ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

MODELO ECONOMÉTRICO PARA LA DEMANDA DE CRÉDITO EN PERSONAS NATURALES EN ECUADOR

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERA MATEMÁTICA

PROYECTO DE INVESTIGACIÓN

CARMEN BELÉN CACHAGO LLUGLLUNA
carmen.cachago@epn.edu.ec

Directora: DR. ADRIANA UQUILLAS ANDRADE
adriana.uquillas@epn.edu.ec

QUITO, FEBRERO DE 2023

DECLARACIÓN

Yo CARMEN BELÉN CACHAGO LLUGLLUNA, 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.

Carmen Belén Cachago Llugllluna

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por CARMEN BELÉN CACHAGO LLUGLLUNA, bajo mi supervisión.

Dr. Adriana Uquillas Andrade
Director del Proyecto

AGRADECIMIENTOS

Agradezco profundamente a mis padres Marco y Rosario, porque con su sacrificio constante he tenido la oportunidad de ser a futuro una profesional. Además, me supieron enseñar que el trabajo dignifica y nos ayuda a salir de las más duras adversidades juntos.

Agradezco a mis hermanos Erika y Marco que me han protegido y estado allí en todo momento. A los miembros de mi familia, en general, que han colaborado incondicionalmente y de corazón.

A mi directora de tesis, Adriana Uquillas, considerando que fue muy paciente en el proceso y cuyo apoyo ha sido útil en conocimiento como profesional.

A mis amigos, dentro y fuera de la universidad. Lisbeth, con quien he podido contar en todo momento por ser mi mejor amiga. Shirley, quien es un gran ser humano y cuya compañía ha hecho que el camino universitario sea grato de vivir. Andrea, quien es una gran inspiración por su dedicación a mejorarse a sí misma como profesional y con quien también tuve el agrado de compartir buenas experiencias.

A tantos docentes de la Escuela Politécnica Nacional, que jamás les faltó el esfuerzo por compartir su conocimiento y experiencias vividas en sus carreras. A la facultad y a toda la universidad que cada día forja grandes profesionales en el país.

DEDICATORIA

A mis padres, son las personas más importantes en mi vida y cualquier logro cumplido lo celebraré junto a ellos.

A mis amigas Lisbeth, Andrea y Shirley. La experiencia universitaria es de las más difíciles que he vivido, pero vivirla junto a cada una de ellas la volvió inolvidable y maravillosa.

Índice general

Índice de cuadros

Índice de figuras

Resumen

El presente trabajo se ha planteado bajo objetivos que buscan la inclusión financiera como una herramienta de apoyo económico para la población ecuatoriana y para el Estado. Se plantea un modelo econométrico basado en información obtenida a partir un organismo oficial nacional. Este es la Encuesta Nacional Empleo, Desempleo y Subempleo (ENEMDU), que aporta a la descripción de las características demográficas y socioeconómicas del país durante un periodo específico en el tiempo.

A través de modelos logísticos y con corrección de sesgo, se intenta reconocer las variables que las personas naturales presentan y les genera la necesidad de obtener un crédito, en general cualquier tipo de crédito. Puesto que en la actualidad no existen muchos modelos sobre la demanda de crédito, se ha trabajado bajo análisis hechos en otros países, todo con su respectiva referencia.

Considerando que la oferta y demanda de crédito van de la mano, se intenta reconocer qué es lo que representa y genera en la población la necesidad de obtener un crédito, para sugerir cambios que permitan a esta población acceder al mismo. Finalmente, los resultados e interpretaciones pueden servir de un apoyo para quienes deseen trabajar con un enfoque más profundo o diferente. El software utilizado es conocido como R y también se observará el código que genera el modelo final. Además de la bibliografía que ha ayudado a conseguir los objetivos propuestos.

Palabras clave: modelo econométrico, corrección de sesgo, oferta y demanda de crédito, inclusión financiera.

Abstract

The present paper has been proposed, under objectives that seek financial inclusion as a tool of economic support for the Ecuadorian population and for the state. An economic model based on true and reliable information from the National Survey on Employment, Unemployment and Underemployment (ENEMDU) is proposed, which provides a description of the demographic and socioeconomic characteristics of the country over a considerable period of time.

Through the theory on Logistic Models and Logistic Models with bias correction, an attempt is made to recognize the variables that natural persons present and generate the need to obtain a loan, in general any type of credit. Since at present there are not many models on the demand for credit, we have worked under analyzes made in other countries, all with their respective reference.

Considering that the supply and demand for credit go hand in hand, an attempt is made to recognize what the need for credit represents and generates in the population, in order to suggest changes that will allow this population to access it.

The results and interpretations can be supportive for those who wish to work with a deeper or different approach. The software used is known as R and the code that generates the final model will be observed. In addition to the bibliography that has helped to achieve the proposed objectives.

Key-Words: econometric model, correction of bias, supply and demand for credit, financial inclusion.

Capítulo 1

Introducción

El crédito en el Ecuador ha sido tema de investigación en distintas áreas de estudio porque tienen resultados de interés general. Entre estos se encuentran: acreedores a crédito, trabajadores de instituciones y superintendencias bancarias, inversionistas, etc. Mismos que día a día trabajan para obtener el mayor beneficio posible a través del manejo del crédito.

La participación en el crédito es tan esencial para el beneficiario como para el desarrollo de la economía del país. A esta participación se la puede dividir en dos partes esenciales: la Demanda, que es la necesidad de obtener servicios financieros; y la Oferta, que es el acceso a estos servicios. Considerando la Ley de Say[1], el estudio se centra en reconocer los determinantes de la necesidad de crédito para el entendimiento de los canales utilizados y proceder, en lo posible, con cambios sobre el acceso al crédito.

A nivel microeconómico, la utilización de los servicios financieros da lugar a que los hogares puedan obtener oportunidades de negocio, que usualmente están limitadas por falta de capital; también facilita la inversión en capital humano, vital para la reducción de la pobreza y la desigualdad [1]. Es así, que el incentivar la participación al crédito es parte de un plan de desarrollo en la economía del hogar y también del Estado.

En el 2017, al 55 % de los Latinoamericanos les pertenecía una cuenta en una institución financiera. Países como Brasil y Chile se encuentran sobre el promedio de la región, mientras que países como Colombia y Ecuador están por debajo de la media. Si bien ha habido un incremento en la población consumidora de crédito con

[1]La ley de Say es un principio atribuido a Jean-Baptiste Say que indica que la demanda está determinada por la producción, y que solo produciendo se puede generar demanda

respecto a los años anteriores (37 % y 46 % en los años 2011 y 2014 respectivamente); esto ha demostrado que el avance dentro del país permanece por debajo de las expectativas a nivel latinoamericano [2].

Parte de los motivos por los cuales está reducida esta población que participa del crédito, se explica a través de la teoría de racionamiento. Esta teoría considera primordial el estudio de las distintas poblaciones que existen según la utilización y acceso que se tiene al respecto. A continuación, se tiene una ilustración sobre cómo funciona.

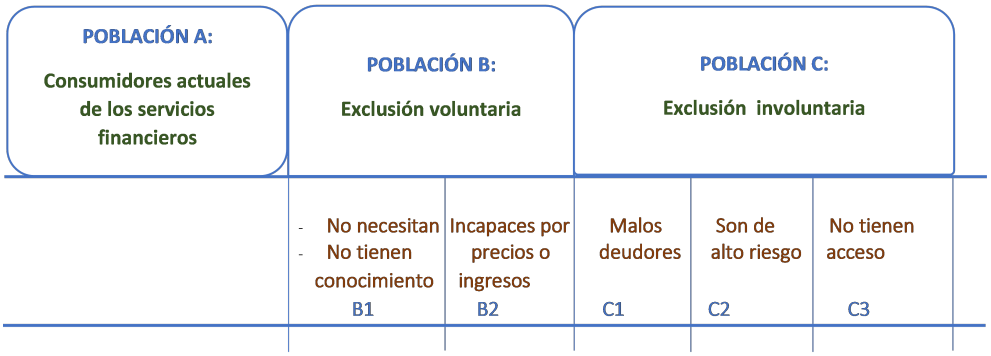

Figura 1.1: Teoría de Racionamiento
Elaborado: autor.

Según la Figura 1.1, la población adulta se puede clasificar en tres grandes grupos: el grupo A está conformado por quienes tienen acceso y usan los créditos. El grupo B, por quienes acceden a los créditos pero deciden voluntariamente no usarlos; ya sea porque no los necesitan (B1), los consideran muy caros, piensan que sus ingresos serán insuficientes para pagarlos o, simplemente, asumen que su solicitud será rechazada. En el subgrupo (B2) están incluidos aquellos que desean y necesitan el crédito pero no lo solicitan. El grupo C comprende a quienes son rechazados debido a que presentan deudas no pagadas o mal historial crediticio (C1) o a quienes las instituciones financieras los consideran muy riesgosos (C2). Aquellos que no tienen acceso al crédito porque no cuentan con una fuente de oferta accesible son el grupo (C3).

La transición de una persona dentro de uno de estos grupos hacia otro representa un cambio dentro del manejo de crédito en el país. Principalmente cuando el grupo al que se aspira ingrese la mayor cantidad de personas es el de la Población A. Esto bajo la condición de arriesgar lo menos posible a las instituciones financieras a trabajar con «malos» deudores.

1.1. Planteamiento del problema

En términos generales, se plantea descubrir las características que presentan los ecuatorianos que necesitan de un crédito. En lo posible, un enfoque sobre estas permite abrir paso a los cambios dentro de las políticas financieras y la concesión de créditos con las que actualmente se está trabajando. Dentro de la Figura 1.1 esta población de interés está representada dentro de los grupos B y C (aunque no completamente, debido a que se prefiere evitar trabajar con personas del grupo (C2)).

Para describir a la población ecuatoriana se ha considerado sus características como ciudadanos y como consumistas. Es decir, variables que representan su situación socioeconómica, demográfica y de consumo. Para ello, se hace uso de la información recogida por medio de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU). Esta encuesta representa un instrumento estadístico importante para estudiar la situación del empleo en el país [3].

1.2. Justificación

A nivel latinoamericano los estudios enfocados en la cuantificación de la demanda de crédito se han centrado, en su mayoría, en empresas. Donde usualmente la demanda se estima a través de agentes heterogéneos, flujo de caja, rentabilidad, tasa de interés, etc.; factores exclusivamente ligados a la posición financiera de una gran empresa[4]. Por otro lado, la necesidad del crédito en personas naturales es un tema cuya bibliografía requiere un especial enfoque y desarrollo.

Un estudio realizado en Ecuador en el año 2000 [5] sobre la demanda de 3 tipos de crédito: hipotecario, de consumo y vivienda, está enfocado en la evolución de la demanda mensual de crédito encontrando los factores que lo determinan. Ciertas limitantes se presentan en tal estudio: la cotización del dólar había tenido un fuerte impacto en tal época, el mercado y la política en general eran diferentes al actual y, sobre todo, no existía evidencia estadística confiable para la causalidad de las variables utilizadas.

Es así, que se sugiere un modelo econométrico para determinar la necesidad de crédito considerando características económicas, regionales y sociales, todo esto antes de la crisis COVID-19. Este estudio podrá ser de utilidad para contrastar estos perfiles encontrados con los perfiles que podrían encontrarse durante y después de la crisis sanitaria y así en estudios futuros promover políticas de inclusión financie-

ra.

1.3. Inclusión Financiera

La inclusión financiera en Ecuador se define como el acceso de las personas y empresas a los servicios que brinda el sistema financiero y es el resultado de factores que afectan su oferta y demanda. Este se ha visto limitada por la ausencia de incentivos de políticas públicas, fondos insuficientes, propiedad compartida, escasa educación financiera y desconfianza en entidades bancarias [6]. Además, conviene considerar características como disponibilidad de entidades financieras, procesos de facilitación de acceso a cuentas bancarias, instrucción sobre el manejo financiero; todo esto para describir por qué la población ecuatoriana no hace uso de estos recursos lo máximo posible y qué es lo que la limita.

Expertos en educación financiera señalan que la inclusión financiera contribuye en los hogares a reducir la vulnerabilidad ante situaciones imprevistas o adversas. Además evita que las personas acudan a estrategias que deterioran aún más la calidad de vida, como los prestamistas informales.[7] Esto a través de una promoción de un acceso asequible, oportuno y adecuado a una amplia gama de productos y servicios financieros regulados a todos los segmentos de la sociedad, mediante la aplicación de enfoques innovadores, incluyendo su sensibilización y educación financiera.

En este contexto, la inclusión financiera puede resumirse en 3 partes esenciales: acceso, uso y calidad [8]. Centrándose así el estudio en la demanda que existe y cómo esta puede limitar el acceso al crédito en el país. Los resultados ayudarán a describir la población que requiere el uso de crédito. Esto, a su vez, permitirá evidenciar la necesidad de inserción de las políticas públicas sobre las investigaciones que quieren fomentar la educación financiera en el país. Los hacedores políticos pueden tomar este, y muchos otros estudios, como las bases para una regulación sobre las políticas financieras enfocada en el bienestar, el dinamismo y la inclusión de la población sobre el uso de los productos que ofrecen las instituciones financieras existentes.

1.4. Objetivos

1.4.1. Objetivo General

Especificar un modelo de regresión Logit que permita describir y predecir la necesidad de crédito de un individuo en el país antes de la crisis COVID-19.

1.4.2. Objetivos Específicos

1. Establecer las características demográficas y socioeconómicas de los individuos que tienen la necesidad de solicitar un crédito.
2. Estimar y validar el modelo idóneo a través de pruebas estadísticas y realizar un backtesting de ajuste del modelo.
3. Segmentar la cartera de clientes recomendando políticas y procedimientos para mejorar la gestión en la concesión del crédito.

Capítulo 2

Marco Teórico

Los modelos estadísticos han ido desarrollándose con mayor frecuencia dentro del sistema económico y financiero por el buen uso de la información que se posee y por los resultados predictivos. A través de los años estos han sido perfeccionados según las necesidades de los estadísticos, entre los más reconocidos se encuentran: Modelos ARMA, Modelo Lineal Generalizado, Random Forest, Modelos Logit, Modelos de Corrección de Sesgo, etc.

Para elegir aquella metodología que se adecúa a los objetivos propuestos, se detallará lo más esencial de la teoría detrás de cada uno de los modelos citados con anterioridad. Para luego describir el proceso por el que se ha trabajado y los resultados que se han obtenido.

2.1. Modelos Lineales Generalizados

El objetivo de la econometría ha sido encontrar un modelo de relación entre variables, con el objeto de cuantificar la magnitud de la dependencia de ellas. Esto ayuda a resolver la duda sobre qué implicaciones existen entre las variables con las que se ha trabajado sobre una específica estadísticamente [9]. Esto está representado de la manera genérica siguiente:

$$y = f(x_1, x_2, ..., x_k, \mu|\beta) \tag{2.1}$$

que trata de explicar el comportamiento económico utilizando información proporcionada por un conjunto de k variables explicativas, así como por una variable aleatoria no observada y, esta se definirá más adelante. μ es la perturbación o error

aleatorios, y en él se recogen los posibles efectos que podrían influir en el comportamiento de la variable dependiente que no están reflejados en las variables explicativas.

Las variables observables constituyen el vector **x**, de dimensión $kx1$, o representado como una fila $x^t = (x_1, x_2, ..., x_k)$ y la relación de dependencia entre la variable y y el vector **x** envolverá, generalmente, un vector de parámetros que denotamos por $\boldsymbol{\beta}$. En una muestra de sección cruzada, diversos agentes económicos de una naturaleza similar proporcionan la información solicitada en un mismo instante de tiempo. Para este caso se utiliza un subíndice i para denotar los valores de las variables correspondientes a la unidad económica i-ésima; cuando se utilizan datos de series temporales se utilizará el subíndice t para denotar las observaciones correspondientes a un mismo instante de tiempo.

Es así, que se dispone de una lista de relaciones:

$$y_i = f(x_{1i}, x_{2i}, ..., x_{ki}, \mu_i|\beta), \quad i = 1, 2, ..., N \tag{2.2}$$

que relacionan los valores correspondientes y_i, x_{1i}, x_{2i}, ..., x_{ki} que componen cada una de las N observaciones muestrales. El modelo anterior está escrito para el caso de una sección cruzada de datos; en el caso de series temporales se tendría:

$$y_i = f(x_{1t}, x_{2t}, ..., x_{kt}, \mu_t|\beta), \quad t = 1, 2, ..., T \tag{2.3}$$

Cuando la función f es una función lineal, el modelo se reconoce como: Modelo de Regresión Lineal Múltiple o Modelo Lineal Generalizado (MLG). En este, los componentes del vector $\boldsymbol{\beta}$ son los coeficientes de las variables explicativas en el modelo lineal.

La variable aleatoria μ_i, a la cual se le referirá como «término de error» del modelo, entra aditivamente en el modelo y no necesita ir acompañada de ningún coeficiente. La variable y se denomina variable endógena, mientras que las variables x_1, x_2, ..., x_k se denominan variables explicativas del modelo. Los coeficientes β_1, β_2, ..., β_k recogen la magnitud del impacto de cada de las variables explicativas sobre la variable endógena.

Salvo algunas excepciones el MLG usualmente incluye un término constante:

$$y_i = \beta_1 + \beta_2 x_{2i} + ... + \beta_k x_{ki}, \quad i = 1, 2, ..., N \tag{2.4}$$

2.1.1. Características del Modelo

- El modelo es estocástico. Se supone que la esperanza matemática del término del error μ_i del modelo es cero; si, por el contrario, se tuviese $E(\mu_i) = a \neq 0$, este sería un efecto constante sobre y_i, y debería incluirse como parte de la constante β_1 en la Ecuación 2.4. Este supuesto propone una esperanza matemática nula para cada una de dichas variables aleatorias, ya correspondan a las distintas observaciones de sección cruzada o a períodos diferentes de tiempo. De modo que se formula $E(\mu) = 0_N$. En consecuencia, la matriz de covarianzas de μ, $Var(\mu)$, es simétrica definida positiva, de dimensión NxN.

$$Var(\mu) = \begin{pmatrix} Var(\mu_1) & Cov(\mu_1, \mu_2) & \dots & Cov(\mu_1, \mu_N) \\ Cov(\mu_2, \mu_1) & Var(\mu_2) & \dots & Cov(\mu_2, \mu_N) \\ \dots & \dots & \dots & \dots. \\ Cov(\mu_N, \mu_1) & Cov(\mu_N, \mu_2) & \dots & Var(\mu_N) \end{pmatrix}$$

  Suponiendo que la varianza es constante, que las perturbaciones son independientes, es posible escribir la Matriz de Varianza-Covarianza de la siguiente forma:

$$Var(\mu) = \sigma_\mu^2 I_N$$

- Los coeficientes del modelo $\beta_1, \beta_2, ..., \beta_k$ son constantes en tiempo. Se mantiene este supuesto y se explicará más adelante el por qué.

- Existe una relación causal entre las variables explicativas hacia la variable endógena.

- Las variables x no son linealmente dependientes. Además, son deterministas.

Los supuestos sobre las perturbaciones o «términos de error» y las distribuciones de las variables endógena y explicativas generan distintas maneras de estimar los modelos [10]. Se asume que las variables explicativas son de tipo determinístico, el modelo se reformula con N observaciones independientes $y_1, y_2, ..., y_N$ procedentes de distribuciones que pertenezcan a las familias exponenciales. Donde una función se dice que pertenece a la familia exponencial si su función de masa de probabilidad (si y es discreta) o su función de densidad (si es y continua) tiene la siguiente forma:

$$f(y_i, \theta_i, \psi) = \exp\left(\frac{y\theta_i - b(\theta_i)}{a(\psi)} + c(y_i, \psi)\right) \tag{2.5}$$

Las funciones $a(.)$, $b(.)$ y $c(.)$ varían dependiendo de la distribución de y_i. El parámetro de interés es θ, también conocido como parámetro canónico. Además, ψ se conoce como parámetro de escala o dispersión. A continuación, una tabla con ejemplos de distribuciones que pertenecen a las familias exponenciales [11]:

	Rango de y	$f(y)$	$\mu(\theta)$	Varianza $V(\mu)$	$\theta(\mu)$
Bernoulli $B(\mu)$	$\{0,1\}$	$\mu^y(1-\mu)^{1-y}$	$\frac{e^\theta}{1+e^\theta}$	$\mu(1-\mu)$	$log(\frac{\mu}{1-\mu})$
Binomial $B(k,\mu)$	$\{0,\cdots,k\}$	$\binom{k}{y}\mu^y(1-\mu)^{k-y}$	$\frac{ke^\theta}{1+e^\theta}$	$\mu(1-\frac{\mu}{k})$	$log(\frac{\mu}{k-\mu})$
Poisson $P(\mu)$	$\{0,1,2,\cdots\}$	$\frac{\mu^y}{y!}e^{-\mu}$	$\exp(\theta)$	μ	$log(\mu)$
Exponencial $Exp(\mu)$	$(0,\infty)$	$\frac{1}{\mu}exp(\frac{-x}{\mu})$	$-\frac{1}{\theta}$	μ^2	$\frac{1}{\mu}$
Normal $N(\mu,\psi^2)$	$(-\infty,\infty)$	$\frac{exp\{-(y-\mu)^2/(2\psi^2)\}}{\sqrt{2\pi}\psi}$	θ	1	μ

Cuadro 2.1: Ejemplos de Distribuciones de Familias Exponenciales

Elaborado: autor

Después de haber especificado la distribución de y, la función de enlace g es el segundo componente en decidir para el MLG. Esta función conecta los predictores de un modelo con el valor esperado de la variable de respuesta (dependiente) de forma lineal. Es decir:

$$g(\mu) = X^t\beta \tag{2.6}$$

donde,

$$\mu_i = E[y_i|x_i] \tag{2.7}$$

A partir de ahora se asume que las perturbaciones μ_i son variables aleatorias independientes idénticamente distribuidas normales con media cero y varianza constante. Esto permitirá establecer características útiles en el modelo más adelante.

$$\mu_i \sim \mathcal{N}(\mu, \sigma^2) \tag{2.8}$$

2.2. Estimación

Suponiendo que la distribución de y pertenece a la familia exponencial, es posible derivar estimaciones máximas de verosimilitud para los coeficientes de un MLG. Además, veremos que, aunque la estimación necesita una aproximación numérica, cada paso de la iteración puede estar dado por un ajuste de mínimos cuadrados ponderado. Dado que los pesos varían durante la iteración, la probabilidad se optimiza mediante un algoritmo de mínimos cuadrados reponderados iterativamente.

2.2.1. Propiedades de la Familia Exponencial

Considerando que se pretende hacer uso del estimador de máxima verosimilitud, para derivar los detalles del algoritmo de máxima verosimilitud se necesita discutir algunas propiedades de la masa de probabilidad o función de densidad $f(.)$ [11]. Por conveniencia, se considera que f es una función de densidad en la siguiente derivación. Sin embargo, las conclusiones también serán válidas para una función de masa de probabilidad.

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(y,\theta,\psi)dy \\ &= \int \frac{\partial}{\partial\theta} f(y,\theta,\psi)dy \\ &= \int \left\{\frac{\partial}{\partial\theta} log f(y,\theta,\psi)\right\} f(y,\theta,\psi)dy \\ &= E\left\{\frac{\partial}{\partial\theta}\ell(y,\theta,\psi)\right\} \end{aligned} \tag{2.9}$$

donde, $\ell(y,\theta,\psi) = log(y,\theta,\psi)$ denota la función de máxima verosimilitud. La deriva de la función ℓ con respecto de θ se denomina función de puntuación (score), para la que se conoce que:

$$E\left\{\frac{\partial^2}{\partial\theta^2}\ell(y,\theta,\psi)\right\} = -E\left\{\frac{\partial^2}{\partial\theta}\ell(y,\theta,\psi)\right\}^2 \tag{2.10}$$

Tomando la primera y segunda derivadas de la ecuación 2.5 y considerando la ecuación 2.10, se tiene como resultado:

$$0 = E\left\{\frac{y-b'(\theta)}{a(\psi)}\right\} \quad \text{y} \quad E\left\{\frac{-b''(\theta)}{a(\psi)}\right\} = -E\left\{\frac{y-b'(\theta)}{a(\psi)}\right\}^2 \tag{2.11}$$

De lo que se puede concluir:

$$\begin{aligned} E[y] &= \mu = b'(\theta) \\ Var[y] &= V[\mu]a(\psi) = b''(\theta)a(\psi) \end{aligned} \tag{2.12}$$

Donde claramente y depende únicamente del parámetro de interés θ. Se asumirá que $a(\psi)$ será idéntica para todas las observaciones.

2.2.2. Estimador de Máxima Verosimilitud

Como se señaló anteriormente, el método de estimación de elección para β es la máxima verosimilitud. Como alternativa, la literatura se refiere a la minimización de la desviación. Se observará durante la siguiente derivación que ambos enfoques son idénticos. Suponiendo que el vector μ tiene una distribución: $\mu \sim \mathcal{N}(\mu, \sigma^2)$. Bajo esta hipótesis, se requiere maximizar la función de verosimilitud $L(\theta)$. Además, se asume que la distribución de $Y_i|x_i$ pertenece a una familia exponencial, es decir, cumple con la ecuación 2.5:

$$\begin{aligned} L(\theta) = L(\theta_1, \theta_2, ..., \theta_N) &= \prod_{i=1}^{N} \Big(f(y_i, \theta_i) \Big) \\ &= \exp \left\{ \sum_{i=1}^{N} \frac{y\theta_i - b(\theta_i)}{a(\psi)} + c(y_i, \psi) \right\} \end{aligned} \tag{2.13}$$

Los métodos de maximización de esta ecuación dependerán del modelo elegido a estimar. Por la naturaleza de la información que se posee se realizará un enfoque sobre 3 tipos de Modelos de Regresión: Logit, Probit y de Corrección de Sesgo.

2.3. Estimador del Modelo Logit

En el Modelo Logit se asume lo siguiente:

- $y_i \sim B(p_i, n_i)$, para $i = 1, 2, ..., N$

- Además:

$$\pi = \pi(x) = P(y = 1|x) = \frac{\exp(\sum_{i=1}^{k} x_i\beta_i)}{1 + \exp(\sum_{i=1}^{k} x_i\beta_i)} \tag{2.14}$$

Mediante transformaciones, la ecuación 2.14 se puede expresar en términos de lo que se conoce como transformación logit, de la forma:

$$log\left\{\frac{\pi(x)}{1-\pi(x)}\right\} = \sum_{i=1}^{k} x_i\beta_i \tag{2.15}$$

La función de verosimilitud para este tipo de modelos será:

$$log(L(\beta)) = \sum_{i=1}^{N} \{y_i log\pi(x) + (1-y_i)log(1-\pi(x))\} \tag{2.16}$$

2.3.1. Ajuste del Modelo Logit:

Aquí se espera encontrar los valores de β. Se inicia con una muestra de tamaño N de las variables aleatorias respuesta y. Es decir, se consideran N variables independientes Bernoulli $y_1, y_2, ..., y_N$[12] . Cada una de estas está asociada a una combinación de k variables explicativas $x_1, x_2, ..., x_k$. Si se denota $x_q = (x_{q1}, x_{q2}, ..., x_{qk})$, con $q = 1, 2, ..., Q$ la q-ésima combinación de los valores de las k variables explicativas, pueden ocurrir los siguientes casos:

1. Para cada individuo muestral existe una combinación diferente de niveles de las k variables explicativas $(Q = N)$. Esto significa que hay una única observación de la variable respuesta y_i en cada combinación de valores de las variables explicativas, y suele ocurrir cuando las variables explicativas son todas continuas.

2. A $(Q \leq n)$. Esto quiere decir que hay más de una observación de la variable respuesta en cada combinación de valores de las variables explicativas.

Denotando por n_q al número de observaciones muestrales con $X = x_q$ y por y_q al número de respuestas $y = 1$ de entre estas N_q observaciones, se dispone de una muestra de Q variables aleatorias independientes y_q con distribuciones binomiales $B(n_q, \pi_q)$, donde $\pi_q = P(y = 1|X = x_q)$. Por lo tanto, $E(y_q) = n_q\pi_q$ y $\sum_{q=1}^{Q} n_q = N$. Observando que y_q representa en general el número de respuestas $y = 1$ en cada $X = x_q$. En el caso en que no hay observaciones repetidas $(Q = N)$ los valores observados y_q corresponden a las respuestas binarias individuales (1 o 0) esto es $n_q = 1$ y por lo tanto la distribución de y_q en lugar de Binomial es Bernoulli.

El modelo de regresión logística muestral es de la forma:

$$\pi_q = \frac{\exp(\sum_{i=1}^{k} \beta_i x_{qi})}{1 + \exp(\sum_{i=1}^{k} \beta_i x_{qi})}$$

donde $x_{q0} = 1, \forall q = 1, 2, ..., Q$. En forma lineal:

$$L_q = log \left\{ \frac{\pi_q}{1 - \pi_q} \right\} = \sum_{i=1}^{k} \beta_i x_{qi} \tag{2.17}$$

y equivalentemente en forma matricial:

$$L = X\beta$$

donde L es el vector $Qx1$ de transformaciones logit $L = (L_1, ..., L_Q)^t$, β es el vector $k+1$ parámetros $\beta = (\beta_0, \beta_1, ..., \beta_k)^t$ y X es la matriz de variables explicativas.

$$X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,i} \cdots x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,i} \cdots x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{q,1} & \cdots & x_{q,i} \cdots x_{q,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{Q,1} & \cdots & x_{Q,i} \cdots x_{Q,k} \end{pmatrix}$$

La función de verosimilitud es el producto de las Q funciones de probabilidad de las Q binomiales independientes y_q dada por:

$$\prod_{q=1}^{Q} \binom{n_q}{y_q} \pi_q^{y_q} (1 - \pi_q)^{(n_q - y_q)}$$

cuyo núcleo (que alcanza el máximo en el mismo punto que la propia función de verosimilitud) es:

$$\prod_{q=1}^{Q} \pi_q^{y_q} (1 - \pi_q)^{(n_q - y_q)}$$

Luego, encontrando el logaritmo del núcleo de la función de verosimilitud,

$$\ell(\beta) = \sum_{q=1}^{Q} \left\{ y_q ln(\pi_q) + (n_q - y_q) ln(1 - \pi_q) \right\}$$

y derivando esta última expresión con respecto a cada parámetro e igualando a cero,

se obtienen las ecuaciones de verosimilitud

$$\frac{\partial}{\partial \beta_i}\ell(\beta) = \sum_{q=1}^{Q} y_q x_{qi} - \sum_{q=1}^{Q} n_q \pi_q x_{qi}, \quad i = 1, 2, \cdots, k$$

Como es conocido, la solución de estas ecuaciones es el estimador de máxima verosimilitud. Como señala Agresti [13], estas ecuaciones son no lineales y requieren de métodos iterativos para su solución. El método que suele ser utilizado en este caso es el de Newton-Raphson, este consiste en resolver en forma iterativa la siguiente expresión:

$$\beta^{(t+1)} = \beta^{(t)} - \mathbf{H}_{(t)}^{-1}\mathbf{u}^{(t)}$$

donde $\mathbf{u} = \frac{\partial}{\partial \beta}\ell(\beta)$, es el vector de primeras derivadas de $\ell(\beta)$, $\mathbf{H} = \frac{\partial^2}{\partial \beta \partial \beta^T}\ell(\beta)$ es la matriz Hessiana, $u^{(t)}$ es el vector u evaluado en $\beta^{(t)}$, $H_{(t)}$ es la matriz Hessiana H evaluada en $\beta^{(t)}$, y finalmente $\beta^{(t)}$ y $\beta^{(t+1)}$ representan la solución anterior y nueva respectivamente. Este procedimiento iterativo continúa hasta que los cambios en $\ell(\beta^{(t)})$ son suficientemente pequeños. El método iterativo presentado anteriormente puede ser escrito como:

$$\beta^{(t+1)} = \beta^{(t)} + (X^T W^{(t)} X)^{-1} X^T (\mathbf{y} - \boldsymbol{\mu}^{(t)})$$

donde $\mathbf{y} = (y_1, y_2, ..., y_Q)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_Q)^T$, $\mu_q = n_q \pi_q^{(t)}$, $\pi_q^{(t)}$ es π_q en $\beta^{(t)}$, W es una matriz diagonal con elementos $n_q \pi_q (1 - \pi_q)$ y $W^{(t)}$ es la matriz W evaluada en $\beta^{(t)}$. La solución de este método iterativo será el estimador de máxima verosimilitud del parámetro β del modelo logit que será denotado por $\hat{\beta}$. Por otro lado, la matriz de información de Fisher del modelo logit está dada por:

$$I_F(\beta) = -E\left(\frac{\partial^2}{\partial \beta \partial \beta^T}\ell(\beta)\right) = X^T W X$$

Este resultado es importante porque de acuerdo con la teoría de máxima verosimilitud la varianza del estimador de máxima verosimilitud está dada por la inversa de esta matriz, esto es:

$$V(\hat{\beta}) = (X^T W X)^{-1}$$

y para estimar esta matriz se considera:

$$\widehat{V(\hat{\beta})} = (X^T \widehat{W} X)^{-1}$$

donde $\widehat{W}$ es la matriz W evaluada en $\hat{\beta}$.

2.4. Estimador del Modelo Probit

Pese a que el Modelo Logit es uno de los más populares al momento de trabajar con respuestas binarias, existen otras opciones, en ciertos casos, más apropiadas. Considérese la función $g\left[\pi(x) = P(y = 1)\right]$, cuya forma:

$$g\left[\pi(x)\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

es conocida como función de enlace (link) g, diferente de la función logística. Según [13] el modelo tiene la forma:

$$\Phi^{-1}\left[\pi(x)\right] = \alpha + \beta x$$

Con función de enlace Φ^{-1}. Estudios proveen que esta función está próxima la la función de densidad de una variable aleatoria Normal con parámetros μ y σ^2 desconocidos. A este caso se le conoce como Modelo Probit.

La función de verosimilitud es la siguiente:

$$L(\beta) = ln\left\{\prod_{i=1}^{N}\left[\Phi\left(\sum_j \beta_j x_{ij}\right)\right]^{y_i}\left[1 - \Phi\left(\sum_j \beta_j x_{ij}\right)\right]^{n_i - y_i}\right\} \tag{2.18}$$

La diferenciación con respecto a β_j da como resultado el siguiente sistema de ecuaciones:

$$\sum_j \frac{n_i\left[y_i - \Phi\left(\sum_j \beta_j x_{ij}\right)\right]}{\Phi\left(\sum_j \beta_j x_{ij}\right)\left[1 - \Phi\left(\sum_j \beta_j x_{ij}\right)\right]}\phi\left(\sum_j \beta_j x_{ij}\right) = 0 \tag{2.19}$$

Con $\phi(.)$ la función de densidad de una normal estándar. Cuando la función de enlace no es de la forma canónica (el cual es el caso para el modelo logit), no existe un proceso para un estimador suficiente. Para resolver tal sistema de ecuaciones existe un método conocido como "Fisher scoring"[14]. La matriz de covarianza asintótica estimada para $\hat{\beta}$ tiene la forma:

$$\widehat{cov}(\hat{\beta}) = (X^T \hat{W} X)^{-1}$$

Para modelos Probit, $\hat{W}$ es la matriz diagonal con elementos:

$$\widehat{w}_i = n_i\left[\phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\right]^2 \Big/ \left\{\Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\left[1 - \Phi\left(\sum_j \hat{\beta}_j x_{ij}\right)\right]\right\}$$

El algoritmo de Raphson-Newton ayuda a calcular el estimador de máxima ve-

rosimilitud con un error estándar mínimo.

2.5. Modelo Logit con Corrección de Sesgo:

Se dice que se estudian datos de eventos raros, cuando existe una desproporción significativa con respecto a los valores de la variable respuesta. Esto es, las variables dependientes binarias «unos» son decenas a miles de veces menos (eventos, como guerras, vetos, casos de activismo político o infecciones epidemiológicas) que «ceros» (no eventos). En muchas literaturas, estas variables han resultado difíciles de explicar y predecir, un problema que parece tener al menos dos fuentes. Primero, los procedimientos estadísticos populares, como la regresión logística, pueden subestimar drásticamente la probabilidad de eventos raros. En segundo lugar, las estrategias de recopilación de datos comúnmente utilizadas son extremadamente ineficaces para los datos de eventos raros.

El miedo a recopilar datos con muy pocos eventos ha llevado a recopilar datos con un gran número de observaciones, pero relativamente pocas variables explicativas y mal medidas. Existen diseños de muestreo más eficientes para hacer inferencias válidas, como muestrear todos los eventos disponibles y una pequeña fracción de los no eventos. Esto permite a los académicos ahorrar hasta el 99 % de sus costos de recopilación de datos (no fijos) o recopilar variables explicativas mucho más significativas. [15]

2.5.1. Utilidad del Modelo

Aunque las propiedades estadísticas de los modelos de regresión lineal son invariantes para la media (incondicional) de la variable dependiente, no ocurre lo mismo con los modelos de variables dependientes binarias. La media de una variable binaria es la frecuencia relativa de eventos en los datos, que, además del número de observaciones, constituye el contenido de información del conjunto de datos.

Dados los recursos fijos, siempre existe una compensación entre recopilar más observaciones e incluir variables mejores o adicionales. En los datos de eventos raros, el miedo a recopilar conjuntos de datos sin eventos (y, por lo tanto, sin variación en y) ha llevado a los investigadores a elegir un gran número de observaciones con pocas variables explicativas y, en la mayoría de los casos, mal medidas. Esta es una opción razonable, dadas las limitaciones percibidas, pero resulta que existen estrate-

gias de recopilación de datos mucho más eficientes. Por ejemplo, los investigadores pueden recopilar todos los «unos» (o todos los disponibles) y una pequeña muestra aleatoria de «ceros» y no perder consistencia o incluso mucha eficiencia en relación con la muestra completa [16].

Considerando que la información real en los datos se encuentra mucho más en los eventos con respuesta $y = 1$ que en $y = 0$, se espera concentrarse en esta parte en particular. Pero, los investigadores deben tener cuidado de evitar el sesgo de selección. Afortunadamente, las correcciones para ello son posibles.

2.5.2. Notación

Como ya se observó en la sección 2.3 la variable de resultado y_i sigue una función de probabilidad de Bernoulli que toma el valor 1 con probabilidad π_i y 0 con probabilidad $1 - \pi_i$.

Entonces π_i varía sobre las observaciones como una función logística inversa de un vector x_i, que incluye una constante y $k - 1$ variables explicativas:

$$\begin{aligned} y_i &\sim Bernoulli(y_i, \pi_i) \\ \pi_i &= \frac{1}{1 + e^{-x_i\beta}} \end{aligned} \tag{2.20}$$

Bernoulli tiene función de probabilidad $P(y_i|\pi_i) = \pi_i^{y_i}(1 - \pi_i)^{(1-y_i)}$. El parámetro desconocido β es un vector $kx1$. Una forma alternativa de definir el mismo modelo es imaginando una variable continua no observada y_i^* distribuida según una densidad logística con media μ_i. Entonces μ_i varía sobre las observaciones como una función lineal de x_i. El modelo sería muy cercano a una regresión lineal si y_i^* fuese observada:

$$\begin{aligned} y_i^* &\sim Logistic(y_i^*, \mu_i) \\ \mu_i &= x_i\beta \end{aligned} \tag{2.21}$$

donde, $Logistic(y_i, \mu_i)$ es la densidad de probabilidad logística de un parámetro,

$$P(y_i^*) = \frac{e^{-(y_i^* - \mu_i)}}{(1 + e^{-(y_i^* - \mu_i)})^2} \tag{2.22}$$

Desafortunadamente, en lugar de observar y_i^*, solo se observa su realización dicotómica, y_i, donde $y_i = 1$, si $y_i^* > 0$ y $y_i = 0$, si $y_i^* \leq 0$. Por ejemplo y para el caso presente, si y_i mide si necesidad un individuo un crédito o no, entonces y_i^* muestra

cuán **propenso** es el individuo para necesitar del crédito. El modelo sigue siendo el mismo porque:

$$\begin{aligned} P(y_i = 1|\beta) = \pi_i &= P(y_i^* > 0|\beta) \\ &= \int_0^\infty Logistic(y_i^*|\mu_i)dy_i^* \\ &= \frac{1}{1+e^{-x_i\beta}} \end{aligned} \tag{2.23}$$

que es exactamente como en la ecuación 2.20. También se conoce que el mecanismo de observación, que convierte la y_i^* continua en la y_i dicotómica, genera la mayor parte del daño. Es decir, según simulaciones por parte de Zeng y King [16] tratando de estimar β a partir de una y^* observada y el modelo 2.21 se descubre que la estimación de máxima verosimilitud de β es aproximadamente insesgada en muestras pequeñas. Los parámetros se estiman por máxima verosimilitud, con la función de verosimilitud formada asumiendo independencia sobre las observaciones:

$$L(\beta|y) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{(1-y_i)} \tag{2.24}$$

Tomando logaritmos y usando la ecuación 2.20, la probabilidad logarítmica se simplifica a:

$$\begin{aligned} ln(L(\beta|y)) &= \sum_{y_i=1} ln(\pi_i) + \sum_{y_i=0} ln(1-\pi_i) \\ &= -\sum_{i=0}^{N} ln(1+e^{(1-2y_i)x_i\beta}) \end{aligned} \tag{2.25}$$

El análisis logit de máxima verosimilitud encuentra el valor de β que da el valor máximo de esta función, que se denomina $\widehat{\beta}$. La matriz de varianza asintótica, $V(\widehat{\beta})$, también se calcula para obtener los errores estándar. Cuando las observaciones se seleccionan aleatoriamente, o aleatoriamente dentro de estratos definidos por algunas o todas las variables explicativas, $\widehat{\beta}$ es consistente y asintóticamente eficiente (excepto en casos degenerados de perfecta colinealidad entre las columnas en X o perfecta discriminación entre ceros y unos).

El hecho de que los «unos» en los datos de eventos raros sean más estadísticamente informativos que los «ceros», se puede ver al estudiar la matriz de varianza.

$$V(\widehat{\beta}) = \left[\sum_{i=1}^{N} \pi_i(1-\pi_i)x_i'x_i\right]^{-1} \tag{2.26}$$

La parte de esta matriz afectada por eventos raros es el factor $\pi_i(1-\pi_i)$. La mayoría de las aplicaciones de eventos raros producen pequeñas estimaciones de $P(y_i = 1|x_i) = \pi_i$ para todas las observaciones. Sin embargo, si el modelo logit tiene algún poder explicativo, la estimación de π_i entre las observaciones para las cuales se dan los eventos raros (es decir, para las cuales $y_i = 1$) generalmente será mayor (y más cercana a 0.5, porque las probabilidades en los estudios de eventos raros son normalmente muy pequeñas [17] que entre las observaciones para las cuales $y_i = 0$. El resultado es que $\pi_i(1-\pi_i)$ generalmente será mayor para «unos» que para «ceros», por lo que la varianza (su inversa) será menor. En esta situación, adicionar «unos» harán que la varianza caiga más y, por lo tanto, son más informativos que los ceros adicionales [18] [19].

2.5.3. Selección de la variable dependiente

La estrategia habitual, como se conoce en econometría, es el muestreo aleatorio, donde todas las observaciones (x, y) se seleccionan al azar, o el muestreo estratificado exógeno, que permite que y se seleccione al azar dentro de las categorías definidas por x. La estrategia consiste en seleccionar y mediante la recopilación de observaciones (al azar o todas las disponibles) para las que $y = 1$ y una selección aleatoria de observaciones para las que $y = 0$.

Un "diseño de muestreo de partes iguales" sugiere que la fracción observada de «unos» en la muestra $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i = 0{,}5$ es óptimo en un número limitado de situaciones y cercano al óptimo en un gran número [18]. La única decisión real, entonces, es cuántos «ceros» recolectar. Si recopilar «ceros» no tuviera ningún costo, deberíamos recopilar tantos como podamos, ya que más datos siempre son mejores. Sin embargo, dado que la contribución marginal al contenido de información de las variables explicativas para cada cero adicional comienza a disminuir a medida que el número de «ceros» sobrepasa al número de «unos», a menudo no queremos recolectar más de (aproximadamente) **dos a cinco veces más «ceros» que «unos»**. [19]

2.5.4. Ajuste del Modelo

Se procede a ponderar los datos para compensar por diferencias en las fracciones de la muestra ($\overline{y}$) y de la población (τ) inducidas por el muestreo basado en la elección. El estimador de máxima verosimilitud de muestreo exógeno ponderado

resultante (según a Manski y Lerman [20]) es relativamente simple. En lugar de maximizar la probabilidad logarítmica en la ecuación 2.25, se maximiza la probabilidad logarítmica ponderada:

$$\begin{aligned} \ell_w(\beta) &= ln(L_w(\beta|y)) \\ &= w_1 \sum_{y_i=1} ln(\pi_i) + w_0 \sum_{y_i=0} ln(1-\pi_i) \end{aligned} \tag{2.27}$$

donde los pesos son $w_1 = \frac{\tau}{\bar{y}}$, $w_0 = \frac{1-\tau}{1-\bar{y}}$, y donde:

$$w_i = w_1 y_i + w_0(1-y_i) \tag{2.28}$$

Por otro lado, aplicando la función exponencial a esta ecuación se obtiene que la función de verosimilitud ponderada:

$$\mathcal{L}_w(\beta) = \prod_{i=1}^{N} \pi_i^{w_i y_i} (1-\pi_i)^{w_0(1-y_i)} \tag{2.29}$$

King y Zeng [16] notan que la función de verosimilitud ponderada dada en la ecuación 2.29 es similar a la del siguiente modelo de regresión binaria:

$$y_i \sim Bernoulli(\pi_i^*) \tag{2.30}$$

$$\pi_i^* = \pi_i^{w_i} \tag{2.31}$$

$$\eta_i = \sum_{j=1}^{k} \beta_i x_{ji} \tag{2.32}$$

Entonces usando la técnica de McCullagh y Nelder [10] en la ecuación 2.31 tenemos que una aproximación sesgo del estimador de máxima verosimilitud de los coeficientes de regresión logística para una muestra dependiente de la variable respuesta es dada por:

$$b = (X^T W_w X)^{-1} X^T W_w \boldsymbol{\xi} \tag{2.33}$$

donde $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_N)^T$ cuyos elementos se definen como:

$$\xi_i = -\frac{1}{2} \left(\frac{\pi_i^{*\prime\prime}}{\pi_i^{*\prime}} \right) Q_{ii}$$

donde $\pi_i^{*\prime} = \frac{\partial \pi_i^*}{\partial \eta_i}$, $\pi_i^{*\prime\prime} = \frac{\partial^2 \pi_i^*}{\partial \eta_i^2}$ y Q_{ii} es el i-ésimo elemento de la diagonal de $Q = X(X^T W_w X)^{-1} X^T$.

Entonces para encontrar la expresión para ξ_i, recordando que:

$$\pi_i^* = \pi_i^{w_i} = \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{w_1}$$

luego se puede obtener que:

- $\pi_i^{*\prime} = \frac{\partial \pi_i^*}{\partial \eta_i} = w_1 \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{w_1} \frac{1}{1+e^{\eta_i}} = w_1 \pi_i^{w_1}(1-\pi_i)$
- $\pi_i^{*\prime\prime} = \frac{\partial^2 \pi_i^*}{\partial \eta_i^2} = w_1 \left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right)^{w_1} \frac{1}{(1+e^{\eta_i})^2}(w_1 - e^{\eta_i}) = w_1 \pi_i^{w_1}(1-\pi_i)(w_1 - (1+w_1)\pi_i)$

Finalmente se obtiene:

$$\xi_i = 0{,}5 Q_{ii}((1+w_1)\pi_i - w_1) \tag{2.34}$$

que nos permite hallar el estimador de máxima verosimilitud b en un modelo de regresión logística ponderado.

2.6. Inferencia

Cuando se hayan estimado los parámetros del modelo de regresión logística múltiple se propone hacer inferencia para extrapolar los resultados muestrales a la población. Se hará énfasis en los contrastes sobre los parámetros y de bondad de ajuste de cada uno de los modelos.

2.6.1. Contrastes de bondad de ajuste

Tomando en cuenta que y_q representa el número de respuestas $y = 1$ (éxitos) en las n_q observaciones correspondientes a la q-ésima combinación de valores de las variables explicativas. Una vez estimados los parámetros, a partir de ellos se estiman las probabilidades $\widehat{\pi_q}$. Por lo tanto, las frecuencias esperadas de respuesta $y = 1$, estimadas bajo el modelo de regresión logística, son en este caso de la forma $\widehat{m_q} = n_q \widehat{\pi_q}$.

Para contrastar la bondad del ajuste global del modelo cuando el número de observaciones n_q en cada combinación de valores de las variables explicativas es grande, se dispone del estadístico chi-cuadrado de Pearson y del estadístico de Wilks de razón de verosimilitudes. Cuando n_q no es suficientemente grande se usará el estadístico de Hosmer y Lemeshow que es una versión modificada del estadístico

chi-cuadrado de Pearson. El test global de bondad de ajuste del modelo de regresión logística múltiple contrasta la hipótesis nula:

$$H_0 : \pi_q = \frac{\exp\left(\sum_{i=1}^{k} \beta_i x_{qi}\right)}{1 + \exp\left(\sum_{i=1}^{k} \beta_i x_{qi}\right)} \quad \forall q = 1, 2, \cdots, Q$$

frente la hipótesis alternativa:

$$H_1 : \pi_q \neq \frac{\exp\left(\sum_{i=1}^{k} \beta_i x_{qi}\right)}{1 + \exp\left(\sum_{i=1}^{k} \beta_i x_{qi}\right)}, \quad \text{para algún} \quad q$$

Test chi-cuadrado de Pearson

El estadístico chi-cuadrado de Pearson de bondad de ajuste a un modelo M de regresión logística es

$$X^2(M) = \sum_{q=1}^{Q} \frac{(y_q - n_q\widehat{\pi_q})^2}{n_q\widehat{\pi_q}(1 - \widehat{\pi_q})} = \sum_{q=1}^{Q} \frac{(y_q - \widehat{m_q})^2}{\widehat{m_q}(n_q - \widehat{m_q})}$$

donde $\widehat{\pi_q}$ es la estimación de máxima verosimilitud de π_q bajo el modelo M y $\widehat{m_q} = n_q\widehat{\pi_q}$ es la estimación de máxima verosimilitud de los valores esperados $m_q = n_q\pi_q$.

Este estadístico tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre el número de parámetros π_q (transformaciones logit muestrales) y el número de parámetros independientes en el modelo. Es decir:

$$X^2(M) \xrightarrow[d]{n_q \to \infty} \chi^2_{Q-(k+1)}$$

Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

$$X^2(M)_{Obs} \geq \chi^2_{Q-(k+1),\alpha}$$

donde $\chi^2_{Q-(k+1),\alpha}$ es el cuantil $1 - \alpha$ de una distribución $\chi^2_{Q-(k+1)}$.

Test chi-cuadrado de razón de verosimilitudes

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste de un modelo de regresión logística múltiple M es de la forma:

$$G^2(M)_{Obs} = 2\left(\sum_{q=1}^{Q}(n_q - y_q)ln(\frac{n_q - y_q}{n_q - \widehat{m_q}}) + \sum_{q=1}^{Q} y_q ln(\frac{y_q}{\widehat{m_q}})\right)$$

El estadístico de Wilks de razón de verosimilitudes tiene distribución asintótica chi cuadrado con grados de libertad obtenidos como la diferencia entre la dimensión

del espacio paramétrico y la dimensión de este espacio bajo la hipótesis nula. Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

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

donde $\chi^2_{Q-(k+1),\alpha}$ es el cuantil $1-\alpha$ de una distribución $\chi^2_{Q-(k+1)}$.

Test de Hosmer y Lemeshow

Cuando no hay un número suficiente de observaciones n_q en cada combinación de valores x_q de las variables explicativas, no se puede asumir la distribución chi-cuadrado de los estadísticos de Pearson y de razón de verosimilitudes como buena. La norma para poder usar estos contrastes es que el 80 % de las frecuencias estimadas bajo el modelo, $\widehat{m_q} = n_q\widehat{\pi_q}$, sean mayores que cinco y todas mayores que uno.

El estadístico de Hosmer y Lemeshow es el estadístico chi-cuadrado de Pearson de bondad de ajuste al modelo después de agrupar adecuadamente los datos en intervalos, de modo que su valor depende fuertemente del número de clases resultantes de la agrupación [21].

Una vez realizado el agrupamiento de las variables explicativas en grupos o clases, si se denota n'_g al número total de observaciones en el g-ésimo grupo, por u_g al número de respuestas $y = 1$ en el g-ésimo grupo, y por $\overline{\pi_g}$ a la probabilidad estimada bajo el modelo de respuesta $y = 1$ para el g-ésimo grupo obtenida como la media de las probabilidades $\widehat{\pi_q}$ de los valores x_q de dicho grupo el estadístico de Hosmer y Lemeshow es de la forma:

$$C = \sum_{g=1}^{G} \frac{(u_g - n'_g\overline{\pi_g})^2}{n'_g\overline{\pi_g}(1-\overline{\pi_g})}$$

y tiene también distribución asintótica chi-cuadrado con $G-2$ grados de libertad.

2.6.2. Medidas globales de bondad de ajuste

Para cuantificar la bondad del ajuste global del modelo se dispone de medidas como la Tasa de clasificaciones correctas y medidas R-cuadrado alternativas al coeficiente R^2 de la regresión lineal.

Tasa de clasificaciones correctas

La tasa de clasificaciones correctas es la proporción de individuos clasificados correctamente por el modelo, obtenida como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N.

Un individuo es clasificado correctamente por el modelo logit cuanto su valor observado de respuesta (1 o 0) coincide con su valor estimado por el modelo. Para asignar respuesta $y = 1$ o $y = 0$ bajo el modelo a los datos se elige un punto de corte (cut-off), $\pi_0 \in (0, 1)$, de modo que a una observación con valor $X = x_q$ se le estima respuesta $Y = 1$ si $\hat{\pi}_q \geq \pi_0$ y se le estima respuesta $y = 0$ cuando $\hat{\pi}_q < \pi_0$.

Como punto de corte para clasificar en «unos» y «ceros» a las observaciones se suele elegir 0.5 aunque es más apropiado elegir la proporción de «unos» en la muestra.

Medidas Tipo R^2

El R^2 de Cox y Snell está dado por:

$$R^2_{CN} = 1 - \left(\frac{V_0}{V_M}\right)^{\frac{2}{N}}$$

donde V_0 es máximo de la verosimilitud bajo el modelo nulo dado sólo por un término constante y V_M es el máximo de la verosimilitud bajo el modelo ajustado con todos los parámetros. Al aumentar el número de parámetros de un modelo, aumenta el máximo de la verosimilitud, y por otro lado, como las probabilidades están entre cero y uno, la medida R^2 toma valores entre 0 y 1.

El valor máximo de esta medida es dado por

$$\text{máx}\, R^2_{CN} = 1 - (V_0)^2$$

que puede estar cerca de 0 cuando se tiene pocos datos. Por esta razón se propone el siguiente coeficiente determinación ajustado denominado de R^2 de Nagelkerke.

$$R^2_N = \frac{R^2_{CN}}{\text{máx}\, R^2_{CN}}$$

Sin embargo, cabe recalcar que las medidas tipo R^2 suelen no ser adecuadas para los modelos logit[22].

2.6.3. Contrastes sobre los parámetros del modelo

Suponiendo que se desea contrastar que un parámetro β_i del modelo de regresión es nulo, esto es la variable X_i no es significativa para el modelo. Por lo tanto, la hipótesis nula del contraste es:

$$H_0 : \beta_i = 0$$

Para lo cual existen los siguientes contrastes:

Contraste de Wald

El estadístico de Wald de este contraste está dado por:

$$W = \frac{\widehat{\beta^2}}{\hat{\sigma^2}(\hat{\beta}_i)}$$

donde $\hat{\sigma^2}(\widehat{\beta})$ es la varianza de $\widehat{\beta}_i$ que se encuentra como el i-ésimo elemento de la diagonal de $\widehat{V(\hat{\beta})}$ que se definió anteriormente. Luego, se rechazará la hipótesis nula al nivel de significación α cuando el valor observado de este estadístico W_{Obs} sea mayor que el cuantil de orden $(1-\alpha)$ de la distribución χ^2_1 denotado por $\chi^2_{1,\alpha}$.

Contrastes condicionales de razón de verosimilitudes

Suponiendo que un modelo de regresión logística M_G se ajusta bien y se desea contrastar si un parámetro β_i es nulo. Denotando por M_P al modelo más simple que resulta al hacer cero este parámetro en M_G, de modo que el modelo particular M_P está anidado en el modelo general M_G. Las hipótesis de este contraste se pueden expresar como:

$$\begin{aligned} &H_0 : \beta_i = 0 \quad (M_P \text{ se verifica}) \\ &H_1 : \beta_i \neq 0 \quad (\text{asumiendo cierto } M_G) \end{aligned} \tag{2.35}$$

Asumiendo que M_G se verifica, el estadístico del test de razón de verosimilitudes para contrastar si M_P se verifica es de la forma:

$$\begin{aligned} G^2(M_P|M_G) &= -2(L_P - L_G) \\ &= -2(L_P - L_S) - 2(-2(L_G - L_S)) \\ &= G^2(M_P) - G^2(M_G) \end{aligned}$$

donde y L_S, L_P y L_G son los máximos de la función de log-verosimilitud bajo la suposición de que se verifican los modelos saturados, M_P y M_G, respectivamente. En este caso se denomina modelo saturado a aquel modelo que tiene tantos parámetros

como observaciones y por lo tanto, como señala Agresti [13], da un ajuste perfecto; obviamente este no es un modelo útil pero sirve para la comparación de modelos. Además, a la cantidad $G^2(MP) = -2(L_P - L_S)$ se le conoce como devianza del modelo M_P y se puede ver a la estadística $G^2(M_P|M_G)$ como la diferencia entre las devianzas de los modelos particular y general.

Se rechazará la hipótesis nula a nivel de significación α cuando el valor observado $G^2_{Obs}(M_P|M_G)$ sea mayor o igual que el cuantil de orden $(1-\alpha)$ de la distribución $\chi^2_{1,\alpha}$.

Existen casos en los que el test de Wald no es tan potente como el test de razón de verosimilitudes, proporcionando a veces resultados no deseables. Por esta razón, es aconsejable usar el test de razón de verosimilitudes en los procedimientos de selección de variables.[23]

2.6.4. Validación y diagnóstico de modelos logit

Una vez contrastado que un modelo se ajusta globalmente bien, se procede a estudiar mediante medidas alternativas la bondad del ajuste observación a observación, así como la naturaleza de la falta de ajuste. Los métodos gráficos suelen ser de mucha ayuda. Otra forma habitual de validar un modelo es el estudio de los residuos que comparan el número observado de éxitos, en cada combinación de valores de las variables explicativas, con su valor ajustado por el modelo.

En el caso de modelos lineales generalizados el estudio de los residuos puede poner de manifiesto si la falta de ajuste se debe a una elección inapropiada de la ligadura o a la falta de linealidad en los efectos de las variables explicativas. También se pueden detectar observaciones influyentes calculando residuos y aproximando el efecto que produce sobre los parámetros al eliminar observaciones simples.

2.6.5. Residuos

En base a los estadísticos X^2 y G^2 se definen dos tipos de residuos en cada combinación de valores x_q de las variables explicativas.

- Residuos de Pearson o residuos estandarizados

$$r_q = \frac{y_q - n_q\widehat{\pi_q}}{(n_q\widehat{\pi_q})(1-\widehat{\pi_q})^{\frac{1}{2}}} \tag{2.36}$$

Observando que el estadístico chi-cuadrado de Pearson se descompone como

$$X^2 = \sum_{q=1}^{Q} r_q^2$$

Una vez estimados los residuos se contrasta su significación estadística mediante el test:

$$H_0 : r_q = 0$$
$$H_1 : r_q \neq 0$$

Bajo la hipótesis nula r_q tiene distribución asintótica normal con media cero y varianza estimada $\widehat{\sigma^2}(r_q) < 1$. Esto significa que los residuos r_q tienen menor variabilidad que una v.a. normal estándar. A pesar de esto, los residuos de Pearson r_q suelen ser tratados como normales estándar, considerándose significativos cuando sus valores absolutos son mayores que dos (falta de ajuste).

Para evitar este problema se definen los residuos de Pearson ajustados que tienen distribuciones asintóticas normales estándar.

$$r_q^s = \frac{\sum_{q=1}^{Q} r_q}{(1 - h_{qq})^{\frac{1}{2}}}$$

donde h_{qq} es el elemento diagonal de la matriz

$$H = \widehat{W^{1/2}} X (X^T \widehat{W} X)^{-1} X^T \widehat{W}^{1/2}$$

con $\widehat{W} = Diag(n_q \widehat{\pi_q}(1 - \hat{\pi}_q))$. Nótese que la matriz H es la equivalente a la matriz hat del modelo de regresión lineal múltiple siendo h_{qq} la influencia (leverage) de la observación q.

Entonces se toma como estadístico del contraste:

$$H_0 : r_q^s = 0$$
$$H_1 : r_q^s \neq 0$$

a r_q^s que bajo la hipótesis nula tiene distribución asintótica normal estándar, o bien su cuadrado que tiene distribución chi-cuadrado con un grado de libertad. Finalmente se rechazará la hipótesis nula (residuo significativamente distinto de cero) al nivel de significación α cuando se verifique:

$$|r_q^s| \geq z_\alpha$$

- Residuos de la devianza o residuos estudentizados

$$d_q = signo(y_q - \hat{m}_q)\left(2\left[y_q ln\left(\frac{y_q}{\hat{m}_q}\right) + (n_q - y_q) ln\left(\frac{n_q - y_q}{n_q - \hat{m}_q}\right)\right]\right)^{\frac{1}{2}}$$

Nótese que el estadístico chi-cuadrado de razón de verosimilitudes se descompone como:

$$G^2 = \sum_{q=1}^{Q} d_q^2$$

De nuevo, bajo la hipótesis nula $H_0 : d_q = 0$, el residuo dq tiene distribución asintótica normal con media cero y varianza estimada $\hat{\sigma}^2(\hat{d}_q) < 1$. Para evitar este problema se definen los residuos de la devianza ajustados o estandarizados.

$$d_q^s = \frac{d_q}{(1 - h_{qq})^{1/2}}$$

que bajo la hipótesis nula del contraste

$$H_0 : d_q = 0$$
$$H_1 : d_q \neq 0$$

tiene distribución asintótica normal estándar. Por lo tanto, se rechazará la hipótesis nula (residuo significativamente distinto de cero) al nivel de significación $1 - \alpha$ cuando se verifique que:

$$|d_q^s| \geq z_\alpha$$

La diferencia entre ambos tipos de residuos es que los de la devianza convergen más rápidamente a la distribución normal que los de Pearson. Como alternativa se pueden usar también los residuos de la devianza modificados:

$$d_q^* = d_q + \frac{1 - 2\hat{\pi}_q}{(n_q \hat{\pi}_q (1 - \hat{\pi}_q))^{1/2}}$$

Que bajo la hipótesis nula $H_0 : d_q = 0$ tienen distribución asintótica normal incluso cuando los tamaños muestrales n_q son pequeños.

2.6.6. Curva ROC

Como una medida de bondad de ajuste se puede calcular la tasa de clasificaciones correctas bajo un punto de corte π_0 (recordando que se predice que el evento ocurre, esto es $y = 1$, cuando $\hat{\pi} > \pi_0$). Sin embargo como señalan Hosmer[21] una

mejor y más completa descripción de la capacidad predictiva del modelo pude ser hecha utilizando la curva característica de operación, usualmente denominada curva ROC (por sus siglas en inglés, Receiver Operating Characteristic).

Esta curva muestra la "sensibilidad" como una función de «1-especificidad» para cada posible punto de corte π_0, donde la sensibilidad se define como la probabilidad de predecir la ocurrencia de un evento cuando este en realidad ocurre esto es $P(\hat{\pi} > \pi_0 | y = 1)$ y la especificidad es la probabilidad de predecir que el evento no va ocurrir cuando este en realidad este no ocurre$P(\hat{\pi} \leq \pi_0 | y = 0)$.

De acuerdo con Agresti [13] la curva ROC es más informativa que una única tasa de clasificaciones correctas porque resume el poder predictivo del modelo para todos los posibles puntos de corte π_0. Además, para una dada especificidad, el mejor poder predictivo corresponde a una sensibilidad alta. Así, entre mejor sea el poder predictivo, mayor será el área bajo la curva ROC. De esta manera, el área bajo la curva ROC es utilizada como una medida del poder predictivo de un modelo. De acuerdo con Hosmer[21] cuando el área bajo la curva ROC es al menos 0,7 el modelo se considera que tiene capacidad de discriminación aceptable.

2.6.7. Coeficiente de Gini

El miembro más tradicional de la familia de la desigualdad de ingresos es el coeficiente de Gini. Esta medida puede definirse de varias formas [24]. En general, el índice de Gini es una función $G : \mathbb{R}_n^+ \to [0, 1]$ que asigna a cada vector de ingreso no negativo un número real entre 0 y 1, que representa el nivel de desigualdad de la muestra. Esta medida es 0 en máxima igualdad y 1 en perfecta desigualdad. La definición formal del índice de Gini es el doble del área entre la línea de igualdad y la curva de Lorenz en el cuadro de la unidad (como se muestra en la figura 2.1).

La línea en 45° representa la perfecta igualdad de ingresos y el área entre esta línea y la curva de Lorenz se llama área de concentración. Puede notarse que a mayor índice de Gini se tiene una mayor desigualdad. Si dos curvas de Lorenz se cruzan entre sí, se recomienda no sacar conclusiones de carácter visual, ya que pueden ser engañosas; es mejor comparar la desigualdad que representan, calculando primero los índices de Gini correspondientes a cada curva. El índice de Gini se puede expresar como

$$G = 2 \int_0^1 (p - L(p)) dp \tag{2.37}$$

tal que $p = F(x)$ es una función de distribución acumulativa de variable no negativa

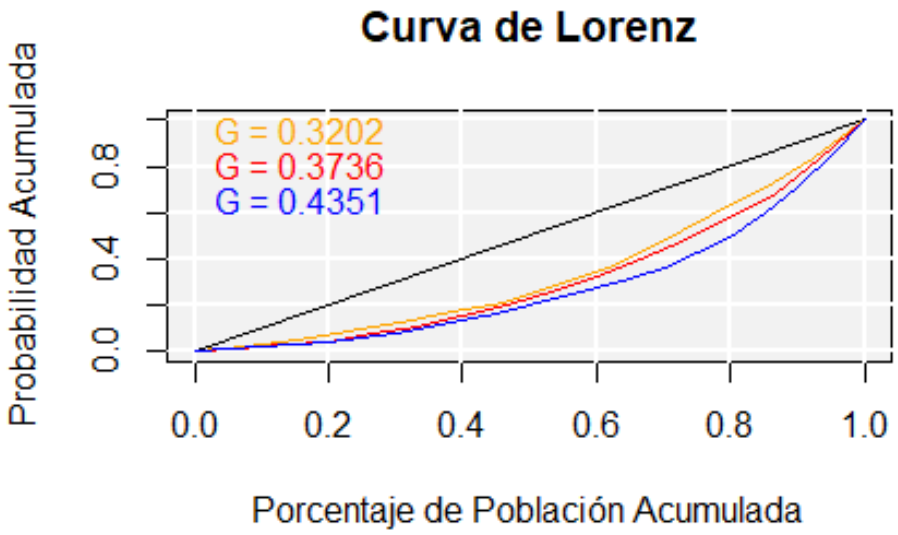

Figura 2.1: Área entre la Curva de Lorenz y la línea de igualdad
Elaborado: autor.

con esperanza positiva y finita μ, $L(p)$ la función de Lorenz dada por:

$$\frac{1}{\mu}\int_0^1 F^{-1}(p) = inf\left\{x|F(x) \geq p : p \in [0,1]\right\} \tag{2.38}$$

El coeficiente de Gini al igual que el indicador ROC, es una medida de qué tan bien el modelo clasifica a individuos correctamente, cuando el punto de corte varía a lo largo del intervalo de la probabilidad pronosticada. Este varía entre 0 y 1, cuanto más cercano a 1 se encuentre, el modelo genera una discriminación mayor. Para calcularlo:

$$G = 1 - \sum_{i=1}^{n}[P_m(s_i) - P_m(s_{i-1})][P_b(s_i) - P_b(s_{i-1})] \tag{2.39}$$

donde:

n: número de intervalos

$P_m(s_i)$: proporción acumulada de negativos para un intervalo i

$P_m(s_{i-1})$: proporción acumulada de negativos para un intervalo anterior al i

$P_b(s_i)$: proporción acumulada de positivos riesgos para un intervalo i

$P_b(s_{i-1})$: proporción acumulada de positivos para un intervalo anterior al i

2.7. Selección de Variables

Dentro de un modelo se sugiere incluir las variables más representativas de la muestra. Para hallarlas existen pruebas que comprueben que existe una relación estadística necesaria para hacer inferencia.

2.7.1. Prueba de Kolmogorov-Smirnov

Esta prueba compara las funciones de distribución empírica de la muestra y la que se desea contrastar. Es aplicable a distribuciones continuas. Para distribuciones continuas, los valores críticos están tabulados para: distribuciones con parámetros especificados, algunas distribuciones con parámetros no especificados (normal, Weibull, gamma, exponencial)[25]. El cálculo es el siguiente:

Observadas $y_1, y_2, \cdots, y_N$, considerar la distribución empírica

$$F_e(x) = \frac{\#(\{i|y_i) \leq x\}}{N}$$

$F_e(x)$ es la proporción de valores observados menores o iguales a x. La hipótesis nula es $H_0 : F_e(x)$ es cercana a $F(x)$. Y el estadístico para esta prueba es:

$$D = max_x|F_e(x) - F(x)|, \quad -\infty < x < \infty$$

Se ordena a $y_1, y_2, \cdots, y_N$, en orden creciente

$$y_{(j)} = j - \text{ésimo valor más pequeño}$$
$$y_{(1)} < y_{(2)} < \cdots < y_{(N)}$$

La distribución empírica de $F_e(x)$ es:

$$F(x) = \begin{cases} 0 & \text{para } x < y_{(1)} \\ \frac{1}{n} & \text{para } y_{(1)} < x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & \text{para } y_{(j)} < x < y_{(j+1)} \\ \vdots & \\ 1 & \text{para } y_{(N)} < x \end{cases}$$

Si definimos lo siguiente:

$$D^+ = \max_{1\leq j\leq N}\left\{F_e(y_{(j)}) - F(y_{(j)})\right\} = \max_{1\leq j\leq N}\left\{\left(\frac{j}{n}\right) - F(y_{(j)})\right\}$$
$$D^- = \max_{1\leq j\leq N}\left\{F(y_{(j)}) - F_e(y_{(j)})\right\} = \max_{1\leq j\leq N}\left\{F(y_{(j)}) - \left(\frac{j-1}{n}\right)\right\}$$

D se puede redefinir como:

$$D = \max_{1\leq j\leq N}\left\{\left(\frac{j}{n}\right) - F(y_{(j)}), F(y_{(j)}) - \left(\frac{j-1}{n}\right)\right\}$$

El procedimiento para contrastar el test K-S será elegir un grado de significación α. Luego p, tal que: $p = P_F(D \geq d)$. La decisión será:

$$p < \alpha : \text{se rechaza } H_0$$
$$p > \alpha : \text{no se rechaza } H_0$$

Los procedimientos para hallar p serán facilitados por el uso del software R, útil para los estudios estadísticos.

2.7.2. Test de Valor de Información:

La prueba de valor de información (VI) es un filtro popular para seleccionar variables predictoras para regresión logística binaria. En muchos libros de texto sobre calificación crediticia se dan pautas para decidir si el IV de un predictor X es lo suficientemente alto para usarlo en el modelado [26]. Por ejemplo, estos textos dicen que $IV > 0{,}3$ muestra que X es un fuerte predictor. Una práctica común al preparar un predictor X es agrupar los niveles de X para eliminar valores atípicos y revelar una tendencia. Pero la IV disminuye a medida que se colapsan los niveles de X.

Este test se utiliza para decidir sobre variables categóricas. Sea X una variable con r categorías diferentes, se define:

- g_i: cantidad de individuos en la i-ésima categoría que en la variable respuesta tienen el valor 1.
- b_i: cantidad de individuos en la i-ésima categoría que en la variable respuesta tienen el valor 0.
- g: cantidad de individuos en la variable respuesta tienen el valor 1.
- b: cantidad de individuos en la variable respuesta tienen el valor 0.

Para calcular el VI, se realiza la siguiente fórmula:

$$IV = \sum_{i=1}^{r} \left(\frac{g_i}{g} - \frac{b_i}{b} \right) ln(\frac{g_i/g}{b_i/b}) \tag{2.40}$$

Para nuestro trabajo, la elección de cuál variable categórica debe ir dentro del modelo se basó en lo siguiente:

- Cuando $IV \in [0,02 - 0,1]$, entonces el predictor solo tiene una relación débil. Por lo cual, será descartada la variable.
- Cuando $IV \in]0,1 - 0,3]$, entonces el predictor tiene una relación de fuerza media. Por lo cual, la variable será considerada dentro del modelo siempre que los resultados sean beneficiarios, es decir, siempre que sea estadísticamente significativa la variable.
- Cuando $IV > 0,3$, entonces el predictor tiene una relación fuerte. Se recomienda incluir esta variable.

Además de las variables mencionadas, algunas de estas fueron modificadas utilizando árboles de decisión.

2.7.3. Árboles de Decisión

Un árbol de decisión es una forma gráfica y analítica de representar todos los eventos (sucesos) que pueden surgir a partir de una decisión asumida en cierto momento. Esto ayuda a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a buscar subgrupos específicos y relaciones que tal vez no se encuentran con estadísticos más tradicionales[23]. Los árboles de decisión son una técnica estadística para la segmentación, la estratificación, la predicción, la reducción de datos y el filtrado de variables, la identificación de interacciones, la fusión de categorías y la discretización de variables continuas.

Existen diferentes tipos de árbol: CHAID, CHAID exhaustivo, CRT y QUEST, según el que mejor se ajuste a los datos. Se ha seleccionado el método CHAID, este consiste en un rápido algoritmo de árbol estadístico y multidireccional que explora datos de forma rápida y eficaz, y crea segmentos y perfiles con respecto al resultado deseado. Permite la detección automática de interacciones mediante Chi-cuadrado.

En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

2.7.4. Método CHAID

El método de detección automática de interacción chi-cuadrado (CHAID) es un algoritmo que realiza la división de múltiples variables mediante el uso de la prueba chi-cuadrado. Además, utiliza el chi-cuadrado de Pearson cuando la variable objetivo es categórica y utiliza el estadístico Chi-cuadrado de razón de verosimilitud como referencia de separación cuando una variable objetivo es continua. El chi-cuadrado se calcula a partir de la tabla de partición rxc compuesta por observaciones n_{ij}. La función del estadístico Chi cuadrado de Pearson se muestra como:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{n_{ij}}$$

La función del estadístico Chi-cuadrado de la razón de verosimilitud se muestra como:

$$\chi^2 = 2 \sum_i \sum_j n_{ij} ln \left(\frac{n_{ij}}{e_{ij}} \right)$$

El estadístico Chi-cuadrado implica que las distribuciones de las variables objetivo para cada categoría de la variable predictora son las mismas. Por tanto, se puede concluir que la variable predictora no afecta la clasificación de las variables objetivo. La magnitud del estadístico de chi-cuadrado para el grado de libertad se puede expresar como un valor p. Cuando el estadístico chi-cuadrado es menor que el grado de libertad, el valor de p aumenta. Como resultado, el uso del estadístico Chi cuadrado como referencia de separación significa que el nodo hijo está formado por la variable predictora con el valor p más pequeño y la separación óptima.

Capítulo 3

Estimación del Modelo

3.1. Descripción de los Datos

El Instituto Nacional de Empleo (INEM), con el propósito de organizar y administrar un sistema permanente de información sobre el comportamiento de la fuerza de trabajo, procede al levantamiento de la Encuesta Permanente de Empleo y Desempleo en el área urbana del Ecuador [3], desde noviembre de 1987 con periodicidad anual. La encuesta contaba con el financiamiento del Banco Central del Ecuador (BCE), el Programa de las Naciones Unidas para el Desarrollo (PNUD) y la asistencia técnica de la Organización Internacional del Trabajo (OIT).

La ENEMDU está diseñada para proporcionar estadísticas sobre los niveles, tendencias y cambios en el tiempo de la población económicamente activa, población económicamente inactiva, el empleo, subempleo y desempleo en Ecuador con representatividad nacional, urbana, rural y cinco ciudades principales para la población de 15 años y más. Estas características ayudan a describir la situación del país, pues nuestro propósito es describir a la población que necesita del crédito.

La base de datos de la encuesta ENEMDU es de libre acceso y se la puede encontrar en la página oficial del INEC [27]. Cuenta con archivos en dos formatos: Excel y CSS. Además, existe documentación como guía para los usuarios que necesiten comprender la recolección de los datos. Es importante recalcar que el país y todo el mundo atravesó por una crisis sanitaria conocida como COVID-19 [28]. Es por esta razón que la información utilizada pertenece a un periodo previo a este acontecimiento. El periodo elegido es Septiembre-2019.

La base de datos se obtuvo a partir de una encuesta realizada a las familias ecuatorianas que han estado colaborado con equipo de encuestadores del INEC. Se cuen-

ta con información de 59,208 familias, de las cuales se recolecto 158 variables. La variable respuesta está representada a través de la pregunta: «¿Usted o algún miembro del hogar tiene planes de endeudamiento en los próximos 3 meses (bancos, financieras, etc.)?» Para esta pregunta existen las únicas dos posibles respuestas: $y = 1$, cuando la respuesta es sí y $y = 0$, cuando la respuesta es no. Las demás preguntas representan las variables descriptoras en el modelo y, para un mejor entendimiento del lector, estas se representan en los anexos al final del trabajo.

Para evitar que existan registros repetidos se tomó la decisión de trabajar únicamente con la información del jefe del hogar, puesto que este representa el estilo de vida de su familia. Es así como los registros se reducen a 17,001.

3.2. Estadística Descriptiva

Algunas de las variables más relevantes de esta muestra se observan, previo a la realización del modelo, mediante un análisis de frecuencias y estadísticos generales.

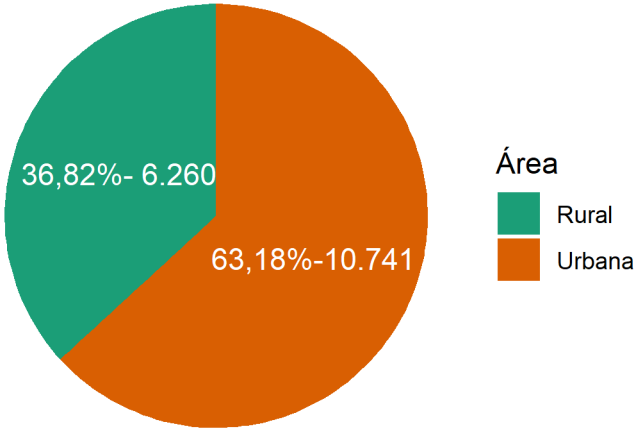

Figura 3.1: Descripción de la población según el área en que viven
Elaborado: autor

- ***area***: se observa en la figura 3.1 que el 36, 82 % de la población pertenece a un área rural mientras que el 63, 18 % pertenece a un área urbana.

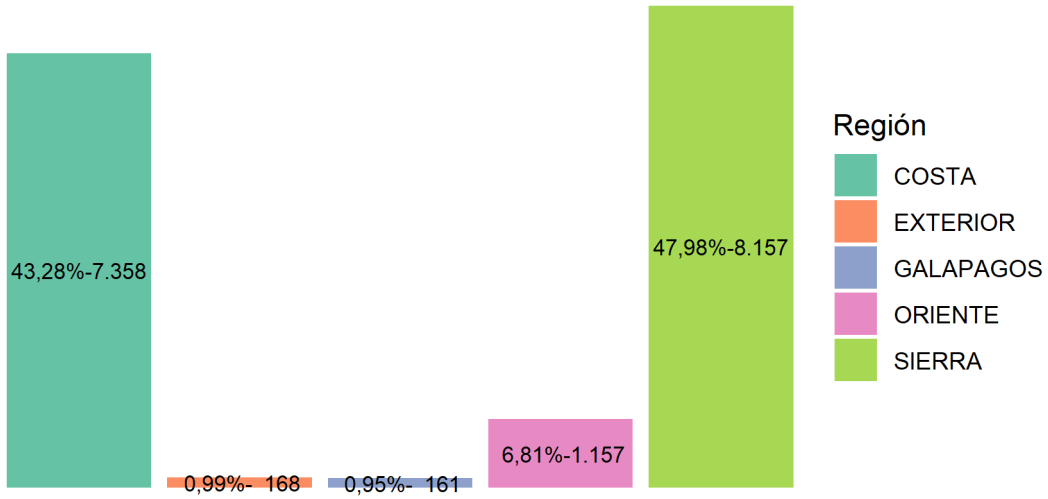

Figura 3.2: Descripción de la población según la provincia en que viven
Elaborado: autor

- ***ciudad***: se observa en la figura 3.2 que en su mayoría provienen de la Sierra, 43, 28 %, luego está de la Costa 47, 96 % y del Oriente con 6,81 %. En cuanto al Exterior y Galápagos con mucho menor representación, menos de 1 %.

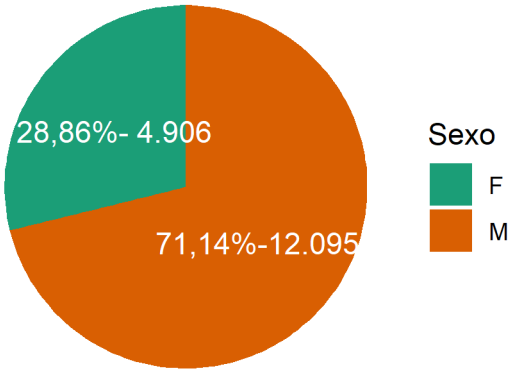

Figura 3.3: Descripción de la población según su sexo
Elaborado: autor

- ***sexo***: en la figura 3.3 existe una mayor proporción de hombres sobre mujeres, con 71,14 % para hombres y 28,86 % para mujeres.

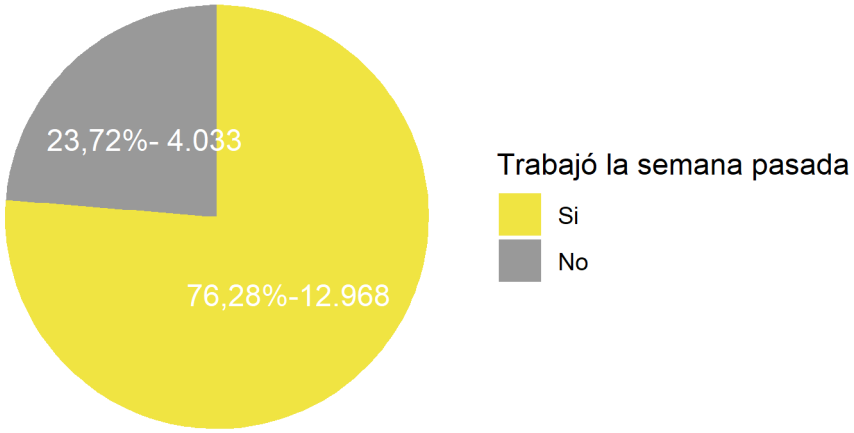

Figura 3.4: Descripción de la población según si trabajo o no la semana pasada
Elaborado: autor

- ***trabajo***: se observa en la figura 3.4 la variable que representa si el individuo tiene un trabajo o, más específicamente, si trabajó la semana pasada. En su mayoría las personas sí trabajaron, un 76,28 % y el resto no, 23,72 %.

- Para el caso de la variable ***categoria***, esta representa la categoría de ocupación de la población trabajadora, es decir, sobre aquellas personas que respondieron que «sí» a la variable ***trabajo***. Se presentan los resultados en la tabla 3.1

Categoría de Ocupación	
Empleado de gobierno	8,18 %
Empleado privado	24,17 %
Empleado terciarizado	11,99 %
Jornalero o peón	5,40 %
Patrono	46,64 %
Cuenta Propia	1,08 %
Trabajador del hogar no remunerado	0,18 %
Trabajador no del hogar no remunerado	0,01 %
Ayudante no remunerado de asalariado/jornalero	2,34 %

Cuadro 3.1: Frecuencia sobre la Categoría de Ocupación

Elaborado: autor

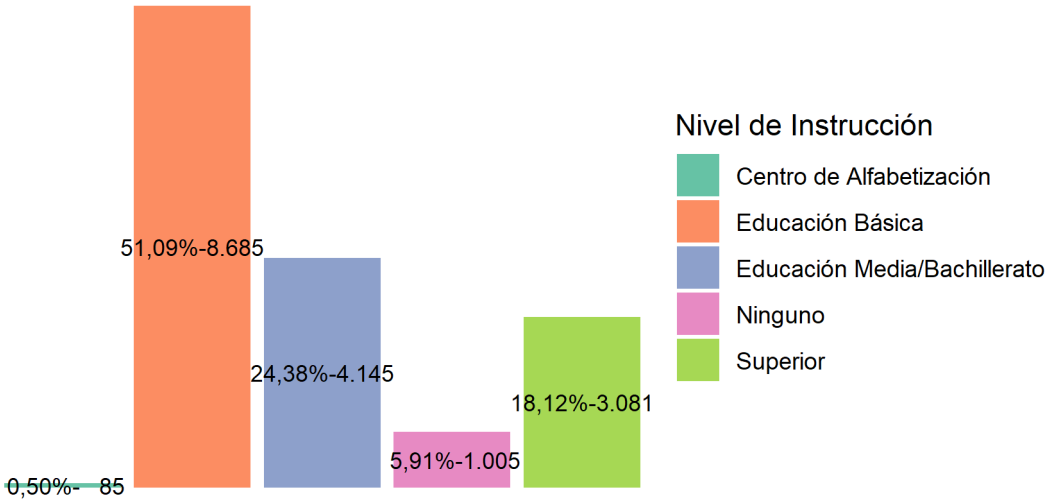

Figura 3.5: Descripción de la población según su Nivel de Instrucción
Elaborado: autor

- En la figura 3.5 se encuentra la variable ***nivel_instruccion*** para designar al Nivel de Instrucción de los individuos. En su mayoría las personas han tenido una instrucción hasta Educación Básica, 51,09 %, seguido por Educación Media, Superior, Ninguna y Centro de Alfabetización.

- La variable de interés es ***variable_respuesta*** que representa si el individuo necesita o no de un crédito. Como se puede observar en la figura 3.6 existe una notable diferencia entre la proporción que dice necesitar con respecto a la que no. Esto ha representado especial interés sobre la elección del modelo a trabajar. Debido a esta situación, y como se explica en el capítulo 2, los modelos de corrección de sesgo trabajan sobre muestras como esta; puesto que se considera que la información de mayor interés se encuentra sobre esta pequeña proporción de individuos.

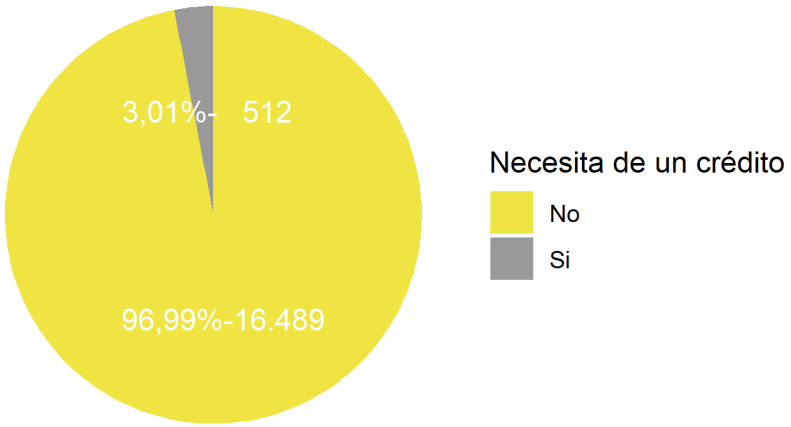

Figura 3.6: Descripción de la población según su **Necesidad de Crédito**
Elaborado: autor

- Además de presentar variables de tipo categórico, a continuación, se observan algunas de las variables continuas más relevantes que se poseen. Estas son: ***edad***, la Edad, ***monto_cuenta_ahorros***, el Monto que posee en su cuenta de ahorros, y ***ingpc***[1], el ingreso per cápita. Se han calculado las estadísticas básicas que se presentan en la tabla:

Estadística	Mínimo	Primer Quintil	Mediana	Media	Segundo Quintil	Máximo	Varianza
Edad	15.00	41.00	53.00	53.32	65.00	99.00	259.87
Monto de Ahorros	2	150	275	6,832	500	999,999	63,396
Ingreso per cápita	0.6	110.0	193.3	293.1	345.8	18,333.3	153,705.2

Cuadro 3.2: Estadísticas básicas de las variables continuas: *edad*, *monto_cuenta_ahorros* y *ingpc*

Elaborado: autor.

- Según estas estadísticas la ***edad*** mínima es de 15 años, lo cual es una condición necesaria para llenar la encuesta ENEMDU, y la edad máxima es 99 años.

[1]El ingreso per cápita tiene una estrecha relación con el ingreso nacional. El ingreso hace referencia a todas las entradas económicas que recibe una persona, una familia, una empresa, una organización, etc. Este se calcula dividiendo el ingreso total del hogar por el número de personas que lo componen.

Es importante recalcar que para la variable ***monto_cuenta_ahorros***, se consideraron únicamente a los individuos que poseen una cuenta de ahorros, estos fueron apenas 784 personas. Más adelante se explicará cómo el poseer o no cuenta de ahorros interviene al momento de demandar/necesitar de crédito. Finalmente, la variable ***ingpc*** tiene un rango de $(0{,}6 - 18,3000)$, con una media de 293,1 que es bastante bajo del salario básico en el Ecuador $400,00.

3.3. Limpieza de Datos

Los datos, como se menciona anteriormente, provienen de la red libre por parte del INEC. Esta información se encuentra separada en 3 partes: variables de consumo, variables de vivienda y variables de persona. Al ser unificadas tales partes, se fueron reduciendo las variables que se repetían. Además, se consideraron ciertos cambios a las variables siguientes:

- ***ciudad:*** Inicialmente esta variable es un código que representa el cantón de donde proviene el individuo. Debido a que esta representación es muy segmentada, o incluso innecesaria, se decidió hacer uso del diccionario de tales códigos y transformar esta variable en la región a la cual pertenece el individuo. Para más tarde analizar si existe una relación en la nueva variable creada con respecto a la necesidad de crédito [29].

- ***anio_aprobado:*** Esta variable es el número de años que ha aprobado el individuo en cuanto a su educación. Se presenta 5,9 % de datos faltantes, y para que el modelo sea lo mayor conservador posible tales datos se reemplazan con el valor 0, es decir, que el individuo no ha estudiado.

- ***lee_escribe:*** Esta variable presenta 5 % de datos faltantes. Considerando que es una cantidad alta de datos faltantes, algunos expertos prefieren no trabajar con la misma, sin embargo, su importancia prevalece debido a que la capacidad de lectura y escritura aún es un problema social en Ecuador. Para que el modelo sea lo más conservador posible, se llenaron los datos con 0 que representan que el individuo no sabe leer ni escribir. Este criterio considera los casos de la manera más realista posible.

- ***ingresos:*** Existen 5,8 % de datos faltantes y, al ser importante, la variable se mantiene dentro del modelo. Se la complementa con el valor 0. Bajo el mismo criterio explicado anteriormente.

- ***retiro_negocio***, ***recibio_especias_alim*** y ***recibio_bienes*** presentan un 0,5 % de datos faltantes, mismos que fueron completados con 0, es decir, que el individuo no realizó tal acción por dicha variable respectivamente. Cada variable tiene su significado. ***retiro_negocio*** representa que el individuo retiro en los último mes un monto económico por el negocio que posee, ***recibio_especias_alim*** que recibió por su trabajo especies o servicios tales como: alimentos, vivienda, vestido, etc. y ***recibio_bienes*** que recibió por su trabajo pago en especies o retiró del negocio o producción bienes o productos para el consumo del hogar.

- ***ingpc:*** Con una cantidad pequeña de datos faltantes, 0,008 %, se completa con el valor de 0.

3.4. Selección de Variables

Inicialmente la encuesta recolecta 158 variables por individuo, de las cuales se han seleccionado las más relevantes de acuerdo con un criterio experto en el tema. Como resultado, se ha trabajado con 59 variables elegidas, y a partir de las cuales se ha decidido utilizar las pruebas mencionadas en el capítulo 2.7.1. Para la selección de las variables continuas, se hace uso de la Prueba de Kolmogorov-Smirnov, cuyos resultados son:

No.	Variable	KS
1	edad	0.1881
2	ingpc	0.1316
3	ingresos	0.0694
4	monto_donacion	0.0653
6	salario_indep	0.0288
7	monto_cuenta_ahorros	0.0143
8	monto_familiares	0.0113
9	sem_sin_trabajo	0.0062

Cuadro 3.3: KS para las variables continuas

Elaborado: autor

Se han seleccionado tales variables dentro del modelo, sin embargo, no todas han sido representativas. Se explicará con más detalle después.

Para la selección de las variables categóricas, se hace uso de la Prueba Valor de Información. Los resultados obtenidos para las 50 variables se muestran, sin embar-

go, solo se observarán los más altos debido a que el resto no serán de uso dentro del modelo estimado. A continuación, tales valores:

No.	Variable	VI
1	dejo_trabajar	1.889
2	motivo_desempleo	1.735
3	situacion_futura	0.325
4	ciudad	0.193
5	condicion_actividad	0.173
6	condic_inact	0.151
7	nivel_instruccion	0.116
8	material_techo	0.099
9	ruc_establ	0.091
10	sect_empleados	0.091
11	tipo_vivienda	0.083
12	estado_vivienda	0.082
13	poblacion_rama	0.081
14	estabilidad	0.080
15	aporte_seguros	0.076

Cuadro 3.4: VI para las variables categóricas

Elaborado: autor

Considerando que el VI para el resto de variables ha sido menor que el valor propuesto para incluir en el modelo, se han incluido las variables anteriores y el resto se decide no incluirlas.

3.4.1. Muestra de Modelamiento y Validación

La validación del modelo es necesaria para evitar el problema de sobreajuste. Esto significa, que el modelo obtenido debe funcionar para muestras o datos distintos de los que se han utilizado. Es decir, que el poder de predicción del modelo estimado no será perdido cuando este sea utilizado con una base de datos diferente.

Muchos estadísticos sugieren trabajar con la regla de Pareto[2]. Entre las variaciones que existen en esta ley, se ha decidido trabajar con una muestra del 70 % de información, para realizar la validación con el resto (30 %), esto debido a que se requiere trabajar con la mayor cantidad de datos posibles.

[2]El principio de Pareto, también conocido como la regla del 80-20 y ley de los pocos vitales, describe el fenómeno estadístico por el que en cualquier población que contribuye a un efecto común, es una proporción pequeña la que contribuye a la mayor parte del efecto.[30]

Por último, se ha elegido un nivel de significancia $\alpha = 0{,}05$ para comprobar la hipótesis nula en la prueba de significancia de los parámetros.

3.4.2. Creación de Variables

Para las variables cuyas opciones eran muchas categorías, se consideró propicio disminuir tales categorías utilizando el Software AnswerTree[3], los resultados se presentan a continuación:

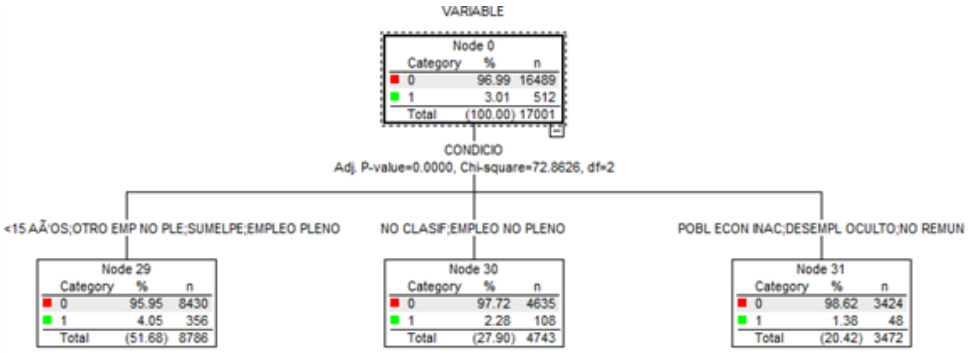

Figura 3.7: Árbol de Decisión para la variable ***condicion_actividad***
Elaborado: autor.

En la figura 3.7 se observa que para la variable ***condicion_actividad*** fue necesario agrupar sus categorías en 3 principales. Es sí que se reducen: *menores de 15 años*, *otro empleo no pleno*, *subempleo por insuficiencia de ingresos* y *empleo adecuado pleno* se vuelve una categoría. *Empleo no clasificado* y *Otro empleo no pleno* se vuelven otra categoría y, finalmente, *Población Económicamente Inactiva*, *Desempleo oculto* y *Empleo no remunerado* se vuelven una tercera categoría.

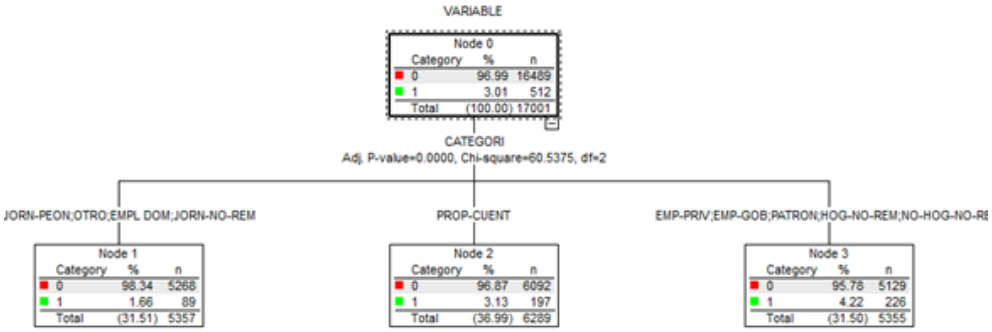

Figura 3.8: Árbol de Decisión para la variable ***categoria_ocupacion***
Elaborado: autor.

[3]Este programa proporciona herramientas gráficas para la preparación de datos y la construcción de árboles de decisión.

En la figura 3.8 se observa que para la variable ***categoria_ocupacion*** fue necesario agrupar sus categorías en 3 principales. Es sí que se reducen: *Jornalero o peón*, *otro* , *Empleado(a) Doméstico(a)* y *Ayudante no remunerado de asalariado/jornalero* se vuelve una categoría. *Cuenta Propia* se mantiene como una categoría. Una tercera categoría será: *Empleado de gobierno*, *Empleado privado*, *Patrono*, *Trabajador del hogar no remunerado* y *Trabajador no del hogar no remunerado*.

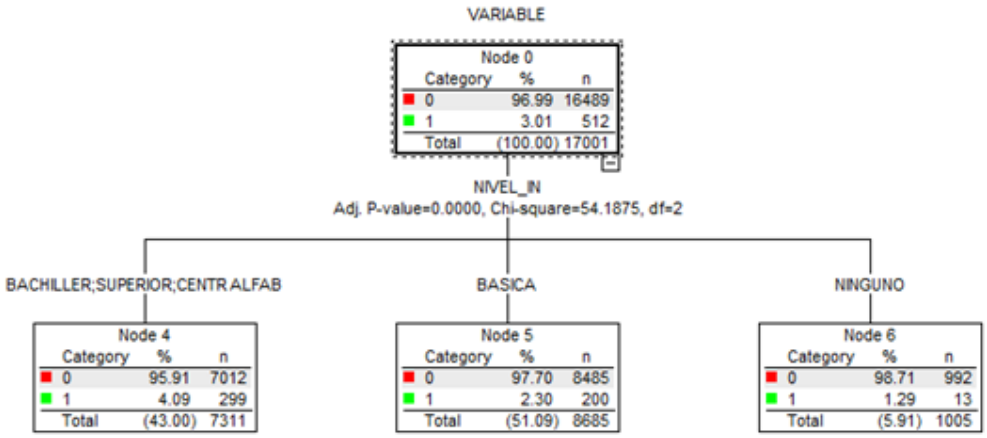

Figura 3.9: Árbol de Decisión para la variable ***nivel_instruccion***
Elaborado: autor.

Para la figura 3.9 fue necesario crear una categoría con: *Bachiller*, *Superior* y *Centro de Alfabetización*; mientras que las otras dos categorías se mantuvieron: *Básica* y *Ninguno*.

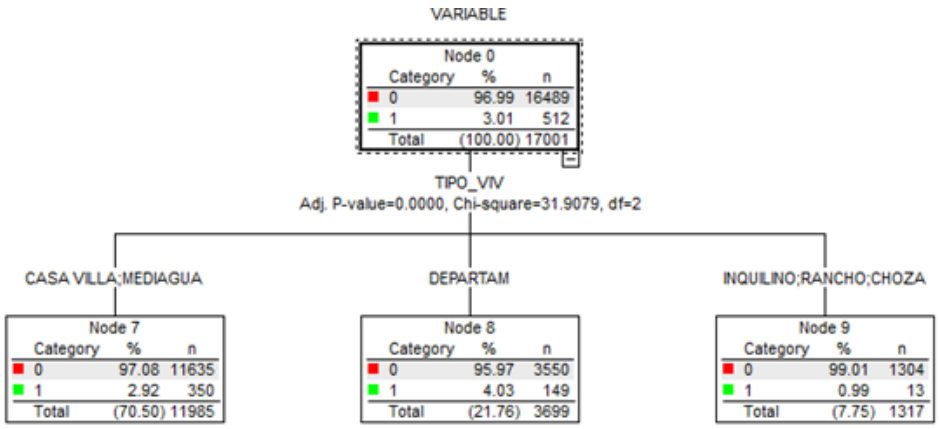

Figura 3.10: Árbol de Decisión para la variable ***tipo_vivienda***
Elaborado: autor.

Para la figura 3.10 fue necesario crear una categoría con: *Casa Villa*, *Media Agua*. La categoría *Departamento* se mantuvo y se creó la última categoría con *Inquilino*, *Rancho* y *Choza*

Capítulo 4

Modelo Probit

Considerando el método de selección de variables, el Modelo Probit estimado es el siguiente:

Coefficients	Estimación	Sd.standar	Valor z	P-valor	
Intercepto	-1.350e+00	3.601e-01	-3.750	0.000177	***
ciudadORIENTE	4.619e-01	1.685e-01	2.741	0.006120	**
ciudadSIERRA	3.156e-01	1.122e-01	2.812	0.004922	**
sexoM	1.681e-01	1.273e-01	1.320	0.186909	**
edad	-1.504e-02	4.107e-03	-3.663	0.000250	***
lee_escribe1	7.002e-02	1.121e-01	0.625	0.532247	***
trabajo_1sem1	-4.568e-01	1.988e-01	-2.298	0.021546	***
monto_cuenta_ahorros	9.353e-04	3.285e-04	2.847	0.004410	**
condicion_actividadINAC NO REM	-8.368e+00	1.455e+02	-0.058	0.954134	***
ingpc	3.125e-04	1.337e-04	2.337	0.019447	**
situacion_futuraMEJOR	6.968e-01	1.264e-01	5.511	3.56e-08	***

Cuadro 4.1: Estimación Probit para la demanda de Crédito

Elaborado: autor

Se observan los resultados en la tabla 4.1, con sus respectivos valores y significancia[1]. En contraste con el modelo lineal, en los modelos probit y logit los parámetros no corresponden al efecto marginal sobre la variable dependiente de un cambio en una de las variables de control.

La interpretación de los valores estimados es particularmente distinta a la usual en los modelos lineales. En general un valor β_i es la pendiente y mide el cambio en la variable respuesta ocasionado por un cambio unitario en x_i, es decir, dice cómo

[1]Significancia de los códigos: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1

el logaritmo de las probabilidades a favor de necesitar un crédito cambia a medida que la variable x_i cambia en una unidad. De esto, se puede concluir que:

- Si la edad aumenta en un punto porcentual, el individuo tiene un 1.5 % menos de probabilidad de necesitar un crédito.
- Para el caso de la región, se puede observar que, en la Sierra y Galápagos, los coeficientes son positivos, lo que significa que si el individuo pertenece a estas regiones, tiene más probabilidad de necesitar un crédito (46 % para el caso del Oriente y 31 % para el caso de la Sierra).
- Aquel individuo cuya expectativa de la situación a futuro sea mejor que la actual, (considerando que tuvo tal expectativa al momento de realizarse la encuesta) este tiene 69 % más de probabilidad de necesitar un crédito.
- Cuando el individuo aumenta su ingreso per cápita en un 1 %, su probabilidad de necesitar un crédito aumenta 0.3 %.

4.1. Bondad de Ajuste

Los estadísticos de bondad de ajuste y sus respectivos valores se presentan a continuación:

Test	**Estadístico**	**P-valor**
Devianza	813.6174	0.000
Pearson	3466.15	0.000
Hosmer y Lemeshow	3400	0.000

Cuadro 4.2: Pruebas de Bondad de Ajuste Modelo Probit

Elaborado: autor

Entonces para un nivel de significación del 5 % se acepta la hipótesis nula de que el modelo se ajusta bien a los datos observados. Esto permite concluir que la hipótesis de la normalidad de los residuos no se rechaza y, de este modo, las conclusiones en el modelo son válidas, así como las debidas interpretaciones que esto implica.

A continuación, se observa la Matriz de Confusión, o Tabla de Clasificación. Esta permite ver las predicciones aciertas con respecto a las estimaciones del modelo.

El valor de R-cuadrado será:

	Observación	
Estimación	**0**	**1**
0	13182	409
1	9	1

Cuadro 4.3: Tabla de Clasificación del Modelo Probit

Elaborado: autor

Los valores del cuadro 4.3 se obtuvieron con los datos de la muestra de validación. Se observa que, en la diagonal de la matriz, las predicciones que coinciden con los valores observados de la variable respuesta. Estos valores son conocidos como «verdaderos positivos» y «verdaderos negativos».

El coeficiente de *GINI* es: 0.4445753. El área bajo la curva *AUC* es: 0.6996184. Este último representa el poder de predicción del modelo, el cual es estadísticamente bueno.

4.1.1. Normalidad en los residuos

Las pruebas sobre los residuos han dado como resultado:

Prueba	Estadístico	p_valor
Prueba de Jarque Bera	616769327	0.000
Prueba de Pearson chi-square	1221751	0.000

Cuadro 4.4: Tabla de Clasificación del Modelo Probit

Elaborado: autor

Se consideran también medidas de diagnóstico, como los residuales de Pearson y de la devianza presentados en la Figura 4.1.

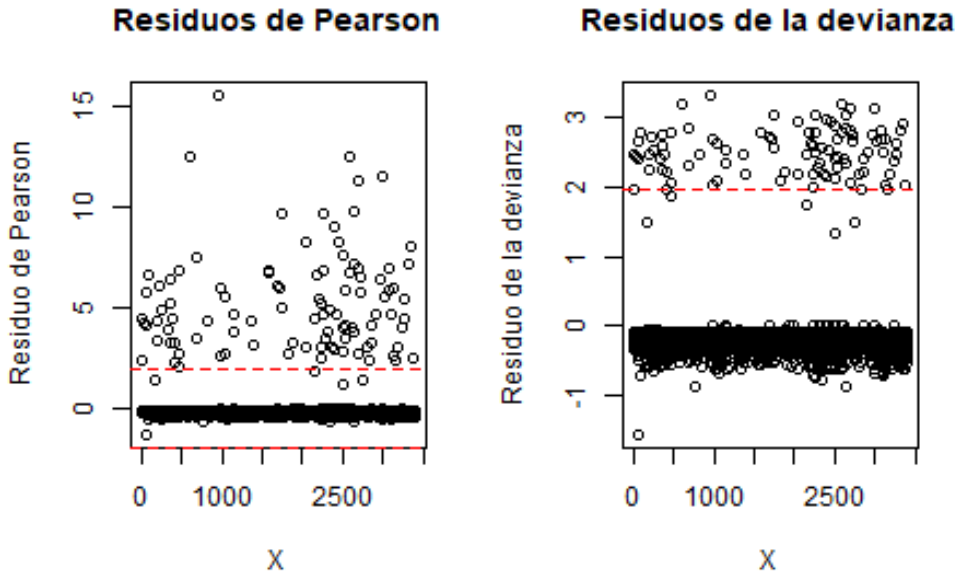

Figura 4.1: Residuos de Pearson y de la devianza para el Modelo Probit
Elaborado: autor.

Se pueden notar que algunas observaciones pueden ser consideradas valores outlier. Se procede a observar la gráfica de la curva ROC y gráfica de las distribuciones de acumulación empíricas junto con el estadístico $K - S$. El estadístico K_S muestra la distancia máxima entre las dos curvas:

La curva ROC ayuda a reconocer el poder de predicción del modelo. Además, está calculada el estadístico J de Youden, que es un estadístico que captura el rendimiento de una prueba de diagnóstico dicotómica. En este caso, el modelo es aceptable y se procede a las pruebas de multicolinealidad.

4.1.2. Multicolinealidad

Se presenta a continuación la prueba de multicolinealidad. Esta representa la correlación entre las variables predictoras y la variable respuesta. Como supuesto, no debe existir correlación.

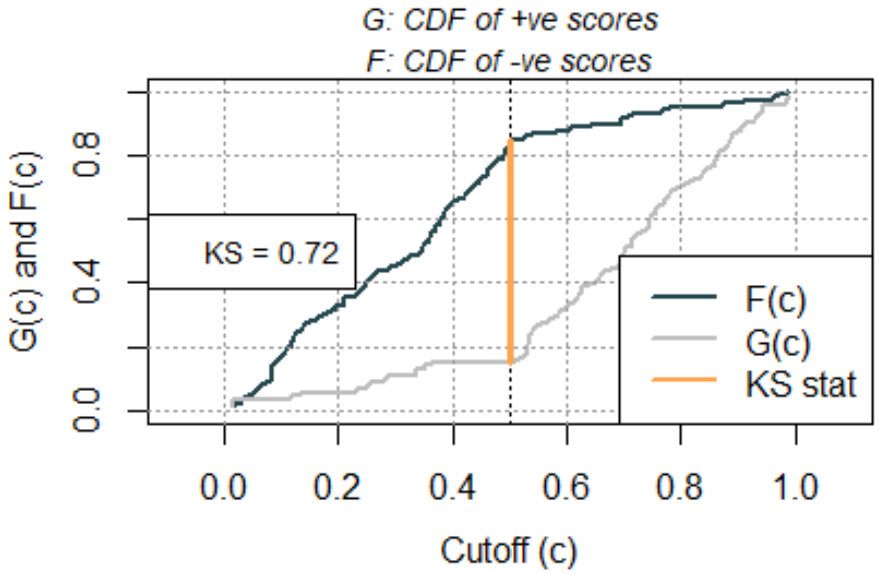

Figura 4.2: Distribución Acumulada Modelo Probit
Elaborado: autor.

	GVIF	**Df**	$GVIF^{(1/(2*Df))}$
ciudad	1.058576	4	1.007141
sexo	1.065775	1	1.032364
edad	1.325190	1	1.151169
lee_escribe	1.226192	1	1.107335
trabajo_1sem	2.413320	1	1.553486
monto_cuenta_ahorros	1.122822	1	1.059633
condicion_actividad	3.025629	3	1.202641
ingpc	1.288927	1	1.135309
situacion_futura	1.040837	2	1.010057

Cuadro 4.5: Factor GVIF para los parámetros estimados del modelo probit

Elaborado: autor

Como se observa en el cuadro 4.5 los valores no sobrepasan los límites, lo que prueba que no existe multicolinealidad, es decir, no hay correlación entre las variables predictoras y la variable respuesta.

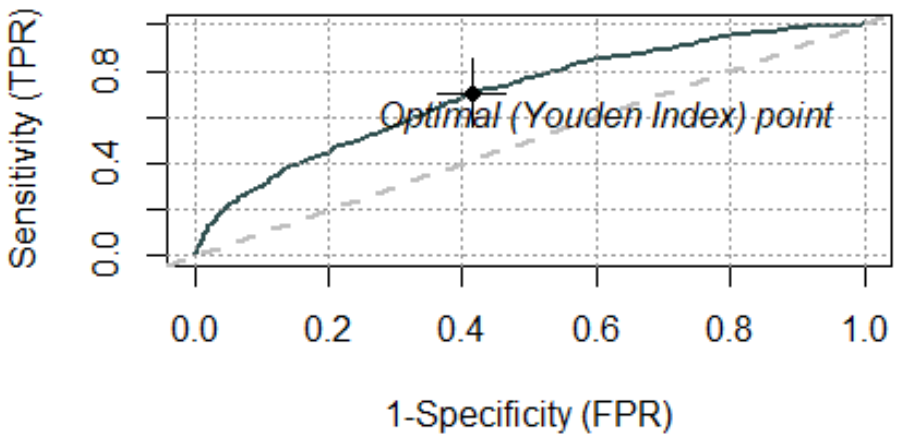

Figura 4.3: Curva ROC del Modelo Probit
Elaborado: autor.

Capítulo 5

Modelo Logit

Considerando el método de selección de variables, el Modelo Logit estimado es el siguiente:

Coefficients	**Estimación**	**Sd.standar**	**Valor z**	**P-valor**	
Intercepto	-2.365e+00	6.541e-01	-3.615	0.000300	***
ciudadORIENTE	1.045e+00	3.574e-01	2.925	0.003447	**
ciudadSIERRA	8.457e-01	2.489e-01	3.398	0.000678	***
sexoM	2.806e-01	2.564e-01	1.095	0.273709	**
edad	-3.055e-02	8.310e-03	-3.676	0.000237	***
lee_escribe1	1.048e-01	2.321e-01	0.45 1	0.651650	***
trabajo_1sem1	-9.278e-01	4.006e-01	-2.316	0.020551	***
monto_cuenta_ahorros	4.613e-04	3.342e-04	1.381	0.167433	***
condicion_actividadINAC NO REM	-1.234e+00	6.231e-01	-1.980	0.047693	***
condicion_actividadSUB MEN15	-3.036e-01	3.969e-01	-0.765	0.444280	***
ingpc	5.827e-04	2.376e-04	2.453	0.014182	***
situacion_futuraMEJOR	-1.475e+00	2.479e-01	5.952	2.65e-09	***

Cuadro 5.1: Estimación Logit para la demanda de Crédito

Elaborado: autor

Se observan los resultados en la tabla 5.1, con sus respectivos valores y significancia[1]. Como se mencionó anteriormente, en los modelos logit los parámetros corresponden al efecto marginal sobre la variable dependiente de un cambio en una de las variables de control.

Es importante precisar que muchas veces se buscan estimar probabilidades y probabilidades condicionadas, en lugar de interpretaciones directas sobre las varia-

[1]Significancia de los códigos: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1

bles predictoras. Sin embargo, se intenta explicar qué representan las estimaciones de los coeficientes del modelo.

- Los signos de los coeficientes se mantienen con respecto al Modelo Probit, en la mayoría de las estimaciones.
- El coeficiente de un predictor continuo es el cambio estimado en el logaritmo natural de las probabilidades para el evento de referencia por cada incremento de una unidad en el predictor. Por ejemplo, Por cada unidad porcentual de adicional del ingreso per cápita, el logaritmo de la razón de probabilidad de necesitar el crédito aumenta 0.0006491.
- Para predictores continuos, la interpretación de las probabilidades puede ser más significativa que la interpretación de la relación de probabilidades.
- Para las variables factores el coeficiente es el cambio estimado en el logaritmo natural de las probabilidades cuando se cambia del nivel de referencia al nivel del coeficiente. Por ejemplo, la variable *lee_escribe*, el coeficiente es 0.1048, entonces un cambio en la variable de sabe leer y escribir a no sabe ninguna, hace que el logaritmo natural de las probabilidades del evento aumente en 0.1048.

5.1. Bondad de Ajuste

Los estadísticos de bondad de ajuste y sus respectivos valores se presentan a continuación:

Test	**Estadístico**	**P-valor**
Devianza	813.6174	0.000
Pearson	3466.15	0.000
Hosmer y Lemeshow	3400	0.000

Cuadro 5.2: Pruebas de Bondad de Ajuste Modelo Logit

Elaborado: autor

Entonces para un nivel de significación del 5 % se acepta la hipótesis nula de que el modelo se ajusta bien a los datos observados.

A continuación, se observa la Matriz de Confusión, o Tabla de Clasificación.

	Observación	
Estimación	**0**	**1**
0	8277	118
1	4914	292

Cuadro 5.3: Tabla de Clasificación del Modelo Logit

Elaborado: autor

Los valores del cuadro 5.3 se obtuvieron con los datos de la muestra de validación.

El coeficiente de *GINI* es: 0.4445753. El área bajo la curva *AUC* es: 0.7222876. Este último representa el poder de predicción del modelo, el cual es estadísticamente bueno.

Se pueden notar que algunas observaciones pueden ser consideradas valores outlier. Se procede a observar la gráfica de la curva ROC y la distribución acumulada con el estadístico $K - S$.

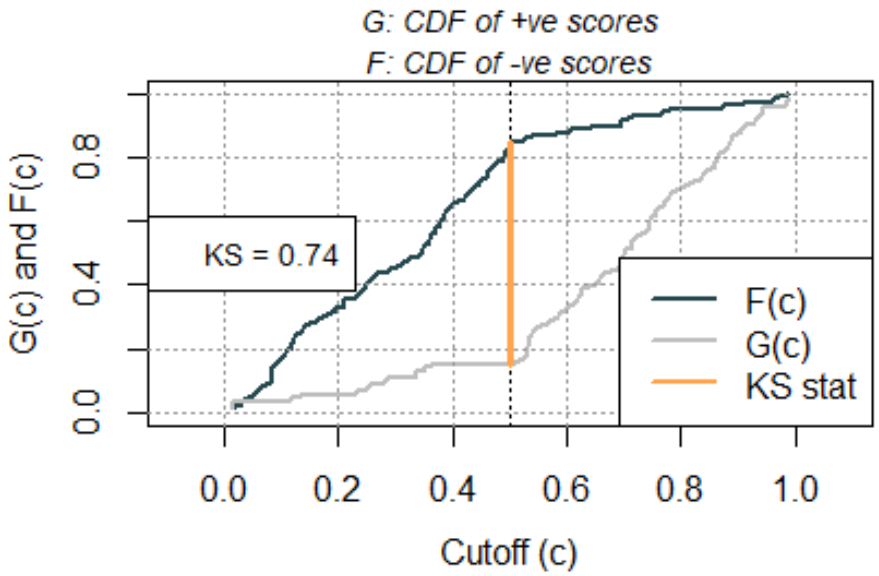

Figura 5.1: : Distribución Acumulada Modelo Logit
Elaborado: autor.

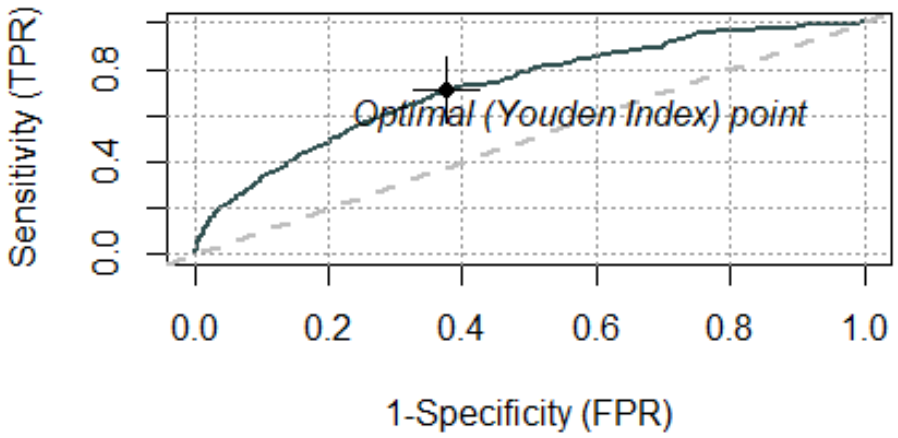

Figura 5.2: Curva ROC del Modelo Logit
Elaborado: autor.

La curva ROC ayuda a reconocer el poder de predicción del modelo. En este caso, el modelo es aceptable y se procede a las pruebas de multicolinealidad.

5.1.1. Multicolinealidad

Se presenta a continuación la prueba de multicolinealidad. Esta representa la correlación entre las variables predictoras y la variable respuesta. Como supuesto, no debe existir correlación.

	GVIF	**Df**	$GVIF^{(1/(2*Df))}$
ciudad	1.063033	4	1.007670
sexo	1.061740	1	1.030408
edad	1.273320	1	1.128415
lee_escribe	1.206188	1	1.098266
trabajo_1sem	1.968085	1	1.402885
monto_cuenta_ahorros	1.162742	1	1.078305
condicion_actividad	2.479379	3	1.163386
ingpc	1.304562	1	1.142174
situacion_futura	1.047247	2	1.011608

Cuadro 5.4: Factor GVIF para los parámetros estimados del Modelo Logit

Elaborado: autor

Como se observa en el cuadro 5.4 los valores no sobrepasan los límites, lo que prueba que no existe multicolinealidad, es decir, no hay correlación entre las variables predictoras y la variable respuesta.

Capítulo 6

Modelo Logit con corrección de Sesgo

Pueden notarse que los resultados de los Modelos Probit y Logit son muy parecidos. Además, tiene un fuerte poder de predicción para los casos negativos, es decir, cuando un individuo no necesita de crédito. Pero para el interés del trabajo, se necesita que el modelo sea capaz de trabajar sobre los individuos que sí necesitan de crédito. Por la naturaleza de la información, existe un sesgo que ha afectado a los modelos anteriores y que se pretende corregir en esta sección.

Como se observa en el capítulo 2.5, es necesario trabajar sobre una muestra que considere la diferencia de proporciones en la variable respuesta. Para ello, se ha considerado toda la población cuya respuesta ha sido que sí necesita de crédito y se ha tomado, en forma aleatoria, **5 veces más** a la población cuya respuesta fue que no.

Las estimaciones fueron las siguientes:

Coefficients	**Estimación**	**Sd.standar**	**Valor z**	**P-valor**	
(Intercept)	-4.825e+00	4.901e-01	-9.845e+00	<2e-16	***
edad	2.122e-02	5.802e-03	3.657e+00	0.000255	***
ingpc	-8.545e-04	2.645e-04	-3.230e+00	0.001236	**
situacion_futuraMEJOR	-1.622e+00	1.909e-01	-8.496e+00	<2e-16	***
lee_escribe1	4.704e-01	1.418e-01	3.317e+00	0.000911	***
trabajo_1sem1	8.739e-01	3.315e-01	2.636e+00	0.008391	**
condicion_actividadINAC NO REM	1.832e+00	4.462e-01	4.105e+00	4.03e-05	***

condicion_actividadNO CLAS NI PLENO	8.562e-01	2.495e-01	3.432e+00	0.000600	***
ciudadGALAPAGOS	-5.067e+05	5.573e-01	9.091e+05	<2e-16	***
ciudadORIENTE	-1.061e+00	2.270e-01	-4.676e+00	2.93e-06	***
ciudadSIERRA	-8.885e-01	1.513e-01	-5.872e+00	4.30e-09	***

Cuadro 6.1: Estimación Logit con corrección de sesgo para la demanda de Crédito

Elaborado: autor.

Es importante considerar que las probabilidades estimadas en el contexto de eventos raros tienden a ser bajas, e incluso puede darse el caso de que ninguna probabilidad exceda de 0,5 [16]. Las estimaciones de los coeficientes, si bien son significativos [1], difieren en los signos de las estimaciones de algunos de los parámetros de los modelos Logit y Probit. Es importante ser cuidadoso en las interpretaciones cuando esto ocurre.

6.1. Bondad de Ajuste

Los estadísticos de bondad de ajuste y sus respectivos valores se presentan a continuación:

Test	**Estadístico**	**P-valor**
Devianza	751.35	0.000
Pearson	830.33	0.000
Hosmer y Lemeshow	5400	0.000

Cuadro 6.2: Pruebas de Bondad de Ajuste Modelo Logit con corrección de sesgo

Elaborado: autor

Entonces para un nivel de significación del 5 % se acepta la hipótesis nula de que el modelo se ajusta bien a los datos observados.

[1] Significancia de los códigos: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1

A continuación, se observa la Matriz de Confusión, o Tabla de Clasificación.

	Observación	
Estimación	**0**	**1**
0	1093	101
1	171	1707

Cuadro 6.3: Tabla de Clasificación del Modelo Logit con corrección de sesgo

Elaborado: autor

Los valores del cuadro 6.3 se obtuvieron con los datos de la misma muestra, únicamente para este caso, debido a que se posee una cantidad pequeña de datos cuya variable respuesta es sí.

El coeficiente de *GINI* es: 0.5832. El área bajo la curva *AUC* es: 0.7223. Este último representa el poder de predicción del modelo, el cual es estadísticamente bueno.

Se pueden notar que algunas observaciones pueden ser consideradas valores outlier. Se procede a observar la gráfica de la curva ROC y la distribución acumulada con el estadístico $K - S$.

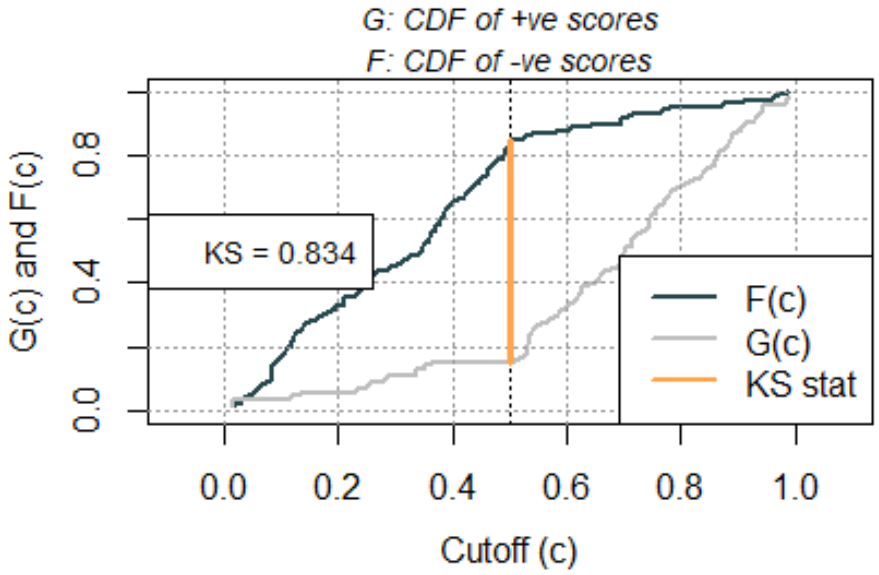

Figura 6.1: : Distribución Acumulada Modelo Logit con corrección de sesgo
Elaborado: autor.

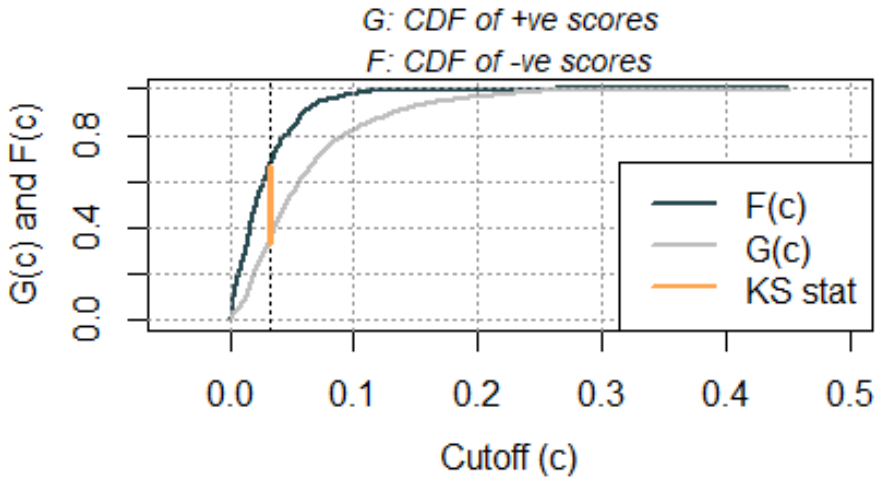

Figura 6.2: Curva ROC del Modelo Logit con corrección de sesgo
Elaborado: autor.

La curva ROC ayuda a reconocer el poder de predicción del modelo. En este caso, el modelo es aceptable y se procede a las pruebas de multicolinealidad.

6.1.1. Multicolinealidad

Se presenta a continuación la prueba de multicolinealidad. Esta representa la correlación entre las variables predictoras y la variable respuesta. Como supuesto, no debe existir correlación.

	GVIF	**Df**	$GVIF^{(1/(2*Df))}$
ciudad	1.14302	4	1.027670
sexo	1.092751	1	1.050408
edad	1.653320	1	1.238415
lee_escribe	1.634188	1	1.075266
trabajo_1sem	1.968085	1	1.272885
condicion_actividad	2.479379	3	1.163386
ingpc	1.304562	1	1.142174
situacion_futura	1.047247	2	1.011608

Cuadro 6.4: Factor GVIF para los parámetros estimados del Modelo Logit con corrección de sesgo

Elaborado: autor

Como se observa en el cuadro 6.4 los valores no sobrepasan los límites, lo que prueba que no existe multicolinealidad, es decir, no hay correlación entre las variables predictoras y la variable respuesta. Esto permite validar los supuestos principales del modelo, es decir, que no existe relación de dependeica de las variables predictoras.

6.2. Elección del Modelo

La decisión sobre qué modelo toma en cuenta la figura 6.3 con las curvas ROC para los 3 estimadores y en el cuadro 6.5 se presenta el área bajo la curva ROC, punto de corte basado en el punto más cercano a la esquina superior izquierda y la especificidad, AIC y tasa de clasificaciones correctas bajo este punto de corte.

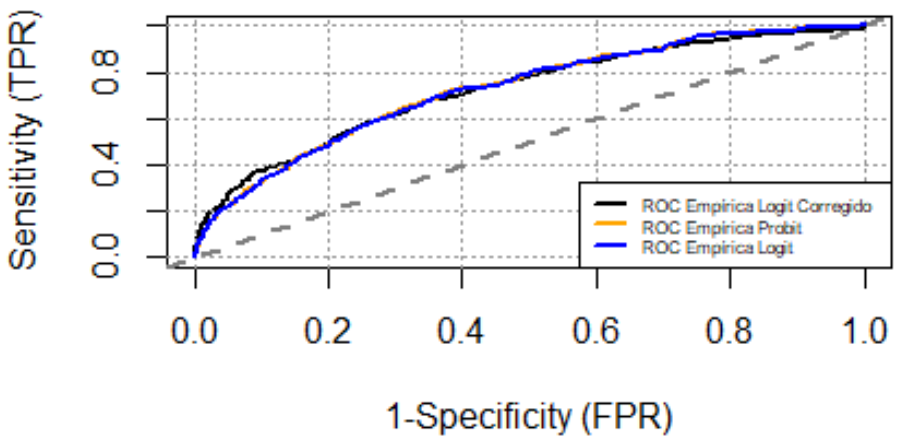

Figura 6.3: Curva ROC de los Modelos Probit, Logit y Logit con corrección de sesgo
Elaborado: autor.

La elección del mejor modelo puede ser el Modelo Logit con corrección de Sesgo si se toma en consideración los valores ***AUC*** y ***Tasa de calificaciones correctas***.

Modelo	AUC	Punto de corte	AIC	Tasa de clasificaciones correctas
Probit	0.7342539	0.02545957	845.62	0.6300272
Logit	0.7452876	0.02457219	845.76	0.6270127
Logit Corregido	0.8342683	0.03166267	212.02	0.866667

Cuadro 6.5: Factor GVIF para los parámetros estimados del Modelo Logit con corrección de sesgo

Elaborado: autor

Capítulo 7

Conclusiones

- En la presente investigación se analizan los determinantes que influyen en la probabilidad de los ecuatorianos de demandar crédito, a través de modelos logísticos, las variables consideradas que describen la situación socioeconómica del país y extraídas de la Encuesta Nacional de Empleo, Desempleo y Subempleo del año 2019 del INEC. Se evidencia que tales variables influyen sobre los individuos al momento de demandar crédito.

- Una de las variables cuya significancia destaca en los 3 modelos es *situacion_futura*, esta determina que mientras exista mejor expectativa en el individuo sobre la situación futura del país, menor es la probabilidad de que necesite de un crédito. Este hecho sugiere que la necesidad de crédito de la población ecuatoriana está condicionada a los cambios socieconómicos del país, lo cual se debe considerar primordial puesto que el período post-Covid en el Ecuador, y en todos los países, tendrá muchos cambios sobre la economía del país.

- La probabilidad de demandar crédito aumenta cuando los individuos pertenecen a las regiones Sierra y Oriente. Si bien esto ayuda a un enfoque más específico sobre donde está la población de interés, es importante reconocer que el proceso tiene alcances mayores y que, sobretodo, se debe continuar con las demás partes de la inclusión financiera que son: uso y calidad de los productos financieros. Es decir, si existe población que demanda crédito en estos sectores, hay que investigar si pudieron acceder a este y que tan beneficioso les resultó.

- El análisis económico sobre la demanda y uso de crédito tiene una bibliografía amplia en distintos países, sin embargo, es muy reciente y merece atención

por parte de los investigadores de las políticas de inclusión financiera, especialmente en Ecuador.

- Los resultados en las estimaciones del modelo Logit y el modelo Probit son similares en cuanto a signos y cuantificaciones de los parámetros.Por otro lado, las estimaciones del modelo Logit con corrección de sesgo tiene resultados más precisos y un enfoque sobre la población de interés, recordando que este modelo trabaja sobre la población que requiere de crédito y, como objetivo planteado, merece atención y estudio para reconocer los motivos por los cuales ha podido o no acceder a este.

Capítulo 8

Recomendaciones

- Cabe recalcar que los resultados que se obtuvieron reflejan las características de la población en un período limitado de tiempo, lo que conduce a la necesidad de promover la investigación a futuro para posibles comparaciones y, sobretodo, porque la población ecuatoriana enfrenta cambios resultantes de la pandemia actual vivida.

- Los planes de incidencia que suelen ser sugeridos sobre la oferta de cartera crediticia consideran necesario estudiar la demanda de esta, ya que esta influye sobre la posición en las personas sobre hacer, o no, uso de los recursos financieros. Para ello, se sugieren estudios que den importancia a las poblaciones , no solo que necesitan de crédito, sino también que deciden realizarlo y logran obtenerlo. Para conocer esta característica se puede profundizar la entrevista o encuesta a la población con preguntas específicas. Un ejemplo más preciso puede observarse en el estudio realizado en Perú [31], el cual considera las razones por las cuales las personas, pese a necesitar crédito, deciden no solicitar uno.

- Las políticas sobre el manejo de los recursos financieros influyen sobre la respuesta por parte de la población. Desde las tasas de crédito existentes, hasta la distribución geográfica de las instituciones financieras han sido relevantes a la hora de solicitar un crédito. Para promover la inclusión financiera es necesario considerar cambios positivos tales como: taller de información sobre poblaciones con menores recursos, expansión de las instituciones financieras en zonas con mayor población (no únicamente sobre ciudades capitales), entre otras. Cambios que, por supuesto, deberán realizarse considerando los resultados de diversos estudios que con el tiempo se sugiere realizar con una mayor

participación política.

Bibliografía

[1] Demirgüç-Kunt, A. y Klapper, L. Measuring Financial Inclusion: The Global Findex Database. *Policy Research Working Paper*, 6025:2–5, 2012.

[2] A. Demirgüç-Kunt, L. Klapper, D. Singer, S. Ansar, and J. Hess. *La base de datos Global Findex 2017: Medición de la inclusión financiera y la revolución de la tecnología financiera*. Banco Internacional de Reconstrucción y Fomento, Estados Unidos, 2018.

[3] D. Rivadeneira, D. Sandoval, D. Zambonino, A. Albán, and C. Garcés. Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) Documento Metodológico. *INEC - Instituo Nacional de Estadística y Censos*, 2:3, 2019.

[4] L. Vera. Determinantes de la demanda de crédito. Una estimación con un modelo mensual de series de tiempo para Venezuela. *Escuela de Economía de la Universidad Central de Venezuela y Centro de Estudios Latinoamericanos de la Universidad de Oxford*, pages 110–115, 2003.

[5] E. Veintimilla. Estimación Econométrica de una función de demanda de crédito para el Ecuador: período enero 1990-diciembre 1997. *Revista Cuestiones Económicas; Vol 16, No 2 (Año 2000).*, 16:72–74, 2000.

[6] Z. Maoz and B. Russett. The little data book on financial inclusion 2018. *Banco Mundial*, page 12–15, 2018.

[7] Asobanca. `https://asobanca.org.ec/educacion-financiera/la-inclusion-financiera-genera-un-ganar-ganar/`, 2019. [En línea].

[8] S. Claessens. Access to financial services: A review of the issues and public policy objectives. *World Bank Research Observer*, 21:207–240, 2006.

[9] A. Novales. *Econometria (2a. ed.).* McGraw Hil, Madrid, 1993.

[10] J. Nelder and R. Wedderburn. *Generalized Linear Models*. Journal of the Royal Statistical Society A., Estados Unidos, 1972.

[11] M. Müller. Generalized Linear Models. *Fraunhofer Institute for Industrial Mathematics (ITWM)*, pages 3–10, 2004.

[12] E. Lay. *Modelo Logístico para eventos raros: aplicación para predecir el incumplimiento de pago en una empresa de productos de belleza*. PhD thesis, Departamento de Estadística e Investigación Operativa. Universidad de Granada, Septiembre 2015.

[13] A. Agresti. *Categorical Data Analysis*. Wiley, New York, 2013.

[14] R. A. Fisher. *The Logic of Inductive Inference*. Blackwell Publishing for the Royal Statistical Society, Reino Unido, 2009.

[15] Z. Maoz and B. Russett. Normative and structural causes of democratic peace, 1946–86. *American Political Science Review 87*, 3:624–638, 1993.

[16] G. King and L. Zeng. Logit regression in rare events data. *Center for Basic Research in the Social Sciences*, pages 145–153, 2001.

[17] N. Beck, G. King, and L. Zeng. Improving quantitative studies of international conflict: A conjecture. *American Political Science Review*, 94:1–15, 2000.

[18] G. Imbens. Improving quantitative studies of international conflict: A conjecturean efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, 60:1187–1214, 1992.

[19] S. Cosslett. Efficient estimation of discrete choice models. *MIT Press. MA*, 60:90–94, 1981.

[20] C. Manski and S Lerman. The estimation of choice probabilities from choice based samples. *Econometrica 45*, 45:1977–1988, 1998.

[21] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons Inc., New York, 2013.

[22] P. Ryan. *Modern Regression Methods*. Wiley, New York, 2013.

[23] V Berlanga, J. Rubio, and R. Vilà. Cómo aplicar árboles de decisión en spss. *Revista d'innovació i recerca en Educació*, pages 68–70, 2013.

[24] Mirzaeia S., G. Mohtashami, and M. Amini. A comparative study of the gini coefficient estimators based on the linearization and u-statistics methods. *Revista Colombiana de Estadística*, 40:205–221, 2017.

[25] P. Kisbye. Test de kolmogorov-smirnov. *La Facultad de Matemática, Astronomía, Física y Computación*, pages 4–15, 2010.

[26] S. Finlay. *Credit Scoring, Response Modelling and Insurance Rating: A Practical Guide to Forecasting Consumer Behaviour*. Palgrave Macmillan, New York, 2010.

[27] Instituto Nacional de Estadística y Censos. `https://www.ecuadorencifras.gob.ec/institucional/comunicado-oficial-enemdu/`, 2019. [En línea].

[28] Organización Mundial de la Salud. `https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/advice-for-public/q-a-coronaviruses`, 2019. [En línea].

[29] Censo Nacional de Instituciones Educativas. `http://web.educacion.gob.ec/CNIE/pdf/Anexo%20con%20Codificacion.pdf`, 2019. [En línea].

[30] Ley de Pareto. `https://es.wikipedia.org/wiki/Principio_de_Pareto`, 2020. [En línea].

[31] J. Alvarado and M. Pintado. *Necesidad, demanda y obtención de crédito en el sector agropecuario en el Perú*. n IV Censo Nacional Agropecuario 2012: Investigaciones para la toma de decisiones en políticas públicas, Lima, 2017.

[32] B. Haewon. Chi-square automatic interaction detection modeling for predicting depression in multicultural female students. *Department of Speech Language Pathology*, 12:179–181, 2017.

[33] L. Cayuela. *Modelos Lineales Generalizados (GLM)*. IREC, New York, 2009.

[34] E. Perez. *Algoritmo de random forest aplicado a la detección de fraude en el sistema bancario ecuatoriano*. PhD thesis, Escuela Politécnica Nacional, Diciembre 2019.

Apéndice A

Tabla de Variables

No.	Variable	Tipo	Descripción
1	area	categórica	Área en que vive
2	ciudad	categórica	Código de la ciudad en que vive, existen 547 ciudades diferentes.
3	hogar	categórica	Hogar
4	sexo	categórica	Sexo: Masculino o Femenino
5	edad	numérica	Edad
6	estado_civil	categórica	Estado civil
7	anio_aprobado	categórica	Es el año aprobado en relación con el nivel de instrucción
8	lee_escribe	factor	Sabe leer y escribir. Si=1, No = 0
9	etnia	categórica	Se refiere a como se identifican las personas según sus culturas y constumbres.
10	trabajo_1sem	factor	Trabajo la semana pasada. Si=1, No = 0
11	busco_trabajo	categórica	Durante las últimas cuatro semanas hizo alguna gestión para buscar trabajo
12	condic_inact	categórica	Condición de inactividad. De los que no trabajan
13	dejo_trabaj	categórica	Motivos por los que dejó de trabajar

14	sem_sin_trabajo	numérica	Número de semanas sin trabajar
15	rama_actividad	categórica	Se refiere a la actividad a la que se dedica el negocio
16	grupo_ocupacion	categórica	Grupo de ocupación. Es la tarea o actividad específica que desarrolla o desarrolló el trabajador dentro del establecimiento.
17	categoria_ocupacion	categórica	Es la relación de dependencia en la que una persona ejerce su trabajo
18	estabilidad	categórica	Se refiere a la estabilidad de una persona ocupada en una empresa o establecimiento, en el que actualmente está trabajando.
19	sitio_trabajo	categórica	Corresponde al sitio o lugar de trabajo, donde la persona realiza su actividad productiva.
20	reg_contab	categórica	Se refiere a si el negocio o empresa lleva un registro contable completo.
21	ruc_establ	categórica	Se refiere a si el negocio o dependencia cuenta con uno de los requisitos de funcionamiento, el Registro Único de Contribuyentes.
22	aporte_seguros	categórica	A cuál de las siguientes formas de seguridad social aporta actualmente.
23	ingresos	numérica	Monto en dinero que recibió por la venta de los productos, bienes o servicios de su negocio o establecimiento.
24	retiro_negocio	factor	Retiró de su negocio o tomó de lo que produce o vende, bienes, servicios o productos para el consumo del hogar. Si=1, No=0.

25	recibio_especias_alim	factor	Recibió por su trabajo especies o servicios tales como: alimentos, vivienda, vestido, etc. Si=1, No=0.
26	salario_indep	numérica	Ingreso de salario independiente.
27	recibio_bienes	factor	Recibió por su trabajo pago en especies o retiró del negocio o producción bienes o productos para el consumo del hogar, Si=1, No=0.
28	recibio_cuenta_ahorros	factor	Recibió ingresos por concepto de intereses por: cuenta de ahorros, corrientes, préstamos a terceros, hipotecas; bonos por acciones; arriendo de casas, edificios, terrenos, maquinaria, etc. Si=1, No=0.
29	monto_cuenta_ahorros	numérica	Recibió ingresos por concepto de intereses por: cuenta de ahorros, corrientes, préstamos a terceros, hipotecas; bonos por acciones; arriendo de casas, edificios, terrenos, maquinaria, etc.
30	recibio_pensiones	factor	Recibió ingresos por concepto de pensión por: jubilación, orfandad, viudez, invalidez, enfermedad, divorcio, cesantía, etc. Si=1, No=0.
31	monto_donacion	numérica	Recibió dinero o especies por regalos o donaciones de personas o instituciones que vivan dentro del país.
32	recibio_familiares	factor	Recibió dinero o especies enviado por parte de familiares o amigos que vivan en el exterior. Si=1, No=0.

33	monto_familiares	numérica	Recibió dinero o especies enviado por parte de familiares o amigos que vivan en el exterior.
34	recibe_bono_desarrollo	numérica	Recibe el BONO DE DESARROLLO HUMANO. Si=1, No=0.
35	monto_bono_desarrollo	numérica	Cuanto recibió por el bono de desarrollo humano.
36	recibe_bono_cuidar	factor	Recibe el BONO POR EL CUIDADO BRINDADO A UNA PERSONA DISCAPACITADA DEL HOGAR. Si=1, No=0.
37	motivo_desempleo	categórica	El principal motivo por el que usted está sin trabajo.
38	condicion_actividad	categórica	Condición de actividad
39	sect_empleados	catgórcia	Sectorización de empleados (15 y más)
40	empleo	factor	Población con empleo.
41	desempleo	factor	Población desempleada.
42	poblacion_ocupacion	categórica	Grupo de Ocupación CIUO8 (población ocupada de 15 años y más)
43	poblacion_rama	categórica	Rama de Actividad CIIU4 (población ocupada de 15 años y más)
44	ingpc	numérica	Ingreso per cápita.
45	pobreza	factor	Pobreza
46	epobreza	factor	Extrema pobreza. NO INDIGENTE=0, INDIGENTE=1
47	nivel_instruccion	categórica	Nivel de instrucción
48	situacion_economica	categórica	Con relación al mes anterior, la situación económica de su hogar es: Mejor=1, Igual=2, Peor=3
49	situacion_futura	categórica	Cómo cree usted que será la situación económica de su hogar dentro de los próximos 3 meses es Mejor=1, Igual=2, Peor=3

50	variable_respuesta	factor	Usted o algún miembro del hogar tiene planes de endeudamiento en los próximos 3 meses (bancos, financieras, etc.) Si=1, No=2
51	via_vivienda	categórica	Vía de acceso principal a la vivienda
52	tipo_vivienda	categórica	Tipo de vivienda
53	material_techo	categórica	Material techo
54	estado_techo	categórica	Estado del techo. Bueno=1, Regular=2, Malo=3
55	material_piso	categórica	Material piso
56	estado_piso	categórica	Estado del piso. Bueno=1, Regular=2, Malo=3
57	estado_paredes	categórica	Estado de las paredes. Bueno=1, Regular=2, Malo=3
58	material_paredes	categórica	Material de las paredes.
59	estado_vivienda	categórica	Forma de tenencia de la vivienda.

Cuadro A.1: Variables utilizadas para el modelo

Elaborado: autor.

Apéndice B

Código en R

[1]

```
# librerias ----
library(haven)
library(data.table)
library(tidyverse)
library(dplyr)
library(stats)
library(party)
library(partykit)
library(rpart)
library(ROSE)
library(Momocs)
library(rpart.plot)
library(xlsx)
library(caTools)
library(ResourceSelection)
library(plotROC)
library(ggplot2)
library(ROCit)
library(nortest)
library(tsoutliers)
library(normtest)
library(Zelig)
library(logistf)
library(pedometrics)

# Seleccion de Variables ----
```

[1] El código completo y las bases de datos pueden ser solicitados para mejor comprensión, en caso de necesitarlos para una investigación.

```r
# Funcion KS
TestKS <- function(x, y){
  if(class(x)!="character"){
    vars <- data.frame(y,x)
    vars_e <- subset(vars,subset=vars[,1]==1)
    vars_f <- subset(vars,subset=vars[,1]==0)
    ks <- suppressWarnings(ks.test(vars_e[,2],vars_f[,2],alternative="two.sided"))
    ks <- round(as.numeric(ks$statistic),4)
  } else{
    ks <- 0
  }
  return(ks)
}

# Valor de informacion (IV)
TestVI <- function(x,y){
  if(class(x)=="character"){
    tc <- table(y,x)
    f1 <- tc[1,]
    f2 <- tc[2,]
    aux1 <- ifelse(f1/sum(f1)==0,0.001,ifelse(f1/sum(f1)==1,0.999, f1/sum(f1)))
    aux2 <- ifelse(f2/sum(f2)==0,0.001,ifelse(f2/sum(f2)==1,0.999, f2/sum(f2)))
    wof <- log(aux2/aux1)
    wof <- ifelse(wof==-Inf,0,wof)
    VI <-   sum(((f2/sum(f2))-(f1/sum(f1)))*wof)
  }else{
    VI <- 0
  }
  return(VI)
}

dnum <- datos[, c("edad", "anio_aprobado", "sem_sin_trabajo",
                  "ingresos", "salario_indep", "monto_cuenta_ahorros",
                  "monto_donacion", "monto_familiares", "ingpc")]
dcat <- datos[,- c("edad", "anio_aprobado", "sem_sin_trabajo",
                   "ingresos", "salario_indep", "monto_cuenta_ahorros",
                   "monto_donacion", "monto_familiares", "ingpc")]

VarDep <- dcat$variable_respuesta
KS <- sapply(dnum, TestKS, y =VarDep)
KS <- sort(KS, decreasing = T)
```

```
68 d.KS <- data.frame(names(KS), KS)
69 colnames(d.KS) <- c("Variable", "KS")
70 row.names(d.KS) <- NULL
71
72 VI <- sapply(dcat, TestVI, y =VarDep)
73 VI <- sort(VI, decreasing = T)
74 d.VI <- data.frame(names(VI), VI)
75
76 colnames(d.VI) <- c("Variable", "VI")
77 row.names(d.VI) <- NULL
78 d.VI <- as.data.table(d.VI)
79
80 # Guardar
81 write.xlsx(list("KS_Var" = d.KS), file = "KS.xlsx")
82 write.xlsx(list( "VI" = d.VI), file = "VI.xlsx")
83
84 # Creacion de variables con arboles ----
85 datos$categoria_ocupacion <- ifelse(is.na(datos$categoria_ocupacion)==
       T,"DESEMP",
86                                     ifelse(datos$categoria_ocupacion==
        "PROP-CUENT", "PROP-CUENT",
87                                            ifelse(datos$categoria_
       ocupacion=="EMPLEAD DOM"|
88                                                   datos$categoria_
       ocupacion=="JORN-NO-REM"|
89                                                   datos$categoria_
       ocupacion=="JORN-PEON", "EMPLEADO JORN","GOB-PRIV-HOG")))
90
91 # condicion_actividad:
92 datos$condicion_actividad <- ifelse(datos$condicion_actividad=="EMPLEO
        PLENO", "EMPLEO PLENO",
93                                     ifelse(datos$condicion_actividad==
        "EMPLEO NO PLENO"|
94                                              datos$condicion_actividad
       == "NO CLASIF", "NO CLAS NI PLENO",
95                                            ifelse(datos$condicion_
       actividad=="DESEMPL OCULTO"|
96                                                   datos$condicion_
       actividad=="NO REMUN"|
97                                                   datos$condicion_
       actividad=="POBL ECON INAC", "INAC NO REM","SUB MEN15")))
98 # nivel_instruccion
99 datos$nivel_instruccion <- ifelse(datos$nivel_instruccion== "BACHILLER
       "|
```

```
                                    datos$nivel_instruccion=="SUPERIOR
    "|
                                    datos$nivel_instruccion=="CENTR
    ALFAB", "BACHILLER O SUP",datos$nivel_instruccion)
# tipo_vivienda
datos$tipo_vivienda <- ifelse(datos$tipo_vivienda== "CASA VILLA"|
                                datos$tipo_vivienda=="MEDIAGUA", "CASA
    /VILLA",
                              ifelse(datos$tipo_vivienda=="DEPARTAM",
    "DEPARTAM","INQUILINO"))

# Muestra de modelamiento y validacion ----
set.seed (12345)
sample <- sample.split(  datos$variable_respuesta , SplitRatio = 0.2)
mod <-setDT ( subset ( setDF ( datos ) , sample == TRUE ) )
val <-setDT ( subset ( setDF ( datos ) , sample == FALSE ) )

# Modelo Probit ----
fitprob <- glm(
  variable_respuesta ~
  ciudad+
  sexo+
  edad+
  lee_escribe+
  trabajo_1sem+
  grupo_ocupacion+
  categoria_ocupacion+
  ingresos+
  monto_cuenta_ahorros+
  condicion_actividad  +
  ingpc+
  situacion_futura,
  family = binomial ( link = "probit") ,data = mod )
summary(fitprob)

# Residuos
devianza <-sum( residuals ( fitprob , type ="deviance") ^2)
devianza_p <- 1 - pchisq( devianza , fitprob$df.null - fitprob$df.
    residual )

pearson <-sum(residuals(fitprob , type ="pearson") ^2)
pearson_p <-1 - pchisq( pearson , fitprob$df.null - fitprob$df.
    residual )

```

```r
hoslem.test(mod$variable_respuesta, fitted(fitprob),g=10)

rp <- rstandard( fitprob, type = "pearson" )
rd <- rstandard( fitprob, type = "deviance" )
png("residuos_probit.png", width = 450, height = 300, units = "px",
    pointsize = 12)
par( mfrow = c( 1, 2 ) )
plot( rp,main="Residuos de Pearson",
      ylab="Residuo de Pearson", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
plot(rd,main="Residuos de la devianza",
      ylab="Residuo de la devianza", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
dev.off()

# Predicciones
res <- predict(fitprob, val, type="response")
res <- ifelse(res > 0.5, 1, 0) # Punto de corte en 0.5

# Matriz de confusion
mc <- table(res, val$variable_respuesta)
mc[1,1] # Verdaderos positivos
mc[2,2] # Verdaderos negativos
mc[1,2] # Falsos positivos
mc[2,1] # Falsos negativos

# Tasas clasificadas correctamente / Pseudo R cuadrado
(mc[1,1]+mc[2,2])/sum(mc)

# Curva ROC
dres <- data.frame(pred=predict(fitprob, val, type="response"), var=
    val$variable_respuesta)
ROC <- rocit(score=dres$pred, class=dres$var)
plot(ROC,legend = FALSE)

# AUC
ROC$AUC

#GINI
ROC$gini

# Prueba de Kolmogorov - Smirnov
lillie.test( res )

```

```r
# Prueba de Jarque Bera
jb.norm.test(res)

#Prueba Pearson
pearson.test(res)

# Modelo Logit
fitlogit <- glm(
  variable_respuesta ~
    ciudad+
    sexo+
    edad+
    #anio_aprobado+
    lee_escribe+
    trabajo_1sem+
    #grupo_ocupacion+
    #categoria_ocupacion+
    #ingresos+
    monto_cuenta_ahorros+
    condicion_actividad  +
    ingpc+
    situacion_futura,
  family = binomial ( link = "logit") ,data = mod )
summary(fitlogit)

# Residuos
devianza <-sum( residuals ( fitlogit , type ="deviance") ^2)
devianza_p <- 1 - pchisq( devianza , fitlogit$df.null - fitlogit$df.
    residual )

pearson <-sum(residuals(fitlogit , type ="pearson") ^2)
pearson_p <-1 - pchisq( pearson , fitlogit$df.null - fitlogit$df.
    residual )

hoslem.test(mod$variable_respuesta, fitted(fitlogit),g=10)

rp <- rstandard( fitlogit, type = "pearson" )
rd <- rstandard( fitlogit, type = "deviance" )
png("residuos_logit.png", width = 450, height = 300, units = "px",
    pointsize = 12)
par( mfrow = c( 1, 2 ) )
plot( rp,main="Residuos de Pearson",
      ylab="Residuo de Pearson", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
```

```
plot(rd,main="Residuos de la devianza",
     ylab="Residuo de la devianza", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
dev.off()

# Predicciones
res <- predict(fitlogit, val, type="response")
res <- ifelse(res > 0.5, 1, 0) # Punto de corte en 0.5

# Matriz de confusion
mc <- table(res, val$variable_respuesta)
mc[1,1] # Verdaderos positivos
mc[2,2] # Verdaderos negativos
mc[1,2] # Falsos positivos
mc[2,1] # Falsos negativos

# Tasas clasificadas correctamente / Pseudo R cuadrado
(mc[1,1]+mc[2,2])/sum(mc)

# Cura ROC
dres <- data.frame(pred=predict(fitlogit, val, type="response"), var=
    val$variable_respuesta)
ROC <- rocit(score=dres$pred, class=dres$var)
plot(ROC,legend = FALSE)
ksplot(ROC)

# AUC
ROC$AUC

#GINI
ROC$gini

# Prueba de Kolmogorov - Smirnov
lillie.test( res )

# Prueba de Jarque Bera
jb.norm.test(res)

#Prueba Pearson
pearson.test(res)

# Modelo Logit----
fitlogit <- glm(
  variable_respuesta ~
```

```
    ciudad+
    sexo+
    edad+
    lee_escribe+
    trabajo_1sem+
    monto_cuenta_ahorros+
    condicion_actividad  +
    ingpc+
    situacion_futura,
  family = binomial ( link = "logit") ,data = mod )
summary(fitlogit)

# Residuos
devianza <-sum( residuals ( fitlogit , type ="deviance") ^2)
devianza_p <- 1 - pchisq( devianza , fitlogit$df.null - fitlogit$df.
    residual )

pearson <-sum(residuals(fitlogit , type ="pearson") ^2)
pearson_p <-1 - pchisq( pearson , fitlogit$df.null - fitlogit$df.
    residual )

hoslem.test(mod$variable_respuesta, fitted(fitlogit),g=10)

rp <- rstandard( fitlogit, type = "pearson" )
rd <- rstandard( fitlogit, type = "deviance" )
png("residuos_logit.png", width = 450, height = 300, units = "px",
    pointsize = 12)
par( mfrow = c( 1, 2 ) )
plot( rp,main="Residuos de Pearson",
      ylab="Residuo de Pearson", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
plot(rd,main="Residuos de la devianza",
     ylab="Residuo de la devianza", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
dev.off()

# Predicciones
res <- predict(fitlogit, val, type="response")
res <- ifelse(res > 0.5, 1, 0) # Punto de corte en 0.5

# Matriz de confusion
mc <- table(res, val$variable_respuesta)
mc[1,1] # Verdaderos positivos
mc[2,2] # Verdaderos negativos
```

```
mc[1,2] # Falsos positivos
mc[2,1] # Falsos negativos

# Tasas clasificadas correctamente / Pseudo R cuadrado
(mc[1,1]+mc[2,2])/sum(mc)

# Cura ROC
dres <- data.frame(pred=predict(fitlogit, val, type="response"), var=
    val$variable_respuesta)
ROC <- rocit(score=dres$pred, class=dres$var)
plot(ROC,legend = FALSE)
ksplot(ROC)

# AUC
ROC$AUC

#GINI
ROC$gini

# Prueba de Kolmogorov - Smirnov
lillie.test( res )

# Prueba de Jarque Bera
jb.norm.test(res)

#Prueba Pearson
pearson.test(res)

# Muestra para Estimador KZ
set.seed(1234)
muestra1 <- datos[datos$variable_respuesta==1,]
muestra2 <- sample_n(datos[datos$variable_respuesta==0,],512*5)
muestra <- rbind(muestra1,muestra2)
muestra <- as.data.table(muestra)
muestra$variable_respuesta <- as.numeric(muestra$variable_respuesta)
muestra[muestra$variable_respuesta==2,]$variable_respuesta<-0
fit.kz <- zelig(variable_respuesta ~
                  edad+
                  ingpc+
                  sexo+
                  situacion_futura+
                  trabajo_1sem+
                  condicion_actividad+
                  ciudad,
```

```
                data = muestra, model = "relogit", tau = 512/17001 ,
    case.control = "weighting")
summary(fit.kz)

# Residuos
devianza <-sum( residuals ( fit.kz , type ="deviance") ^2)
devianza_p <- 1 - pchisq( devianza , fit.kz$df.null - fit.kz$df.
    residual )

pearson <-sum(residuals(fit.kz , type ="pearson") ^2)
pearson_p <-1 - pchisq( pearson , fit.kz$df.null - fit.kz$df.residual
    )

hoslem.test(muestra$variable_respuesta, fitted(fit.kz),g=10)

rp <- rstandard( fit.kz, type = "pearson" )
rd <- rstandard( fit.kz, type = "deviance" )
png("residuos_sesgo.png", width = 450, height = 300, units = "px",
    pointsize = 12)
par( mfrow = c( 1, 2 ) )
plot( rp,main="Residuos de Pearson",
      ylab="Residuo de Pearson", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
plot(rd,main="Residuos de la devianza",
     ylab="Residuo de la devianza", xlab="X")
abline(h=c(1.96,-1.96),lty=2,col=2)
dev.off()

# Predicciones
predicciones <-predict(fit.kz, muestra )
res <- ifelse(predicciones > 0.5, 1, 0) # Punto de corte en 0.5

# Matriz de confusion
mc <- table(res, muestra$variable_respuesta)
mc[1,1] # Verdaderos positivos
mc[2,2] # Verdaderos negativos
mc[1,2] # Falsos positivos
mc[2,1] # Falsos negativos

# Tasas clasificadas correctamente / Pseudo R cuadrado
(mc[1,1]+mc[2,2])/sum(mc)

# Cura ROC
```

```
dres <- data.frame(pred=predict(fit.kz, muestra, type="response"), var
   =muestra$variable_respuesta)
ROC <- rocit(score=dres[[1]], class=dres$var)
plot(ROC,legend = FALSE)
ksplot(ROC)

# AUC
ROC$AUC

#GINI
ROC$gini

# Prueba de Kolmogorov - Smirnov
lillie.test( res )

# Prueba de Jarque Bera
jb.norm.test(res)

#Prueba Pearson
pearson.test(res)

viflog <-car::vif( fit.kz$zelig() )

# ROC de los 3 Modelos
plot(ROC, col = c(1,"gray50"),
     legend = FALSE, YIndex = FALSE)
lines(ROC_probit$TPR ~ ROC_probit$FPR,
      col = "orange", lwd = 2)
lines(ROC_logit$TPR ~ ROC_logit$FPR,
      col = "blue", lwd = 2)
legend( cex=0.5, col = c("black", "orange", "blue"), "bottomright",
       c("ROC Empirica Logit Corregido", "ROC Empirica Probit", "ROC
   Empirica Logit"), lwd = 2)
```

Listing B.1: Código utilizado