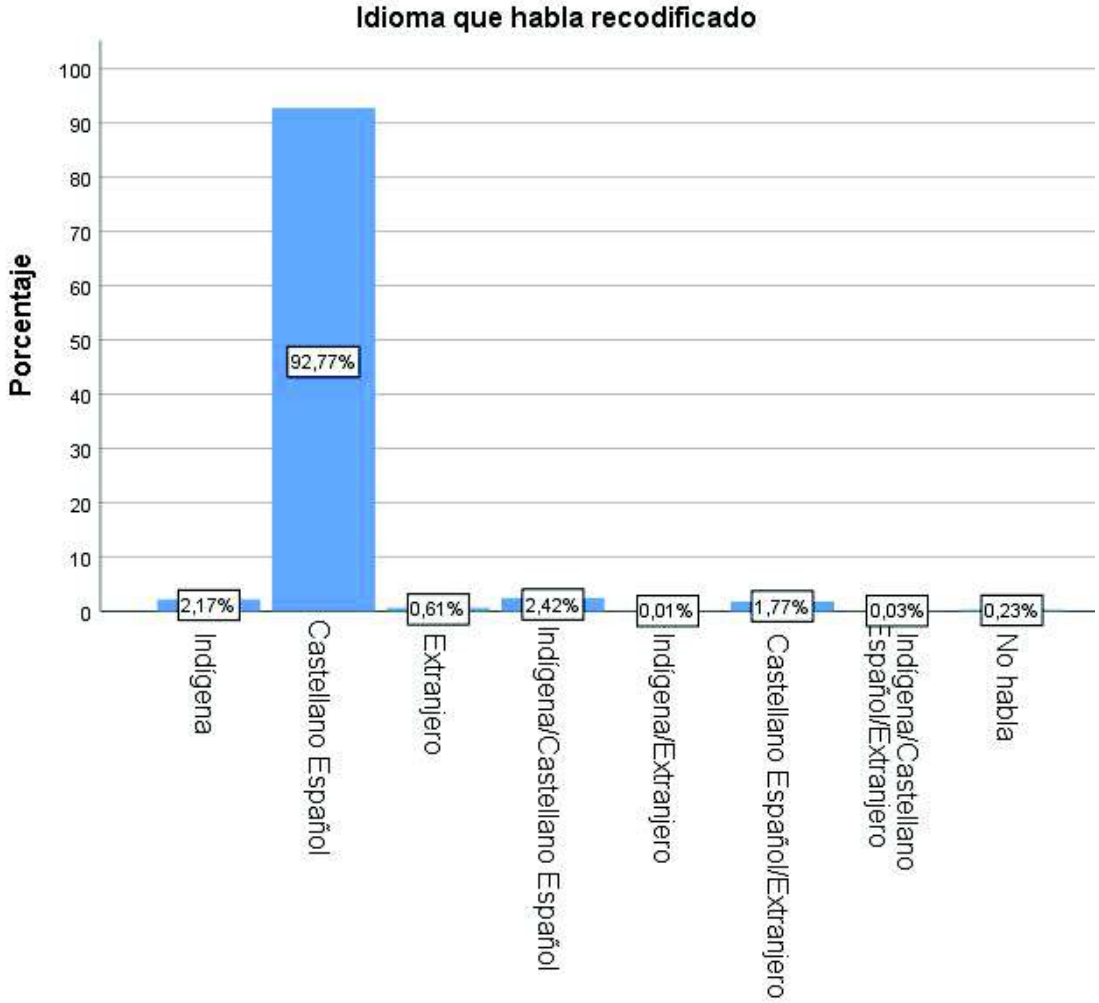


Figura 3.26: Diagrama del idioma

- Etnia : es una variable nominal que representa la autoidentificación de una persona, "para captar la etnicidad de las personas a partir del Censo 2001 y en el proceso de homologación que se realizó con el INEC para definir a la variable étnica, se ha establecido utilizar el concepto de autoidentificación . Esto significa que las personas autocalifican su pertenencia étnica".[9]

  Además, según SIISE, "Etnia se refiere a los valores y prácticas culturales que distinguen a los grupos humanos. Los miembros de un grupo étnico se ven a sí mismos como diferentes a otros grupos. El concepto alude, en general, a dos dimensiones: un conjunto compartido de características culturales y sociales (lengua, fe, residencia, etc.) y un sentido compartido de identidad o tradición".

  Por lo tanto, se han definido las siguientes categorías:

  - Indígena
  - Afroecuatoriano/Afrodescendiente, Negro, Mulato
  - Montubio
  - Mestizo
  - Blanco

Autoidentificación étnica recodifica

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Indígena	611500	6,6	6,6	6,6
	Afroecuatoriano/Afrodescendiente, Negro, Mulato	627043	6,8	6,8	13,4
	Montuvio	653105	7,1	7,1	20,5
	Mestizo	6780173	73,5	73,5	94,0
	Blanco	549295	6,0	6,0	100,0
	Total	9221116	100,0	100,0	

Figura 3.27: Tabla de frecuencias de la etnia

En las figuras 3.27 y 3.28 se tiene que la mayoría de la población con el 73,5 % se autoidentifican como mestizos, luego para las etnias de los grupos indígena y el grupo de afroecuatoriano/afrodescendiente, negro, mulato tienen el 6,6 % y 6,8 % respectivamente de la población. Además hay 7,1 % de la población que se consideran montuvios y el 6,0 % como blancos.

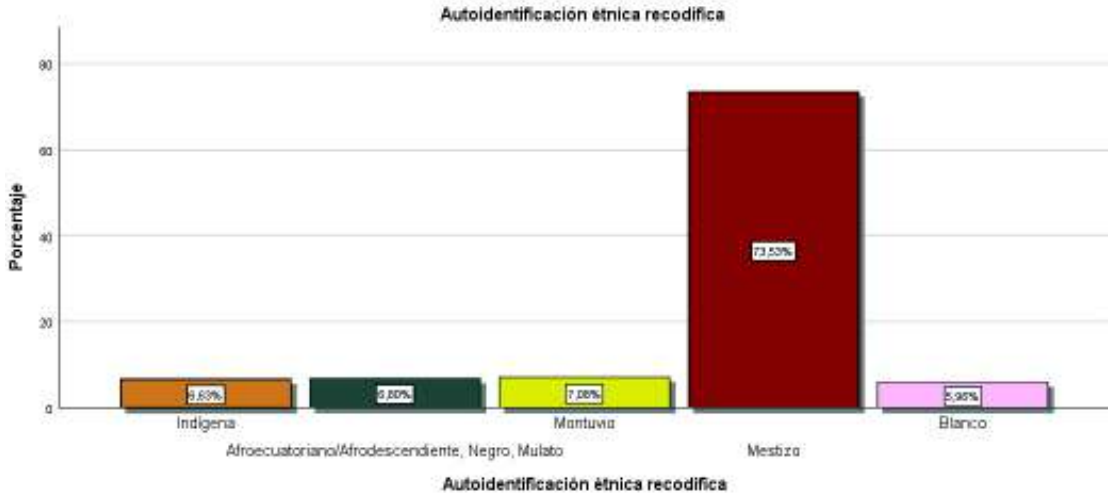


Figura 3.28: Diagrama de la etnia

La correlación para las variables cualitativas son el coeficiente de correlación de Pearson que es una medida de asociación lineal especialmente apropiada para estudiar la relación entre variables de intervalo o razón, y el coeficiente de correlación de Spearman que también es una medida de asociación lineal, pero para variables ordinales.

Sin embargo, "Ambos coeficientes poseen escasa utilidad para estudiar las pautas de relación presentes en una tabla de contingencia típica, dado que lo habitual es utilizar las tablas de contingencia para cruzar variables de tipo nominal, o a lo sumo, de tipo ordinal con solo unos pocos niveles".[10]

Las tablas de contingencia o tablas cruzadas realizadas con la variable de estudio están expuestas en el apéndice A.

Una vez generada la estructura de variables constituida por la variable dependiente Y y X_k con $k = 1, ..., m$ variables independientes, se procederá con la etapa de modelización, dado que ya se ha depurado la base de datos.

El modelo va a ser generado por medio del paquete estadístico R, en el cual también se instala el entorno de R-Studio(ver apendice B) que es donde se va a programar el código, se ajusta el modelo usando la polrfunción del paquete MASS. Además, "Polr" significa Regresión Lineal de Proporciones Proporcionales.

El paquete MASS se lo puede encontrar en la CRAN de R y se procede a instalar en R, (MASS significa Modern Applied Statistics with S , un libro de WN Venables y BD Ripley. R es una implementación de código abierto de S.)[48]

El código del modelo se encuentra expuesto en anexos A, al igual que los resultados, con lo cual se plantea el modelo en la siguiente sección.

3.3. Regresión logística

Sean Y la variable dependiente ordinal con $g \geq 3$ categorías $y_1, ..., y_g$, X es el vector de variables explicativas $(X_1, X_2, ..., Xm)$ con $k = 1, ..., m$ se plantea el modelo lineal:

$$f(\gamma_j(X)) \quad = \quad f(P(Y \leq y_j|X)) = \alpha_j - \beta^T X$$

donde: $j = 1, ..., g-1$, f es conocida como la función de enlace la cual para el presente proyecto la función de enlace es: $logit$, $\alpha_j - \beta^T X$ es el predictor lineal en el cual α_j y $\beta = (\beta_1, ..., \beta_m)^T$ son los parámetros a estimar.

Por medio del paquete R se obtiene el modelo, primero para una mejor explicación se considera: $y_j = 2$, se tiene entonces:

$$\begin{aligned} f(\gamma_2(X)) &= logit(P(Y \leq y_2|X)) \\ &= \alpha_2 - \beta^T X \\ &= \alpha_2 - (\beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m) \\ &= -1.0479 - 0.26Region2 - 0.14Region3 + ... + 0.49Etnia4 + 0.60Etnia5 \end{aligned}$$

En el lado derecho del signo igual vemos un modelo lineal simple con una pendiente β, y una intersección que cambia según y_j y α_j, donde y_j es el nivel de una categoría ordenada con los niveles de $j = 1, 2, ..., g-1$. Por como se ha definido la variable Y, se tiene $g = 4$ categorías por lo tanto en este caso, $j = 1, 2, 3$ no se debe confundir la notación de la respuesta de la variable Y con la de la categoría, para que quede claro se tiene por ejemplo lo siguiente, cuando $Y = 0$ estamos en la primer categoría, es decir: $y_1 = 1$ es la categoría "Ninguno". Entonces vemos que tenemos una intercepción diferente dependiendo del nivel de interés.

Ahora una pregunta intereresante es: ¿Por qué j solo se extiende a $g-1$?

Esto se debe a que se esta modelando la probabilidad de estar en una categoría igual (o inferior) en lugar de estar en categorías por encima de la misma.

Por ejemplo, $P(Y \leq 2)$ significa la probabilidad de ser un individuo con el nivel de educación "Ninguno" o "Primaria" en lugar de ser "Secundaria" o las otras categorías superiores. Por lo tanto, se utiliza los niveles como límites. En este modelo, el nivel más alto devuelve una probabilidad de 1, es decir, $P(Y \leq y_4) = P(Y \leq 4) = 1$, por lo que no lo modelamos.

Variable Categórica	Descripción	β_k	OR
Region2	Sierra	-0.2612	0.7702
Region3	Oriente	-0.1419	0.8677
Region4	Insular	-0.4193	0.6575
Area2	Area rural	-0.6088	0.5440
Sexo2	Mujer	0.1947	1.2150
GrupoEdadRec1	Edad entre 13 y 18	0.8756	2.4002
GrupoEdadRec2	Edad entre 19 y 25	1.9706	7.1751
GrupoEdadRec3	Edad entre 26 y 40	1.6542	5.2290
GrupoEdadRec4	Edad mayor a 41	1.0345	2.8136
TieneSeguroP2	No	-0.2639	0.7680
TieneDiscapacidad2	No	0.6847	1.9831
LeerEscribir2	No	-4.9420	0.0071
InternetU6M2	No	-0.8909	0.4103
ComputadorU6M2	No	-1.0807	0.3393
ComoTrabajaRec2	Empleado/a u obrero/a privado	-1.1002	0.3328
ComoTrabajaRec3	Jornalero/a o peón	-2.5380	0.0790
ComoTrabajaRec4	Patron/a	-0.8620	0.4223
ComoTrabajaRec5	Socio/a	-0.9475	0.3877
ComoTrabajaRec6	Cuenta propia	-1.6918	0.1842
ComoTrabajaRec7	Trabajador/a no remunerado	-1.6420	0.1936
ComoTrabajaRec8	Empleado/a doméstico/a	-2.3669	0.0938
ComoTrabajaRec10	No trabaja	-1.6921	0.1841
EstadoConyugalRec2	Unido/a	-0.3534	0.7023
EstadoConyugalRec3	Separado/a	-0.1681	0.8452
EstadoConyugalRec4	Divorciado/a	0.6095	1.8395
EstadoConyugalRec5	Viudo/a	-0.9414	0.3901
EstadoConyugalRec6	Soltero/a	0.2699	1.3098
EstadoConyugalRec7	No aplica	-1.2581	0.2842
Idioma2	Castellano Español	0.4480	1.5651
Idioma3	Extranjero	0.6332	1.8837
Idioma4	Indígena/Castellano Español	0.2171	1.2424
Idioma5	Indígena/Extranjero	0.5065	1.6594
Idioma6	Castellano Español/ Extranjero	1.3924	4.0245
Idioma7	Indígena / Castellano Español / Extranjero	1.2953	3.6520
Idioma8	No habla	-1.1735	0.3093
Etnia2	Afroecuatoriano /Afrodescendiente, Negro, Mulato	0.1992	1.2204
Etnia3	Montuvio	0.1761	1.1926
Etnia4	Mestizo	0.4944	1.6395
Etnia5	Blanco	0.6055	1.8322

Tabla 3.2: Estimadores del modelo

En la tabla 3.2 se presentan los estimadores β_k del modelo y en la tabla 3.3 se presentan las constantes α_j del modelo.

Descripción	Constante	α_j	Std.error	t	p
Ninguno | Primaria	0 | 1	-7.0015	0.0114	-615.1898	0.00E+00
Primaria | Secundaria	1 | 2	-1.0479	0.0106	-98.4365	0.00E+00
Secundaria | Superior	2 | 3	1.6415	0.0106	154.5749	0.00E+00

Tabla 3.3: Contantes del modelo

Finalmente el modelo queda determinado como sigue:

$$
\begin{aligned}
f(\gamma_1(X)) &= logit(P(Y \leq y_1|X)) = \alpha_1 - \beta^T X \\
f(\gamma_1(X)) &= -7.0015 - \big(- 0.2612 Region2 - 0.1419 Region3 - 0.4193 Region4 \\
&\quad - 0.6088 Area2 + 0.1947 Sexo2 + 0.8756 GrupoEdadRec1 + 1.9706 GrupoEdadRec2 \\
&\quad + 1.6542 GrupoEdadRec3 + 1.0345 GrupoEdadRec4 - 0.2639 TieneSeguroP2 \\
&\quad + 0.6847 TieneDiscapacidad2 - 4.9420 LeerEscribir2 - 0.8909 InternetU6M2 \\
&\quad - 1.0807 ComputadorU6M2 - 1.1002 ComoTrabajaRec2 - 2.5380 ComoTrabajaRec3 \\
&\quad - 0.8620 ComoTrabajaRec4 - 0.9475 ComoTrabajaRec5 - 1.6918 ComoTrabajaRec6 \\
&\quad - 1.6420 ComoTrabajaRec7 - 2.3669 ComoTrabajaRec8 - 1.6921 ComoTrabajaRec10 \\
&\quad - 0.3534 EstadoConyugalRec2 - 0.1681 EstadoConyugalRec3 \\
&\quad + 0.6095 EstadoConyugalRec4 - 0.9414 EstadoConyugalRec5 \\
&\quad + 0.2699 EstadoConyugalRec6 - 1.2581 EstadoConyugalRec7 \\
&\quad + 0.4480 Idioma2 + 0.6332 Idioma3 + 0.2171 Idioma4 + 0.5065 Idioma5 \\
&\quad + 1.3924 Idioma6 + 1.2953 Idioma7 - 1.1735 Idioma8 + 0.1992 Etnia2 \\
&\quad + 0.1761 Etnia3 + 0.4944 Etnia4 + 0.6055 Etnia5 \big)
\end{aligned}
$$

De forma análoga se tiene para los niveles siguientes cambiando solamente la constante, es decir:

$$
\begin{aligned}
f(\gamma_2(X)) &= logit(P(Y \leq y_2|X)) = \alpha_2 - \beta^T X \\
&= -1.0479 - \big(- 0.2612 Region2 - 0.1419 Region3 - 0.4193 Region4 - 0.6088 Area2 \\
&\quad + 0.1947 Sexo2 + 0.8756 GrupoEdadRec1 + 1.9706 GrupoEdadRec2 + \cdots + \\
&\quad + 1.3924 Idioma6 + 1.2953 Idioma7 - 1.1735 Idioma8 + 0.1992 Etnia2 \\
&\quad + 0.1761 Etnia3 + 0.4944 Etnia4 + 0.6055 Etnia5 \big)
\end{aligned}
$$

$$
\begin{aligned}
f(\gamma_3(X)) &= logit(P(Y \leq y_3|X)) = \alpha_3 - \beta^T X \\
&= 1.6415 - \big(- 0.2612 Region2 - 0.1419 Region3 - 0.4193 Region4 - 0.6088 Area2 \\
&\quad + 0.1947 Sexo2 + 0.8756 GrupoEdadRec1 + 1.9706 GrupoEdadRec2 + \cdots + \\
&\quad + 1.3924 Idioma6 + 1.2953 Idioma7 - 1.1735 Idioma8 + 0.1992 Etnia2 \\
&\quad + 0.1761 Etnia3 + 0.4944 Etnia4 + 0.6055 Etnia5 \big)
\end{aligned}
$$

Capítulo 4

Validación y resultados

4.1. Validación del modelo

Una vez que se ha ajustado el modelo a los datos y se tienen los valores estimados de los distintos parámetros, el siguiente paso en la modelación es verificar que el modelo es adecuado.

- Primero: se va a verificar que se cumple con el supuesto de las rectas paralelas.
- Segundo: se comprueba que los coeficientes de las variables independientes son estadísticamente distintos de cero.
- Tercero: se realizan las pruebas globales del modelo, dado que es un modelo de regresión logística se distingue entre medidas de asociación y eficiencia predictiva, medidas de bondad ajuste y las pseudo R^2.

Así, el modelo que se ha obtenido en el capítulo 3, se lo va a validar con el programa SPSS, además se determinan las tres medidas de bondad de ajuste que ofrece el SPSS y alguna otra de interés.

4.1.1. Test de líneas paralelas

El modelo de regresión logit ordinal presentado en el capítulo 2, que se puede ver en la ecuación (2.12) se ha desarrollado bajo el supuesto del test de líneas paralelas o proportional odds que implica que los parámetros β_k, con $k = 1, .., m$ que se estiman para cada variable regresora de Y, son iguales para todas las categorías de la variable dependiente.

El estadístico de prueba es igual a menos dos veces el logaritmo de la razón de verosimilitud de los estimadores, esto es:

$$-2l(\widehat{\beta}_0) - 2l(\widehat{\beta}_1)$$

Bajo la hipótesis nula, este estadístico tiene asintóticamente una distribución chi-cuadrado (χ^2) con $(g-2)m$ grados de libertad.

La regla de decisión: se fija el nivel de significancia α y se rechaza la hipótesis nula cuando el estadístico calculado es mayor que el valor de la distribución con un nivel de significancia α y con $(g-2)m$ grados de libertad.

Prueba de líneas paralelas[a]

Modelo	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Hipótesis nula	2006022,067			
General	437412,735[b]	1568609,33[c]	62	,182

La hipótesis nula indica que los parámetros de ubicación (coeficientes de inclinación) son los mismos entre las categorías de respuesta.

a. Función de enlace: Logit.

b. El valor de log-verosimilitud no se puede aumentar más después del número máximo de subdivisión por pasos.

Figura 4.1: Prueba de líneas paralelas

Así la hipótesis nula de esta prueba es que los coeficientes de regresión β_k son los mismos entre las categorías de respuesta, se puede ver en la figura 4.1 que, por el valor p obtenido no se rechaza la hipótesis nula con un nivel de significancia del 10 %, por lo tanto hay evidencia de que la función de enlace logit es apropiada.(Norusis, 2005)[31]

Los incumplimientos de estos supuestos pueden provocar una formulación incorrecta del modelo y para evitarlo se requieren los tests de bondad de ajuste del modelo que se explican en las siguiente sección para su correcta interpretación.

Ahora, surge la pregunta: ¿Qué pasaría si la hipótesis no se cumple?

Como en muchos supuestos se tendría que los estimadores son sesgados e ineficientes.Sin embargo estos modelos han sido criticados porque generalmente en su mayoría la prueba de líneas paralelas no suele cumplirse ya que suelen tener una tendencia a rechazar la hipótesis nula e incluso en los casos donde el supuesto sí se cumple (ver Harrell 2001 p. 335). [22] Así estos modelos pueden pasar otros tipos de análisis que ayudan a determinar que el modelo sigue siendo bueno.

También cuando dicha hipótesis no se cumple se han desarrollado varias soluciones, una de ellas es realizar la prueba de wald para determinar que variables explicativas no cumplen con esta prueba y proceder a calcular sus propios estimadores, para contrastar la hipótesis de líneas paralelas de cada una de las variables regresoras, en caso de que alguna de las variables cumpla la condición de líneas paralelas, tendríamos que ajustar un modelo logístico ordinal con proporcionalidad parcial, de lo contrario, se puede realizar otro modelo, como por ejemplo los modelos logit ordenados generalizados.

4.1.2. Pruebas individuales de los estimadores

Se verifica que en el modelo de regresión logit ordinal los estimadores sean significativos, es decir, es la prueba de significancia de los estimadores de los coeficientes de los regresores.Esta prueba es conocida como el test de Wald.

Como se puede ver en las tablas 4.1 y 4.2 que se presentan a continuación muestran los resultados de dicha prueba.

Descripción	**Const.**	α	**Std. Error**	**t value**	**p value**
Ninguno | Primaria	0 | 1	-7.0015186	0.011381070	-615.18982	0.00E+00
Primaria | Secundaria	1 | 2	-1.0479457	0.010645901	-98.43654	0.00E+00
Secundaria | Superior	2 | 3	1.6415108	0.010619520	154.57486	0.00E+00

Tabla 4.1: Test de Wald para los estimadores α_j del modelo

Se verifica que los estimadores son significativos por su valor p, que rechaza la hipótesis nula, de acuerdo a lo explicado en el apéndice B.

Categorías	β	Std. Error	t value	p value
Region2	-0.2612	0.00178242	-146.51942	0.00E+00
Region3	-0.1419	0.004042205	-35.11102	4.58E-270
Region4	-0.4193	0.018254749	-22.96666	1.00E-116
Area2	-0.6088	0.001848058	-329.41528	0.00E+00
Sexo2	0.1947	0.001740333	111.8976	0.00E+00
GrupoEdadRec1	0.8756	0.005597054	156.43434	0.00E+00
GrupoEdadRec2	1.9706	0.005855652	336.53177	0.00E+00
GrupoEdadRec3	1.6542	0.005973577	276.92332	0.00E+00
GrupoEdadRec4	1.0345	0.006045622	171.10851	0.00E+00
TieneSeguroP2	-0.2639	0.002662834	-99.11681	0.00E+00
TieneDiscapacidad2	0.6847	0.003689299	185.58013	0.00E+00
LeerEscribir2	-4.9420	0.005536023	-892.70223	0.00E+00
InternetU6M2	-0.8909	0.003255079	-273.68526	0.00E+00
ComputadorU6M2	-1.0807	0.003189289	-338.86433	0.00E+00
ComoTrabajaRec2	-1.1002	0.004114151	-267.41394	0.00E+00
ComoTrabajaRec3	-2.5380	0.005303234	-478.58027	0.00E+00
ComoTrabajaRec4	-0.8620	0.007133621	-120.83797	0.00E+00
ComoTrabajaRec5	-0.9475	0.010782458	-87.87377	0.00E+00
ComoTrabajaRec6	-1.6918	0.004305925	-392.91046	0.00E+00
ComoTrabajaRec7	-1.6420	0.009806552	-167.43982	0.00E+00
ComoTrabajaRec8	-2.3669	0.007041304	-336.14951	0.00E+00
ComoTrabajaRec10	-1.6921	0.004115878	-411.10919	0.00E+00
EstadoConyugalRec2	-0.3534	0.002514693	-140.52571	0.00E+00
EstadoConyugalRec3	-0.1681	0.004156735	-40.44961	0.00E+00
EstadoConyugalRec4	0.6095	0.006437863	94.67395	0.00E+00
EstadoConyugalRec5	-0.9414	0.005569366	-169.03722	0.00E+00
EstadoConyugalRec6	0.2699	0.002610063	103.39343	0.00E+00
EstadoConyugalRec7	-1.2581	0.006416538	-196.07299	0.00E+00
Idioma2	0.4480	0.007907132	56.65355	0.00E+00
Idioma3	0.6332	0.012701494	49.85477	0.00E+00
Idioma4	0.2171	0.00828422	26.20269	2.48E-151
Idioma5	0.5065	0.096809717	5.23147	1.68E-07
Idioma6	1.3924	0.010097566	137.89563	0.00E+00
Idioma7	1.2953	0.044184805	29.31521	6.64E-189
Idioma8	-1.1735	0.022626147	-51.86647	0.00E+00
Etnia2	0.1992	0.006139437	32.44246	6.92E-231
Etnia3	0.1761	0.006232621	28.25715	1.16E-175
Etnia4	0.4944	0.005354193	92.3347	0.00E+00
Etnia5	0.6055	0.006216787	97.40209	0.00E+00

Tabla 4.2: Test de Wald para los estimadores β_k del modelo

4.1.3. Bondad de ajuste

Para empezar a analizar el ajuste del modelo el SPSS nos da la prueba de ajuste global del modelo, en el cual se plantea como hipótesis nula que el modelo sin la inclusión de las variables explicativas es adecuado.

Así se puede ver en la figura 4.2 que para el ajuste global del modelo, la hipótesis nula se rechaza con un nivel de significancia inferior al 1 %.

Información de ajuste de los modelos

Modelo	Logaritmo de la verosimilitud -2	Chi-cuadrado	gl	Sig.
Sólo intersección	7346137,898			
Final	2006022,067	5340115,831	31	,000

Función de enlace: Logit.

Figura 4.2: Ajuste global del modelo

El SPSS da también el resultado de la prueba de Hosmer-Lemeshow, en esta prueba las probabilidades predichas se dividen en 10 grupos basados en los deciles para estructurar una tabla de 2X10; sobre esta tabla se comparan predichos con las frecuencias observadas y se calcula una Chi-cuadrada de Pearson. El criterio aplicado será que los valores más pequeños (y sin significancia) son indicativos de un buen ajuste del modelo a los datos.

Bondad de ajuste

	Chi-cuadrado	gl	Sig.
Pearson	2775948,478	116948	,254
Desvianza	1901763,588	116948	,138

Función de enlace: Logit.

Figura 4.3: Bondad de ajuste

En la figura 4.3 los valores p de las pruebas de falta de ajuste son mayores que 0.05, lo que permite concluir que con la ecuación se alcanza un buen ajuste, debido a que las pruebas de bondad de ajuste no dan evidencia para considerar que hay falta de ajuste (se acepta la hipótesis nula pues el valor p es mayor que el nivel de significancia), lo que significa que las probabilidades de ocurrencia de los valores de la variable dependiente que se estiman según el modelo para las diferentes combinaciones de las independientes, no divergen significativamente de la frecuencia con la cual ocurren en la muestra los valores de la variable dependiente para estas combinaciones.

4.1.4. Pseudo R cuadrado

El coeficiente de determinación R^2 en el modelo de regresión lineal es un buen indicador del nivel de ajuste del modelo a los datos. Cuando el modelo tiene un buen ajuste, el valor de R^2 se aproxima a 1; en sentido contrario, cuando el ajuste es malo, el valor de R^2 se aproxima a cero. En el caso de los modelos de regresión con variable dependiente categórica también existen, en la literatura, propuestas de estadísticos R^2; sin embargo, no tienen las mismas características.

Los estadísticos denominados *pseudo* R^2 pretenden ser intentos de medir la fuerza de la asociación entre las variables, deben interpretarse cuidadosamente y en relación a otras pruebas. A continuación se presentan tres coeficientes estadísticos llamados *pseudo* R^2.

- **Pseudo - R^2** : **Cox y Snell**(1989)

  Compara el modelo final (modelo con los parámetros estimados) respecto al modelo nulo (modelo que tiene solamente la constante) con ello pretende indicar el grado de mejora del ajuste del modelo, para obtener esto se realiza mediante la estimación del logaritmo de la razón de verosimilitud del modelo nulo con la del modelo completo.

  El valor máximo que se puede obetener para este estadístico es menor a 1, incluso si se obtuviera un modelo "perfecto" y aunque su interpretación es dificultosa, pretende explicar la cantidad de varianza.

- **Pseudo - R^2** : **Nagelkerke**(1991)

  Es una modificación del coeficiente estadístico de Cox y Snell para que varíe entre 0 y 1, es decir puede tomar el valor de 1, por lo tanto su valor será más alto. En el supuesto caso de que si el modelo final alcanza un valor igual a 1 de verosimilitud entonces este *pseudo* R^2 sería igual a 1.

  Este *pseudo* R^2 es el estadístico más frecuentemente informado según la tesis doctoral de Jacinto Pallarés Mestre .[33]

- **Pseudo - R^2** : **McFadden**(1974)

  Este estadístico también se encuentra basado en la comparación del logaritmo de la verosimilitud del modelo nulo y el modelo final, el cual pretende reflejar la variabilidad explicada y el grado de mejora de ajuste del modelo final respecto al modelo nulo.

En general no hay un acuerdo sobre cuál de estos y/o otros estadísticos (R^2 de Cragg y Uhler, Mckelvey y Zavoina, Aldrich y Nelson, etc) *pseudo* R^2 es el mejor, además es muy complicado que estos coeficientes estadísticos proporcionen un valor cercano a 1, los investigadores prefieren no reportarlos; en el libro de Hosmer y Lemeshow se tiene este mismo criterio. [16]

En la actualidad la utilidad y aplicación de todos estos coeficientes ha dado lugar a varios estudios del tema e incluso algunos autores sugieren su uso con fines para la selección del modelo.

En el modelo obtenido los resultados de estos estadísticos se muestran en la figura 4.4 en la cual podemos ver que para el estadístico Nagelkerke se tiene aproximadamente 0.6, lo cual es bueno ya que generalmente estos valores suelen ser muy bajos en este tipo de modelos, así se puede decir que es un buen indicador para el modelo que se ha obtenido.

Pseudo R cuadrado

Cox y Snell	,515
Nagelkerke	,572
McFadden	,314

Función de enlace: Logit.

Figura 4.4: Pseudo R^2

Finalmente, debido a que los resultados de la prueba de paralelismo y las pruebas de bondad de ajuste, resultaron ser satisfactorias y los pseudo R^2 en general resultaron relativamente aceptables se puede concluir que el modelo estimado es adecuado.

4.2. Matrices de confusión

Analizar la precisión en la clasificación por medio de las matrices de confusión, es también interesante, para evaluar el modelo.

Podemos clasificar a el individuo i de la muestra en aquella categoría j para la cual la probabilidad:

$$P(Y = y_j | X = X_i) = \gamma_j(X = X_i) - \gamma_{j-1}(X = Xi)$$

Se crea y analiza la matriz de confusión (tabla de contingencia de las frecuencias observadas y esperadas), las cuales fueron creadas por medio del paquete R. Los resultados son expuestos en las tablas 4.3 y 4.4 resultados de acuerdo al modelo, y en las tablas 4.5 y 4.6 resultados de acuerdo a la base testing.

Modelo	**Actual**			
Predictivo	0	1	2	3
0 | Ninguno	173925	136808	0	0
1 | Primaria	160998	2759456	906444	84560
2 | Secundaria	11585	548607	1252102	444995
3 | Superior	1196	20636	252282	623301

Tabla 4.3: Matriz confusión de la base del modelo

Modelo	Registros
Total aciertos	4808784
Total matriz	7376895
Porcentaje aciertos	65.19 %

Tabla 4.4: Porcentaje de aciertos en el modelo

La matriz de confusión de la tabla 4.3 clasificó correctamente el 65 % de los individuos.

Testing	**Actual**			
Predictivo	0	1	2	3
0 | Ninguno	43217	34067	0	0
1 | Primaria	40289	689783	226129	21118
2 | Secundaria	2876	137521	313955	111715
3 | Superior	308	5241	62622	155380

Tabla 4.5: Matriz confusión de la base testing

Testing	Registros
Total aciertos	1202335
Total matriz	1844221
Porcentaje aciertos	65.19 %

Tabla 4.6: Porcentaje de aciertos en el testing

La matriz de confusión de la tabla 4.5 clasificó correctamente el 65 % de los individuos.

Los resultados de las tablas precedentes, nos dan el mismo acierto de lo cual se puede concluir que el modelo se puede generalizar.

4.3. Resultados del modelo

Del modelo que se obtuvo en el capítulo 3, se pueden obtener varias conclusiones con la interpretación de los *OddsRatios* (OR) las cuales se presentan en las siguientes tablas.

Variable (Vs)	**V.Categórica: descripción**	***OR***	**Interpretación**
Regiones del Ecuador (Región1 : Costa)	Region2:Sierra	0.7702	Una persona que vive en la región Sierra es 22.98 % menos probable que tenga un mayor nivel de estudio, que alguien que vive en la región Costa (conservando el resto de las variables constantes).
	Region3:Oriente	0.8677	Una persona que vive en la región del Oriente es 13.23 % menos probable que tenga un mayor nivel de estudio, que alguien que vive en la región Costa (conservando el resto de las variables constantes).
	Region4:Insular	0.6575	Una persona que viva en la región Insular es 34.25 % menos probable que tenga un mayor nivel de estudio, que alguien que vive en la región Costa (conservando el resto de las variables constantes).

Tabla 4.7: Interpretación del modelo para las regiones del Ecuador

Variable (Vs)	V.Categórica: Descripción	*OR*	Interpretación
Area urbana o rural(Area 1: Area urbana)	Area2: Area rural	0.5440	Una persona que viva en el área rural es 45.6 % menos probable que tenga un mayor nivel de estudio, que alguien que vive en la zona urbana (conservando el resto de las variables constantes).
Sexo(Sexo1: Hombre)	Sexo2: Mujer	1.2150	Una mujer es 1.21 mas probable que tenga un mayor nivel de estudio, que un hombre (conservando el resto de las variables constantes).
Edad cumplida (GrupoEdadRec0: Edad entre 5 y 12)	GrupoEdadRec1: Edad entre 13 y 18	2.4002	Una persona que tenga entre 13 y 18 años es 2.4 más probable que tenga un mayor nivel de estudio, que una persona entre 5 y 12 años de edad, conservando el resto de las variables constantes.
	GrupoEdadRec2: Edad entre 19 y 25	7.1751	Una persona que tenga entre 19 y 25 años es 7.18 más probable que tenga un mayor nivel de estudio, que una persona entre 5 y 12 años de edad, conservando el resto de las variables constantes.
	GrupoEdadRec3: Edad entre 26 y 40	5.2290	Una persona que tenga entre 26 y 40 años es 5.23 más probable que tenga un mayor nivel de estudio, que una persona entre 5 y 12 años de edad, conservando el resto de las variables constantes.
	GrupoEdadRec4: Edad mayor a 41	2.8136	Una persona que tenga entre 26 y 40 años es 2.81 más probable que tenga un mayor nivel de estudio, que una persona entre 5 y 12 años de edad, conservando el resto de las variables constantes.

Tabla 4.8: Interpretación del modelo para el área, el sexo y grupos de edad en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
Tiene seguro de salud privado (TieneSeguroP1 : Si)	TieneSeguroP2: No	0.7680	Una persona que no tenga seguro es 23.2 % menos probable que tenga un mayor nivel de estudios frente a una persona que si tiene algún seguro de salud privado, conservando el resto de variables constantes.
Tiene discapacidad permanente por más de un año (Tiene Discapacidad1 : Si)	Tiene Discapacidad2 : No	1.9831	Una persona que no tenga discapacidad frente a una que tenga, es 1.98 veces más probable que tenga un mayor nivel de estudio, conservando el resto de variables constantes.(También se puede decir que una persona que no tenga discapacidad es 98.31 % más probable de que tenga un mayor nivel de estudios)
Sabe leer y escribir (LeerEscribir1: Si)	LeerEscribir2: No	0.0071	Una persona que no sabe leer y escribir es 99.29 % menos probable que tenga un mayor nivel de estudio frente a una persona que si (conservando el resto de variables constantes).
En los últimos 6 meses ha utilizado Internet (InternetU6M1 : Si)	InternetU6M2 : No	0.4103	Una persona que no use internet es 58.97 % menos probable que tenga un mayor nivel de estudio frente a una persona que si usa, conservando el resto de variables constantes.
En los útlimos 6 meses ha utilizado Computadora (ComputadorU6M1 : Si)	Computador U6M2 : No	0.3393	Una persona que no usa una computadora es 66.07 % menos probable que tenga un mayor nivel de estudios frente a una que si usa, conservando el resto de variables constante.

Tabla 4.9: Interpretación del modelo para el poseer seguro privado, discapacidad, saber leer y escribir, usar internet y usar computadora en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
ComoTrabajaRec (ComoTrabaja-Rec1: Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales)	ComoTrabajaRec2: Empleado/a u obrero/a privado	0.3328	Una persona que trabaja como empleado/a u obrero/a privado es 66.72 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec3: Jornalero/a o peón	0.0790	Una persona que trabaja como Jornalero/a o peón es 92.1 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec4: Patron/a	0.4223	Una persona que trabaja como patron/a es 57.1 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec5: Socio/a	0.3877	Una persona que trabaja como socioo/a o peón es 61.23 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).

Tabla 4.10: Interpretación del modelo para como trabaja en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
ComoTrabajaRec (ComoTrabaja-Rec1: Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales)	ComoTrabajaRec6: Cuenta propia	0.1842	Una persona que trabaja por cuenta propia es 81.58 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec7: Trabaja-dor/a no remunera-do	0.1936	Una persona que trabaja como trabajador/a no remunerado es 80.64 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec8: Emplea-do/a domésti-co/a	0.0938	Una persona que trabaja como Empleado/a doméstico/a es 90.62 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).
	ComoTrabajaRec10: No trabaja	0.1841	Una persona que no trabaja es 81.59 % menos probable que tenga un mayor nivel de estudio con respecto a una persona identificada con la variable ComoTrabajaRec1 (conservando el resto de las variables constantes).

Tabla 4.11: Interpretación del modelo para como trabaja en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
Estado Conyugal (EstadoConyugalRec1: Casado/a)	EstadoConyugalRec2: Unido/a	0.7023	Una persona que se encuentra unido/a es 29.77 % menos probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).
	EstadoConyugalRec3: Separado/a	0.8452	Una persona que se encuentra separado/a es 15.48 % menos probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).
	EstadoConyugalRec4: Divorciado/a	1.8395	Una persona que es divorciada es 1.84 más probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).
	EstadoConyugalRec5: Viudo/a	0.3901	Una persona que es viudo/a es 60.99 % menos probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).
	EstadoConyugalRec6: Soltero/a	1.3098	Una persona que es soltera es 1.31 más probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).
	EstadoConyugalRec7: No aplica	0.2842	Las personas que no aplican ningún estado conyugal son el 71.58 % menos probable que tenga un mayor nivel de estudio con respecto a una persona casada (conservando el resto de las variables constantes).(Lo cual tiene sentido ya que Las personas en esta categoría son niños y adolescentes)

Tabla 4.12: Interpretación del modelo para el estado conyugal en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
Idioma que habla recodificado (Idioma1: Indígena)	Idioma2: Castellano Español	1.5651	Una persona que hable castellano español es 1.57 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma3: Extranjero	1.8837	Una persona que hable un idioma extranjero es 1.88 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma4: Indígena/ Castellano Español	1.2424	Una persona que hable indígena/castellano español es 1.24 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma5: Indígena/ Extranjero	1.6594	Una persona que hable indígena/extranjero es 1.66 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma6: Castellano Español/ Extranjero	4.0245	Una persona que hable castellano español/extranjero es 4.02 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma7: Indígena/ Castellano Español/ Extranjero	3.6520	Una persona que hable indígena/ castellano español / extranjero es 3.65 más probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.
	Idioma8: No habla	0.3093	Una persona que no habla es 69.07 % menos probable que tenga un mayor nivel de estudio que una que habla indígena, conservando el resto de las variables constantes.

Tabla 4.13: Interpretación del modelo para el idioma que se habla en el Ecuador

Variable (Vs)	V.Categórica: descripción	*OR*	Interpretación
Auto identificación étnica recodifica (Etnia1: Indígena)	Etnia2: Afroecuatoriano / Afrodescendiente, Negro, Mulato	1.2204	Una persona que se autoidentifca como afroecuatoriano/a es 1.22 más probable que tenga un mayor nivel de estudio que una persona que se autoidentifica como indígena, conservando el resto de las variables constantes.
	Etnia3: Montuvio	1.1926	Una persona que se autoidentifca como montuvio/a es 1.19 más probable que tenga un mayor nivel de estudio que una persona que se autoidentifica como indígena, conservando el resto de las variables constantes.
	Etnia4: Mestizo	1.6395	Una persona que se autoidentifca como mestiza/o es 1.64 más probable que tenga un mayor nivel de estudio que una persona que se autoidentifica como indígena, conservando el resto de las variables constantes.
	Etnia5: Blanco	1.8322	Una persona que se autoidentifca como blanco/a es 1.83 más probable que tenga un mayor nivel de estudio que una persona que se autoidentifica como indígena, conservando el resto de las variables constantes.

Tabla 4.14: Interpretación del modelo para la autoidentificación por etnia en el Ecuador

Se obtuvo varios análisis del modelo, algunos son más sobresalientes como los siguientes:

- Una persona que viva en la región Insular es 34.25 % menos probable que tenga un mayor nivel de estudio que alguien que vive en la región Costa, conservando el resto de las variables constantes.
- Si se vive en el área urbana o rural, si influye para alcanzar un nivel de educación, ya que una persona que viva en el área rural es 45.6 % menos probable que tenga un

mayor nivel de estudio, que alguien que vive en la zona urbana.

- Una mujer es 1.21 mas probableque tenga un mayor nivel de estudio, que un hombre.
- La edad influye en la educación, se encontró que para los grupos de edad de 13-18, 19-25 y 26-40, son 2.4, 7.18 y 5.23 respectivamente son más probables de que tengan un mayor nivel de educación que una persona entre 5-12 años de edad, conservando el resto de las variables constantes.
- Una persona que no posee un tipo de seguro privado es 23.2 % menos probable de que tenga un mayor nivel de estudios frente a una persona que si, consrvando el reto de variables constantes.
- Una persona que no sufre de algún tipo de discapacidad es 1.98 veces más probable que tenga un mayor nivel de estudio frente a una que si, conservando el resto de las variables constantes.
- Una persona analfabeta es 99.29 % menos probable que tenga un mayor nivel de estudio frente a una persona que si sabe leer y escribir, conservando el resto de las variables constantes.
- Una persona que no usa internet y computador son 58.97 % y 66.07 % respectivamente, menos probable de que tengan un mayor nivel de estudios frente alguna que si los usa, conservando el resto de las variables constantes.
- Una persona que trabaja como jornalero y empleado doméstico son 92.1 % y 90.62 % respectivamente, menos probable de que tengan un mayor nivel de educación frente a un empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial y Juntas Parroquiales, conservando el resto de variables constantes.
- Una persona que es divorciada o soltera son 1.84 y 1.31 respectivamente, más probable de que tengan un mayor nivel de estudio frente a una persona casada, conservando el resto de las variables constantes.
- Una persona que habla castellano español y un idioma extranjero es 4.02 más probable que tenga un mayor nivel de educación frente a una persona que habla indígena, conservando el resto de las variables constantes.
- Una persona que se autoidentifica como mestizo/a o blanco/a son 1.64 y 1.83 respectivamente, más probable de que tengan un mayor nivel de estudios que una autoidentificada como indígena, conservando el resto de las variables constantes.

Capítulo 5

Conclusiones y comentarios

En el presente proyecto de investigación se tienen las siguientes conclusiones y comentarios:

De acuerdo a la teoría del modelo se determinó lo siguiente:

- Se ha considerado la parametrización:

$$log\left(\frac{P(Y \leq y_j|X)}{1-P(Y \leq y_j|X)}\right) = \alpha_j - \beta^T X, \quad j = 1, ..., g-1,$$

  es decir, se tiene a los estimadores beta con el signo menos, lo cual se debe a la desigualdad, por lo que se puede también trabajar con el signo más en donde de esa forma se tendría que cambiar a la desigualdad. Además los paquetes estadísticos SPSS y R utilizan la parametrización que se ha desarrolado.

De acuerdo a los resultados interpretativos del modelo se determinó lo siguiente:

- Los determinantes más influyentes en la población ecuatoriana de acuerdo al censo del INEC del año 2010, son: región, área, sexo, edad, seguro privado, discapacidad, saber leer y escribir, usar el internet, usar la computadora, el tipo de trabajo(status socio-económico), el estado conyugal, el idioma y la etnia.

- Se encontró que el poseer algún tipo de discapacidad influye para alcanzar un nivel de educación, pues una persona que no tenga discapacidad es 1.98 veces más probable que tenga un mayor nivel de educación frente a una que si, conservando el resto de las variables constantes.

- Una mujer es 1.21 mas probable que tenga un mayor nivel de estudio, que un hombre. Se puede ver como las mujeres a través de la historia ha ido ocupando un lugar en la educación, en la investigación, etc. pues la tasa neta de asistencia al bachillerato que

se presenta en la tabla 1.3 la de la mujer con la del hombre son bastante similares, por lo que es un referente para decir que de alguna forma se encuentran a la par y sin embargo las mujeres pueden tener mayor nivel de estudios.

- Una persona analfabeta es 99.29 % menos probable que tenga un mayor nivel de estudio frente a una persona que si sabe leer y escribir, conservando el resto de variables constantes. Este puede ser un factor importante, ya que puede dar la idea que para mejorar el nivel de la educación, es fundamental mejorar la educación desde los niveles más bajos(primaria) ya que el leer y escribir es lo primero que se aprende.

- Una persona que hable además del español un idioma extranjero es 4.02 más probable que tenga un mayor nivel de estudio frente a una persona que habla indígena, conservando las variables constantes. De aquí se debería pensar en la importancia de inculcar la convivencia e interacción entre las diferentes culturas aprovechando la diversidad cultural que existe en nuestro país y también en otros países.

- El tipo de trabajo al que se dedica una persona nos da una idea del estatus socioeconómico, el cual influye para una buena educación, ya que si una persona que trabaja como jornalero/a o empleado/a doméstico/a son 92.1 % y 90.62 % respectivamente, menos probable de que tengan un mayor nivel de educación frente a un empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial y Juntas Parroquiales, conservando el resto de variables constantes.

De acuerdo a la validación del modelo se determinó lo siguiente:

- Cuando el supuesto de líneas paralelas no se cumple, se puede realizar un modelo logit proporcional o incluso un modelo logit generalizado.

Apéndice A

Anexos

A.1. Tablas de Datos

Se presenta una tabla de las 14 variables que se ha obtenido para realizar el presente estudio.

Variable	Etiqueta de Información	Nivel de medición
Region	Regiones del Ecuador	Nominal
Area	Area urbana o rural	Nominal
Sexo	Cuá es el Sexo	Nominal
GrupoEdadRec	Edad cumplida	Nominal
TieneSeguroP	Tiene seguro de salud privado	Nominal
TieneDiscapacidad	Tiene discapacidad permanente por más de un año	Nominal
LeerEscribir	Sabe leer y escribir	Nominal
InternetU6M	En los últimos 6 meses ha utilizado Internet	Nominal
ComputadorU6M	En los útlimos 6 meses ha utilizado Computadora	Nominal
ComoTrabajaRec	En el lugar indicado trabaja o trabajó como	Nominal
EstadoConyugalRec	Estado Conyugal	Nominal
Idioma	Idioma que habla recodificado	Nominal
Etnia	Autoidentificación étnica recodifica	Nominal
NivelEducativo	Nivel Educativo	Ordinal

Tabla A.1: Información de las variables

Variable	Categoría	Etiqueta
Región	1	Costa
	2	Sierra
	3	Oriente
	4	Región Insular
Área	1	Urbana
	2	Rural
Sexo	1	Hombre
	2	Mujer
GrupoEdadRec	0	Edad entre 5 y 12
	1	Edad entre 13 y 18
	2	Edad entre 19 y 25
	3	Edad entre 26 y 40
	4	Edad mayor a 41
TieneSeguroP	1	Si
	2	No
TieneDiscapacidad	1	Si
	2	No
LeerEscribir	1	Si
	2	No
InternetU6M	1	Si
	2	No
ComputadorU6M	1	Si
	2	No
ComoTrabajaRec	1	Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales
	2	Empleado/a u obrero/a privado
	3	Jornalero/a o peón
	4	Patron/a
	5	Socio/a
	6	Cuenta propia
	7	Trabajador/a no remunerado
	8	Empleado/a doméstico/a
	10	No trabaja

Tabla A.2: Valores de las variables

Variable	Categoría	Etiqueta
EstadoConyugalRec	1	Casado/a
	2	Unido/a
	3	Separado/a
	4	Divorciado/a
	5	Viudo/a
	6	Soltero/a
	7	No aplica
Idioma	1	Indígena
	2	Castellano Español
	3	Extranjero
	4	Indígena/Castellano Español
	5	Indígena/Extranjero
	6	Castellano Español/Extranjero
	7	Indígena/Castellano Español/Extranjero
	8	No habla
Etnia	1	Indígena
	2	Afroecuatoriano/Afrodescendiente, Negro, Mulato
	3	Montuvio
	4	Mestizo
	5	Blanco
NivelEducativo	0	Ninguno
	1	Primaria
	2	Secundaria
	3	Superior

Tabla A.3: Valores de las variables

A.2. Pruebas chi-cuadrado

En el SPSS se realiza la prueba a cada variable siendo estadísticamente significativas.

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Regiones del Ecuador * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Regiones del Ecuador*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Regiones del Ecuador	Costa	202144	2061995	1527721	607827	4399687
	Sierra	212569	2014478	1341087	789814	4357948
	Oriente	19475	249237	138169	40250	447131
	Insular	206	6409	6557	3178	16350
Total		434394	4332119	3013534	1441069	9221116

Figura A.1: Tabla cruzada de región

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	59482,752[a]	9	,000
Razón de verosimilitud	61026,925	9	,000
Asociación lineal por lineal	11,609	1	,001
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 770,23.

Figura A.2: Prueba para región

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Área urbana o rural * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Área urbana o rural*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Área urbana o rural	Area urbana	160971	2331132	2210426	1211884	5914413
	Area rural	273423	2000987	803108	229185	3306703
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	700152,619[a]	3	,000
Razón de verosimilitud	725651,002	3	,000
Asociación lineal por lineal	684849,941	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 155774,20.

Figura A.3: Tabla cruzada y prueba para el área

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Cúal es el Sexo * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Cúal es el Sexo*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Cúal es el Sexo	Hombre	183243	2176638	1490721	664232	4514834
	Mujer	251151	2155481	1522813	776837	4706282
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	15891,981[a]	3	,000
Razón de verosimilitud	15937,496	3	,000
Asociación lineal por lineal	966,860	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 212687,57.

Figura A.4: Tabla cruzada y prueba para el sexo

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Edad cumplida * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Edad cumplida*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Edad cumplida	Edad entre 5 y 12	13907	1643435	116389	0	1773731
	Edad entre 13 y 18	13122	376747	850951	50724	1291544
	Edad entre 19 y 25	25375	320758	550790	416489	1313412
	Edad entre 26 y 40	67066	773648	841422	552754	2234890
	Edad mayor a 41	314924	1217531	653982	421102	2607539
Total		434394	4332119	3013534	1441069	9221116

Figura A.5: Tabla cruzada para la edad

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	3151220,13[a]	12	,000
Razón de verosimilitud	3333709,656	12	,000
Asociación lineal por lineal	182874,551	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 60842,85.

Figura A.6: Prueba para la edad

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Tiene seguro de salud privado * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Tiene seguro de salud privado*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Tiene seguro de salud privado	Si	16188	266268	319185	334419	936060
	No	418206	4065851	2694349	1106650	8285056
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	365764,914[a]	3	,000
Razón de verosimilitud	318890,353	3	,000
Asociación lineal por lineal	319619,438	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 44096,49.

Figura A.7: Tabla cruzada y prueba para el seguro privado

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Tiene discapacidad permanente por más de un año * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Tiene discapacidad permanente por más de un año*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Tiene discapacidad permanente por más de un año	Si	112486	290103	118111	42263	562963
	No	321908	4042016	2895423	1398806	8658153
Total		434394	4332119	3013534	1441069	9221116

Figura A.8: Tabla cruzada para la discapacidad

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	349834,822[a]	3	,000
Razón de verosimilitud	234799,340	3	,000
Asociación lineal por lineal	172701,893	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 26520,41.

Figura A.9: Prueba para la discapacidad

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Sabe leer y escribir * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Sabe leer y escribir*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Sabe leer y escribir	Si	77673	4113945	3013534	1441069	8646221
	No	356721	218174	0	0	574895
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	4585850,68[a]	3	,000
Razón de verosimilitud	2166748,092	3	,000
Asociación lineal por lineal	1391041,477	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 27082,51.

Figura A.10: Tabla cruzada y prueba para si sabe leer y escribir

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
En los últimos 6 meses ha utilizado Internet * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada En los últimos 6 meses ha utilizado Internet*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
En los últimos 6 meses ha utilizado Internet	Si	4499	485350	1029756	1081008	2600613
	No	429895	3846769	1983778	360061	6620503
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2389108,33[a]	3	,000
Razón de verosimilitud	2390831,753	3	,000
Asociación lineal por lineal	2241159,568	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 122511,28.

Figura A.11: Tabla cruzada y prueba para el uso de internet

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
En los útlimos 6 meses ha utilizado Computadora * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada En los últimos 6 meses ha utilizado Computadora*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
En los útlimos 6 meses ha utilizado Computadora	Si	5795	853807	1228798	1166956	3255356
	No	428599	3478312	1784736	274113	5965760
Total		434394	4332119	3013534	1441069	9221116

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2036520,76[a]	3	,000
Razón de verosimilitud	2135916,653	3	,000
Asociación lineal por lineal	1943384,022	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 153355,31.

Figura A.12: Tabla cruzada y prueba para el uso de la computadora

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
En el lugar indicado trabaja o trabajó como * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada En el lugar indicado trabaja o trabajó como*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
En el lugar indicado trabaja o trabajó como	Empleado/a u obrero/a del Estado, Gobierno, Municipio, Consejo Provincial, Juntas Parroquiales	3763	53063	136714	267288	460828
	Empleado/a u obrero/a privado	20943	362274	630978	489559	1503754
	Jornalero/a o peón	59378	367239	131237	7582	565436
	Patron/a	2947	36147	49265	48078	136437
	Socio/a	949	12359	18952	16970	49230
	Cuenta propia	97060	575867	417615	157291	1247833
	Trabajador/a no remunerado	4091	25799	24265	9282	63437
	Empleado/a doméstico/a	9952	96079	57720	4859	168610
	No trabaja	235311	2803292	1546788	440160	5025551
Total		434394	4332119	3013534	1441069	9221116

Figura A.13: Tabla cruzada para como trabaja

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	1717413,72[a]	24	,000
Razón de verosimilitud	1600661,098	24	,000
Asociación lineal por lineal	757916,236	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 2319,16.

Figura A.14: Prueba para como trabaja

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Estado Conyugal * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Estado Conyugal*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Estado Conyugal	Casado/a	148012	998037	793264	574219	2513532
	Unido/a	97508	673826	597556	161972	1530862
	Separado/a	21802	148433	137069	51148	358452
	Divorciado/a	4468	35740	46547	52080	138835
	Viudo/a	64168	137700	45897	15394	263159
	Soltero/a	85561	807118	1367820	586256	2846755
	No aplica	12875	1531265	25381	0	1569521
Total		434394	4332119	3013534	1441069	9221116

Figura A.15: Tabla cruzada para el estado conyugal

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	2585272,79[a]	18	,000
Razón de verosimilitud	2901407,709	18	,000
Asociación lineal por lineal	125401,234	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 6540,32.

Figura A.16: Prueba para el estado conyugal

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Idioma que habla recodificado * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Idioma que habla recodificado*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Idioma que habla recodificado	Indígena	50146	112080	32851	5082	200159
	Castellano Español	342012	4032239	2868296	1311548	8554095
	Extranjero	1881	21964	18887	13165	55897
	Indígena/Castellano Español	26449	138596	48525	9253	222823
	Indígena/Extranjero	102	400	156	74	732
	Castellano Español/Extranjero	414	19924	42239	100324	162901
	Indígena/Castellano Español/Extranjero	58	773	915	1164	2910
	No habla	13332	6143	1665	459	21599
Total		434394	4332119	3013534	1441069	9221116

Figura A.17: Tabla cruzada para el idioma

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	724398,509[a]	21	,000
Razón de verosimilitud	477695,599	21	,000
Asociación lineal por lineal	55780,788	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 34,48.

Figura A.18: Prueba para el idioma

Resumen de procesamiento de casos

	Casos					
	Válido		Perdido		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Autoidentificación étnica recodifica * Nivel_Educativo	9221116	100,0%	0	0,0%	9221116	100,0%

Tabla cruzada Autoidentificación étnica recodifica*Nivel_Educativo

Recuento

		Nivel_Educativo				
		Ninguno	Primaria	Secundaria	Superior	Total
Autoidentificación étnica recodifica	Indígena	91099	369859	127445	23097	611500
	Afroecuatoriano/Afrodescendiente, Negro, Mulato	33099	319168	227309	47467	627043
	Montubio	58382	383187	173010	38526	653105
	Mestizo	236291	3044329	2287754	1211799	6780173
	Blanco	15523	215576	198016	120180	549295
Total		434394	4332119	3013534	1441069	9221116

Figura A.19: Tabla cruzada para la etnia

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	440441,948[a]	12	,000
Razón de verosimilitud	430395,213	12	,000
Asociación lineal por lineal	319513,823	1	,000
N de casos válidos	9221116		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 25876,53.

Figura A.20: Prueba para la etnia

A.3. Código en R

Se presenta las líneas de código que se consideran más importantes, para obtener el modelo.

Figura A.21: Se eliminan los registros pérdidos o null

```
### PASO 1: CARGAR LA BASE ###
Base_Tesis <- read.spss('1. BASE_TOTAL.sav',use.value.labels = TRUE,
                        max.value.labels = TRUE, to.data.frame = TRUE)

###Paso 2: Eliminar los nulos de las variables:
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$Region),]  ##12139014
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$Area),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$Sexo),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$GrupoEdadRec),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$TieneCedula),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$TieneSeguroP),] #11593331
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$TieneDiscapacidad),] #11057061
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$LeerEscribir),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$CelularU6M),] #10722867
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$InternetU6M),] #10030534
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$ComputadorU6M),] #10011538
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$ActividadPasada),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$HaTrabajadoRec),] #9716397
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$ComoTrabajaRec),] #9582273
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$HorasTrabajoRec),] #9563950
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$ComoTrabajaRec),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$HorasTrabajoRec),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$DondeRealizaTrabajoRec),] #9559781
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$EstadoConyugalRec),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$SeguridadSocialRec),] #9350327
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$Idioma),]
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$Etnia),] #9316840
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$GrupoOcupacionRec),] #9221116
Base_Tesis <- Base_Tesis[!is.na(Base_Tesis$NivelEducativo),] #9221116
#Se tiene 9221116 registros
```

Figura A.22: Se hace la partición de bases 80/20

```
#SE TOMA EL 80% PARA EL MODELO Y EL 20% PARA TESTING:
set.seed(500)
registros <- createDataPartition(Base_Tesis$NivelEducativo, p = 0.80, list = FALSE)
Base_Modelo <- Base_Tesis[registros,]
Base_Testing <- Base_Tesis[-registros,]

### GUardar en SPSS las Bases
library(haven)
write_sav(Base_Modelo, "E:/Paty/Data Final v3/2. Base_Modelo.sav")
write_sav(Base_Testing, "E:/Paty/Data Final v3/3. Base_Testing.sav")
```

Figura A.23: Se transforman las variables a factores de la base modelo, para que lea correctamente las categorías de cada variable

```
### Transformar a factor las variables categoricas
Base_Modelo$Region            <- as.factor(Base_Modelo$Region)
Base_Modelo$Area              <- as.factor(Base_Modelo$Area)
Base_Modelo$Sexo              <- as.factor(Base_Modelo$Sexo)
Base_Modelo$GrupoEdadRec      <- as.factor(Base_Modelo$GrupoEdadRec)
Base_Modelo$TieneCedula       <- as.factor(Base_Modelo$TieneCedula)
Base_Modelo$TieneSeguroP      <- as.factor(Base_Modelo$TieneSeguroP)
Base_Modelo$TieneDiscapacidad       <- as.factor(Base_Modelo$TieneDiscapacidad)
Base_Modelo$LeerEscribir            <- as.factor(Base_Modelo$LeerEscribir)
Base_Modelo$CelularU6M              <- as.factor(Base_Modelo$CelularU6M)
Base_Modelo$InternetU6M             <- as.factor(Base_Modelo$InternetU6M)
Base_Modelo$ComputadorU6M           <- as.factor(Base_Modelo$ComputadorU6M)
Base_Modelo$ActividadPasada         <- as.factor(Base_Modelo$ActividadPasada)
Base_Modelo$HaTrabajadoRec          <- as.factor(Base_Modelo$HaTrabajadoRec)
Base_Modelo$ComoTrabajaRec          <- as.factor(Base_Modelo$ComoTrabajaRec)
Base_Modelo$HorasTrabajoRec          <- as.factor(Base_Modelo$HorasTrabajoRec)
Base_Modelo$DondeRealizaTrabajoRec   <- as.factor(Base_Modelo$DondeRealizaTrabajoRec)
Base_Modelo$EstadoConyugalRec        <- as.factor(Base_Modelo$EstadoConyugalRec)
Base_Modelo$SeguridadSocialRec       <- as.factor(Base_Modelo$SeguridadSocialRec)
Base_Modelo$Idioma                   <- as.factor(Base_Modelo$Idioma)
Base_Modelo$Etnia                    <- as.factor(Base_Modelo$Etnia)
Base_Modelo$GrupoOcupacionRec        <- as.factor(Base_Modelo$GrupoOcupacionRec)
Base_Modelo$NivelEducativo    <- factor(Base_Modelo$NivelEducativo, ordered =  TRUE)
```

Figura A.24: Se transforman las variables a factores de la base testing, para que lea correctamente las categorías de cada variable

```
### Transformar a factor las variables categoricas del testing
Base_Testing$Region            <- as.factor(Base_Testing$Region)
Base_Testing$Area              <- as.factor(Base_Testing$Area)
Base_Testing$Sexo              <- as.factor(Base_Testing$Sexo)
Base_Testing$GrupoEdadRec      <- as.factor(Base_Testing$GrupoEdadRec)
Base_Testing$TieneCedula       <- as.factor(Base_Testing$TieneCedula)
Base_Testing$TieneSeguroP      <- as.factor(Base_Testing$TieneSeguroP)
Base_Testing$TieneDiscapacidad       <- as.factor(Base_Testing$TieneDiscapacidad)
Base_Testing$LeerEscribir            <- as.factor(Base_Testing$LeerEscribir)
Base_Testing$CelularU6M              <- as.factor(Base_Testing$CelularU6M)
Base_Testing$InternetU6M             <- as.factor(Base_Testing$InternetU6M)
Base_Testing$ComputadorU6M           <- as.factor(Base_Testing$ComputadorU6M)
Base_Testing$ActividadPasada         <- as.factor(Base_Testing$ActividadPasada)
Base_Testing$HaTrabajadoRec          <- as.factor(Base_Testing$HaTrabajadoRec)
Base_Testing$ComoTrabajaRec          <- as.factor(Base_Testing$ComoTrabajaRec)
Base_Testing$HorasTrabajoRec          <- as.factor(Base_Testing$HorasTrabajoRec)
Base_Testing$DondeRealizaTrabajoRec   <- as.factor(Base_Testing$DondeRealizaTrabajoRec)
Base_Testing$EstadoConyugalRec        <- as.factor(Base_Testing$EstadoConyugalRec)
Base_Testing$SeguridadSocialRec       <- as.factor(Base_Testing$SeguridadSocialRec)
Base_Testing$Idioma                   <- as.factor(Base_Testing$Idioma)
Base_Testing$Etnia                    <- as.factor(Base_Testing$Etnia)
Base_Testing$GrupoOcupacionRec        <- as.factor(Base_Testing$GrupoOcupacionRec)
Base_Testing$NivelEducativo    <- factor(Base_Testing$NivelEducativo, ordered =  TRUE)
```

Figura A.25: Se obtiene el modelo a partir de la base modelo

```
#### PASO 1: CORRER TODOS LOS PAQUETES:
library('dplyr')    #Filtrar, ordenar, resumen variables, muestras aleatorias, etc
library(foreign)
require(ggplot2)
require(MASS)
require(Hmisc)
require(reshape2)
library(car)
library('svMisc')
library('caret')
library('magrittr')
library('olsrr')
library('descr')

##CARGAR DESDE EL ARCHIVO GUARDADO EN R
load('Bases.RData') #data previamente depurada

## semilla segun el numero sale un diferente modelo:
set.seed(1)
memory.limit(size=6000000)
#MODELO #0.65 acierto
Modelo <- polr(NivelEducativo ~ Region + Area  + Sexo
               +GrupoEdadRec+TieneSeguroP + TieneDiscapacidad
               + LeerEscribir +InternetU6M + ComputadorU6M
               + ComoTrabajaRec +EstadoConyugalRec+Idioma+ Etnia,
               data = Base_Modelo,method = "logistic", Hess=TRUE)
#RESUMEN DEL MODELO
summary(Modelo)

#PARA SACAR EL P-VALOR DE VADA VARIABLE
(ctable <- coef(summary(Modelo)))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
```

Figura A.26: Se obtiene los aciertos en la base modelo

```
############### RESULTADOS EN LA BASE DE MODELO ######################

#PseudoR2(Modelo, c("McFadden", "Nagel"))
#PREDICCION
Modelo_predictions <- predict(Modelo, Base_Modelo)
#ERROR
mean(factor(Modelo_predictions, ordered =  TRUE)  == Base_Modelo$NivelEducativo)
#MATRIZ DE CONFUSION
table(predicted = Modelo_predictions,actual = Base_Modelo$NivelEducativo)

#PREDICCIONES CON PROBABILIDAD: SE ADICIONAN A LA BASE
Probabilidades_Modelo <- predict(Modelo, type="probs")
Base_Modelo$prediccion <- predict(Modelo, Base_Modelo)
```

Figura A.27: Se obtiene los aciertos en la base testing

```
################## RESULTADOS EN LA BASE DE TESTING #################

#PREDICCION
Testing_predictions <- predict(Modelo, Base_Testing)
#ERROR
mean(factor(Testing_predictions, ordered =  TRUE)  == Base_Testing$NivelEducativo)
#MATRIZ DE CONFUSION
table(predicted = Testing_predictions,actual = Base_Testing$NivelEducativo)
#PREDICCIONES CON PROBABILIDAD: SE ADICIONAN A LA BASE
#Base_Testing$probs <- predict(Modelo, type="probs")
```

A.4. Resultados del modelo

Descripción	Const.	α	Std. Error	t value	p value
Ninguno | Primaria	0 | 1	-7.0015186	0.011381070	-615.18982	0.00E+00
Primaria | Secundaria	1 | 2	-1.0479457	0.010645901	-98.43654	0.00E+00
Secundaria | Superior	2 | 3	1.6415108	0.010619520	154.57486	0.00E+00

Tabla A.4: Resultados de R para las constantes del modelo

Categorías	β	Std. Error	t value	p value
Region2	-0.2612	0.00178242	-146.51942	0.00E+00
Region3	-0.1419	0.004042205	-35.11102	4.58E-270
Region4	-0.4193	0.018254749	-22.96666	1.00E-116
Area2	-0.6088	0.001848058	-329.41528	0.00E+00
Sexo2	0.1947	0.001740333	111.8976	0.00E+00
GrupoEdadRec1	0.8756	0.005597054	156.43434	0.00E+00
GrupoEdadRec2	1.9706	0.005855652	336.53177	0.00E+00
GrupoEdadRec3	1.6542	0.005973577	276.92332	0.00E+00
GrupoEdadRec4	1.0345	0.006045622	171.10851	0.00E+00
TieneSeguroP2	-0.2639	0.002662834	-99.11681	0.00E+00
TieneDiscapacidad2	0.6847	0.003689299	185.58013	0.00E+00
LeerEscribir2	-4.9420	0.005536023	-892.70223	0.00E+00
InternetU6M2	-0.8909	0.003255079	-273.68526	0.00E+00
ComputadorU6M2	-1.0807	0.003189289	-338.86433	0.00E+00
ComoTrabajaRec2	-1.1002	0.004114151	-267.41394	0.00E+00
ComoTrabajaRec3	-2.5380	0.005303234	-478.58027	0.00E+00
ComoTrabajaRec4	-0.8620	0.007133621	-120.83797	0.00E+00
ComoTrabajaRec5	-0.9475	0.010782458	-87.87377	0.00E+00
ComoTrabajaRec6	-1.6918	0.004305925	-392.91046	0.00E+00
ComoTrabajaRec7	-1.6420	0.009806552	-167.43982	0.00E+00
ComoTrabajaRec8	-2.3669	0.007041304	-336.14951	0.00E+00
ComoTrabajaRec10	-1.6921	0.004115878	-411.10919	0.00E+00
EstadoConyugalRec2	-0.3534	0.002514693	-140.52571	0.00E+00
EstadoConyugalRec3	-0.1681	0.004156735	-40.44961	0.00E+00
EstadoConyugalRec4	0.6095	0.006437863	94.67395	0.00E+00
EstadoConyugalRec5	-0.9414	0.005569366	-169.03722	0.00E+00
EstadoConyugalRec6	0.2699	0.002610063	103.39343	0.00E+00
EstadoConyugalRec7	-1.2581	0.006416538	-196.07299	0.00E+00
Idioma2	0.4480	0.007907132	56.65355	0.00E+00
Idioma3	0.6332	0.012701494	49.85477	0.00E+00
Idioma4	0.2171	0.00828422	26.20269	2.48E-151
Idioma5	0.5065	0.096809717	5.23147	1.68E-07
Idioma6	1.3924	0.010097566	137.89563	0.00E+00
Idioma7	1.2953	0.044184805	29.31521	6.64E-189
Idioma8	-1.1735	0.022626147	-51.86647	0.00E+00
Etnia2	0.1992	0.006139437	32.44246	6.92E-231
Etnia3	0.1761	0.006232621	28.25715	1.16E-175
Etnia4	0.4944	0.005354193	92.3347	0.00E+00
Etnia5	0.6055	0.006216787	97.40209	0.00E+00

Tabla A.5: Resultados de R de las variables categóricas del modelo

Apéndice B

Apéndice

B.1. Programas estadísticos

B.1.1. Software libre R

La siguientes información se obtuvo de la página oficial de R [1]

R es un entorno de software libre para computación estadística y gráficos. Compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS. Para descargar R , elija su espejo CRAN preferido .

Así R es un lenguaje y entorno de libre disposición para la computación estadística y los gráficos que proporciona son una amplia variedad de técnicas estadísticas y gráficas: modelado lineal y no lineal, pruebas estadísticas, análisis de series de tiempo, clasificación, agrupamiento, etc.

Se puede visitar el CRAN, que es una red de servidores ftp y web de todo el mundo que almacena versiones idénticas y actualizadas de código y documentación para R. Utilice el espejo CRAN más cercano para minimizar la carga de red.

El presente proyecto se realizó con RStudio, que es un entorno de desarrollo integrado (IDE) para R. Es software libre con licencia GPLv3 y se puede ejecutar sobre distintas plataformas (Windows, Mac, or Linux) o incluso desde la web usando RStudio Server.Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para el trazado, el historial, la depuración y la administración del espacio de trabajo.

[1] página web: https://cran.r-project.org/

B.1.2. Programa SPSS

SPSS son las siglas de Statistical Package for the Social Sciences, que en su traducción al castellano quedaría como "Paquete Estadístico para las Ciencias Sociales".

Es un programa o software estadístico que se emplea muy a menudo en las ciencias sociales y, de un modo más específico por las empresas y profesionales de investigación de mercados. Ello quiere decir que este software estadístico resultará de gran utilidad a la hora de llevar a cabo una investigación.[49]

Además, el SPSS fue muy útil para la investigación ya que el programa permite recodificar las variables y registros según las necesidades del usuario, esto quiere decir que si se necesita se puede expresar alguna variable de otra forma (recodificación de variables). También se tiene la ventaja que SPSS posee la capacidad para trabajar con bases de datos de gran tamaño.

B.2. Prueba Chi-cuadrado

Valor p

El valor p es una probabilidad que mide la evidencia en contra de la hipótesis nula. Las probabilidades más bajas proporcionan una evidencia más fuerte en contra de la hipótesis nula.Se utiliza el valor p para determinar si se puede o no puede rechazar la hipótesis nula, que dice que no existe ninguna asociación entre dos variables categóricas.

Interpretación

Para determinar si las variables son independientes, compare el valor p con el nivel de significancia. Por lo general, un nivel de significancia (denotado como (α o alfa) de 0.05 funciona adecuadamente. Un nivel de significancia de 0.05 indica un riesgo de 5 % de concluir que existe una asociación entre las variables cuando no hay una asociación real.

Valor $p \leq \alpha$: Las variables tienen una asociación estadísticamente significativa (Rechazar H_0) Si el valor p es menor que o igual al nivel de significancia, usted rechaza la hipótesis nula y concluye que hay una asociación estadísticamente significativa entre las variables.

Valor $p > \alpha$: No se puede concluir que las variables están asociadas (No se puede rechazar H_0) Si el valor p es mayor que el nivel de significancia, usted no puede rechazar la hipótesis nula, porque no hay suficiente evidencia para concluir que las variables están asociadas.

B.3. Test de Wald

El Test de Wald es un contraste de hipótesis donde se puede ver la coherencia de afirmar un valor concreto de un parámetro de un modelo probabilístico, dado que se tiene un modelo previamente seleccionado y ajustado. Así pués es un test más general.

Se aplica después de haber obtenido un modelo (una distribución cualquiera, una regresión simple, una regresión logística, etc) y luego se procede a hacer contraste de hipótesis sobre uno o varios parámetros: Por ejemplo, la media de la normal es 10, la pendiente de la recta es 0, el coeficiente principal de una regresión logística es algún valor, etc.

La fórmula del contraste es:

$$W = \frac{(\widehat{\theta} - \theta_0)^2}{Var(\widehat{\theta})}$$

donde:

W se compara con una distribución chi-cuadrada

$\widehat{\theta}$:estimaciones de los parámetros

θ_0: valores de los parámetros propuestos

Se observa que es un valor de la distancia entre lo observado y lo esperado, en un contraste de hipótesis se está valorando si lo que se ve, es o no muy distante de lo que se espera, en el caso de que sea cierta la hipótesis nula.

Es más usual para contrastar si es cero o no un determinado coeficiente que multiplica a una variable independiente en una regresión. Si el $p - valor$, como siempre, es menor que 0.05, se rechaza esa hipótesis nula que afirma que ese coeficiente es cero, y se entiende entonces que ese coeficiente no es cero y que, por lo tanto, el modelo es útil para representar una determinada relación. Si, por el contrario, el $p - valor$ es mayor que 0.05 eso significa que el valor del coeficiente podría ser perfectamente cero y estar viendo lo que vemos, por lo tanto, esa variable no influye a la hora de determinar la variable dependiente del modelo de regresión.

Alternativamente, la diferencia puede ser comparada con una distribución normal . En este caso el resultado es:

$$W = \frac{(\widehat{\theta} - \theta_0)^2}{error(\widehat{\theta})}$$

donde: $error(\widehat{\theta})$ es el error estándar de la estimación de máxima verosimilitud.

El error estándar para MLE puede ser estimado por $\frac{1}{\sqrt{I_n(MLE)}}$, donde I_n es la información de Fisher del parámetro.

En el caso multivariado, una prueba sobre varios parámetros a la vez se lleva a cabo utilizando una matriz de varianza.[23]

Un uso común para esto es llevar a cabo una prueba de Wald en una variable categórica por recodificación como diversas variables dicotómicas.

Bibliografía

[1] John H Aldrich and Forrest D Nelson. Linear probability, logit and probit models (quantitative applications in social sciences, vol. 45). beverly hills, 1984.

[2] T. Arnold and J. Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3:34-35, 2008.

[3] Gary S Becker. Human capital columbia university press. *New York*, 1964.

[4] Alfonso Castro. Regresión lineal. *Monografías de Matemática y Estadística, Quito*, 2008.

[5] J. S. Cramer. *Logit Models from Economics and Other Fields*. Cambridge University Press, Cambridge, 2003.

[6] Ministerio de Educación (s/f). Ministerio de Educación (s/f). Información legal. *Reuperado de https://educacion.gob.ec/wp-content/uploads/downloads/2012/09/A1_Base_Legal_11.pdf*, 4ta Edición, 2012.

[7] Instituto Nacional de Estadística y Censos (INEC). El Censo Informa: Educación. Censo de Población y Vivienda, 2010. *Recuperado de http://www.ecuadorencifras.gob.ec/wp-content/descargas/Presentaciones/capitulo_educacion_censo_poblacion_vivienda.pdf* , 2010.

[8] Sistema Integrado de Indicadores Sociales del Ecuador(SIISE). Definiciones del SIISE (Ficha metológica). *Recuperado de http://www.siise.gob.ec/siiseweb/PageWebs/glosario/ficglo_areare.htm*, 2019.

[9] Sistema Integrado de Indicadores Sociales del Ecuador(SIISE). Definiciones del SIISE (Ficha metológica). *Recuperado de http://www.siise.gob.ec/siiseweb/PageWebs/glosario/ficglo_etnlen.htm*, 2019.

[10] Santiago de la Fuente Fernández. Tablas de contingencia. *Análisis de variables categóricas, Universidad Autónoma de madrid (UNAM). Recuperado de http://www.estadistica.net/ECONOMETRIA/CUALITATIVAS/CONTINGENCIA/tablas-contingencia.pdf*, pages 10–11, 2011.

[11] Organización Mundial de la Salud (OMS). World Health Organization. *Recuperado de http: https: //www. who. int/ gender-equity-rights/ en/*, 2018.

[12] Sistema Nacional de Nivelación y Admisión (SNNA). Brecha en el acceso a educación superior e incremento de la oferta académica. *Ayuda memoria*, 2018.

[13] Banco Central del Ecuador. Cuentas nacionales trimestrales del Ecuador:Resultados de las variables macroeconómicas. *Recuperado de http: //contenido. bce. fin. ec/ home1/ estadisticas/ cntrimestral/ CNTrimestral. jsp*, 1, 2014,julio-octubre 2015,enero-marzo.

[14] Registro Civil del Ecuador. Estado Civil en Ecuador. *Recuperado de http: //www. ecuadorlegalonline. com/ consultas/ registro-civil/ estado-civil/*, 2017.

[15] Correa Delgado. Constitución Política del Estado (2008). Corporación de Estudios y Publicaciones. SEMPLADES. *Recuperado de http: //www. planificacion. gob. ec/ wp-content/ uploads/ downloads/ 2017/ 04/ Informe-a-la-Nacion. pdf*, Informe a la nación 2007-2017, 2017.

[16] Hosmer DW. and Lemeshow S. *Applied logistic regression*. A Wiley-Interscience Publication, New York, USA, 2000.

[17] The Analysis Factor. Logistic Regression Models for Multinomial and Ordinal Variables. *Recuperado de https: //www. theanalysisfactor. com/ logistic-regression-models-for-multinomial-and-ordinal-variables/*, 2018.

[18] Ronald Quispe Flores. *Regresión logística ordinal aplicado al estudio de la gravedad de lesiones por accidente de tránsito en la región Madre de Dios, 2010–2014*. PhD thesis, Universidad Nacional Mayor de San Marcos, Perú, Lima, 2016.

[19] Carlos Freile. Hitos de la historia de la educación en el Ecuador (siglos XVI-XX). *Publicación Trimestral del Instituto de Enseñanza y Aprendizaje de la Universidad San Francisco de Quito*, En revista para el aula. Número 13:4–6, 2015.

[20] Kleinbaum David G and Klein Mitchel. *Logistic regression: a self-learning text*. Editorial Springer Science & Business Media, 2002.

[21] Damodar N. Gujarati and Dawn C. Porter. *Econometría*. McGraw-Hill/Interamericana editores, S.A. DE C.V., México, D. F., 2010.

[22] Frank E. Harrell. Jr. 2001. regression modeling strategies.

[23] Frank E. Harrell. Regression modeling strategies. *as implemented in R package 'rms' version*, 3(3), 2014.

[24] Soledad Herrera. La importancia de la educación en el desarrollo: La teoría del capital humano y el perfil edad - ingresos por nivel educativo en Viedma y Carmen de patagones, Argentina. *Revista Pilquen • Sección Ciencias Sociales, Recuperado de* `file:///C:/Users/Usuario/Downloads/Dialnet-LaImportanciaDeLaEducacionEnElDesarrollo-3641304.pdf`, 2010.

[25] Xing Liu and Hari Koirala. Ordinal Regression Analysis: Using Generalized Ordinal Logistic Regression Models to Estimate Educational Data. *Recuperado de* `http://digitalcommons.wayne.edu/jmasm/vol11/iss1/21`, Voll.11:Iss.1, Article 21, 2012.

[26] F. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, page page 6878, 1951.

[27] Peter McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.

[28] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[29] McFadden. *Conditional Logit Analysis of Qualitative Choice Behaviour*. Frontiers in Econometrics, Academic Press, Nueva York, 1973.

[30] Jacob Mincer. *Schooling, Experience and Earnings*. University Press for National Bureau of Economics Research, USA, New York, 1974.

[31] MJ Norusis. Spss 13.0 statistical procedures companion. chicago: Spss, 2005.

[32] Gabriela Ossenbach. Políticas educativas en el Ecuador en el periodo 1944-1983. *En Estudios Interdisciplinarios de América Latina y el Caribe (EIAL), Recuperado de* `http://eial.tau.ac.il/index.php/eial/article/view/1048/1080`, vol.10, No 1, 1999.

[33] Jacinto Pallarés Mestre. *La metodología cuantitativa aplicada al estudio de la reincidencia en menores infractores*. PhD thesis, Universitat Jaume I, 2016.

[34] Organización para la Cooperación y el Desarrollo Económicos (OCDE). Escuelas y calidad de la enseñanza. *Recuperado de* `https://revistas.ucm.es/index.php/RCED/article/viewFile/RCED92921103O1A/18060`, 1991.

[35] Organización para la Cooperación y el Desarrollo Económicos (OCDE). Perspectivas económicas para América Latina. *Recuperado de* `http://www.planeducativonacional.unam.mx/CAP_00/Text/00_05a.html` , 2009.

[36] Carlos Daniel Rivera Chacón. *Estructura de Grupos Abelianos Ordenados*. EPN, Ecuador, Quito, 2018.

[37] Francisco Adrián Briones Rugel. *La educación en el Ecuador, situación y propuestas del sistema de vouchers educativos como alternativa*. PhD thesis, Repositorio de la Escuela Superior Politécnica del Litoral. Artículo Tesis de Grado. Recuperado de https://www.dspace.espol.edu.ec/handle/123456789/16995, Ecuador, Guayaquil, 2011.

[38] Nicolás Bajo Santos. Educación, economía global y mercado laboral. *Anuario Jurídico y Económico Escurialense, Recuperado de file:///C:/Users/Usuario/Downloads/Dialnet-EducacionEconomiaGlobalYMercadoLaboral-1143078.pdf*, 2005.

[39] SENESCYT. Rendición de cuentas 2015. Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación . *Recuperado de http://www.senescyt.gob.ec/rendicion2015/assets/presentaci%C3%B3n-rendici%C3%B3n-de-cuentas.pdf*, 2015.

[40] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley and Sons, 2012.

[41] Significados.com. Idioma. *Recuperado de https://www.significados.com/idioma/*, 2018.

[42] M Silva et al. Ordinal logistic regression models: application in quality of life studies. *Cadernos de Saúde Pública (CSP)*, 2008.

[43] M. R. Sumonkanti, D. y Rajwanur. Application of ordinal logistic regression analysis in determining risk factors of child malnutrition in Bangladesh. *Nutrition Journal 10:124*, 2011.

[44] Luna Tamayo. *Políticas educativas en el Ecuador, 1950-2010: Las acciones del Estado y las iniciativas de la sociedad.* PhD thesis, Universidad Nacional de Educación a Distancia (UNED), España - Madrid, 2014.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[46] Lester Thurow. Educación e igualdad económica. *Educación y sociedad*, 2:159–172, 1983.

[47] Tue Tjur. Coefficients of determination in logistic regression models—a new proposal: The coefficient of discrimination. *The American Statistician*, 63(4):366–372, 2009.

[48] William N Venables and Brian D Ripley. Tree-based methods. In *Modern Applied Statistics with S*, pages 251–269. Springer, 2002.

[49] Bienvenido Visauta Vinacua. Análisis estadístico con spss para windows. *Estadística Multivariante*, 1997.

[50] Strother H Walker and David B Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, pages 167–179, 1967.

[51] Terán Najas Rosemarie y Guadalupe Soasti. La educación laica y el proyecto educativo velasquista en el Ecuador, 1930-1950. *En Revista Ecuatoriana de Historia Procesos. Número 23. Quito.*, I semestre 2006.

[52] Chen C. y Hughes J. Using Ordinal Regression Model to Analyze Student Satisfaction Questionnaires. *IR Applications*, Vol. 1., 2004.

[53] Álvaro Sáenz y Diego Palacios. La dimensión demográfica en la historia ecuatoriana. *Nueva historia del Ecuador, Ensayos generales I. Corporación Editora Nacional. Quito.*, Vol.12, 1983.