UNIVERSIDAD DE NARIÑO

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

DEPARTAMENTO DE MATEMÁTICAS Y ESTADÍSTICA

ANÁLISIS MEDIANTE UN MODELO DE REGRESIÓN LOGÍSTICA DE LOS FACTORES QUE INFLUYEN EN EL PUNTAJE QUE LOS INDIVIDUOS OBTIENEN EN LAS PRUEBAS SABER-11

Autor:

García P. Jácome J.

San Juan de Pasto, Octubre 2018

UNIVERSIDAD DE NARIÑO

FACULTAD DE CIENCIAS EXACTAS Y NATURALES

DEPARTAMENTO DE MATEMÁTICAS Y ESTADÍSTICA

ANÁLISIS MEDIANTE UN MODELO DE REGRESIÓN LOGÍSTICA DE LOS FACTORES QUE INFLUYEN EN EL PUNTAJE QUE LOS INDIVIDUOS OBTIENEN EN LAS PRUEBAS SABER-11.

Tutor: Mg. Alvaro Bravo

Autor:

PEDRO PABLO GARCÍA LASSO

JOSÉ LIZARDO JÁCOME CABRERA

San Juan de Pasto Octubre 2018

NOTA DE RESPONSABILIDAD

Las ideas y conclusiones aportadas en este trabajo de grado son responsabilidad de los autores.

Articulo 1 del Acuerdo No. 324 de octubre 11 de 1966, emanado por el Honorable Concejo Directivo de la Universidad de Nariño

Nota de Aceptación

Firma del Presidente del jurado

Firma del Jurado

Firma del Jurado

San Juan de Pasto Octubre 2018

Resumen

Mediante este proyecto se trato de utilizar el modelo de regresión logística como una de las técnicas estadísticas multivariadas de más frecuente uso en las últimas décadas, se consideró cuestiones de tipo técnico como, numero de sujetos necesarios para aplicarla, situaciones en las que esta recomendado su uso, tipo de variables a las que se puede aplicar, la interpretación de los resultados, etc. Aplicamos el modelo de regresión logística para explicar si algunos factores influyen en el puntaje obtenido en las pruebas SABER-11. La conclusión es que se obtuvo fue que el modelo explica en un buen porcentaje la variable respuesta, además también puede predecir en base a las variables predictoras dicho puntaje.

Palabras claves: Pruebas saber, modelo, estadístico, regresión, logística, indice, significancia, intervalo.

Abstract

Through this project, we tried to use the logistic regression model as one of the multivariate statistical techniques most frequently used in recent decades. We considered technical issues such as the number of subjects needed to apply it, situations in which it is recommended. use, type of variables to which it can be applied, interpretation of results, etc. We apply the logistic regression model to explain if some factors influence the score obtained in the SABER-11 tests. The conclusion is that it was obtained that the model explains in a good percentage the response variable, in addition it can also predict based on the predictor variables said score.

Keywords: Tests know, model, statistical, regression, logistics, index, significance, interval.

Índice general

Índice de cuadros

Capítulo 1

INTRODUCCIÓN

La educación es uno de los temas más preocupantes a nivel mundial ya que su buen desempeño hace que los países surjan. En algunos países en los que, al desarrollo educativo, le ha sido inyectado un mayor presupuesto e inversión se ha visto reflejada en un notable desarrollo. En nuestro entorno la preocupación por la educación no viene de los estamentos gubernamentales, sino que está a cargo de los padres que quieren un mejor futuro para sus hijos. Ellos, incluso, invierten en el sistema privado de educación para mejorar la educación de los hijos y no por gusto, sino por los escasos cupos que ofrece y de los que disponen las universidades estatales. La probabilidad de obtener un cupo en las universidades estatales es muy baja, por este motivo se han creado centros en los cuales se prepara a los estudiantes para presentar los exámenes de estado que sirven como requisito para entrar a la educación superior en Colombia.

En la actualidad hay una gran cantidad de centros educativos que ofrecen cursos de preparación para presentar las pruebas ICFES, SABER-11, y que siempre han estado a la

vanguardia de los cambios que estas han tenido a lo largo de su existencia, cuyo fin es el de ayudar a los estudiantes que recién egresan de bachillerato a obtener un cupo en la educación superior.

Por otro lado, el interés de este trabajo es hacer un análisis estadístico con el cual se pueda determinar las variables que están relacionadas con los resultados que obtienen los estudiantes en las pruebas; para ello, se realiza una encuesta a un grupo de estudiantes que tomaron el curso de preparación para las pruebas y así determinar cuáles de ellas tienen mayor correlación.

Para este fin el trabajo consta de cuatro capítulos, en el primer capítulo se da una introducción, se plantea el problema, se da una justificación del porqué del trabajo, se plantean los objetivos tanto general como específicos y se menciona la metodología aplicar. En el segundo capítulo se refiere al marco teórico, en el cual se menciona una breve historia de las pruebas de estado para el ingreso a la educación superior, también el propósito de los centros pre-ICFES, además de conceptos estadísticos, los cuales ayudan al desarrollo del trabajo, como modelos estadísticos y el centro del trabajo que es el modelo de REGRECIÓN LOGÍSTICA.

El tercer capitulo se centra en el análisis de los datos y la obtención del modelo y el cuarto capítulo están las conclusiones que el modelo arrojó.

1.1. Planteamiento Del Problema

Supondríamos que después de terminar el bachillerato cuando han transcurrido por lo menos trece años de estudio, nos referimos a los jóvenes que han cursado el ciclo de acuerdo a lo requerido por el ministerio de educación nacional, estarían en capacidad de presentar dicha prueba y obtener un puntaje regular que le permita ingresar a la educación superior; sin embargo, los jóvenes se inscriben a distintos cursos pre-ICFES, que se ofrecen en diferentes academias, para obtener un mejor resultado en las Pruebas Saber. Se hace necesario, entonces, preguntar los motivos que justifican la existencia de estos cursos.

Después de estos trece años que los estudiantes pasan en una institución educativa, se esperaría que no necesiten una preparación adicional para enfrentar la prueba saber 11. Sin embargo, cada día es más evidente la persistencia de los cursos pre-ICFES como herramienta de los colegios y familias para potenciar los resultados de sus hijos. Se justifica la búsqueda de estas herramientas en el deseo de estudiantes y familias de obtener un buen puntaje, ya que éste determinará si se accede o no, y cómo se accede a la educación superior. Se configura así la prueba estandarizada *SABER-11* como una medición clasificatoria de aspirantes

Las empresas educativas que ofrecen sus servicios como preparación para el examen de estado, tienen un profundo conocimiento en la estructura de la prueba, como son: el tipo de preguntas y cómo resolverlas en un determinado tiempo. En una institución educativa, sus docentes están más preocupados por cumplir con los estándares, que en preparar a sus estudiantes para enfrentar dicha prueba. En las instituciones donde se dictan estos cursos

se prepara a los estudiantes exclusivamente para enfrentar dicha prueba, haciendo varios simulacros.

Los exámenes que presentan los jóvenes en las diferentes materias tienen la forma de la prueba estatal; las notas pasan a segundo plano, siendo más importante el análisis que deben hacer a las preguntas que se les propone.

Además de la preparación de la escuela tradicional, hay que sumarle otros aspectos como: el colegio donde está estudiando o ha terminado el bachillerato, si tiene modalidad y qué tipo de modalidad es, también se debe tener en cuenta la situación socio económica, lugar donde vive, educación de los padres, como también la situación laboral de los mismos; todos estos aspectos conllevan a hacer una pregunta ¿Qué aspectos de los estudiantes influyen para que ellos obtengan puntajes óptimos en las pruebas *SABER-11*?

1.2. Objetivos

1.2.1. Objetivo General

Analizar qué factores influyen en el puntaje que se obtienen en las pruebas **SABER-11**.

1.2.2. Objetivos Específicos

- Hacer una descripción estadística de los datos
- Comprobar si el puntaje de las pruebas **SABER-11** tiene alguna relación con la situación social del individuo que es objeto de investigación
- Formular un modelo de regresión logística para determinar qué factores tiene mayor influencia en el puntaje que obtienen en las pruebas **SABER-11**.

1.3. Justificación

Cuando se llega la hora de matricularse al grado once, se espera que los jóvenes tengan claro que va a hacer después de terminar de estudiar, que va a hacer de su futuro, que carrera escoger, la universidad donde va a cursar sus estudios superiores. Pero para ello sabemos que hay un ítem que le permite seguir o no estudiando, llamado de puntaje Icfes, que según diríamos mide el conocimiento adquirido durante la etapa estudiantil. en la mayoría de las ocasiones solamente nos preguntamos cual fue el puntaje que obtuvo, que número figura en su tarjeta y con ello si le alcanza para para seguir estudiando, pero no tenemos en cuenta que factores han tenido influencia para obtener dicho puntaje. En el trabajo que vamos a

desarrollar investigaremos que factores influyen para obtener un buen puntaje, como la edad, el sexo, lugar de procedencia, la escolaridad de los padres, el tipo de bachillerato que cursó, la profesión que desea para su vida, que influyen para obtener un buen puntaje.

1.4. Metodología

La población que se va a analizar es un grupo de 60 (sesenta) estudiantes escogidos al azar en quienes se va a observar características tales como: lugar de procedencia, estrato socioeconómico, colegio donde terminó, el tipo de bachillerato que curso y la preparación académica de los padres. Se quiere observar si estas variables tienen alguna incidencia en el puntaje.

Mediante una encuesta se quiere analizar el puntaje obtenido con respecto a cada una de las variables a considerar. En este caso, la variable dependiente o variable respuesta será el puntaje global, el cual se categorizará con las expresiones "alto" para los puntajes superiores a 310 y "bajo" para el caso contrario y las variables independientes serán los puntajes obtenidos en cada una de las materias evaluadas y los aspectos mencionados anteriormente que se encuentran en la encuesta. La encuesta se va a entregar a un grupo de 60 (sesenta) estudiantes que se encuentran matriculados en la institución pre- ICFES José Alfredo Peña en la ciudad de Pasto, quienes nos van a facilitar los datos de manera voluntaria.

Capítulo 2

Marco Teórico

2.1. Breve Historia De Las Pruebas Saber

En Colombia, el origen de estas pruebas se encuentra en una institución que ha logrado un reconocimiento y ha marcado una impronta muy profunda en el modo de ser educativo del país. El Servicio Nacional de Pruebas del ICFES ha introducido en la vida colombiana un modo de ser evaluativo. No es temeroso afirmar que la evaluación en nuestro país no sería lo que es sin la influencia del Servicio Nacional de Pruebas. Su labor a partir de mediados de la década de los años sesenta se ha constituido, y no es exagerado decirlo, en un hito del desarrollo educativo y social del país. Por esta razón, su surgimiento y desarrollo acredita constituirse en un objeto de investigación: de investigación evaluativa, educativa y social.

Una investigación rigurosa siempre debe ir a las fuentes y respetarlas. Las fuentes son, sin duda, origen de sentido, de sabiduría, de pertinencia.

En 1968 se crea el ICFES y una de sus dependencias, el Servicio Nacional de Pruebas (SNP), realiza los primeros exámenes nacionales. El SNP surgió a partir de la reestructuración realizada al Servicio de Admisión Universitaria y Orientación Profesional. En este primer examen se evaluó: aptitud matemática, aptitud verbal, razonamiento abstracto, relaciones espaciales, ciencias sociales y filosofía, química, física, biología e inglés.

En 1980 se reglamentan los exámenes de estado para ingreso a la educación superior. A partir de este año, los resultados obtenidos se convierten en requisito para el ingreso a cualquier programa de pregrado dentro del territorio nacional. Desde 1980 y hasta 1999 el examen incluyó nueve pruebas, agrupadas en cinco áreas. Los resultados se entregaban a cada estudiante por prueba, por área, promedio de las pruebas, y puntaje total, obtenido como la sumatoria de los puntajes de las cinco áreas. Sin embargo, a partir del año 2000 el examen de estado para ingreso a la educación superior presenta una transformación, cambiando su enfoque de contenidos a un enfoque por competencias. El fenómeno de la globalización lleva tras de sí otro tipo de exigencias sociales, políticas, culturales y económicas, que, unido con los nuevos propósitos educativos del país, forzaron la implementación de un nuevo tipo de evaluación.

A partir del año 2005 el examen de estado para ingreso a la educación superior tiene los siguientes propósitos:

- Servir como un criterio para el Ingreso a la Educación Superior.

- Informar a los estudiantes acerca de sus competencias en cada una de las áreas evaluadas, con el ánimo de aportar elementos para la orientación de su opción profesional.

- Apoyar los procesos de autoevaluación y mejoramiento permanente de las instituciones escolares.

- Constituirse en base e instrumento para el desarrollo de investigaciones y estudios de carácter cultural, social y educativo.

- Servir de criterio para otorgar beneficios educativos.

2.2. Propósito De Los Centros Preicfes

Esta estrategia educativa busca promover en los estudiantes el desarrollo de habilidades de pensamiento y competencias, que les permita asumir con seguridad la prueba **SABER-11** y obtener desempeños superiores en ella. Durante el desarrollo del llamado "curso Pre-icfes", se realizan sesiones de trabajo con expertos, que las refuerzan en diferentes niveles de dificultad:

- Comprensión lectora

- Análisis de gráficos

- Habilidades de pensamiento

- Desarrollo de competencias

Estos programas no pretenden enseñarle al estudiante lo que no aprendió durante más de seis años. Su objetivo es que sirvan como recorderis y refuerzo de todos los conocimientos

adquirido. Los cursos de este tipo pretenden desarrollar competencias, destrezas y habilidades en cada tipo de pregunta. El estudiante tiene el conocimiento. Lo que se potencializa en los pre- ICFES es el manejo del tiempo, la comprensión de lectura, el razonamiento lógico y la aplicación de sus conocimientos dentro de un contexto determinado. Y es que presentar hoy los ICFES no es lo mismo que en años anteriores. En el 2000, la prueba que se realiza en todo el país cambió su formato. El tipo de preguntas presentadas no son de las que puede responder de memoria, pues se utiliza el método de selección múltiple en el que el estudiante evaluado debe aplicar sus conocimientos dentro de un contexto. Las áreas de estudio tampoco son las mismas. A las tradicionales biologías, matemáticas, lengua castellana, física y química; se sumaron historia, idiomas y filosofía, áreas del conocimiento que hoy son la piedra en el zapato de los estudiantes de once. De este grupo de preguntas también hacen parte las de libre elección como medios de comunicación, medio ambiente y violencia y sociedad. Antes la calificación más baja de los estudiantes era en matemáticas, física y química. Pero desde el 2000, la mayoría se raspan es en filosofía por mala comprensión lectora. En esta área se debe poner mucho énfasis en los pre-ICFES, sin dejar a un lado las materias tradicionales. Y es que, hasta la fecha y el día de la prueba cambiaron. El examen que se realizaba en agosto, ahora se lleva a cabo en octubre. Ya no es durante un fin de semana normal, sino domingo y lunes, pues los grupos religiosos cristianos ganaron una tutela al Estado en la que pedían respetar el sábado como su día sagrado y de descanso.

2.3. Modelados Estadísticos

Un modelo es una representación formal de un sistema real, con el que se pretende aumentar su comprensión, hacer predicciones y ayudar a su control. Los modelos pueden ser fisicos (descritos por variables medibles), análogos (diagrama de flujo) y simbólicos (matemáticos, lingüísticos, esquemáticos). Los modelos matemáticos o cuantitativos son descritos por un conjunto de símbolos y relaciones lógicomatemáticas. Para la construcción de un buen modelo es necesario contar con leyes (por ejemplo, físicas) que describan el comportamiento del sistema. También es importante la experiencia, la intuición, la imaginación, la simplicidad y la habilidad para seleccionar el subconjunto más pequeño de variables. El primer paso es establecer el problema en forma clara y lógica delimitando sus fronteras; luego viene la recogida y depuración de datos; el diseño del experimento; las pruebas de contrastes; la verificación del modelo y la validación de las hipótesis. Por ejemplo, un análisis de sensibilidad determinara el grado de influencia en la solución del modelo, debida a variaciones en los parámetros. Un modelo debe ser una buena aproximación al sistema real, debe incorporar los aspectos importantes del sistema y debe resultar fácil de comprender y manejar.

Un factor muy importante es que haya una alta correlación entre lo que predice el modelo y lo que actualmente ocurre en el sistema real.

Los tipos de modelos mas utilizados son los de regresión múltiple si las variables explicativas son cuantitativas y todo radica en la escogencia de estas variables, para escoger las variables adecuadas se debe tener en cuenta la correlación de ellas con la variable respuesta.

En contrapartida, si la variable "a explicar" es cualitativa, el marco de la regresión múltiple ya no es apropiado. El modelo utilizado más corrientemente es el modelo logit, que expresa la probabilidad de observar tal modalidad de la variable a explicar en función de las variables explicativas, cualitativas y eventualmente cuantitativas.

2.4. Modelo De Regresión Logística

Con el nombre de modelos de regresión se incluyen un conjunto de técnicas estadísticas que tratan de explicar cómo se modifica la variable dependiente o resultado, cuando cambian otra u otras variables, denominadas independientes o predictoras. Lo que caracteriza en principio a las distintas clases de modelos de regresión es la naturaleza de la variable dependiente; así, con variables continuas la clase de modelos de regresión lineal es la más utilizada; con variables dicotómicas lo es el modelo de regresión logística.

La regresión logística (RL) es uno de los instrumentos estadísticos más expresivos y versátiles de que se dispone para el análisis de datos en clínica y epidemiología. Su origen se remonta a la década de los sesenta (Confield, Gordon y Smith 1961); su uso se universaliza y expande desde principios de los ochenta debido, especialmente, a las facilidades informáticas con que se cuenta desde entonces. En los últimos años se ha verificado una presencia muy marcada de esta técnica, tanto en la literatura orientada a tratar temas metodológicos como en los artículos científicos biomédicos. Fiel reflejo de esta tendencia es que el empleo de la RL suponía el 32 % de los artículos publicados por American Journal of Epidemiology de 1986 a 1990 y el 68 % de los que aparecieron en el mencionado quinquenio en New England

Journal of Medicine, con lo cual quedó ubicada en el quinto puesto, solo superada por cuatro técnicas convencionales: t de Student, prueba Chi-cuadrado, análisis de la varianza y prueba de Fisher. La evaluación de los artículos publicados en Medicina Clínica entre 1962 y 1992 refleja una escasa utilización de los análisis multivariante, aunque se aprecia una tendencia al alza.

En general, parece observarse el uso de diversos análisis multivariantes, como la regresión logística, la regresión de Cox y otros análisis de supervivencia. El conocimiento de estas técnicas permitiría al lector la compresión de los artículos publicados en revistas médicas, con el fin de obtener el máximo beneficio de la lectura y ser capaz de evaluar el mérito, validez y las conclusiones de la investigación publicada y posteriormente, decidir si las mismas son aplicables a su propia práctica y experiencia.

- ¿QUÉ ES LA REGRESIÓN LOGÍSTICA?

Los métodos de regresión de variable dependiente cualitativa abarcan diferentes modelos que tratan de explicar y predecir una característica cualitativa a partir de los datos de otras variables conocidas, bien cuantitativas o cualitativas que actúan como variables explicativas. La característica que se quiere explicar puede ser: a) una cualidad que puede únicamente tomar dos modalidades (modelos binomiales), son las más utilizadas, b) una cualidad que puede tomar más de dos modalidades diferentes, exhaustivas y mutuamente excluyentes (modelos multinomiales), c) una característica con varias modalidades que presentan entre ellas un orden natural (modelos ordenados) y d) la característica a explicar corresponde a una decisión que puede suponer decisiones encadenadas (modelos anidados).

Como es conocido, el concepto de regresión hace referencia a la ley o fórmula matemática que traduce la relación entre variables correlacionadas. Generalmente cuando se quiere poner una variable en función de otra (o de otras), se acude al bien conocido recurso de la regresión lineal (simple o múltiple). Esta función utiliza normalmente el método de mínimos cuadrados y funciona fluidamente desde el punto de vista aritmético. Pero cuando la variable a explicar sólo puede tomar dos valores, es decir, la ocurrencia o no de un cierto proceso, al evaluar la función para valores específicos de las variables independientes se obtendrá un número que será diferente de 1 y de O (los valores posibles de la variable dependiente), lo cual carece de todo sentido. En este caso, la regresión lineal debe ser descartada, en cambio la RL se ajusta adecuadamente a esta situación. Mediante la RL se pretende es la probabilidad de que ocurra el hecho en cuestión como función de ciertas variables que se presumen relevantes o influyentes.

Por lo tanto, la RL consiste en obtener una función logística de las variables independientes que permita clasificar a los individuos en uno de los dos grupos establecidos por los dos valores de la variable dependiente. La función logística es aquella que halla, para cada individuo según los valores de una serie de variables (Xi), la probabilidad (p) de que presente el efecto estudiado. Una transformación logarítmica de dicha ecuación, a la que se le llama logit, consiste en convertir la probabilidad (p) en odds. De aquí surge la ecuación de la regresión logística, que es parecida a la ecuación de la regresión lineal múltiple.

- ¿DÓNDE Y CUÁNDO APLICARLA?

La RL se utiliza cuando queremos investigar si una o varias variables explican una variable dependiente que toma un carácter cualitativo. Este hecho es muy frecuente en medicina ya

que constantemente intentamos dar respuesta a preguntas formuladas en base a la presencia o ausencia de una determinada característica que no es cuantificable, sino que representa la existencia o no de un efecto de interés, como por ejemplo el desarrollo de un «evento cardiovascular», «un paciente hospitalizado muere o no antes del alta», «se produce o no un reingreso», «un paciente desarrolla o no nefropatía diabética», etc. Una de las ventajas de la RL es que permite el manejo de múltiples variables independientes (también llamadas covariables) con un número reducido de casos. Freeman (1987) ha sugerido que el número de sujetos debe ser superior a $(10)(k+1)$, donde k es el número de covariables. Pero hay que tener en cuenta que el tamaño de la muestra necesaria es inherente al tipo de estudio que se realiza. Como hemos mencionado anteriormente la RL tiene una doble función: explicativa y predictiva. Podemos usarla con finalidad descriptiva siendo posible ofrecer una descripción elocuente y útil, basándonos en una información reducida; un ejemplo clásico es cuando la probabilidad que se estima puede interpretarse como una tasa de prevalencia o de incidencia que dependa de una variable continua. Aunque hay estudios que ejemplarizan este enfoque hay que reconocer que esta variante ha sido poco explotada. Su utilización en la predicción es el uso más frecuente y extendido, enmarcado en los diferentes tipos de estudios, ya sean típicamente prospectivos con finalidad pronóstica (epidemiología clínica), estudios prospectivos con finalidad analítica (cohortes), estudios caso-control (riesgo atribuible) y en los ensayos clínicos. Quisiéramos en este punto resaltar que la RL es un instrumento muy útil para facilitar el tratamiento cuantitativo de los datos, pero no podemos aislarlo del diseño del estudio, so pena de cometer errores que nos conducirían a conclusiones erróneas. Hay que destacar que además de predecir riesgos, la RL puede servir para estimar la fuerza de la asociación de cada factor de riesgo de una manera independiente, es decir, eliminando la posibilidad de que un factor confunda

el efecto de otro.

- ¿CÓMO INTERPRETARLA?

Cuando se realiza una RL lo que se pretende es estimar los parámetros de la ecuación $(\beta_0, \beta_1, \beta_2, ...\beta_k)$ de la función que se pretende evaluar:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

Donde Z es el logaritmo neperiano (Ln) de la odds de padecer la enfermedad, el desenlace o el resultado que se está estudiando; β_0 es la ordenada en el origen de la función de regresión, $\beta_1, \beta_2, ...\beta_k$ representan los coeficientes de la pendiente de la recta y $X_1, X_2, ...X_k$ son las variables independientes o factores de riesgo. Si los datos se ajustan de manera satisfactoria a este modelo, se tendrá la fortuna de poder explicar la relación entre las variables independientes y la respuesta de una manera muy sencilla. Los coeficientes β_i expresan el logaritmo neperiano del odds ratio (OR) para cada factor de riesgo X_i. Por tanto, el OR se estima a partir de la fórmula:

$$OR = antilog(\beta_i) = e^{\beta_i}$$

Una vez que hemos construido nuestro modelo de **RL**, debemos primero analizar los coeficientes de regresión (β_i) de cada variable independiente para obtener sus **OR** y luego confeccionar el valor predictivo de cada variable independiente o bien del modelo en su conjunto.

Ahora se debe plantear dos objetivos:

1. Conocer la fuerza de asociación, a través de los OR, de cada uno de los factores de riesgo con el efecto estudiado de una manera independiente, es decir, eliminando la posibilidad de que un factor confunda el efecto de otro. Una vez obtenidos los coeficientes de regresión logística ($\beta_i = \beta$) de cada una de las variables del modelo.

   Para saber la fuerza de asociación (medida en OR) en el modelo de RL, sólo necesitamos calcular su antilogaritmo, o lo que es lo mismo hallar su exponencial,

   $$OR = antilog(\beta_i) = e^{\beta_i}.$$

   Hoy en día están disponibles diversos paquetes estadísticos (SAS, LIMDEP, SPSS) que facilitan estos cálculos. Uno de los más utilizados es el SPSS al que haremos referencia en cuanto a sus salidas en este artículo. Este programa nos permite obtener los coeficientes de regresión β_i(**B**), los errores estándar de los coeficientes (**SE**), el nivel de significación (**Sig**) de cada coeficiente a través del estadístico de Wald, el coeficiente de correlación parcial (**R**) que es una forma de ver la influencia de cada una de las variables independientes por separado con la variable dependiente, y los exponenciales de los coeficientes que como sabe son los **OR** de cada variable independiente con sus intervalos de confianza al 95 % o al nivel que previamente se haya estipulado

2. Confeccionar el valor predictivo de cada variable independiente o bien del modelo en su conjunto.

   Abordaremos ahora como obtener el valor predictivo del riesgo asociado, para ello partiremos de la ecuación siguiente:

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_k X_k$$

$$Z = Ln(odds)$$

Capítulo 3

Análisis Estadístico y Desarrollo Del Modelo

3.1. Análisis estadístico SPSS

Antes de enfocarnos en hacer una RL para determinar si los factores influyen en el puntaje, se debe establecer el grado de correlación que tienen las variables predictoras con la variable respuesta El análisis estadístico de la asociación (relación, covarianza, correlación) entre variables representa una parte básica del análisis de datos en cuanto que muchas de las preguntas e hipótesis que se plantean en los estudios que se llevan a cabo en la práctica implican analizar la existencia de relación entre variables.

La existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de las mismas; esta asociación entre variables no debe entenderse como una cuestión de todo o

nada, sino como un continuo que iría desde la ausencia de relación (independencia) al nivel máximo de relación entre ellas.

Este grado máximo se plasmaría en una relación determinista, esto es, el caso en que, a partir del valor de un sujeto cualquiera en una variable, se puede afirmar cuál será su valor en la otra variable.

Existen diferentes índices estadísticos orientados a resumir cuantitativamente la asociación entre dos variables categóricas. Aquí nos vamos a centrar en los dos siguientes:

El índice ji-cuadrado de Pearson (x^2):

$x^2 = \sum \frac{(O_i - e_i)^2}{e_i}$

donde O_i representa a cada frecuencia observada y e_i representa a cada frecuencia esperada.

· El índice x^2 toma el valor 0 cuando dos variables son independientes, siendo mayor que 0 cuando exista asociación entre ellas, tanto mayor cuanto más intensa sea esa correlación. Ahora bien, no tiene un límite máximo, lo cual supone una dificultad a nivel interpretativo.

· Sí que puede utilizarse para comparar la asociación entre variables en tablas de contingencia del mismo tamaño (I x J) y con el mismo n. · Muchos de los estadísticos que se han propuesto a posteriori a fin de evaluar la asociación entre variables categóricas se basan en el índice x^2.

El coeficiente phi de Pearson (ϕ):

$$\phi = \sqrt{\frac{x^2}{n}}$$

Puede oscilar entre 0 y $\sqrt{q-1}$, siendo q el número de modalidades de la variable que tenga menos de ellas.

En tablas de contingencia de 2 x 2 oscila entre 0 y 1, por lo que suele utilizarse en esta circunstancia principalmente en la práctica, caso en el que se han extendido las normas interpretativas sugeridas por Cohen a la hora de evaluar la intensidad de la asociación (tamaño del efecto) para este coeficiente:

$\phi \leq 0,3$= nivel bajo de asociación

$0,3 \leq \phi \leq 0,5$ = nivel medio de asociación

$\phi \geq 0,5$ = nivel alto de asociación.

El coeficiente de contingencia de Cramer (V de Cramer):

$$V = \sqrt{\frac{x^2}{n(q-1)}}$$

El coeficiente V de Cramer oscila entre 0 (independencia) y 1, de modo que cuanto más próximos a 1 sean los valores, ello indicará mayor intensidad en la asociación de las variables.

Para hacer un análisis estadístico de las variables a considerar en el modelo se realiza tablas de contingencia en las cuales se determina el grado de relación que existe entre las

variables predictoras y la variable respuesta. De ellas se elije las de mayor correlación.

Las variables que se van a utilizar son de tipo cualitativo y están codificadas de la siguiente forma:

Variable dependiente:

Puntaje$\leq$ 310 bajo código asignado 0.

Puntaje $\geq$ 310 alto código asignado 1

Codificaciones de variables categóricas

			Codificación de parámetro		
		Frecuencia	(1)	(2)	(3)
O.PADRE	0 N.A.	6	1.000	.000	.000
	1 Empleado	6	.000	1.000	.000
	2 Independiente	40	.000	.000	1.000
	3 Profecional	8	.000	.000	.000
ESTRATO	1	37	1.000	.000	.000
	2	15	.000	1.000	.000
	3	7	.000	.000	1.000
	4	1	.000	.000	.000
COLEGIO	privado	5	1.000		
	oficial	55	.000		
PDENCIA	0 Provincia	50	1.000		
	1 Pasto	10	.000		
EDAD.COD	2 16-20	17	1.000		
	1 mas de 20	43	.000		

Cuadro 3.1: CODIFICACIONES DE VARIABLES CATEGÓRICAS

A continuación, relacionamos las tablas cruzadas entre cada una de las variable predictoras y la variable respuesta.

Tabla cruzada PUNTAJE*EDAD COD

			EDAD COD		
			2	1	Total
PUNTAJE	bajo	Recuento	11	19	30
		% dentro de EDAD.COD	64.7 %	44.2 %	50.0 %
	Alto	Recuento	6	24	30
		% del total	35.3 %	55.8 %	50.0 %
Total		Recuento	17	43	60
		% del total	100 %	100 %	100.0 %

Cuadro 3.2: Tabla cruzada PUNTAJE*EDAD COD

	Valor	gl	Sign asint (btral)
Chi-cuadrado de Pearson	2.052	1	.152
Razón de verosimilitud	2.075	1	.150
N de casos válidos	60		

Cuadro 3.3: chi-cuadrado

		Valor	Significación aproximada
Nominal por Nominal	Phi	.373	.304
	V de Cramer	.373	.304
N de casos válidos		60	

Cuadro 3.4: Medidas simétricas

Tabla cruzada PUNTAJE*ESTRATO

			ESTRATRO				
			1	2	3	4	Total
PUNTAJE	bajo	Recuento	23	5	2	0	30
		% dentro de ESTRATO	62.2 %	33.3 %	28.6 %	0.0 %	50.0 %
	Alto	Recuento	14	10	5	1	30
		% dentro de ESTRATO	37.8 %	66.7 %	71.4 %	100.0 %	50.0 %
Total		Recuento	37	15	7	1	60
		% dentro de ESTRATO	100 %	100 %	100 %	100 %	100 %

Cuadro 3.5: Tabla cruzada PUNTAJE*ESTRATO

	Valor	gl	Sig asin(bital)
Chi-cuadrado de Pearson	6.142	3	.105
Razón de verosimilitud	6.625	3	.085
N de casos válidos	60		

Cuadro 3.6: chi-cuadrado

		Valor	Sign. aproximada
Nominal por Nominal	Phi	.320	.105
	V de Cramer	.320	.105
N de casos válidos		60	

Cuadro 3.7: Medidas Simétricas

Tabla cruzada PUNTAJE*O.PADRE

			O.PADRE				
			NA	empleado	indpte	profnal	Total
PUNTAJE	bajo	Recuento	3	2	24	1	30
		% dtro de O.PADRE	50.0 %	33.3 %	60.0 %	12.5 %	50.0 %
	alto	Recuento	3	4	16	7	30
		% dtro de O.PADRE	50.0 %	66.7 %	40.0 %	87.5 %	50.0 %
Total		Recuento	6	6	40	8	60
	% del total	dtro de O.PADRE	100 %	100 %	100 %	100 %	100 %

Cuadro 3.8: Tabla cruzada PUNTAJE*O.PADRE

	Valor	gl	Sign asint (btral)
Chi-cuadrado de Pearson	6.767	3	.080
Razón de verosimilitud	7.352	3	.061
N de casos válidos	60		

Cuadro 3.9: chi-cuadrado

		Valor	Significación aproximada
Nominal por Nominal	Phi	.336	.080
	V de Cramer	.336	.080
N de casos válidos		60	

Cuadro 3.10: Medidas simétricas

Tabla cruzada PUNTAJE*PDENCIA

			PDENCIA		
			provincia	Pasto	Total
PUNTAJE	bajo	Recuento	28	2	30
		% dentro de PDENCIA	56.0 %	20.0 %	50.0 %
	alto	Recuento	22	8	30
		% dentro de PDENCIA	44.0 %	80.0 %	50.0 %
Total		Recuento	50	10	60
		% dentro de PDENCIA	100.0 %	100.0 %	100.0 %

Cuadro 3.11: Tabla cruzada PUNTAJE*PDENCIA

	Valor	gl	Sign asint (btral)
Chi-cuadrado de Pearson	4.320	1	.038
Razón de verosimilitud	4.577	1	.032
N de casos válidos	60		

Cuadro 3.12: chi-cuadrado

		Valor	Significación aproximada
Nominal por Nominal	Phi	.268	.038
	V de Cramer	.268	.038
N de casos válidos		60	

Cuadro 3.13: Medidas simétricas

Tabla cruzada PUNTAJE*COLEGIO

			COLEGIO		
			privado	oficial	Total
PUNTAJE	bajo	Recuento	1	29	30
		% dentro de COLEGIO	20.0 %	52.7 %	50.0 %
	alto	Recuento	4	26	30
		% dentro de COLEGIO	80.0 %	47.3 %	50.0 %
Total		Recuento	5	55	60
		% dentro de COLEGIO	100 %	100 %	100 %

Cuadro 3.14: Tabla cruzada PUNTAJE*COLEGIO

	Valor	gl	Sign asint (btral)
Chi-cuadrado de Pearson	1.964	1	.161
Razón de verosimilitud	2.091	1	.148
N de casos válidos	60		

Cuadro 3.15: chi-cuadrado

		Valor	Significación aproximada
Nominal por Nominal	Phi	-.181	.161
	V de Cramer	.181	.161
N de casos válidos		60	

Cuadro 3.16: Medidas simétricas

3.2. Modelo de regresión logística

Como nuestro modelo corresponde a una regresión logística multivariada, miramos la necesidad primero de realizar un análisis bivariado entre cada una de las variables independientes (edad, sexo, estrato, educación de padre, educación de la madre, ocupación del padre, ocupación de la madre, tipo de bachillerato, lugar de procedencia, vocación) con la variable respuesta (puntaje) para escoger las de mayor significancia y poderlas emplear en nuestro modelo, para ello vamos a tomar las que sean menores o iguales a .25 grados de significancia.

Las siguientes tablas son obtenidas mediante el programa SPSS estadístico, en ellas nos muestra la significancia para hacer la mejor elección de las variables predictoras.

La regresión logística binaria es la técnica estadística que tiene como objetivo comprobar hipótesis o relaciones cuando la variable dependiente (resultado) es una variable binaria (dicotómica, dummy), es decir, que tiene solo dos categorías.

Sobre la bondad del modelo:

- Significación de chi-cuadrado del modelo en la prueba ómnibus:

Si la significación es menor de 0,05 indica que el modelo ayuda a explicar el evento, es decir, las variables independientes explican la variable dependiente.

- R-cuadrado de Cox y Snell, y R-cuadrado de Nagelkerke: Indica la parte de la varianza de la variable dependiente explicada por el modelo. Hay dos R-cuadrados en la regresión logística, y ambas son válidas. Se acostumbra a decir que la parte de la variable dependiente explicada por el modelo oscila entre la R-cuadrado de Cox y Snell y la R-cuadrado de Nagelkerke. Cuanto más alto es la R-cuadrado más explicativo es el modelo, es decir, las variables independientes explican la variable dependiente.

- Porcentaje global correctamente clasificado:

  Este porcentaje indica el número de casos que el modelo es capaz de predecir correctamente. En base a la ecuación de regresión y los datos observados, se realiza una predicción del valor de la variable dependiente (valor pronosticado). Esta predicción se compara con el valor observado. Si acierta, el caso es correctamente clasificado. Si no acierta, el caso no es correctamente clasificado. Cuantos más casos clasifica correctamente (es decir coincide el valor pronosticado con el observado) mejor es el modelo, más explicativo, por tanto, las variable independientes son predictoras del evento o variable dependiente. Si es modelo clasifica correctamente más del 50 % de los casos, el modelo se acepta.

Sobre la relación de las variables independientes con la variable dependiente:

- Significación de b: si es menor de 0,05 esa variable independiente explica la variable dependiente

- Signo de b: indica la dirección de la relación. Por ejemplo, a más nivel educativo mayor probabilidad que suceda el evento.

- Exp(b) exponencial de b: indica la fortaleza de la relación. Cuanto más alejada de 1 está más fuerte es la relación. Para comparar los exponenciales de b entre sí, aquellos que son menores a 1 deben transformarse en su inverso o recíproco, es decir, debemos dividir 1 entre el exponencial de b (pero solo cuando sean menores a 1).

- Análisis del modelo logístico binario:

Codificación de variable dependiente

Se muestra la codificacion que se dio para la variable respuesta en la cual se tiene en cuenta dos categorias.

Valor original	Valor interno
Bajo	0
Alto	1

Cuadro 3.17: Variable Dependiente

Se muestra la codificacion que se dio para las variables predictoras algunas de ellas tienen mas de dos categorias y el programa SPCS las categoriza obteniendo una variables dummy.

CODIFICACIONES DE VARIABLES CATEGÓRICAS

Codificaciones de variables categóricas

			Codificación de parámetro		
		Frecuencia	(1)	(2)	(3)
O.PADRE	0 N.A.	6	1.000	.000	.000
	1 Empleado	6	.000	1.000	.000
	2 Independiente	40	.000	.000	1.000
	3 Profecional	8	.000	.000	.000
ESTRATO	1	37	1.000	.000	.000
	2	15	.000	1.000	.000
	3	7	.000	.000	1.000
	4	1	.000	.000	.000
COLEGIO	privado	5	1.000		
	oficial	55	.000		
PDENCIA	0 Provincia	50	1.000		
	1 Pasto	10	.000		
EDAD.COD	2 16-20	17	1.000		
	1 mas de 20	43	.000		

Cuadro 3.18: CODIFICACIONES DE VARIABLES CATEGÓRICAS

Edad*puntaje

En estas se relaciona la significación que hay entre la variable puntaje y la variable predictora edad, que fue escogida por ser una de las variables de mayor correlación.

- Observamos que la significación chi-cuadrado es menor que 0.25.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que la variable predictora explica en un 55 % a la variable respuesta.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	4.011	1	.045
	Bloque	4.011	1	.045
	Modelo	4.011	1	.045

Cuadro 3.19: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	79.167	.065	.086

Cuadro 3.20: Resumen Del Modelo

			Pronosticado		
			PUNTAJE		Porcentaje
	Observado		bajo	alto	correcto
Paso 1	PUNTAJE	bajo	14	16	46.7
		alto	11	19	63.3
	Porcentaje global				55.0

Cuadro 3.21: Tabla de Clasificación

- **ESTRATO*PUNTAJE**

En estas se relaciona la significación que hay entre la variable puntaje y la variable predictora estrato, que fue escogida por ser una de las variables de mayor correlación.

- Observamos que la significación chi-cuadrado es mayor que 0.25.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que la variable predictora explica en un 55 % a la variable respuesta.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	2.236	2	.327
	Bloque	2.236	2	.327
	Modelo	2.236	2	.327

Cuadro 3.22: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	80.942	.037	.049

Cuadro 3.23: Resumen Del Modelo

			Pronosticado		
			PUNTAJE		Porcentaje
	Observado		bajo	alto	correcto
Paso 1	PUNTAJE	bajo	28	2	93.3
		alto	25	5	16.7
	Porcentaje global				55.0

Cuadro 3.24: Tabla de Clasificación

- **O.PADRE*PUNTAJE**

  En estas se relaciona la significación que hay entre la variable puntaje y la variable predictora ocupación del padre, que fue escogida por ser una de las variables de mayor correlacion.

  Observamos que la significación chi-cuadrado es 0.061.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que la variable predictora explica en un 63.3 % a la variable respuesta.

- **O.PADRE*PUNTAJE**

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	7.352	3	.061
	Bloque	7.352	3	.061
	Modelo	7.352	3	.061

Cuadro 3.25: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	75.825	.115	.154

Cuadro 3.26: Resumen Del Modelo

		Pronosticado			
			PUNTAJE		Porcentaje
	Observado		bajo	alto	correcto
Paso 1	PUNTAJE	bajo	24	6	53.3
		alto	16	14	46.7
	Porcentaje global				63.3

Cuadro 3.27: Tabla de Clasificación

- **PDENCIA*PUNTAJE**

  En estas se relaciona la significación que hay entre la variable puntaje y la variable predictora procedencia, que fue escogida por ser una de las variables de mayor correlacion.

  Observamos que la significación chi-cuadrado es 0.032.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que la variable predictora explica en un 60.0 % a la variable respuesta.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	4.577	1	.032
	Bloque	4.577	1	.032
	Modelo	4.577	1	.032

Cuadro 3.28: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	78.601	0.73	0.098

Cuadro 3.29: Resumen Del Modelo

		Pronosticado			
			PUNTAJE		Porcentaje
	Observado		bajo	alto	correcto
Paso 1	PUNTAJE	bajo	28	2	93.3
		alto	22	8	26.7
	Porcentaje global				60.0

Cuadro 3.30: Tabla de Clasificación

- **COLEGIO*PUNTAJE**

  En estas se relaciona la significación que hay entre la variable puntaje y la variable predictora colegio, que fue escogida por ser una de las variables de mayor correlacion.

  Observamos que la significación chi-cuadrado es 0.032.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que la variable predictora explica en un 60.0 % a la variable respuesta.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	2.091	1	0.148
	Bloque	2.091	1	0.148
	Modelo	2.091	1	0.148

Cuadro 3.31: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	81.086	0.034	0.46

Cuadro 3.32: Resumen Del Modelo

		Pronosticado			
	Observado		PUNTAJE bajo	alto	Porcentaje correcto
Paso 1	PUNTAJE	bajo	29	1	96.7
		alto	26	4	13.3
	Porcentaje global				55.0

Cuadro 3.33: Tabla de Clasificación

3.3. Regresión Logística Multivariada

El objetivo de esta técnica estadística es expresar la probabilidad de que ocurra un hecho como función de ciertas variables, supongamos que son $k(k \geq 1)$, que se consideran potencialmente influyentes. La regresión logística, al igual que otras técnicas estadísticas multivariadas, da la posibilidad de evaluar la influencia de cada una de las variables independientes sobre la variable respuesta y controlar el efecto del resto. Tendremos, por tanto, una variable dependiente, llamémosla Y, que puede ser dicotómica o politómica (en este trabajo nos referiremos solamente al primer caso) y una o más variables independientes, llamémoslas X.

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

La forma analítica en que la probabilidad objeto de interés se vincula con las variables explicativas es la siguiente.

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\ldots\beta_k X_k)}}$$

Esta expresión es la que se conoce como función logística; donde exp denota la función exponencial y $\beta_0, \beta_1, \beta_2 ... \beta_k$ son los parámetros del modelo. Al producir la función exponencial valores mayores que 0 para cualquier argumento, p tomará solo valores entre 0 y 1.

Si β es positiva (mayor que 0) entonces la función es creciente y decreciente en el caso contrario. Un coeficiente positivo indica que p crece cuando lo hace la variable.

CODIFICACIONES DE VARIABLES CATEGÓRICAS

Codificaciones de variables categóricas

			Codificación de parámetro		
		Frecuencia	(1)	(2)	(3)
O.PADRE	0	6	1.000	.000	.000
	1	6	.000	1.000	.000
	2	40	.000	.000	1.000
	3	8	.000	.000	.000
ESTRATO	1	37	1.000	.000	.000
	2	15	.000	1.000	.000
	3	7	.000	.000	1.000
	4	1	.000	.000	.000
COLEGIO	privado	5	1.000		
	oficial	55	.000		
PDENCIA	0	50	1.000		
	1	10	.000		
EDAD.COD	2	17	1.000		
	1	43	.000		

Cuadro 3.34: CODIFICACIONES DE VARIABLES CATEGÓRICAS

A continuación se presenta los resultados de la regresión logística y los coeficientes de las variables y la constante del modelo.

Observamos que la significación chi-cuadrado es 0.040.

- Se observa los índices de calcificación

- R cuadrado Cox y Snell y R cuadrado de Nagelkerke.

- También se tiene la tabla de clasificación en la cual muestra que las variables predictoras explican en un 71.7 % a la variable respuesta.

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	17.576	9	0.040
	Bloque	17.576	9	0.040
	Modelo	17.576	9	0.040

Cuadro 3.35: Pruebas ómnibus de coeficientes de modelo

Paso	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	65.601	0.254	0.339

Cuadro 3.36: Resumen Del Modelo

		Pronosticado			
			PUNTAJE		Porcentaje
	Observado		bajo	alto	correcto
Paso 1	PUNTAJE	bajo	24	6	80.0
		alto	11	19	63.3
	Porcentaje global				71.7

Cuadro 3.37: Tabla de Clasificación

Para nuestro modelo hemos considerado la siguiente función tomando como parámetros: Edad, O.P_3

$$p = \frac{1}{1+e^{-(32{,}864+(-394)edad+(-2{,}081)O.P\_3)}}$$

Se presenta la tabla de las variables de la ecuación de las cuales se selecciona las de mejor significancia

Variables en la ecuación

		Error				95 % C.I.		
	B	estándar	Wald	gl	Sig.	Exp(B)	Inf	Sup
Edad	-.394	.230	2.939	1	.086	.675	.430	1.058
Estro			1.474	3	.688			
Estro1	-22.179	40192.931	0	1	1	0	0	
Estro2	-21.357	40192.931	0	1	1	0	0	
Estro3	-22.593	40192.931	0	1	1	0	0	
O.Padre			3.207	3	.361			
O.Padre1	-1.572	1.468	1.146	1	.284	.208	.012	3.692
O.Padre2	-1.705	1.524	1.251	1	.263	.182	.009	3.606
O.Padre3	-2.081	1.195	3.031	1	.082	.125	.012	1.299
Pdencia	-1.620	1.049	2.385	1	1.122	.198	.025	1.546
Colegio	1.100	1.677	.430	1	.512	3.003	.112	80.351
Constante	32.384	40192.931	0	1	.999	1.159E14		

Cuadro 3.38: Regresión Logística Multivariada

Capítulo 4

Conclusiones y Blibliografía

4.1. Conclusiones

1. Este estudio demuestra la factibilidad de usar la regresión logística para predecir los puntajes que los individuos obtienen en las pruebas saber pro-11. Utilizando variables cualitativas.

2. Se observa que medida que aumente la edad disminuye la probabilidad de obtener un buen puntaje en las pruebas Saber Pro 11.

3. La estabilidad laboral del padre influye como factor para la obtención de un buen puntaje en los jóvenes que buscan ingresar a la universidad pública.

4. Con este estudio podemos observar que los estudiantes de provincia tienen mayor probabilidad de obtener un buen puntaje.

5. No podemos asegurar que los estudiantes que vienen de instituciones públicas

obtengan puntajes altos con referencia a los estudiantes que vienen de colegios privados ya que la muestra no es homogénea.

4.2. BIBLIOGRAFÍA

- https://help.xlstat.com/customer/es/portal/articles/2062460-
- http://www.osso.org.co/docu/tesis/2003/evaluacion/analisis.pdf
- Gabriel Molina y María F. Rodrigo Estadística descriptiva en Psicología Curso 2009-2010
- Barón-López, J.Bioestadística: métodos y aplicaciones
- Solanas, A., Salafranca, L., Fauquet, J. y Núñez, M. I. (2005). Estadística descriptiva en Ciencias del Comportamiento. Madrid: Thompson.
- Silva LC. Excursión a la regresión logística en ciencias de la salud. Madrid:Díaz Santos, 1994:3-11.

ENCUESTA DE RESULTADOS ICFES

Encuesta voluntaria realizada a los estudiantes del intituto preicfres José Alfredo Peña

Señora e Hijos

Se tomaron 60 personas a las cuales se les pregunto lo siguiente:

Sexo:

Masculino_____ Femenino_____

Estrato Socioeconomico _____

Escolaridad del padre _____

Escolaridad de la madre _____

Ocupación del padre _____

Ocupación de la madre _____

Lugar de procedencia : _____

Colegio de Procedencia:

Oficial _____ Privado _____

Modalidad de bachillerato:

Academico _____Técnico _____

Quiere seguir estudiando:

Si _____ No _____

En que tipo de universidad:

Ofícial _____ Privada ___ Cualquiera _____

Puntaje obtenido en las pruebas de icfes de Febrero _____

Puntaje por materia

Lectura critica _____

Matemáticas _____

Ciencias naturales _____

Ciencias Sociales Y ciudadano_____

Ingles_____