ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE CIENCIAS

CONSTRUCCIÓN DE UN MODELO DE ELECCIÓN BINARIA PARA DETERMINAR LA DEMANDA DE CRÉDITO EN EL ECUADOR DURANTE LA CRISIS POR COVID-19

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO MATEMÁTICO

PROYECTO DE INVESTIGACIÓN

PABLO ALEXANDER MOLINEROS NEGRETE
pablo.molineros@epn.edu.ec

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

QUITO, FEBRERO 2021

DECLARACIÓN

Yo PABLO ALEXANDER MOLINEROS NEGRETE, 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.

Pablo Alexander Molineros Negrete

CERTIFICACIÓN

Certifico que el presente trabajo fue desarrollado por PABLO ALEXANDER MOLINEROS NEGRETE, bajo mi supervisión.

Dra. Adriana Uquillas Andrade
Directora del Proyecto

AGRADECIMIENTOS

Quiero agradecer a mis padres, por su amor incondicional, su comprensión, su sacrificio y trabajo, todo se lo debo a ellos.

A mi hermana Jessie, por su compañía, su apoyo y las incontables alegrías.

A la Dra. Adriana Uquillas, por sus valiosas enseñanzas, su tiempo y paciencia durante la realización de este trabajo.

A mis amigos, por permitirme compartir tantos momentos y experiencias inolvidables a su lado.

Muchas gracias a todos.

DEDICATORIA

A mis padres y hermana, por todo su amor y apoyo incondicional.

Índice general

Índice de figuras

Índice de tablas

Resumen

El impacto súbito ocasionado por la pandemia del coronavirus y las distintas medidas de restricción adoptadas por los gobiernos de todo el mundo para contenerla, han ocasionado una crisis económica y social sin precedente alguno. En Ecuador, según cifras del Ministerio de Trabajo, alrededor de 190 mil trabajadores fueron despedidos en el contexto de la cuarentena afectando directamente los ingresos de los hogares y la posibilidad de la falta de recursos suficientes para satisfacer las necesidades básicas. En este sentido, identificar las características de los trabajadores que más necesitan acogerse a un plan de financiamiento es de suma importancia para la planificación y correcta aplicación de políticas públicas que ayuden a mitigar los efectos adversos de la crisis. Por tal razón, el presente trabajo de investigación especifica un modelo de elección binaria que permite identificar los principales determinantes de la demanda de crédito en Ecuador durante la crisis por COVID-19, en base a los datos recogidos por una encuesta elaborada por los profesores del Departamentos de Economía Cuantitativa y del Departamento de Matemática de la Escuela Politécnica Nacional. Se estimaron dos modelos (logit y probit), encontrando que la estimación logística es preferida para explicar el fenómeno de la demanda. Finalmente, los resultados establecen como factores explicativos de la probabilidad de necesitar un crédito a variables demográficas tales como: edad, género, estado civil y número de miembros en el hogar; además, a variables socioeconómicas y financieras tales como: situación laboral, tipo de vivienda, actividad económica, variación de la renta, variación de los gastos, tener o no ahorros y poseer o no deudas.

Palabras clave: COVID-19, crisis económica, determinantes de la demanda de crédito, modelo logit, modelo probit, Ecuador.

Abstract

The sudden impact caused by the coronavirus pandemic and the various restriction measures adopted by governments around the world to contain it, have caused an economic and social crisis without precedent. In Ecuador, according to figures from the Ministry of Labor around 190 thousand workers were laid off in the context of the quarantine, directly affecting household income and the possibility of the lack of sufficient resources to meet basic needs. In this sense, identifying the characteristics of the workers who most need to benefit from a financing plan is extremely important for the planning and correct application of public policies that help mitigate the adverse effects of the crisis. For this reason, this research work specifies a binary choice model that allows identifying the main determinants of the demand for credit in Ecuador during the COVID-19 crisis, based on the data collected by a survey prepared by the professors of the Department of Quantitative Economics and the Department of Mathematics of the National Polytechnic School. Two models (logit and probit) were estimated, finding that the logistic estimation is preferred to explain the demand phenomenon. Finally, the results establish demographic variables such as: age, gender, marital status, and number of members in the household as explanatory factors of the probability of needing a loan; in addition, to socioeconomic and financial variables such as: employment situation, type of dwelling, economic activity, variation in income, variation in expenses, having or not having savings and having or not having debts.

Keywords: COVID-19, economic crisis, determinants of credit demand, logit model, probit model, Ecuador.

Capítulo 1

Introducción

La enfermedad COVID-19, notificada por primera vez en la ciudad de Wuhan (China) el 31 de diciembre de 2019 [1] y catalogada como pandemia por la Organización Mundial de la Salud [2], es el foco de atención en la agenda política de los gobiernos de todo el mundo debido al fuerte impacto económico y social que ha provocado en los últimos meses [3]. Por ejemplo, en China se evidenció una disminución en el consumo y una interrupción en la producción de fábricas y cadenas de suministros a causa de la cuarentena restrictiva impuesta por el Estado [4]. En consecuencia, el PIB de China que representa el 16 % de la economía global cayó durante el primer trimestre del 2020 [5]. En Estados Unidos, más de 10 millones de personas perdieron sus empleos durante el mes de marzo, una cifra histórica que no se registraba desde la crisis financiera de 2008, debido al cierre de trabajos, restricciones de movilidad, prohibiciones de viajes y cancelaciones de eventos [6]. En América Latina, se prevé una caída del 5.2 % en la actividad económica y el aumento de 3.4 puntos porcentuales en la tasa de desempleo, afectando directamente los ingresos de los hogares y la posibilidad de la falta de recursos suficientes para satisfacer las necesidades básicas [7].

En tal sentido, la formulación de políticas públicas es esencial para ayudar a mitigar los efectos de la crisis causada por la enfermedad. Implementar medidas destinadas al financiamiento de crédito por desempleo, subempleo y autoempleo, créditos con bajo interés a pequeñas y medianas empresas, extensiones del pago de préstamos y flexibilidad financiera en general apuntan a reactivar la economía [8]. De modo que, identificar los perfiles de personas que necesitan acogerse a un plan de apoyo económico es de suma importancia para la planificación y correcta aplicación de políticas públicas. Caso contrario, un limitado conocimiento por parte de

los gobiernos para reconocer la demanda de crédito existente en los sectores más vulnerables de la sociedad podría dar mayor apertura al financiamiento informal, a través de prestamistas locales, amigos y familiares [9]. La fácil obtención de capital a través de préstamos informales produce un exceso de endeudamiento en las personas que lo solicitan, debido a que el crédito obtenido es usado para solventar sus gastos, deudas; que a menudo son producto de una deuda anterior; además, las tasas de interés en este tipo de financiamiento son muy elevadas, con lo cual, el beneficiario del crédito no tiene posibilidades de mejorar su rentabilidad poniendo en riesgo el crecimiento de la economía [10].

Una explicación al fenómeno de precisión de la demanda es ampliamente beneficiada por la utilización de modelos de elección binaria, que permiten analizar los factores determinantes en la probabilidad de que un agente económico individual elija un curso de acción de entre dos posibles opciones [11]. En base a la literatura revisada se puede constatar que la mayor cantidad de estudios se basan en el análisis de la probabilidad de acceder a un crédito estudiando variables cuantitativas y cualitativas de los hogares, tales como características sociodemográficas y socioeconómicas.

Tal es el caso de Carballo, *et al*. [12] que proponen un modelo de elección binaria con función de probabilidad logística para establecer los determinantes de la demanda potencial de microcréditos en Argentina, empleando variables socioeconómicas y demográficas obtenidas por la Encuesta de la Deuda Social de la Pontificia Católica de Argentina (EDSA); encontrando que el tipo de empleo es el principal factor que explica dicha probabilidad. Asimismo, los autores señalan que la edad se relaciona positivamente con la probabilidad de necesitar un microcrédito aunque de modo decreciente cuando la edad es mayor. A su vez, que la propensión a solicitar cualquier clase de financiamiento es independiente del nivel educativo alcanzado por el encuestado y que las mujeres tienen una propensión marginal menor de requerir uno.

En esta misma línea, Vizhñay y Samaniego [13], estudian los principales determinantes de acceso al crédito en el Ecuador haciendo uso de un modelo logit en función de características socioeconómicas de los hogares. Por su parte, señalan que el principal determinante de la probabilidad de acceder a un crédito es el hecho de estar bancarizado es decir, poseer una cuenta de ahorros, corriente o ambas en una institución financiera. Además, variables como el estado civil, empleo y estabilidad laboral aumentan significativamente dicha probabilidad, tal es el caso, de un jefe de hogar casado, con trabajo independiente y contrato permanente. Por el contrario,

variables como la edad, sexo, nivel de educación y ser beneficiario de un plan social tienen un impacto negativo, pues un jefe de hogar mujer con instrucción educativa igual o menor a la secundaria y beneficiario de un plan social por parte del Estado reduce la probabilidad de acceder a un crédito.

En otra línea, Díaz [14], emplea un enfoque microeconómico a través de un modelo de variable dependiente discreta con el fin de establecer los determinantes del acceso al microcrédito para emprendedores bolivianos; para dicho propósito el autor hace uso de la Encuesta de Hogares en Bolivia que recoge características sociodemográficas, económicas y financieras. Entre los resultados a destacar se encuentra que el nivel de ingreso se relaciona positivamente con la probabilidad de obtener un crédito. Asimismo, que el historial crediticio y el género son factores importantes que facilitan el acceso a un crédito, siendo mayor la probabilidad cuando el jefe del hogar es mujer.

En relación a la problemática expuesta, se propone la construcción de un modelo de elección binaria que permita identificar los determinantes que inciden en la necesidad de acceso al crédito y facilitar la aplicación de futuras políticas económicas en el Ecuador durante la crisis provocada por la enfermedad COVID-19. Para esto se analizará la información recogida por la Encuesta para evaluar los efectos de la crisis sanitaria sobre los trabajadores y las organizaciones del sector EPS (Economía Popular y Solidaria), elaborada por los profesores del Departamento de Economía Cuantitativa y del Departamento de Matemática de la Escuela Politécnica Nacional. Dicha encuesta incorpora características socioeconómicas, demográficas y financieras de los trabajadores públicos, privados, autónomos y desempleados en Ecuador.

1.1. Justificación

En Ecuador, debido a los altos niveles de propagación y gravedad causados por el brote de COVID-19, mediante Acuerdo Ministerial No.00126-2020 del 11 de marzo de 2020, la Ministra de Salud declara el estado de Emergencia Sanitaria en el sistema de salud del país [15]. A su vez, el 16 de marzo de 2020, por medio del decreto presidencial No. 1017, se declaró el estado de excepción, restricción de movilidad, suspensión de la jornada laboral y cierre de fronteras en el territorio ecuatoriano [16]. Ante este escenario muchas de las actividades se vieron gravemente afectadas frenando el desarrollo socioeconómico del país. El Banco Interamericano de Desarrollo [17] analizó el impacto que ha tenido la paralización de la económica

en la región andina y el Ecuador figura como uno de los países más afectados por el avance del virus COVID-19; además, menciona que se espera que los sectores económicos más afectados sean: minería, comercio, restaurantes y hoteles, transporte y recreación. Según cifras expuestas por el Ministerio de Trabajo [18], alrededor de 190 mil despidos intempestivos se han aplicado por causa de la emergencia sanitaria; trabajadores informales, cuentapropistas y patrones de pequeñas empresas han dejado de percibir ingresos, pues las cuarentenas fueron obligatorias y estrictas. En consecuencia, el apoyo por parte del gobierno a las pequeñas empresas, población informal y sectores más vulnerables es esencial para hacer frente al estado actual.

Por consiguiente, el caso de investigación propuesto ayudará a proporcionar un instrumento para la planificación económica que permita identificar la población y las características asociadas a la misma, que requieran acceder a una línea de préstamo para hacer frente a sus necesidades financieras durante la crisis por COVID-19; a su vez, permitirá una adecuada asignación de presupuesto a los sectores desfavorecidos, incrementando el empleo, la inversión, la productividad y mitigando los efectos adversos del financiamiento informal.

Por otro lado, una de las ramas de la economía más utilizadas para la caracterización de la demanda de crédito es la Econometría, esta disciplina utiliza herramientas de la Matemática y la Estadística para desarrollar un conjunto de modelos, métodos y técnicas que ayudan a estudiar planteamientos económicos. A su vez, la Microeconometría nace de la necesidad de brindar una interpretación económica del comportamiento de los individuos, así como la posibilidad de contrastar estadísticamente hipótesis efectuadas; además, facilita la tarea de identificar características o factores que provocan un comportamiento diferente de los individuos ante un proceso de decisión [19]. Dentro de esta área, los modelos de elección binaria son empleados para estudiar las elecciones que consumidores, hogares, empresas y otros agentes realizan. Es necesario mencionar que, en los últimos años el uso de estos modelos ha ido tomando gran relevancia en diferentes ramas de investigación, tales como: transporte, salud, educación y economía debido a su capacidad de predicción cuando se trabaja con probabilidades de éxito y fracaso [20]; así como, la facilidad de interpretación de sus coeficientes estimados y la amplia disponibilidad de software para su análisis y cálculo [21].

En tal sentido, utilizar un modelo de elección binaria para el estudio de los determinantes de la demanda de crédito ayudaría a identificar y cuantificar el aporte de las variables que inciden sobre la probabilidad de que una persona, dentro de la población en estudio, necesite o no recurrir a una línea de financiamiento. Si bien

otros estudios [13], [22], han aportado valiosa información sobre la problemática en el país, la investigación se orienta en el contexto de la crisis por la enfermedad COVID-19 ampliando y contribuyendo a la literatura existente en el Ecuador.

1.2. Objetivos

1.2.1. Objetivo General

Especificar un modelo de elección binaria para identificar los determinantes en la probabilidad de requerir un crédito por parte de la población objetivo (trabajadores y personas que conforman el sector de la economía popular y solidaria) durante la crisis sanitaria por COVID-19.

1.2.2. Objetivos Específicos

- Determinar la función de distribución asociada (logística o normal) que genere un mejor modelo a partir de criterios conceptuales y prácticos de modelamiento.
- Estimar y validar el modelo idóneo a través de pruebas estadísticas y realizar un backtesting de ajuste del modelo.
- Describir y analizar en qué medida las características socioeconómicas y demográficas de la población objetivo influyen en la propensión de necesitar un crédito.

Capítulo 2

Marco Teórico

En este capítulo se tratarán los fundamentos teóricos y matemáticos detrás de los modelos de elección binaria. Se empezará revisando los principales aspectos del modelo lineal general, continuando con el desarrollo de los modelos lineales generalizados (MLG), los métodos de estimación y técnicas de validación. Finalmente, se explicarán los modelos de elección binaria como caso particular de los MLG.

2.1. Modelo Lineal General

El modelo lineal general o modelo de regresión multivariante es una de las técnicas básicas del análisis econométrico que permite expresar en forma cuantitativa relaciones de dependencia de tipo lineal entre una variable dependiente o endógena, respecto de una o múltiples variables explicativas o exógenas. Gujarati [23] define al modelo de regresión multivariante como el estudio de la dependencia de una variable (variable dependiente) respecto de una o más variables (variables explicativas) con el objetivo de estimar la media o valor promedio poblacional de la primera en términos de los valores conocidos de las segundas.

La forma del modelo lineal general está dada por [24]:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{2.1}$$

donde Y es la variable dependiente o endógena y $x_1, x_2, \ldots, x_k$ son las variables explicativas o exógenas. El término ε se denomina perturbación aleatoria o error aleatorio, y en él se recogen los posibles efectos que podrían influir en el comportamiento de la variable dependiente, y que no están reflejados en las variables expli-

cativas.

Si se extrae una muestra aleatoria de tamaño $n > k+1$ de una población en la cual la variable dependiente Y se relaciona linealmente con las variables explicativas $x_1, x_2, \ldots, x_k$; cada observación de la muestra puede ser expresada por [24]:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2.2}$$

El valor observado de y_i es la suma de dos componentes, una parte determinista y una parte aleatoria ε_i que se asumen independientes idénticamente distribuidas normales con media cero y varianza constante.

$$\varepsilon_i \sim N(0, \sigma^2), \qquad i = 1, \ldots, n \tag{2.3}$$

Sea $\mathbf{y} = (y_1, y_2, \ldots, y_n)^t$ el vector de las observaciones de la variable dependiente, si consideramos el vector $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^t$ formado por los $k+1$ coeficientes y ordenamos las observaciones de las variables explicativas en una matriz X de orden $n \times (k+1)$

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \tag{2.4}$$

el modelo en (2.1) que se aplica a las n observaciones se puede expresar en notación matricial por [24]:

$$\mathbf{y} = X\beta + \mathbf{e} \tag{2.5}$$

donde $\mathbf{e} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^t$ es el vector de errores.

Con este modelo estamos interesados en estimar los coeficientes del vector β a partir de los datos muestrales.

Una característica distintiva del modelo lineal general, es que la variable dependiente Y es de tipo cuantitativo. Las variables explicativas suelen ser de tipo cuantitativo o cualitativo, y pueden ser determinísticas (valores fijos en muestras repetidas) o estocásticas (valores de un vector aleatorio) [25].

Si suponemos que las variables explicativas son de tipo determinístico, el modelo puede reformularse diciendo que tenemos n observaciones independientes $y_1, \ldots, y_n$ procedentes de distribuciones $N(\mu_i, \sigma^2)$ donde la media μ_i es de la for-

ma [26]:

$$\mu_i = x_i^t \beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \qquad i = 1, \ldots, n \tag{2.6}$$

Por otro lado, si las variables explicativas se consideran estocásticas, se suponen n observaciones (y_i, x_i^t) independientes y, dadas las x_i, las Y_i serán independientes con distribución

$$Y_i | x_i \sim N(\mu_i, \sigma^2), \qquad i = 1, \ldots, n \tag{2.7}$$

donde,

$$\mu_i = E(Y_i | x_i) = x_i^t \beta, \qquad i = 1, \ldots, n \tag{2.8}$$

2.2. Modelos Lineales Generalizados (MLG)

Los modelos lineales generalizados (MLG) estudiados por Nelder y Wedderburn [27] extienden la teoría de los modelos lineales, permitiendo la posibilidad de incorporar variables dependientes continuas o categóricas (ordinales o nominales) con distribuciones del error aleatorio no necesariamente homocedásticos (varianza constante).

En un modelo lineal generalizado suponemos que la distribución de las Y_i, condicionadas por las variables explicativas x_i, no necesariamente es normal, sino pertenece a una familia de tipo exponencial, y posiblemente, con un parámetro de escala. En particular, suponemos que la distribución de las $Y_i | x_i$ tiene por función de densidad una familia de tipo exponencial de la forma [28] :

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{\theta_i \, y_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right] \tag{2.9}$$

en donde θ_i se denomina parámetro canónico o natural, ϕ es el parámetro de escala o dispersión, b y c dos funciones que determinan el tipo de familia exponencial. Cuando ϕ es conocido, la función de densidad pertenece a una familia exponencial y cuando ϕ es desconocido, la función de densidad puede o no pertenecer a una familia exponencial [28].

Muchas distribuciones conocidas pertenecen a dicha familia. Por ejemplo, las distribuciones de Poisson, Normal, Gamma y Binomial se pueden escribir de la forma (2.9) como se muestra en la tabla 2.1.

Además, en un modelo lineal generalizado, la manera en que se relacionan la

Distribución	θ	ϕ	$b(\theta)$	$c(y,\phi)$
Poisson $P(\mu)$	$\ln(\mu)$	1	e^{θ}	$-\ln(y!)$
Normal $N(\mu,\sigma^2)$	μ	σ^2	$\frac{\theta}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2}+\ln\left(2\pi\sigma^2\right)\right)$
Gamma $\Gamma(\mu,v)$	$-\frac{1}{\mu}$	v^{-1}	$\ln(-\theta)$	$v\ln(vy)-\ln(y)-\ln(\Gamma(v))$
Binomial $Bi(n,\pi)$	$\ln\left(\frac{\mu}{n-\mu}\right)$	1	$n\ln\left(1+e^{\theta}\right)$	$\ln\binom{n}{y}$

Tabla 2.1: Distribuciones de Poisson, Normal, Gamma y Binomial como miembros de la familia exponencial. **Fuente:** B. Alegre y G. Cahuana, (2020) [20]. **Elaboración:** Autor.

variable dependiente con las variables explicativas no necesariamente es lineal, sino que lo hacen mediante una función g monótona y diferenciable, denominada función de enlace o función *link*, de la forma [28]:

$$g(\mu_i) = x_i^t\beta \qquad i = 1,\ldots,n \tag{2.10}$$

donde,

$$\mu_i = E(Y_i|x_i) \tag{2.11}$$

Como resultado, el modelo quedará totalmente especificado cuando se fije el tipo de familia exponencial para las distribuciones condicionadas $Y_i|x_i$, la función de enlace y la matriz de variables explicativas.

En las distribuciones de las $Y_i|x_i$ se supone que el parámetro canónico es una función de la media; es decir $\theta_i = w(\mu_i)$ siendo [28]:

$$\mu_i = b^{'}(\theta_i) \tag{2.12}$$

y

$$Var(Y_i|x_i) = \phi b^{''}(\theta_i) \tag{2.13}$$

donde $b^{'}$ y $b^{''}$ son las dos primeras derivadas de b. Si $b^{'}$ es una función uno a uno (estrictamente creciente o decreciente) entonces θ_i se puede determinar de forma única a partir de $E(Y_i|x_i) = \mu_i$ y se sigue que [28]:

$$Var(Y_i|x_i) = \phi V(\mu_i) \tag{2.14}$$

donde V es la función de varianza de Y_i que multiplica al parámetro de dispersión.

La función de varianza es fundamental para evaluar el ajuste de los modelos y establecer estimaciones apropiadas [25].

Para cada familia exponencial existe una función de enlace (*link*) natural que es la que iguala al parámetro canónico con el *predictor lineal* $x_i^t\beta$ de la forma [28]:

$$\theta_i = g(\mu_i) = x_i^t\beta \tag{2.15}$$

Tomando el ejemplo de Cabrero y García [26], si las variables $Y_i|x_i$ siguen una distribución binomial $Bi(n_i, p_i)$, su función de densidad será

$$\begin{aligned} f(y_i;\theta_i,\phi) &= \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i} \\ &= \binom{n_i}{y_i} \exp\left[y_i \ \ln\left(\frac{p_i}{1-p_i}\right) + n_i \ \ln(1-p_i)\right] \end{aligned} \tag{2.16}$$

Si comparamos con (2.9), podemos identificar

$$\theta_i = \ln\left(\frac{p_i}{1-p_i}\right), \qquad b(\theta_i) = -n_i \ \ln(1-p_i) \qquad \text{y} \qquad \phi = 1 \tag{2.17}$$

Dado que la media de la distribución binomial $Bi(n_i, p_i)$ es $\mu_i = n_i p_i$, obtenemos

$$\theta_i = \ln\left(\frac{\mu_i}{n_i - \mu_i}\right) \tag{2.18}$$

Luego, la función *link* canónica es $g(\mu) = \ln(\mu/(n-\mu))$.

Finalmente, la ecuación que relaciona la media de la variable dependiente con las variables explicativas $g(\mu_i) = x_i^t\beta$, será

$$\ln\left(\frac{\mu_i}{n_i-\mu_i}\right) = \ln\left(\frac{n_i p_i}{n_i - n_i p_i}\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.19}$$

Por otro lado, si consideramos el caso de que las variables dependientes siguen una distribución Bernoulli

$$Y_i|x_i \sim Bi(1, p_i) \tag{2.20}$$

tendremos un caso particular del ejemplo anterior, en donde la función *link* canónica será $g(\mu) = \ln(\mu/(1-\mu))$ o lo que es lo mismo $g(p) = \ln(p/(1-p))$ dado que $\mu = p$.

Por lo tanto, la ecuación que relaciona la media de la variable dependiente con

las variables explicativas, en este caso, es la misma de antes

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{2.21}$$

por lo que no se suele hacer distinción entre estos dos casos y se habla de la función *link* canónica denominada *logit* [28]:

$$g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) \tag{2.22}$$

En la tabla 2.2 se presentan las funciones de enlace (*link*) y funciones de varianza para algunas distribuciones utilizadas en los MLG.

Familia de Distribuciones	Función *link*	Función de Varianza
Normal	μ	σ^2
Poisson	$\ln(\mu)$	μ
Binomial	$\ln\left(\frac{\mu}{1-\mu}\right)$	$\frac{\mu(1-\mu)}{n}$
Gamma	$-\frac{1}{\mu}$	μ^2

Tabla 2.2: Funciones *link* y funciones de varianza utilizadas por los MLG. **Fuente**: A. Agresti, (2007) [29]. **Elaboración:** Autor.

En el caso Binomial, las funciones de enlace (*link*) más comunes son [29]:

- Logit: $\ln(\mu/(1-\mu))$.
- Probit: $\Phi^{-1}(\mu)$, donde Φ^{-1} es la inversa de la distribución acumulada de la distribución normal estándar.
- Complemento log-log: $\ln(-\ln(1-\mu))$.

2.3. Estimación de los Modelos Lineales Generalizados

Existen diferentes procedimientos para estimar los coeficientes desconocidos de un modelo lineal generalizado a partir de los datos muestrales. De entre esos procedimientos el más empleado es el método de máxima verosimilitud, el cual tiene propiedades de consistencia y eficiencia asintótica [30]. A continuación, se revisan los aspectos más importantes del método de estimación.

2.3.1. Estimador de Máxima Verosimilitud

Si suponemos que las variables $Y_i|x_i$ son independientes e idénticamente distribuidas con distribución perteneciente a la familia exponencial (2.9), y parámetro ϕ conocido, la función de verosimilitud está dada por [26]:

$$\mathcal{L}(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} f(y_i; \theta_i) = \exp\left[\sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right)\right] \tag{2.23}$$

La idea fundamental de este método es tomar como estimación de los parámetros a aquellos valores que maximicen la función de verosimilitud. Dado que el máximo de una función y de su logaritmo se alcanzan en el mismo punto, se determina el máximo del logaritmo de la función de verosimilitud (log-verosimilitud) [26]:

$$\ln \mathcal{L}(\theta_1, \ldots, \theta_n) = \sum_{i=1}^{n}\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right) \tag{2.24}$$

Como ϕ se supone conocido y $\theta_i = w(\mu_i)$, se puede expresar (2.24) de la forma [26]:

$$\begin{aligned}\ln \mathcal{L}(\mu_1, \ldots, \mu_n) &= \sum_{i=1}^{n}\left(\frac{y_i w(\mu_i) - b(w(\mu_i))}{\phi}\right) + \sum_{i=1}^{n} c(y_i, \phi) \\ &= \sum_{i=1}^{n} l_i(\mu_i) + \sum_{i=1}^{n} c(y_i, \phi)\end{aligned} \tag{2.25}$$

Luego, en términos de los β_i y las variables explicativas [26]:

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n}\left(\frac{y_i w(g^{-1}(x_i^t\beta)) - b(w(g^{-1}(x_i^t\beta)))}{\phi}\right) + \sum_{i=1}^{n} c(y_i, \phi) \tag{2.26}$$

Se maximiza la función de log-verosimilitud derivando respecto al parámetro desconocido e igualando a cero. Cabe recalcar que β es un vector y al hablar de la derivada de (2.26) con respecto a β, se refiere al vector de derivadas parciales

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta} = \left(\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_0}, \frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_1}, \ldots, \frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta_k}\right)^t \tag{2.27}$$

el cual se iguala al vector de ceros, dando origen al sistema de ecuaciones de verosimilitud, de $k+1$ ecuaciones con $k+1$ incógnitas $\beta_0, \beta_1, \ldots, \beta_k$.

Derivando (2.26) se obtiene [26]:

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta} = \frac{1}{\phi} \sum_{i=1}^{n} \left(\frac{\partial w(\mu_i)}{\partial \mu_i} \Big|_{\mu_i = g^{-1}(x_i^t \beta)} \right) \mu_i^{'} \left(y_i - \mu_i(\beta) \right) \tag{2.28}$$

Dado que $\mu_i = b^{'}(\theta_i)$ y $\theta_i = w(\mu_i)$ se sigue que:

$$w(\mu_i) = (b^{'})^{-1}(\mu_i) \tag{2.29}$$

Luego, utilizando la fórmula para la derivada inversa:

$$\frac{\partial w(\mu_i)}{\partial \mu_i} = \frac{\partial (b^{'})^{-1}(\mu_i)}{\partial \mu_i} = \frac{1}{b^{''}((b^{'})^{-1}(\mu_i))} = \frac{1}{b^{''}(\theta_i)} = \frac{\phi}{Var(Y_i|x_i)} \tag{2.30}$$

Finalmente, el sistema de ecuaciones de verosimilitud queda determinado por [26]:

$$\frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta} = \sum_{i=1}^{n} \frac{\mu_i^{'}}{Var(Y_i|x_i)} (y_i - \mu_i) = \mathbf{0} \tag{2.31}$$

que normalmente no tiene una solución analítica y se resuelve de forma numérica mediante un método iterativo. Software como RStudio[1] utiliza el método de mínimos cuadrados ponderados IWLS (Iteratively Reweighted Least Squares), también denominado Fisher Scoring. Otras alternativas son el método de Newton-Rhapson y los métodos Quasi-Newton [28].

El estimador de máxima verosimilitud $\hat{\beta}$ obtenido a partir de alguno de los métodos antes mencionados, cuando exista y sea único, tendrá una distribución asintótica normal multivariante [26]

$$\hat{\beta} \sim N(\beta, W) \tag{2.32}$$

siendo la matriz de covarianzas W aproximadamente igual a la inversa de la matriz de información de Fisher [31]:

$$W \approx I^{-1}(\hat{\beta}) \tag{2.33}$$

donde,

$$I(\hat{\beta}) = -E \left[\frac{\partial^2 \ln \mathcal{L}(\hat{\beta})}{\partial \hat{\beta}\, \partial \hat{\beta}^t} \right] \tag{2.34}$$

[1] RStudio es un entorno de desarrollo integrado para el lenguaje de programación R, dedicado a la computación estadística y gráficos.

2.4. Validación de los Modelos Lineales Generalizados

El objetivo del proceso de validación es determinar hasta que punto el modelo explica correctamente los datos muestrales. Para esto, es necesario realizar pruebas de significancia sobre los parámetros estimados, analizar la bondad de ajuste y verificar los supuestos acerca del modelo ajustado.

2.4.1. Pruebas de significancia sobre los parámetros

Una vez obtenidos los estimadores para los parámetros β_i, se realizan test de hipótesis sobre ellos para verificar su significancia. Las hipótesis a probar son de la forma $H_0 : C\beta = c_o$ frente a la alternativa $H_1 : C\beta \neq c_0$, donde C es una matriz de coeficientes de orden $r \times (k+1)$ y c_0 es un vector de coeficientes de orden $r \times 1$. Un caso particular de estas hipótesis, es el contraste [26]:

$$H_0 : \beta_m = \mathbf{0} \quad \text{frente a} \quad H_1 : \beta_m \neq \mathbf{0} \tag{2.35}$$

donde β_m es un vector de dimensión menor o igual a β, es decir, se quiere probar el contraste de ser cero algunas β_i, frente a la alternativa de que todas las β_i son distintas de cero. Se consideran tres tipos de test basados fundamentalmente en la teoría de máxima verosimilitud:

1. Test de Razón de Verosimilitudes.
2. Test de Wald.
3. Test de Score.

Test de Razón de Verosimilitudes

La idea del método se centra en que, para una muestra fija, su verosimilitud $\mathcal{L}$ es una medida de lo bien que "explica" el valor de β los resultados obtenidos. Por tanto, $\sup_{\beta\in\Theta_0} \mathcal{L}(\beta)$ supone un índice acerca de la "mejor explicación" de la muestra que puede obtenerse bajo la hipótesis nula; mientras que $\sup_{\beta\in\Theta} \mathcal{L}(\beta)$ proporciona tal índice entre todos los valores posibles del parámetro [32].

El test de contraste está definido por el estadístico

$$\Lambda = \frac{\sup_{\beta\in\Theta_0} \mathcal{L}(\beta)}{\sup_{\beta\in\Theta} \mathcal{L}(\beta)} = \frac{\mathcal{L}(\tilde{\beta})}{\mathcal{L}(\hat{\beta})} \tag{2.36}$$

donde, $\Theta \subset \mathbb{R}^k$ es el espacio paramétrico y Θ_0 la parte de este espacio definido por la hipótesis nula.

Como mencionan Vélez y García [32], el hecho que $\mathcal{L}(\tilde{\beta})$ sea mucho más bajo que $\mathcal{L}(\hat{\beta})$ significa que la explicación de los resultados, sobre la base de que H_0 es cierta, es mucho peor que la explicación sin tal restricción; parece adecuado entonces rechazar que la distribución de la población verifique H_0. Caso contrario, cuando $\mathcal{L}(\tilde{\beta})$ es próximo a $\mathcal{L}(\hat{\beta})$, no hay razones suficientes para descartar que H_0 sea cierta, puesto que ello no mejoraría sensiblemente la verosimilitud de las observaciones.

La distribución del estadístico de contraste para tamaños de muestra suficientemente grandes se distribuye [32]

$$-2\ln \Lambda = 2\ln \mathcal{L}(\hat{\beta}) - 2\ln \mathcal{L}(\tilde{\beta}) \sim \chi^2_{k+1-q} \tag{2.37}$$

donde, q es la dimensión del espacio paramétrico bajo la hipótesis nula.

Test de Wald

El test de Wald es un método alternativo para probar la significancia de los parámetros y está basado en el estadístico de contraste [26]:

$$Wald = (C\hat{\beta} - c_0)^t \left[C I^{-1}(\hat{\beta}) C^t \right]^{-1} (C\hat{\beta} - c_0) \tag{2.38}$$

donde, $I^{-1}(\hat{\beta})$ es la inversa de la matriz de información de Fisher definida anteriormente.

El estadístico de contraste, bajo la hipótesis nula, se distribuye asintóticamente χ^2_{k+1-q}.

Test de Score

Se define el estadístico de contraste [26]:

$$S = s(\tilde{\beta})^t I^{-1}(\tilde{\beta}) s(\tilde{\beta}) \tag{2.39}$$

donde,

$$s(\beta) = \frac{\partial \ln \mathcal{L}(\beta)}{\partial \beta} \tag{2.40}$$

es la función de score.

Asintóticamente y bajo la hipótesis nula el estadístico de contraste se distribuye como una χ^2_{k+1-q}.

Para muestras grandes, los test de razón de verosimilitudes y Wald muestran resultados similares, sin embargo, para muestras pequeñas pueden diferir. No obstante, la razón de verosimilitudes es preferida por ser un test uniformemente más potente [29].

2.4.2. Bondad de ajuste del modelo

En el proceso de ajuste del modelo se debe evaluar la diferencia entre los datos observados y esperados. Según McCullagh y Nelder [33], el proceso de ajustar un modelo a los datos puede considerarse como una forma de reemplazar el conjunto de observaciones por un conjunto de valores ajustados derivados de un modelo que involucra un número pequeño de parámetros.

Una diferencia pequeña entre los valores observados y los ajustados se puede admitir, en cuanto una diferencia significativa no. En este sentido, lo que se busca es determinar cuántos términos son necesarios en la estructura del modelo para una descripción correcta de los datos. Como menciona Figueroa [34], un número grande de variables explicativas pueden llevar a que un modelo explique bien los datos pero con un aumento de complejidad en su interpretación. Por otro lado, un número pequeño de variables explicativas puede llevar a un modelo de fácil interpretación pero que no se ajuste a los datos.

Dos de los estadísticos más utilizados para contrastar la hipótesis nula de que los datos se adecuan correctamente a un modelo, son el estadístico de Pearson [26]:

$$\lambda = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var(Y_i|x_i)}} \tag{2.41}$$

donde, $\hat{\mu}_i = g^{-1}(x_i^t \hat{\beta})$ es la media estimada y $\widehat{Var(Y_i|x_i)}$ la varianza estimada, y el estadístico de la Devianza (Residual Deviance o Deviance) [26]:

$$D = -2 \sum_{i=1}^{n} [l_i(\hat{\mu}_i) - l_i(y_i)] = \sum_{i=1}^{n} d_i \tag{2.42}$$

donde de manera análoga, $\hat{\mu}_i$ es la media estimada y l_i son las contribuciones de cada uno de los valores muestrales al logaritmo de verosimilitud, definidas en (2.25).

Bajo ciertas condiciones de regularidad la distribución de ambos estadísticos se puede aproximar a una $\chi^2_{n-(k+1)}$. Cuando el número de observaciones es suficientemente grande, los estadísticos D y λ son equivalentes [35].

2.4.3. Adecuación del modelo

Según Figueroa [34], la verificación de la adecuación del modelo es un requisito fundamental que se debe realizar sobre el conjunto de datos para analizar el incumplimiento de los supuestos hechos para el modelo, así como la existencia de observaciones extremas con alguna interferencia desproporcionada en los resultados del ajuste.

Los residuos o residuales son utilizados para verificar la adecuación del modelo, debido a que expresan la diferencia entre una observación y su valor ajustado; además, indican la presencia de valores atípicos que puedan requerir un estudio más detallado. Así, el modelo ajustado debe cumplir que los n residuos de Pearson [26]:

$$r_i^p = \frac{(y_i - \hat{\mu}_i)^2}{\widehat{Var(Y_i|x_i)}} \tag{2.43}$$

deben tener, aproximadamente, media cero y varianza ϕ.

En la práctica, la distribución de los residuos de Pearson suelen ser asimétricos para modelos no normales. Por esta razón, se determinan los n residuos de la devianza, que son los n sumandos d_i del estadístico de la Devianza [26]:

$$r_i^d = signo(y_i - \hat{\mu}_i)\sqrt{d_i} \tag{2.44}$$

para los cuales se admite, bajo el supuesto de que el modelo es adecuado, una distribución $N(0,1)$.

2.5. Modelos de Elección Binaria

En esta sección consideraremos los MLG donde la variable dependiente se mide en una escala binaria. Por ejemplo, la variable puede ser del tipo: si o no, activo o no activo, presente o ausente, necesita crédito o no, etc. En general, se utilizan los términos *éxito* y *fracaso* para cualquiera de las categorías.

Se define la variable aleatoria [35]

$$Z = \begin{cases} 1 & \text{si el resultado es un éxito} \\ 0 & \text{si el resultado es un fracaso} \end{cases} \tag{2.45}$$

con probabilidades $\Pr(Z = 1) = \pi$ y $\Pr(Z = 0) = 1 - \pi$, que es la distribución de Bernoulli $B(\pi)$.

Ahora, si se consideran n variables aleatorias independientes $Z_1, Z_2, \ldots, Z_n$ con $\Pr(Z_j = 1) = \pi_j$, entonces la función de densidad conjunta está dada por [35]:

$$\prod_{j=1}^{n} \pi_j^{z_j}(1-\pi_j)^{1-z_j} = exp\left[\sum_{j=1}^{n} z_j \ln\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \ln(1-\pi_j)\right] \tag{2.46}$$

que pertenece a la familia exponencial (2.9).

Asimismo, para el caso en que todas las π_j's son iguales, se define la variable

$$Y = \sum_{j=1}^{n} Z_j \tag{2.47}$$

de modo que Y es el número de éxitos en n intentos. La variable aleatoria Y tiene una distribución $Bi(n, \pi)$

$$\Pr(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, \ldots, n. \tag{2.48}$$

Además, si se tienen N variables aleatorias independientes $Y_1, Y_2, \ldots, Y_N$ con $Y_i \sim Bi(n_i, \pi_i)$ correspondientes al número de éxitos en N subgrupos como se muestra en la tabla 2.3. La función de log-verosimilitud está dada por [35]:

$$l(y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \ln(1-\pi_i) + \ln\binom{n_i}{y_i} \right] \tag{2.49}$$

	Subgrupos			
	1	2	$\cdots$	N
Éxitos	Y_1	Y_2	$\cdots$	Y_N
Fracasos	$n_1 - Y_1$	$n_2 - Y_2$	$\cdots$	$n_N - Y_n$
Total	n_1	n_2	$\cdots$	n_N

Tabla 2.3: Frecuencias para N distribuciones Binomiales. **Fuente:** A. J. Dobson y A. G. Barnett, (2008) [35]. **Elaboración:** Autor.

El objetivo es describir la proporción de éxitos $P_i = Y_i/n_i$, en cada subgrupo en función de las variables explicativas que caracterizan dicho subgrupo. Dado que $E(Y_i) = n_i\pi_i$, se tiene $E(P_i) = \pi_i$, y se considera el modelo [35]:

$$g(\pi_i) = x_i^t\beta \tag{2.50}$$

donde, x_i es el vector de variables explicativas, β es el vector de parámetros y g es la función de enlace (*link*). Para el caso de observaciones individuales (no repetidas) se considera $n_i = 1$.

El caso más simple es cuando se tiene un modelo lineal general, que para variables binarias se lo conoce como modelo de probabilidad lineal [26]:

$$\pi = x_i^t\beta \tag{2.51}$$

donde, $\pi = \Pr(Y = 1|x_i)$ y la función de enlace g es la función identidad.

Este modelo es utilizado en varias aplicaciones prácticas debido a la facilidad de estimar los parámetros e interpretar los resultados, por otro lado, Wooldridge [36] sostiene que el modelo pone en manifiesto algunas desventajas. El principal inconveniente, es que a pesar de ser π una probabilidad, los valores ajustados de $x_i^t\beta$ pueden ser menores que cero o mayores que uno.

Para asegurarse de que la probabilidad π esté restringida al intervalo $[0, 1]$ se utiliza una función de distribución de probabilidad [35]:

$$\pi = \int_{-\infty}^{t} f(s)ds \tag{2.52}$$

donde, $f(s) \geq 0$ y $\int_{-\infty}^{\infty} f(s)ds = 1$.

Por razones tanto históricas como prácticas, las funciones de distribución de probabilidad que suelen seleccionarse para representar los modelos de elección binaria son la logística y la normal; la primera da lugar a los modelos logit y la última, a los modelos probit [23].

2.5.1. Modelos Logit

Los modelos logit estudiados por Luce [37] y popularizados por McFadden [38], son herramientas que permiten explicar los efectos de las variables explicativas sobre la probabilidad de éxito. Como expresa Train [39], logit es el modelo de elección binaria más simple y de uso más extendido, debido a que su fórmula para las pro-

babilidades de elección tiene una expresión cerrada y es fácilmente interpretable.

El modelo logit de variable dependiente binaria Y con múltiples variables explicativas está dado por [40]:

$$\begin{aligned} \Pr(Y = 1|x_i) = \pi &= \Lambda(x_i^t\beta) \\ &= \frac{\exp(x_i^t\beta)}{1+\exp(x_i^t\beta)} \qquad i = 1, \ldots, k. \end{aligned} \tag{2.53}$$

al despejar la probabilidad π de (2.53) se tiene

$$\ln\left(\frac{\pi}{1-\pi}\right) = x_i^t\beta \tag{2.54}$$

de donde se deduce que la función de enlace (*link*) es la función logit, y de aquí el nombre del modelo.

Los coeficientes β_i en la ecuación (2.54) determinan la tasa de aumento o disminución de la curva logit. Cuando $\beta_i > 0$, π aumenta a medida que aumentan los valores de las variables explicativas x_i, como en la figura 2.1(a). Cuando $\beta_i < 0$, π disminuye a medida que aumentan las x_i, como en la figura 2.1(b). La magnitud de β_i determina que tan rápido aumenta o disminuye la curva.

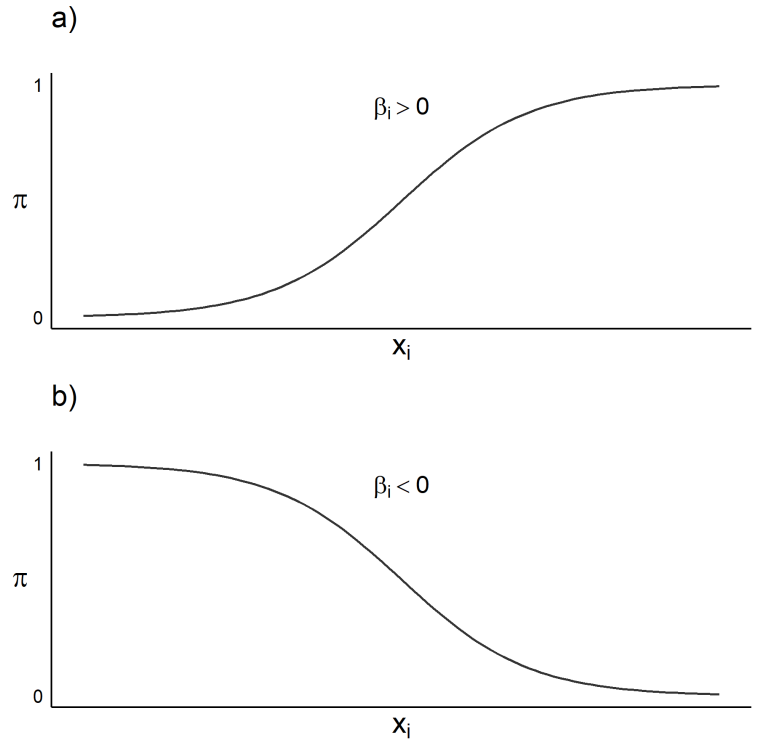


Figura 2.1: Funciones de regresión logística. **Elaboración:** Autor.

Los coeficientes del modelo logit se estiman por el método de máxima verosimi-

litud, donde la función de log-verosimilitud es [35]:

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{N} \left[y_i \ln \Lambda(x_i^t \beta) + (1 - y_i) \ln(1 - \Lambda(x_i^t \beta)) \right] \tag{2.55}$$

El estimador de máxima verosimilitud es consistente y está distribuido normalmente en muestras grandes, por lo que los estadísticos *t-student*[2] y los intervalos de confianza de los coeficientes pueden construirse como en el modelo de regresión clásico [40].

Potencia y limitaciones de logit

La aplicabilidad de los modelos logit se puede resumir a [41]:

1. Simplicidad: La función de distribución de probabilidad logística, es más sencilla al evaluar y permite obtener resultados más eficientes, respecto a otras funciones de distribución.

2. Interpretabilidad: La relación lineal del modelo permite interpretar de manera sencilla los coeficientes, debido a que se la realiza como un ratio de probabilidad.

Los posibles limitantes para la aplicación del modelo logit suelen ser los tamaños de muestra pequeños, que proporcionan estimadores imprecisos [11].

2.5.2. Modelos Probit

Al igual que los modelos logit, el modelo probit relaciona la variable dependiente con las variables explicativas a través de una función de distribución de probabilidad. Los modelos probit para datos binarios fueron presentados por Thurstone [42] bajo el supuesto de que las perturbaciones aleatorias siguen una distribución normal conjunta. Cabe destacar, que según Hann y Soyer [43], el análisis probit es más apropiado para el diseño de experimentos, en función a que su procedimiento permite medir la relación entre la potencia de un estímulo y la proporción de casos que presentan una repuesta a este.

[2]El estadístico *t* permite analizar si el valor observado es lo suficientemente cercano al valor hipotético, como para rechazar o no la hipótesis planteada [31].

El modelo probit de variable dependiente binaria Y con múltiples variables explicativas está dado por [40]:

$$\begin{aligned} \Pr(Y = 1|x_i) = \pi &= \Phi(x_i^t \beta) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i^t \beta} e^{-\frac{s^2}{2}} ds \end{aligned} \tag{2.56}$$

Luego,

$$\Phi^{-1}(\pi) = x_i^t \beta \tag{2.57}$$

donde, la función de enlace (*link*) es la función probit.

De la misma manera, los coeficientes se calculan mediante el método de máxima verosimilitud, que da lugar a estimadores eficientes (con varianza mínima). Para muestras grandes, el estimador de máxima verosimilitud es consistente y se distribuye normalmente [40].

La función de log-verosimilitud para el modelo probit está dada por [31]:

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{N} \left[y_i \ln \Phi(x_i^t \beta) + (1 - y_i) \ln(1 - \Phi(x_i^t \beta)) \right] \tag{2.58}$$

Potencia y limitaciones de probit

Las principales limitaciones al aplicar un modelo probit suelen ser:

1. Complejidad en el cálculo de la función de distribución de probabilidad normal, ya que debe obtenerse mediante una integral, dificultando la manipulación algébrica de los parámetros y su interpretación [44].

2. Inconsistencia en los estimadores de máxima verosimilitud para tamaños de muestra pequeños [43].

Por otro lado, si se cuenta con un gran número de observaciones, el supuesto de distribución normal de los errores podría primar como criterio para la elección del modelo [11].

2.5.3. Modelos Logit y Probit

Medina [44] destaca que la diferencia entre los modelos logit y probit es fundamentalmente operativa, ya que la forma de las curvas es muy similar. Discrepan,

únicamente, en la rapidez con que las curvas se aproximan a los valores extremos, como se ilustra en la figura 2.2. Así, la función logística es más achatada que la normal al alcanzar, esta última, más rápidamente los valores extremos 0 y 1.

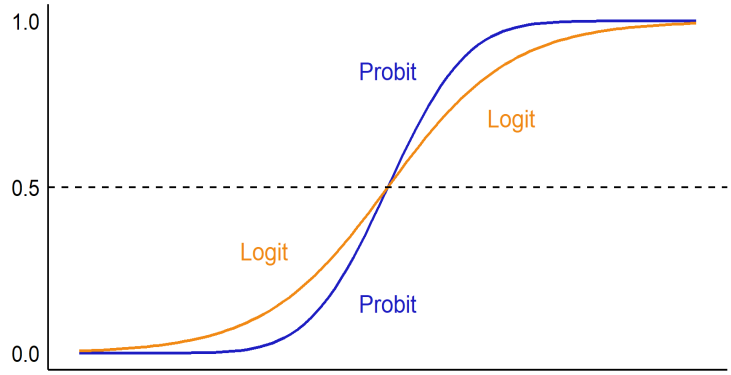


Figura 2.2: Comparación de las funciones de distribución entre el modelo logit y probit. **Elaboración:** Autor.

Lo anterior significa que la probabilidad condicional π se aproxima a 0 o a 1 con una tasa menor en el modelo logit, en comparación con el probit.

Con relación a las estimaciones que proporcionan los dos modelos, difieren cuando la muestra contiene pocos valores de y_i iguales a una de las dos alternativas posibles. Sin embargo, ésta no es una situación habitual, obteniéndose en general estimaciones muy similares entre los dos modelos [44]. Gujarati [23] sostiene que en la práctica, muchos investigadores eligen el modelo logit debido a su comparativa simplicidad matemática.

2.5.4. Odds Ratio

Para una probabilidad de éxito π, el *odds* de éxito se define como [29]:

$$\begin{aligned} odds &= \frac{\pi_i}{1-\pi_i} \\ &= \frac{\Pr(y_i = 1|x_i)}{1-\Pr(y_i = 1|x_i)} \end{aligned} \tag{2.59}$$

Si el $odds > 1$, es más probable que se obtenga un éxito que un fracaso, en otras palabras, para el individuo i, la opción $y_i = 1$ es más probable (mayor beneficio) que la opción $y_i = 0$. Si el $odds < 1$, la interpretación sería la contraria. Y si el $odds = 1$, ambas opciones son igual de probables, es decir, el individuo es indiferente ante ambas opciones [45].

De la misma forma, si se obtiene un $odds = 4$, un éxito es cuatro veces más probable que un fracaso; mientras un $odds = 1/4$, significa que un fracaso es cuatro veces más probable que un éxito.

Los *odds* para los modelos logit y probit son respectivamente:

$$\begin{aligned} odds &= \frac{\Lambda(x_i^t\beta)}{1-\Lambda(x_i^t\beta)} \\ &= \exp(x_i^t\beta) \end{aligned} \tag{2.60}$$

y

$$odds = \frac{\Phi(x_i^t\beta)}{1-\Phi(x_i^t\beta)} \tag{2.61}$$

En muchas ocasiones, puede ser ilustrativo comparar el *odds* entre dos situaciones diferentes. Por ejemplo, si se toma de referencia a un individuo j en el cual se fijan todas las variables contenidas en x_i en su valor medio y se compara su *odds* con el de otro individuo i que difiere de j en el valor de una o más variables explicativas. Se construye así el *odds-ratio* de estos individuos como [29]:

$$odds\text{-}ratio = \frac{odds_i}{odds_j} = \frac{\dfrac{\pi_i}{1-\pi_i}}{\dfrac{\pi_j}{1-\pi_j}} \tag{2.62}$$

Cuando $odds\text{-}ratio > 1$, las probabilidades de éxito para el individuo i son mayores que para el individuo j, es decir, el individuo i tiene una preferencia mayor por la opción $y_i = 1$ frente a la $y_i = 0$ que el individuo j. Siguiendo el mismo razonamiento, se deduce la interpretación para valores de *odds-ratio* menores o iguales a uno. Medina [44] destaca que el cálculo de los *odds-ratio* facilita la interpretación de los coeficientes estimados cuando se aplica al caso concreto de calcular la variación en la preferencia de un individuo frente a una alternativa en concreto.

2.5.5. Validación y contraste de hipótesis

Si bien, los valores estimados difieren dada una especificación del modelo (logit o probit), el desarrollo teórico de los procesos de validación y contrastes de hipótesis son los mismos, utilizando en cada caso la función de distribución correspondiente.

A continuación se describen los test de significancia de los parámetros, contrastes de bondad de ajuste, pruebas de multicolinealidad y medidas de influencia.

Significancia estadística de los parámetros estimados

La distribución de los parámetros estimados por el método de máxima verosimilitud para una muestra de tamaño n es aproximadamente [44]

$$\sqrt{n}(\beta - \hat{\beta}) \sim N(0, I^{-1}(\hat{\beta})) \tag{2.63}$$

Existen varias formas de estimar la inversa de la matriz de información. El software RStudio calcula directamente la matriz de varianzas y covarianzas de $\hat{\beta}$, así como los errores estándar.

En tal situación, se aplican los test basados en la teoría de máxima verosimilitud vistos anteriormente. Sin embargo, dada su disponibilidad en los distintos paquetes econométricos, el test de Wald es el más utilizado para contrastar la hipótesis de nulidad de los parámetros.

Para un coeficiente cualquiera β_i, se verifica que bajo la hipótesis nula $H_0 : \beta_i = 0$, el estadístico w de Wald es [44]:

$$w = \frac{\widehat{\beta_i}^2}{\widehat{Var(\widehat{\beta_i})}} \sim \chi_1 \tag{2.64}$$

En RStudio el contraste está expresado aproximadamente a una Z[3] con su correspondiente p-valor.

Para contrastar la hipótesis conjunta de que todos los coeficientes del modelo son significativamente próximos a cero

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{2.65}$$

se utiliza el estadístico de razón de verosimilitudes. Para ello, se compara el valor de la función de verosimilitud, de un modelo cuya única variable explicativa sea la constante, con el valor de la función de verosimilitud del modelo estimado incluyendo todas las variables explicativas.

Si denotamos a la función de verosimilitud, del modelo cuya única variable explicativa es la constante, por $\mathcal{L}(0)$, y al modelo estimado con todas las variables explicativas por $\mathcal{L}(\hat{\beta})$, el estadístico de razón de verosimilitudes se define como [36]:

$$RV = 2\left[\ln \mathcal{L}(\hat{\beta}) - \ln \mathcal{L}(0)\right] \tag{2.66}$$

[3]La variable Z está distribuida normalmente con media igual a cero y varianza igual a uno.

El estadístico RV se distribuye asintóticamente, bajo la hipótesis nula, como una ji-cuadrada con $n-(k+1)$ grados de libertad.

Medidas de bondad de ajuste del modelo

A continuación se describen los contrastes más utilizados en la literatura econométrica para medir la bondad de ajuste global del modelo.

1. Estadístico de la Devianza

El estadístico de la Devianza D se define como una función del logaritmo de la función de verosimilitud del modelo seleccionado y la del modelo saturado. El modelo saturado es aquel que se ajusta perfectamente a los datos, es decir, los valores predichos por el modelo coinciden con los valores observados, y tiene tantos parámetros desconocidos como observaciones diferentes de las variables explicativas [46].

La hipótesis nula que contrasta el estadístico de la Devianza es, que el modelo seleccionado estima perfectamente los datos observados. La Devianza tiene la siguiente expresión [47]:

$$D = -2 \ln \left[\frac{\widehat{\mathcal{L}}_C}{\widehat{\mathcal{L}}_F} \right] \tag{2.67}$$

donde, $\widehat{\mathcal{L}}_C$ es el valor de la función de verosimilitud estimada del modelo seleccionado y $\widehat{\mathcal{L}}_F$ es el valor correspondiente a la función de verosimilitud estimada del modelo saturado.

El estadístico así construido tiene distribución asintótica ji-cuadrada con grados de libertad igual a la diferencia entre la dimensión del espacio paramétrico y la dimensión del espacio bajo la hipótesis nula [44]. El test coincide con el test de razón de verosimilitudes para contrastar la significancia de los parámetros.

2. Índice de cociente de verosimilitudes

De acuerdo con Medina [44], la función de verosimilitud también puede utilizarse para obtener un estadístico, similar al coeficiente de determinación R^2 en regresión lineal, conocido como "pseudo-R^2 o R^2 de McFadden". Este estadístico compara la función de verosimilitud de dos modelos: uno correspondiente al modelo estimado que incluye todas las variables explicativas y otro sería el modelo

cuya única variable explicativa es la constante. El estadístico se define como [35]:

$$\text{pseudo-}R^2 = 1 - \frac{\ln \mathcal{L}(\widehat{\beta})}{\ln \mathcal{L}(0)} \tag{2.68}$$

El estadístico toma valores en el intervalo de 0 a 1 de forma que [44]:

- Valores próximos a 0 se obtendrán cuando $\ln \mathcal{L}(\widehat{\beta})$ sea muy cercano a $\ln \mathcal{L}(0)$. En tal situación, las variables incluidas en el modelo son poco significativas, pues la estimación de los coeficientes no mejora el error que se comete si dichos coeficientes fueran nulos. En consecuencia, la capacidad del modelo será reducida.

- Cuanto mayor sea la capacidad explicativa del modelo, mayor será el valor de $\ln \mathcal{L}(\widehat{\beta})$ sobre el valor de $\ln \mathcal{L}(0)$, y más se aproximará el cociente de verosimilitudes al valor de 1.

3. Estadístico de Pearson

El estadístico de Pearson o ji-cuadrado de Pearson (χ^2), cuantifica las medidas de error por medio de la diferencia entre el valor observado y el estimado. Se define como [47]:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1 - \widehat{\pi}_i)} \tag{2.69}$$

Tiene la misma distribución asintótica que el estadístico de la Devianza, es decir, una ji-cuadrada con los mismos grados de libertad.

Como expresa Collett [47], tanto para aplicar el test basado en la Devianza D como para el estadístico χ^2 tienen que verificarse que el número de observaciones para cada combinación de las variables explicativas sea grande, es por ello, por lo que estos métodos no se aplican en el caso de variables exógenas continuas o modelos no agrupados de Bernoulli.

4. Estadístico de Hosmer Lemeshow

Otra medida global de la exactitud predictiva, no basada en el valor de la función de verosimilitud sino en la predicción real de la variable dependiente, es el contraste diseñado por Hosmer y Lemeshow [48]. El estadístico C_g se basa en la agrupación de las probabilidades estimadas $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_N$, bajo el modelo de regresión. La

idea principal es que, el primer grupo estará conformado aproximadamente por las N/G observaciones cuyas probabilidades predichas sean más pequeñas, el segundo por las siguientes N/G más pequeñas y así sucesivamente. Los puntos de corte así generados se denominan deciles de riesgo d_i. La tabla 2.4 muestra las frecuencias esperadas y observadas en cada uno de los grupos, utilizadas para el cálculo del estadístico C_g.

	Respuesta			
	$Y = 1$		$Y = 0$	
Grupos	Observado	Esperado	Observado	Esperado
$\hat{\pi}_j < d_1$	o_{11}	e_{11}	o_{01}	e_{01}
$d_1 \leq \hat{\pi}_j < d_2$	o_{12}	e_{12}	o_{02}	e_{02}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$d_9 \leq \hat{\pi}_j < d_{10}$	o_{1G}	e_{1G}	o_{0G}	e_{0G}
Total	o_1	e_1	o_0	e_0

Tabla 2.4: Frecuencias observadas y esperadas para el cálculo del estadístico C_g. **Fuente:** T. Iglesias, (2013) [46]. **Elaboración:** Autor.

El número de individuos observados para los que ocurrió el suceso y para los que no ocurrió, en cada uno de los grupos es respectivamente [46]:

$$o_{1g} = \sum_{j=1}^{n_g} y_j, \qquad o_{0g} = \sum_{j=1}^{n_g} (1 - y_j) \tag{2.70}$$

donde, n_g es el número de observaciones en el grupo g.

Análogamente, el número esperado de individuos para los que ocurrirá el suceso y para los que no, se denotan por

$$e_{1g} = \sum_{j=1}^{n_g} \hat{\pi}_j, \qquad e_{0g} = \sum_{j=1}^{n_g} (1 - \hat{\pi}_j) \tag{2.71}$$

El estadístico C_g se obtiene comparando estos valores observados y esperados de la siguiente forma

$$C_g = \sum_{j=0}^{1} \sum_{g=1}^{G} \frac{(o_{jg} - e_{jg})^2}{e_{jg}} \tag{2.72}$$

Hosmer y Lemeshow [48], demostraron que cuando el número de variables explicativas más uno es menor que el número de grupos $(k+1) < G$, bajo la hipótesis

del modelo logístico, C_g tiene una distribución asintótica χ^2_{G-2}.

5. Porcentaje de aciertos estimados en el modelo

Una forma intuitiva de resumir los resultados de un modelo de elección binaria es mediante una tabla de clasificación. Hosmer y Lemeshow [48] mencionan que esta tabla es el resultado de la clasificación cruzada de la variable endógena, y una variable dicotómica cuyos valores se derivan de las probabilidades estimadas.

Para obtener la variable dicotómica se define un punto de corte c, y se compara cada probabilidad estimada con c. Si la probabilidad estimada excede c, entonces la variable dicotómica toma el valor de 1; caso contrario, toma el valor de 0.

$$\widehat{Y}_i = \begin{cases} 1 & \text{si} \quad \hat{\pi}_i > c \\ 0 & \text{si} \quad \hat{\pi}_i \leq c \end{cases} \tag{2.73}$$

El valor habitual que se le asigna al punto de corte es $c = 0.5$.

De acuerdo con Medina [44], la elección del punto de corte igual a 0.5 no siempre es la mejor alternativa, debido a que si una muestra presenta desequilibrios entre el número de unos y el de ceros, la elección del punto de corte $c = 0.5$ podría conducir a no predecir ningún uno o ningún cero.

Con cualquier valor que se fije para el punto de corte se cometerán dos errores: habrá ceros que se clasifiquen incorrectamente como unos y unos que se clasifiquen incorrectamente como ceros. En consecuencia, el valor que debe tomar el punto de corte depende de la distribución de los datos y de la importancia relativa de cada tipo de error.

Una vez seleccionado el punto de corte, se contabiliza el porcentaje de aciertos para decidir si la bondad del ajuste es alta o no. A partir de este recuento se construye la tabla 2.5 de clasificación.

		Valor real de Y_i	
		$Y_i = 0$	$Y_i = 1$
Valor predicho de $\widehat{Y}_i$	$\widehat{Y}_i = 0$	P_{00}	P_{01}
	$\widehat{Y}_i = 1$	P_{10}	P_{11}

Tabla 2.5: Tabla de clasificación de aciertos. **Fuente:** E. Medina, (2003) [44]. **Elaboración:** Autor.

Donde, P_{00} y P_{11} corresponden a predicciones correctas (valores de cero clasifi-

cados como cero y valores de uno clasificados como uno), mientras que P_{01} y P_{10} corresponderán a predicciones erróneas (valores de cero clasificados como uno y valores de uno clasificados como cero). A partir de estos valores se definen los siguientes índices descritos en la tabla 2.6.

Índice	Definición	Expresión
Tasa de aciertos	Cociente entre las predicciones correctas y el total de predicciones	$\frac{P_{00}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$
Tasa de errores	Cociente entre las predicciones incorrectas y el total de predicciones	$\frac{P_{01}+P_{10}}{P_{00}+P_{01}+P_{10}+P_{11}}$
Sensibilidad	Razón entre los valores 1 correctos y el total de valores 1 observados	$\frac{P_{11}}{P_{01}+P_{11}}$
Especificidad	Razón entre los valores 0 correctos y el total de valores 0 observados	$\frac{P_{00}}{P_{00}+P_{10}}$
Tasa de falsos negativos	Proporción entre la frecuencia de valores 0 incorrectos y el total de valores 0 observados	$\frac{P_{10}}{P_{00}+P_{10}}$
Tasa de falsos positivos	Proporción entre la frecuencia de valores 1 incorrectos y el total de valores 1 observados	$\frac{P_{01}}{P_{01}+P_{11}}$

Tabla 2.6: Índices para medir la bondad de ajuste. **Fuente:** E. Medina, (2003) [44]. **Elaboración:** Autor.

6. Curva ROC

La curva ROC (Receiver Operating Characteristic) se define como una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según varía el punto de corte [49]. Si se van modificando los puntos de corte y se representa la sensibilidad en el eje de las ordenadas frente al complementario de la especificidad en el eje de las abscisas tenemos la curva ROC (Figura 2.3).

De modo que un gráfico ROC representa el equilibrio relativo entre la razón de verdaderos positivos y razón de falsos positivos. El punto $(0, 1)$ representa una clasificación perfecta, mientras que un punto a lo largo de la recta $y = x$ representa una clasificación totalmente aleatoria [50].

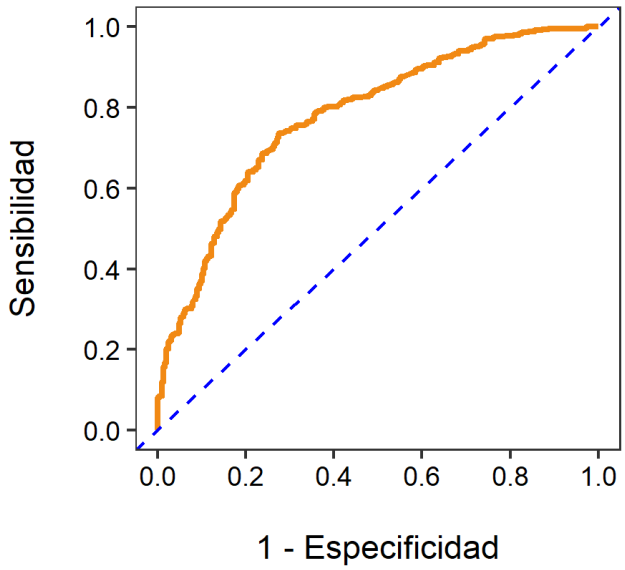


Figura 2.3: Curva ROC. **Elaboración:** Autor.

Mediante la curva ROC se generan índices que miden el rendimiento de un clasificador a lo largo del rango de la probabilidad pronosticada, definiendo como rendimiento a la capacidad de clasificar correctamente las observaciones. Uno de los más utilizados es el área bajo la curva ROC conocido como AUROC (Area Under ROC Curve). Según Fawcett [51] el modelo con mayor capacidad predictiva es aquel que tiene una curva ROC más cercana al punto $(0, 1)$, este coincidirá con el modelo que deje una mayor área entre la curva y la diagonal que se supone elegir aleatoriamente.

Por lo tanto, el índice AUROC tomará valores entre 0 y 1, donde valores cercanos a 1 indican un alto rendimiento del modelo. Anderson [52] manifiesta que un valor para AUROC de 0 implica que las predicciones del modelo son perfectamente erróneas, un valor de 0.5 implica que el modelo realiza una predicción aleatoria y un valor de 1 implica que el modelo realiza una predicción perfecta; además afirma que un valor superior a 0.7 se considera adecuado.

7. Coeficiente de GINI

El coeficiente de GINI al igual que el indicador AUROC, 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 [50]. El coeficiente de GINI varía entre 0 y 1, cuanto más cercano a 1 se encuentre, el modelo genera una

discriminación mayor.

El coeficiente de GINI se puede calcular por [53]:

$$\text{GINI} = 1 - \sum_{i=2}^{I} [N(i) + N(i-1)] [P(i) + P(i-1)] \tag{2.74}$$

donde,

I: Número de intervalos,

$N(i)$: Porcentaje acumulado de negativos hasta el intervalo i, y

$P(i)$: Porcentaje acumulado de positivos hasta el intervalo i.

De la misma forma, se puede obtener el coeficiente de GINI mediante la igualdad

$$\text{GINI} = 2\text{AUROC} - 1 \tag{2.75}$$

de donde tenemos que el coeficiente de GINI es dos veces el área comprendida entre la curva ROC y la recta $y = x$.

8. Prueba de Kolmogorov-Smirnov

Siguiendo el trabajo de Arnold y Emerson [54], el test de Kolmogorov-Smirnov es una prueba de bondad de ajuste del tipo no paramétrico[4], mediante la cual se contrasta la hipótesis de si dos muestras aleatorias independientes provienen de distribuciones continuas idénticas.

Si consideramos una muestra de tamaño n, $x_1, x_2, \ldots, x_n$ de una variable aleatoria X, la distribución acumulada empírica de X tiene la siguiente expresión:

$$ecdf(x) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 1 & \text{si} \quad x_i \leq x \\ 0 & \text{si} \quad \text{c.c} \end{cases} \tag{2.76}$$

Así, para una muestra $x_1, x_2, \ldots, x_{n_1}$ de tamaño n_1 de una variable aleatoria continua X con función de distribución acumulada F_1; y una muestra $y_1, y_2, \ldots, y_{n_2}$ de tamaño n_2 de una variable aleatoria continua Y con función de distribución acumulada F_2. La prueba de Kolmogorov-Smirnov contrasta las hipótesis:

$$H_0 : F_1(x) = F_2(x) \ \ \forall x \quad \text{frente a} \quad H_1 : F_1(x) \neq F_2(x) \tag{2.77}$$

[4]Las pruebas no paramétricas no realizan suposiciones a priori sobre la distribución de los datos.

y, el estadístico *K-S* para contrastar la H_0 se define por:

$$K\text{-}S = \max_{z} |ecdf_1(z) - ecdf_2(z)| \tag{2.78}$$

donde, $ecdf_1$ y $ecdf_2$ son las funciones de distribución acumuladas empíricas de X e Y respectivamente.

De la expresión anterior, el estadístico *K-S* es la distancia máxima entre $ecdf_1$ y $ecdf_2$, y su valor oscila en el intervalo $(0, 1)$, donde valores cercanos a 0 indican que las distribuciones de X e Y son idénticas, y valores cercanos a 1 indican que las distribuciones de X e Y difieren. De manera que, el estadístico *K-S* es utilizado como una medida de divergencia entre las distribuciones de dos variables aleatorias continuas [55].

9. Estadísticos para comparar la bondad del ajuste entre modelos alternativos

Como criterio general de elección entre distintos modelos se presenta aquel con un mayor valor de la función de verosimilitud. Sin embargo, este criterio podría conducir a errores ya que no se toma en cuenta el número de variables explicativas, ni el número de observaciones incluidas [44]. Por este motivo existen varios estadísticos basados en la función de verosimilitud que incorporan dichos conceptos. Entre los más importantes se encuentran:

- El criterio de información de Akaike: Corrige el valor de la función de verosimilitud por el número de parámetros del modelo. El estadístico se define como [35]:

$$AIC = \frac{2k}{n} - \frac{2\widehat{\mathcal{L}}}{n} \tag{2.79}$$

  donde,

  k es el número de parámetros del modelo,

  $\widehat{\mathcal{L}}$ es el valor de la función de verosimilitud, y

  n es el número de observaciones de la muestra.

  Según el criterio de Akaike, será preferible aquel modelo que presente un valor de AIC menor.

- Criterio de información Bayesiano: Tiene en cuenta explícitamente el tamaño de la muestra [35]:

$$BIC = \frac{k\ln(n)}{n} - \frac{2\widehat{\mathcal{L}}}{n} \tag{2.80}$$

Al igual que el estadístico AIC es preferible aquel modelo que presenta un valor de BIC menor.

- Criterio de información de Hannan-Quinn: Se presenta como una alternativa a los criterios antes descritos. Se define por [44]:

$$HQC = \frac{2k\ln(\ln(n))}{n} - \frac{2\widehat{\mathcal{L}}}{n} \tag{2.81}$$

Según este criterio también serán preferibles aquellos modelos que presenten una valor del estadístico HQC menor.

Multicolinealidad

Castro [56] define a la multicolinealidad como el problema de que una variable explicativa incluida en el modelo de regresión sea aproximadamente una combinación lineal de las demás variables, dando como resultado una fuerte correlación. Además, cuando existe una correlación alta se reduce la precisión de los coeficientes estimados, es decir, se incrementa significativamente la varianza de los coeficientes β_i de las variables explicativas.

Para analizar la multicolinealidad se calculan los VIF (Factor de Inflación de la Varianza) de los parámetros estimados. RStudio calcula los factores de inflación a partir de la matriz de covarianzas de los parámetros estimados usando el método propuesto por Davis *et al*. [57], el cual se basa en la matriz de correlación de la matriz de información X.

$$VIF_i = \frac{Var(\widehat{\beta}_i)}{Var(\widehat{\beta}_i^*)} = \frac{1}{1-R_i^2} \tag{2.82}$$

donde,

$$Var(\widehat{\beta}_i^*) = \frac{\sigma^2}{\sum_{j=1}^{N}(x_{ij}-x_j)} \tag{2.83}$$

es la varianza óptima en caso de que no haya correlación entre las variables explicativas,

$$Var(\widehat{\beta}_i) = \frac{\sigma^2}{\sum_{j=1}^{N}(x_{ij}-x_j)(1-R_i^2)} \tag{2.84}$$

es la varianza de un estimador cualquiera, y R_i^2 es el coeficiente de determinación de la regresión auxiliar de la variable x_i sobre el resto de las variables explicativas.

Así, el VIF aumenta considerablemente si la correlación múltiple de la variable x_i es alta, mientras que si la correlación múltiple es baja el VIF será cercano a 1.

Según Moreno [58], valores de $VIF_i > 5$ están asociados a $R_i^2 > 0.8$; en cuyo caso se puede considerar que las consecuencias sobre el modelo pueden ser relevantes. Sin embargo, Neter *et al*. [59] sugieren que la multicolinealidad es severa si $VIF_i > 10$.

Por otro lado, Fox y Weisberg [60] señalan que el VIF no es la forma adecuada de evaluar la inflación de la varianza en variables categóricas. Los autores proponen el factor de inflación de la varianza generalizada (GVIF), que al igual como ocurre con el VIF, valores próximos a 1 indican ausencia de multicolinealidad.

Medidas de influencia: valores extremos

Como menciona Medina [44], cuando se examina la idoneidad del modelo es importante considerar la posible presencia de valores extremos (outliers) que puedan alterar el ajuste de los datos. Para ello se analizan los residuos de Pearson y residuos de la Devianza descritos anteriormente.

Como complemento al análisis de los residuos, se utilizan indicadores que cuantifican la influencia que cada observación ejerce sobre la estimación de los coeficientes o sobre las predicciones hechas a partir de los mismos [49]. Como resultado, cuanto más altos son estos indicadores, mayor es la influencia que ejerce una observación en la estimación del modelo y debe considerarse como dato atípico.

Uno de los indicadores más utilizados es la Distancia de Cook que cuantifica el cambio en los residuos cuando una observación es excluida del cálculo de los coeficientes de regresión. Se define por [49]:

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{y}_j - \widehat{y}_{j(i)})^2}{ks^2} \tag{2.85}$$

donde,

$\widehat{y}_{j(i)}$ es el valor de respuesta ajustado obtenido al excluir i, y

$s^2 = (n-k)^{-1}\mathrm{e}^t\mathrm{e}$ es el error cuadrático medio de la regresión del modelo.

Se considera que una observación es influyente si $D_i > 1$.

Capítulo 3

Estimación del Modelo

Una vez establecido el marco teórico se procede a describir la metodología empleada en la construcción del modelo de elección binaria. Se empezará realizando la depuración de la base de datos, un estudio descriptivo de la muestra y la selección de variables explicativas. Por último, se procederá a la estimación y validación simultánea de modelos; por un lado, un modelo logit y por otro un modelo probit.

3.1. Datos y Variables

La base de datos obtenida a través de la encuesta para evaluar los efectos de la crisis sanitaria en Ecuador, inicialmente cuenta con 3868 registros transversales y 42 variables que describen las características demográficas, socioeconómicas y financieras de los trabajadores públicos, privados, autónomos y desempleados. La descripción de las variables en la base de datos se encuentra en el Anexo A.1.

Esta encuesta tiene representatividad a nivel nacional y se levantó de manera digital con la ayuda del cuestionario electrónico diseñado por la Dirección de Gestión de la Información y Procesos de la Escuela Politécnica Nacional, como también a través de la red social Facebook en el período abril - mayo de 2020.

El universo de personas del estudio se limita al público que se interesó en responder la encuesta, por tal motivo la base de datos posee un sesgo de selección que podría ser tratado en posteriores trabajos de investigación a través de la aplicación de técnicas estadísticas como el *propensity score matching* o la utilización de datos censales.

3.1.1. Depuración de la base de datos

Sobre la muestra se realiza un estudio de cada variable, para ello, se analizan los registros inconsistentes, se examinan los datos faltantes, se sustituyen los valores incorrectos y se modifican los tipos de datos (numérico, factor o carácter) según sea el caso.

Las variables que fueron sujetas al proceso de depuración se presentan a continuación:

- `Pais`: Dada la baja representatividad (1.96 % del total de la muestra) e inconsistencias con el estado migratorio y los años de residencia en el país, se eliminaron los registros cuyo país de origen sea distinto a *Ecuador*.

- `N_Instruccion`: Se excluyeron los registros de los encuestados que no contaban con ningún nivel de instrucción, debido a su nula representatividad en la muestra (0.08 %).

- `Miemb_hog`: Se corrigieron los valores concernientes al número de personas en el hogar, se estableció un mínimo de 1 persona para aquellos registros con 0 y se fijó una cota[1] de 8 personas para los registros con valores superiores.

- `Sit_laboral`: Se eliminaron varios registros de los encuestados cuya situación laboral actual es *desempleado*, debido a que presentaban inconsistencias con su situación laboral en el contexto de la enfermedad COVID-19.

- `Act_Econom`: Al presentar un porcentaje alto de registros en la categoría *Otro* se realizó una exhaustiva revisión de las actividades económicas en dicha categoría, con el objetivo de englobarlas en una actividad económica existente. Asimismo, los registros vacíos se los incluyó en una nueva categoría denominada *Sin actividad por desempleo*, debido a que correspondían a encuestados que no especificaron su actividad económica por encontrarse desempleados.

- `Ing_durCOVID`: Algunos registros presentaban inconsistencias con respecto al incremento y disminución de los ingresos durante la cuarentena. Por ejemplo, distintos encuestados indicaban que sus ingresos habían aumentado y disminuido a la vez, confundiendo la variación de su renta. Por lo tanto, se optó por excluir dichas observaciones de la muestra.

[1] La cota se estableció en base al promedio de personas por hogar a nivel nacional del Censo de Población y Vivienda 2010 [61].

- `Efect_GastosCOVID`: De manera similar a lo ocurrido con la variable de ingresos, ciertos registros mostraban incongruencias con la variación de los gastos; por un lado, los trabajadores indicaban una disminución en sus gastos durante la cuarentena y por otro mencionaban un aumento de los mismos. En consecuencia, dichos registros se descartaron para el análisis de la demanda.

- Coerción[2] a tipo factor de las variables: `Rango_edad`, `Monto_Dcredt`, `Couta_DPag`, `Ing_antCOVID`, `Aumt_IngCOVID`, `Dismy_IngCOVID`, `Gast_antCOVID`, `Aumt_GastCOVID`, y `Dismy_GastCOVID`.

Finalmente, la base de datos para el desarrollo del trabajo contiene 3683 registros, que posteriormente se dividirá en dos grupos, uno de modelamiento y otro de validación.

3.1.2. Características de la demanda de crédito en Ecuador

La demanda de crédito estará definida por la variable `Neces_prestam` que corresponde a la pregunta de la encuesta: *Actualmente, dadas las circunstancias de la crisis sanitaria ¿usted necesitaría un préstamo?*. La variable toma el valor de 1 si el encuestado necesita un préstamo y 0 caso contrario.

Del total de la muestra, el 56.53 % de los encuestados respondieron que necesitan acceso a crédito, mientras que el restante 43.47 % respondió que no requiere del mismo (Figura 3.1).

Los encuestados son, a su vez, un 42.36 % desempleados, un 23.75 % empleados públicos, un 20.23 % empleados privados y un 13.66 % trabajadores autónomos o pertenecientes a organizaciones del sector EPS (Figura 3.2). Además, el 14.8 % se dedican a la enseñanza, el 11.54 % a actividades profesionales como: consultoría, contabilidad, publicidad, auditoría, secretaría y telecomunicaciones, el 10.05 % a actividades de servicio como: carpintería, sastrería, seguridad, peluquería y belleza, y el 5.08 % al comercio por menor. Las demás categorías son menos representativas en la muestra (Figura 3.3). Finalmente, dentro del grupo de los trabajadores que necesitan acceso a crédito, el 42.42 % requieren de financiamiento urgente (Figura 3.4).

La tabla 3.1 muestra la distribución por edades de los encuestados. La primera mayoría corresponde al grupo de trabajadores entre 35 a 44 años, seguido de los

[2]La coerción es una característica de los lenguajes de programación que permite, implícita o explícitamente, convertir un elemento de un tipo de datos en otro [62].

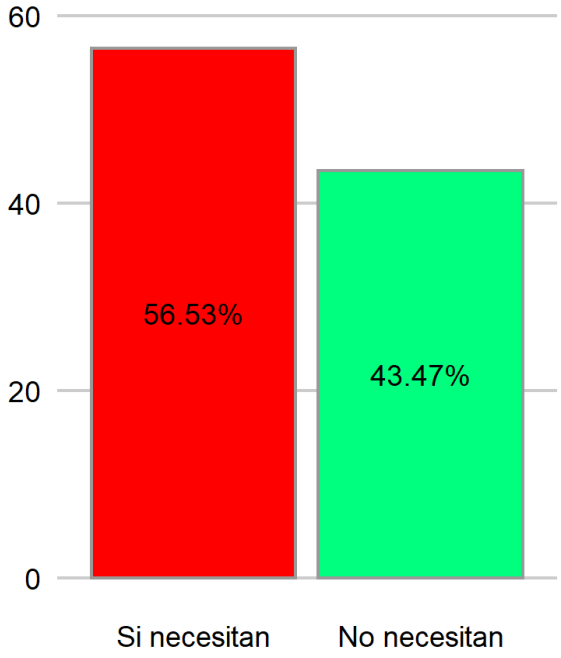


Figura 3.1: Demanda de crédito (en porcentaje) durante la cuarentena. **Elaboración:** Autor.

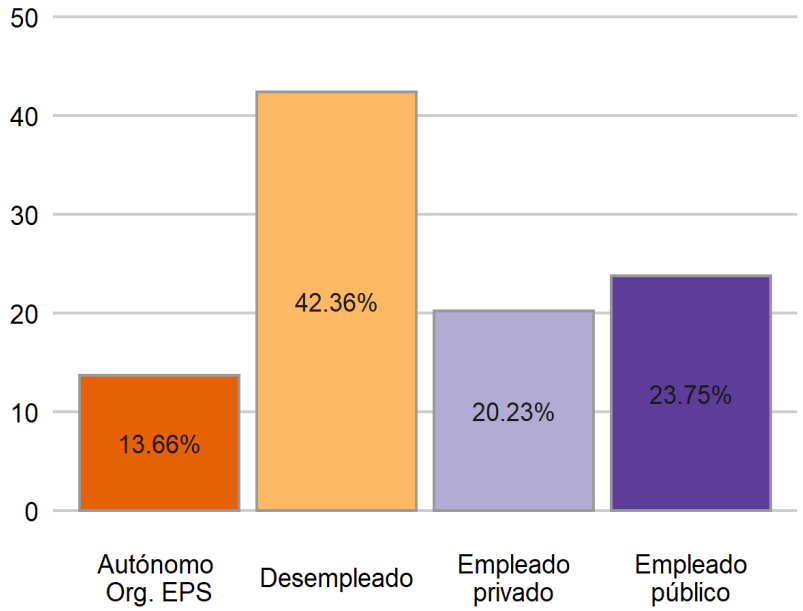


Figura 3.2: Situación laboral de los trabajadores. **Elaboración:** Autor.

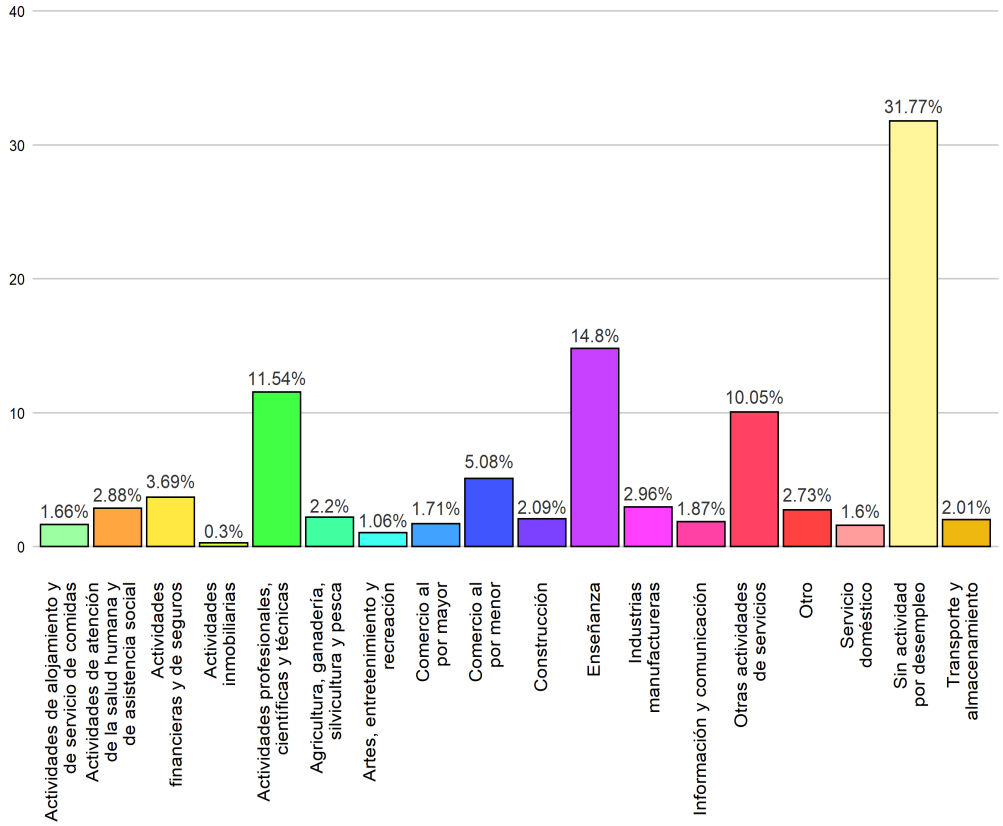


Figura 3.3: Distribución de los encuestados por actividad económica. **Elaboración:** Autor.

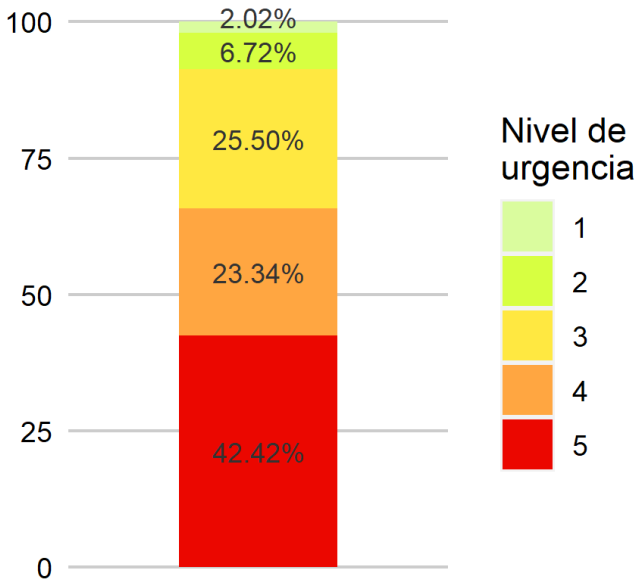


Figura 3.4: Urgencia de la necesidad de crédito. **Elaboración:** Autor.

trabajadores entre 18 a 34 años. Al analizar la demanda de crédito, se observa que las personas que lo necesitan más son los trabajadores entre 35 y 60 años (Figura 3.5).

De 18 a 24 años	24.74 %
De 25 a 34 años	23.43 %
De 35 a 44 años	24.87 %
De 45 a 60 años	22.59 %
Mayor a 60 años	4.37 %
Total	100 %

Tabla 3.1: Distribución de los encuestados por edad. **Elaboración:** Autor.

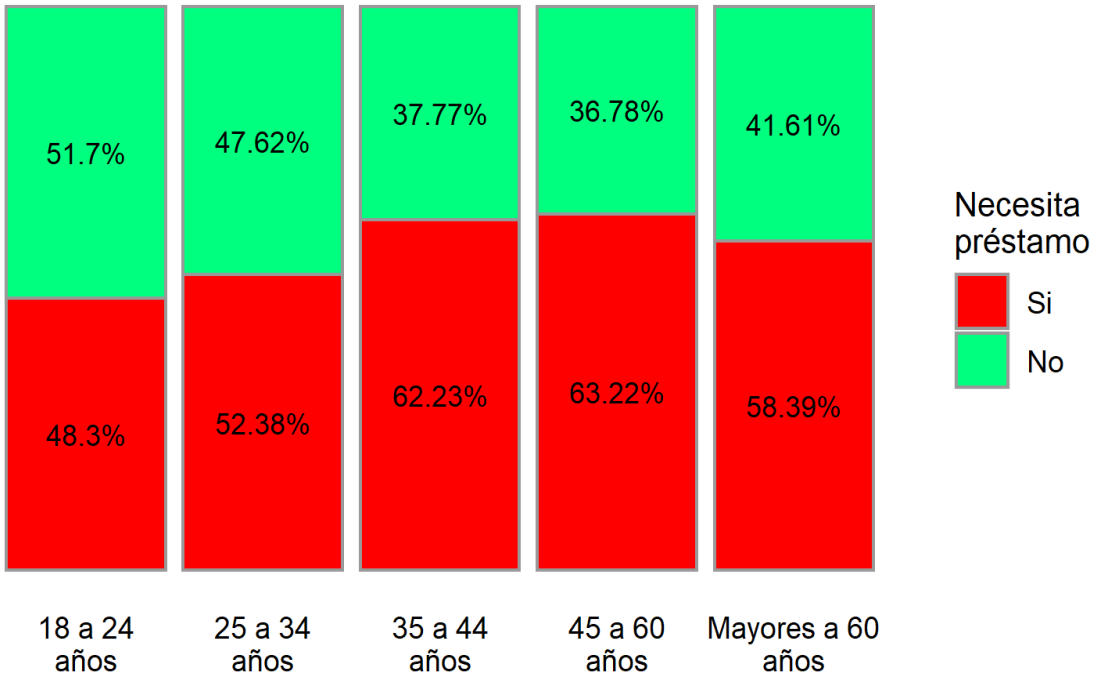


Figura 3.5: Demanda de crédito (en porcentaje) por edad. **Elaboración:** Autor.

Hombres y mujeres se encuentran igualmente distribuidos en la muestra. Los resultados indican que el 58.57 % de los hombres necesitan de financiamiento frente a un 54.53 % de las mujeres encuestadas (Figura 3.6).

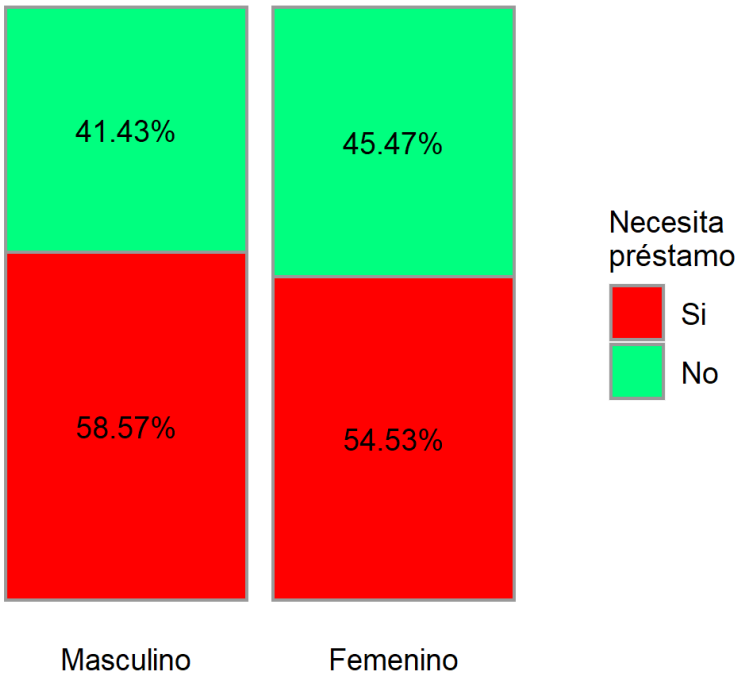


Figura 3.6: Demanda de crédito (en porcentaje) por género. **Elaboración:** Autor.

En cuanto a la situación económica y financiera, la tabla 3.2 muestra que el 62.29 % de los trabajadores han visto reducida su renta durante la cuarentena y el 7.14 % no han percibido ingresos, de ellos el 64.39 % y el 76.05 % necesitan de financiamiento, respectivamente (Figura 3.7). Más aún, del grupo de trabajadores que su renta disminuyó, el 39.68 % perdió sus ingresos en totalidad y el 20.31 % de los casos su salario se redujo a la mitad (Figura 3.8).

Aumentado	1.17 %
Disminuido	62.29 %
No estoy percibiendo ingresos durante la cuarentena	7.14 %
Se han mantenido	29.41 %
Total	100 %

Tabla 3.2: Variación de la renta durante la cuarentena. **Elaboración:** Autor.

En esta misma línea, el 73.58 % de los encuestados posee deudas, y de ellos el 64.06 % necesita de un préstamo, mientras el 67.20 % no posee ahorros, y de ellos el 66.87 % necesita de financiamiento (Figura 3.9).

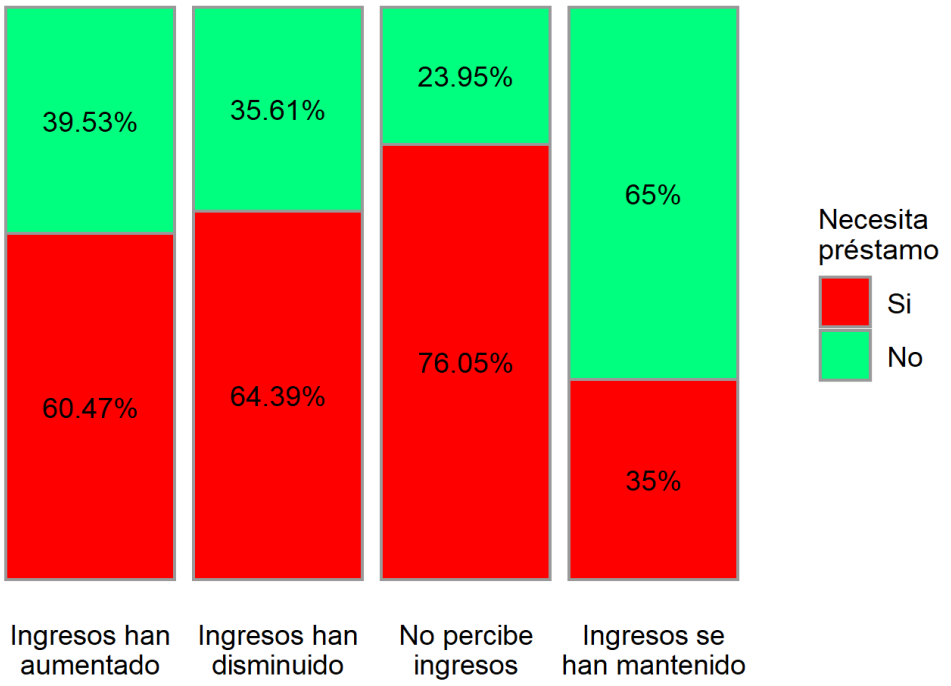


Figura 3.7: Demanda de crédito (en porcentaje) por variación de la renta. **Elaboración:** Autor.

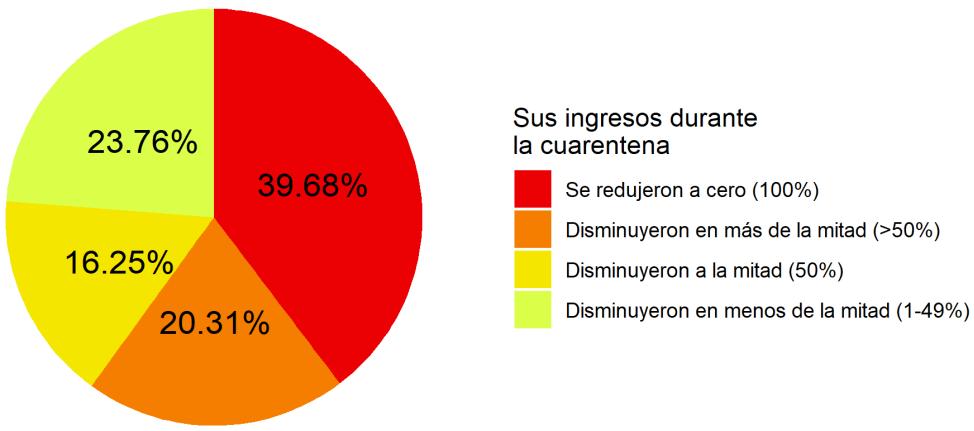


Figura 3.8: Disminución de la renta durante la cuarentena. **Elaboración:** Autor.

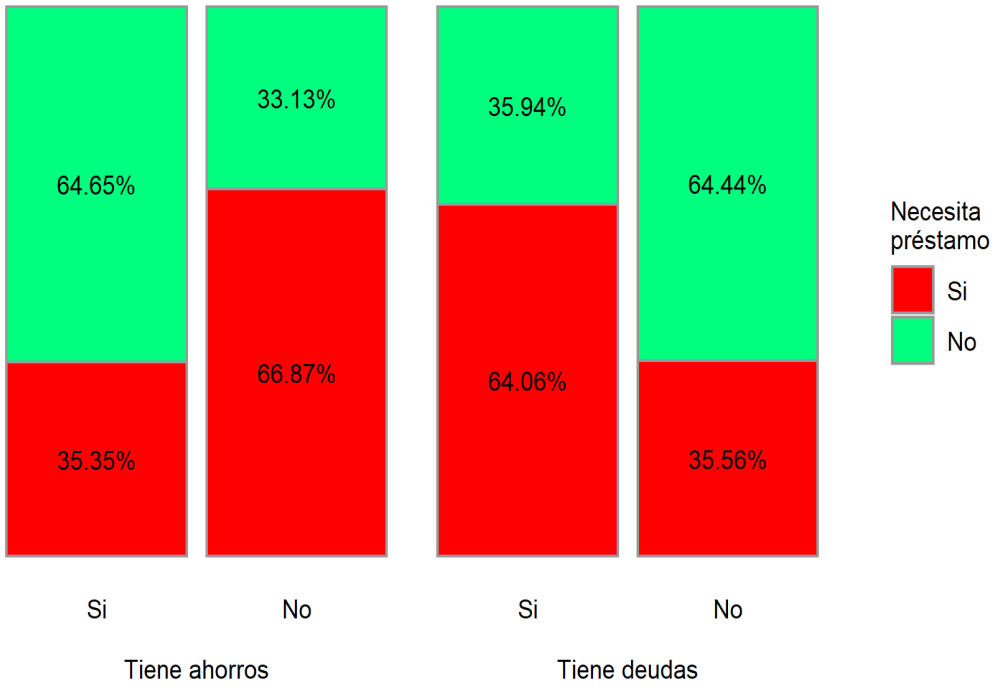


Figura 3.9: Demanda de crédito por ahorro y por deuda. **Elaboración:** Autor.

3.1.3. Selección de variables

Para la selección de las variables explicativas que formarán parte del modelo, se emplea una metodología secuencial en etapas basada en el algoritmo de inducción de reglas CHAID[3] que permite especificar las principales características que determinan la necesidad de crédito. Este procedimiento se lo realiza con el software AnswerTree[4].

En la tabla 3.3 se muestran las variables escogidas por el algoritmo que presentan un mayor grado de asociación con la variable demanda de crédito definida anteriormente.

Una vez realizada la selección de variables que serán parte del modelo se divide la base de datos en dos submuestras, con el objetivo de desarrollar el modelo con la primera (muestra de modelamiento), y validarlo con la segunda (muestra de validación).

[3]El método CHAID se detalla en el Anexo A.2.

[4]AnswerTree es una herramienta de software que permite realizar clasificaciones basadas en árboles de decisión.

Variable	Chi-square	P-value
Tien_ahorros	328.2252	0.0000
Ing_durCOVID	302.7512	0.0000
Tien_deudas	236.6293	0.0000
Sit_laboralCOVID	235.9307	0.0000
Act_Econom	199.2203	0.0000
Sit_laboral	168.9617	0.0000
Efect_GastoCOVID	73.3310	0.0000
Miemb_hog	73.1137	0.0000
Est_civil	68.3879	0.0000
Rango_edad	54.3731	0.0000
Mpercp_ing	51.0643	0.0000
Reg_tenenc	39.4252	0.0000
Genero	6.1155	0.0134

Tabla 3.3: Variables seleccionadas con el método CHAID. **Elaboración:** Autor.

La submuestra para la estimación del modelo corresponde al 80 % de la muestra original, es decir, consta de 2946 registros. Por otro lado, la submuestra para la validación del modelo corresponde al 20 % de la muestra original, es decir, consta de 737 registros. Esta partición se la realiza de forma aleatoria utilizando la función `sample.split()` del software RStudio.

3.2. Modelo Logit

3.2.1. Estimación

Para la estimación del modelo se utiliza el software RStudio. Luego de probar con varios modelos efectuando los correspondientes contrastes de significancia individual sobre los parámetros y en base a los criterios de información para comparar entre modelos alternativos (reportados en la subsección 2.5.5) se retiene el siguiente[5]:

[5]El código para la estimación y validación de los modelos se encuentra en el Anexo B.1

	Estimate	Odds	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.8313	0.1602	0.3034	-6.04	0.0000	***
Tien_ahorrosSi	-0.9636	0.3815	0.0961	-10.02	0.0000	***
Tien_deudasSecRegulado_Banca	1.0369	2.8205	0.1273	8.15	0.0000	***
Tien_deudasSecRegulado_Otro	1.0809	2.9474	0.1803	6.00	0.0000	***
Tien_deudasSec_NoRegulado	1.2670	3.5503	0.1769	7.16	0.0000	***
Tien_deudasAmbosSectores	1.5493	4.7081	0.1712	9.05	0.0000	***
Ing_durCOVIDNo estoy percibiendo ingresos durante la cuarentena	0.6944	2.0026	0.2431	2.86	0.0043	**
Ing_durCOVIDDisminuido	0.7846	2.1915	0.1045	7.51	0.0000	***
Efect_GastoCOVIDAumentado	0.4386	1.5505	0.0906	4.84	0.0000	***
Rango_edadDe 25 a 34 años	0.5284	1.6963	0.1445	3.66	0.0003	***
Rango_edadDe 35 a 44 años	1.0751	2.9302	0.1613	6.66	0.0000	***
Rango_edadDe 45 a 60 años	1.0813	2.9486	0.1639	6.60	0.0000	***
Rango_edadMayor a 60 años	0.9188	2.5063	0.2594	3.54	0.0004	***
Sit_laboralAutónomo u Organizaciones de la EPS	-0.4414	0.6431	0.1918	-2.30	0.0213	*
Sit_laboralEmpleado Público	-1.0861	0.3375	0.1551	-7.00	0.0000	***
Sit_laboralEmpleado Privado	-0.9610	0.3825	0.1649	-5.83	0.0000	***
Act_EconomSector 2	1.1792	3.2516	0.2195	5.37	0.0000	***
Act_EconomSector 3	0.7401	2.0961	0.2202	3.36	0.0008	***
Act_EconomSector 4	0.7569	2.1318	0.2263	3.34	0.0008	***
Miemb_hogDe 4 a 5 Miembros	0.3246	1.3835	0.1017	3.19	0.0014	**
Miemb_hogMás de 5 Miembros	0.7275	2.0698	0.1305	5.57	0.0000	***
Reg_tenencPropio	-0.3700	0.6908	0.0955	-3.87	0.0001	***
GeneroEstcivilHombre otro	-0.0395	0.9613	0.1408	-0.28	0.7789	
GeneroEstcivilMujer casada	-0.4193	0.6575	0.1476	-2.84	0.0045	**
GeneroEstcivilMujer otro	-0.2604	0.7708	0.1378	-1.89	0.0588	.

[1]Nivel de significancia: ‘ *** ’ 0.000 , ‘ ** ’ 0.001, ‘ * ’ 0.01, ‘ . ’ 0.05, ‘ ’ 0.1

[2]Las variables categóricas tienen como característica de referencia a los trabajadores que no poseen ahorros, no tienen deudas, sus ingresos durante la cuarentena se han aumentado, sus gastos durante la cuarentena se han disminuido o se han mantenido, con edades entre los 18 y 24 años, desempleados, pertenecientes al sector económico 1, de 1 a 3 miembros en el hogar, tipo de vivienda arrendada y hombres solteros.

Tabla 3.4: Regresión logística para el modelo de demanda de crédito. **Elaboración:** Autor.

Se puede observar que se realizó algunos cálculos con las variables, por ejemplo, se categorizó la variable `Tien_deudas`, se agruparon las categorías *Aumentado* y *Se han mantenido* de la variable `Ing_durCOVID` y las categorías *Disminuido* y *Se han mantenido* de la variable `Efect_GastoCOVID`, se discretizó la variable `Miemb_hog` y se segmentó la variable `Act_Econom`. Además, se realizó la interacción entre las variables `Genero` y `Est_civil`. Ver Anexo A.3.

En la tabla 3.4 se presenta la estimación de los coeficientes, odds ratios, errores

estándar, valor del estadístico de Wald y el p-valor asociado; se puede observar que todos los coeficientes son significativos al 95 % de confianza, exceptuando el coeficiente de la variable `GeneroEstcivil` para la categoría *Hombre otro*.

La interpretación de los coeficientes se la realiza a través de los odds ratios de la siguiente manera:

La probabilidad de necesitar un crédito es 2.62 veces menor cuando el trabajador tiene ahorros, respecto al trabajador que no los tiene, manteniéndose constantes las demás variables.

Cuando el trabajador mantiene obligaciones de deuda en el sector regulado por la banca o por otra entidad, su probabilidad de necesitar un crédito es 2.82 y 2.95 veces mayor a cuando no tiene ningún tipo de deuda respectivamente, ceteris paribus. Asimismo, esta probabilidad se duplica cuando el trabajador mantiene obligaciones de deuda en los sectores no regulados y en ambos sectores.

En el caso de los trabajadores que han dejado de percibir ingresos como aquellos que su salario disminuyó durante la crisis sanitaria, la probabilidad de necesitar un crédito en relación a los trabajadores que aumentaron sus ingresos es 2 y 2.19 veces mayor respectivamente, si el resto de variables no cambia.

En cuanto a los trabajadores que sus gastos aumentaron durante la cuarentena, son 1.55 veces más propensos a necesitar de financiamiento que aquellos trabajadores que sus gastos se mantuvieron o disminuyeron, si el resto de factores no cambia.

De la misma forma, la probabilidad de necesitar un crédito aumenta 1.7 veces cuando la edad del trabajador se encuentra entre 25 a 34 años, en comparación a los trabajadores de entre 18 y 24 años, manteniéndose constantes las demás variables. A su vez, dicha probabilidad se duplica para los trabajadores de entre 35 a 60 años.

Por otro lado, aquellos trabajadores que se encuentran empleados, tanto en el sector privado como público, la probabilidad de necesitar financiamiento es 2.61 y 2.96 veces menor respectivamente, que aquellos que se encuentran desempleados, si el resto de variables no cambia. Algo semejante ocurre con los trabajadores autónomos y pertenecientes a las organizaciones de la EPS, su probabilidad de demandar un crédito es 1.55 veces menor con respecto a los trabajadores desempleados.

De acuerdo con la segmentación que se encuentra en el Anexo A.3 para la variable `Act_Econom`, la probabilidad de necesitar crédito para los trabajadores en el sector 2 es 3.25 veces mayor, con respecto a los trabajadores del sector 1. Por su parte, dicha probabilidad para el sector 3 y para el sector 4 es 2.1 y 2.13 veces mayor respectivamente, ceteris paribus.

En el caso de los trabajadores cuyos hogares están conformados por 4 o 5 personas, su probabilidad de necesitar un crédito es 1.38 veces mayor que aquellos hogares conformados de 1 a 3 miembros. En cuanto a los hogares de los trabajadores con más de 5 miembros dicha probabilidad aumenta 2.07 veces, si el resto de variables no cambia.

Cuando los trabajadores poseen vivienda propia, la probabilidad de necesitar financiamiento es 1.45 veces menor que aquellos trabajadores que residen en una vivienda arrendada, si las otras variables permanecen sin cambio.

Finalmente, cuando la persona es mujer y se encuentra casada, la probabilidad de necesitar un crédito es 1.52 veces menor con respecto a un hombre casado, si el resto de factores no cambia.

3.2.2. Pruebas de bondad de ajuste

Se empieza estudiando los test de la Devianza, Pearson y Hosmer Lemeshow; cabe señalar que, se dispone de una muestra grande y en su totalidad de variables categóricas, por lo cual la aplicación de las pruebas de la Devianza y Pearson son correctas, los estadísticos y su p-valor asociado se presentan el la tabla 3.5.

Prueba	Estadístico	p-valor
Devianza	3186.431	0.0000
Pearson	2960.888	0.0000
C_g	2946.000	0.0000

Tabla 3.5: Estadísticos de la Devianza, Pearson y Hosmer Lemeshow para el modelo de regresión logística. **Elaboración:** Autor.

Considerando un nivel de significacia $\alpha = 0.05$, los tres procedimientos arrojan p-valores inferiores, por lo tanto resultan significativos. En base a estos resultados, se puede asegurar que el modelo seleccionado se ajusta correctamente a los datos.

Del mismo modo, el coeficiente de determinación de McFadden para el modelo de regresión logística es igual a 0.2132. Valores entre 0.2 y 0.4 son considerados como un ajuste robusto y estadísticamente significativo [63].

Por otro lado, para medir la capacidad de discriminación del modelo se utiliza la prueba de Kolmogorov-Smirnov que calcula la máxima distancia entre las distribuciones de acumulación empíricas de la razón de verdaderos positivos (TPF) y la

razón de falsos positivos (FPF); la prueba supone un mayor poder predictivo cuanto mayor es la diferencia entre estas dos curvas [64].

En la figura 3.10 se muestra el gráfico de las distribuciones de acumulación empíricas junto con el estadístico *K-S*. En este caso, el valor de 0.457 indica una alta divergencia entre las distribuciones, permitiendo concluir que el modelo discrimina acertadamente.

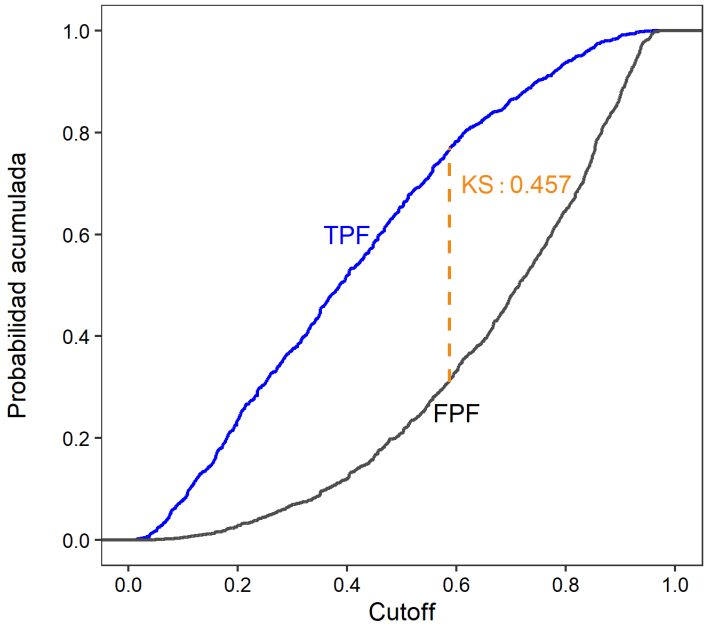


Figura 3.10: Distribución acumulada y estadístico *K-S*. **Elaboración:** Autor.

Sumando a esto, se observa que en la figura 3.11 el comportamiento de la curva ROC tiene una tendencia alejada de la recta $y = x$. En consecuencia, se puede afirmar que el modelo tiene una buena capacidad de discriminación.

De manera analítica, se calculan los índices de AUROC y GINI para avalar el resultado anterior; la tabla 3.6 muestra los valores obtenidos para el modelo estimado. Valores del índice de GINI entre 0.4 y 0.6 sugieren una desigualdad muy grande [65].

Medida	Logit
AUROC	0.798
GINI	0.596

Tabla 3.6: Índices de AUROC y GINI para la regresión logística. **Elaboración:** Autor.

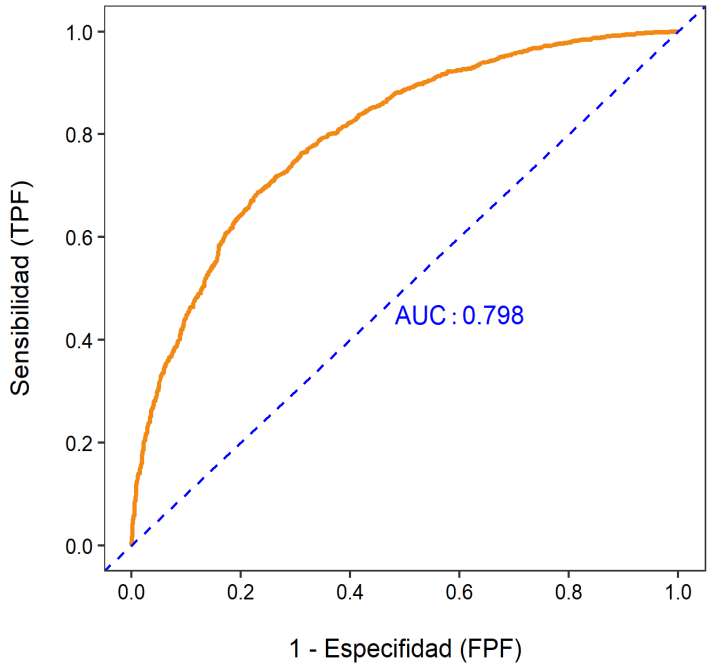


Figura 3.11: Curva ROC para el modelo logit. **Elaboración:** Autor.

En cuanto a la asertividad del modelo, se construye la tabla de clasificación tomando en cuenta el punto de corte donde se maximiza el *K-S*; para el presente modelo se obtuvo un punto de corte de 0.587.

Si se analiza la columna Ajuste de la tabla 3.7, se tiene un 72.37 % de asertividad global, por lo tanto, se puede concluir que los resultados de este modelo son bastante buenos.

Estimado	Observado		Total	Ajuste
	0	1		
0	1010	510	1520	76.86 %
1	304	1122	1426	68.75 %
Total	1314	1632	2946	72.37 %

Tabla 3.7: Tabla de clasificación del modelo logit. **Elaboración:** Autor.

Finalmente, al analizar el criterio de información de Akaike, el modelo presenta un valor de AIC = 3236.431.

3.2.3. Multicolinealidad

Con el fin de analizar la existencia de multicolinealidad en el modelo logit se calculan los GVIF de los parámetros estimados.

	GVIF	Df	GVIF^(1/(2*Df))
Tien_ahorros	1.08	1	1.04
Tien_deudas	1.62	4	1.06
Ing_durCOVID	1.66	2	1.14
Efect_GastoCOVID	1.04	1	1.02
Rango_edad	2.46	4	1.12
Sit_laboral	3.60	3	1.24
Act_Econom	2.04	3	1.13
Miemb_hog	1.17	2	1.04
Reg_tenenc	1.08	1	1.04
GeneroEstcivil	1.50	3	1.07

Tabla 3.8: Factor GVIF para los parámetros estimados del modelo logit. **Elaboración:** Autor.

La tabla 3.8 indica que el factor GVIF más alto es 3.60 asociado con el coeficiente de la variable `Sit_laboral`, seguido de la variable `Rango_edad` que tiene un factor generalizado de inflación de la varianza igual a 2.46 y de la variable que representa a la actividad económica con un GVIF = 2.04. Estas variable podrían analizarse detenidamente en el proceso de ajuste del modelo, sin embargo, al observar que todos los valores de la tabla 3.8 son inferiores a 5 y los errores estándar de los parámetros de la tabla 3.4 no se aprecian grandes, se puede concluir que no hay problemas de multicolinealidad.

3.2.4. Adecuación del modelo

Para analizar la presencia de observaciones atípicas, se exploran los residuos de Pearson y de la Devianza, mediante el gráfico de ajuste del modelo.

En la figura 3.12 se muestran los residuos de Pearson aplicados al modelo estimado. Si se consideran aquellos cuyo valor absoluto es superior a 2, se obtiene que el 4.34 % de los errores son significativos. De igual forma, en la figura 3.13 se presentan los residuos de la Devianza. Si se examinan aquellos con valor absoluto mayor a 2,

se obtiene que tan solo el 1.6 % de los errores corresponden a valores significativos.

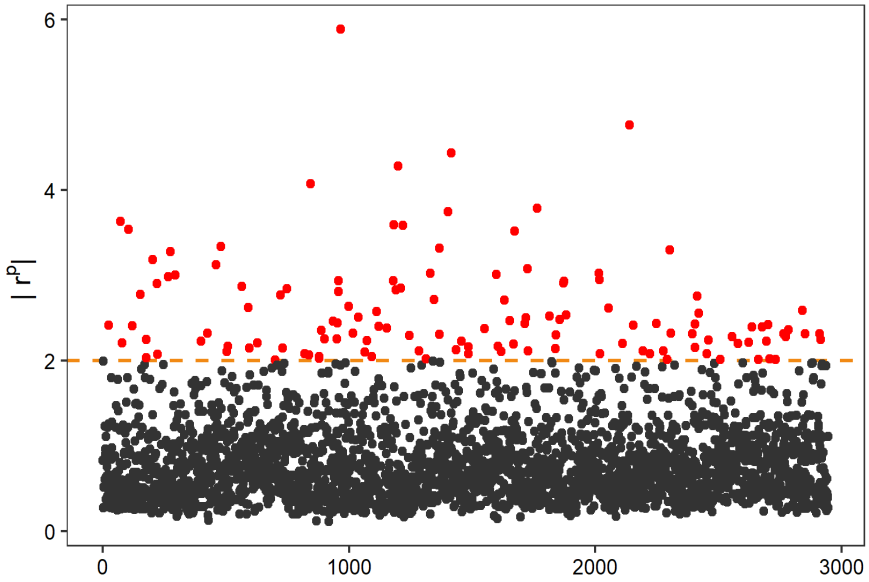


Figura 3.12: Residuos de Pearson del modelo logit estimado. **Elaboración:** Autor.

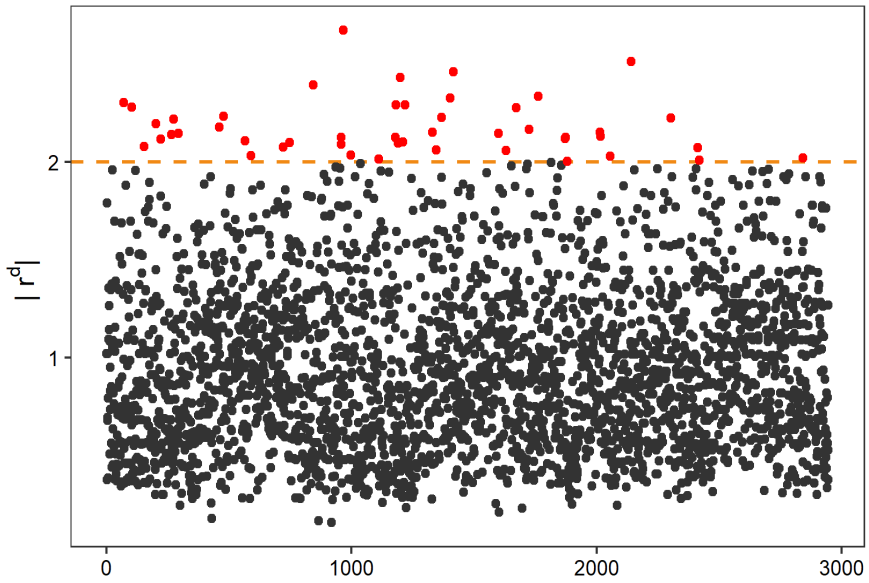


Figura 3.13: Residuos de la Devianza del modelo logit estimado. **Elaboración:** Autor.

Por otra parte, la prueba basada en las distancias de Cook para el modelo logit muestra que ningún valor es influyente[6].

[6]Las pruebas de Distancia de Cook para los diferentes modelos se las realiza mediante la función `cooks.distance()` de RStudio en el Anexo B.2

3.3. Modelo Probit

3.3.1. Estimación

Procediendo de manera análoga, se realiza la estimación para el modelo probit con las variables antes descritas.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.0886	0.1785	-6.10	0.0000	***
Tien_ahorrosSi	-0.5758	0.0574	-10.03	0.0000	***
Tien_deudasSecRegulado_Banca	0.6170	0.0750	8.22	0.0000	***
Tien_deudasSecRegulado_Otro	0.6472	0.1064	6.08	0.0000	***
Tien_deudasSec_NoRegulado	0.7586	0.1039	7.30	0.0000	***
Tien_deudasAmbosSectores	0.9169	0.0995	9.22	0.0000	***
Ing_durCOVIDNo estoy percibiendo ingresos durante la cuarentena	0.4126	0.1410	2.93	0.0034	**
Ing_durCOVIDDisminuido	0.4712	0.0624	7.55	0.0000	***
Efect_GastoCOVIDAumentado	0.2636	0.0536	4.91	0.0000	***
Rango_edadDe 25 a 34 años	0.3163	0.0856	3.70	0.0002	***
Rango_edadDe 35 a 44 años	0.6352	0.0947	6.70	0.0000	***
Rango_edadDe 45 a 60 años	0.6417	0.0965	6.65	0.0000	***
Rango_edadMayor a 60 años	0.5321	0.1522	3.50	0.0005	***
Sit_laboralAutónomo u Organizaciones de la EPS	-0.2567	0.1123	-2.29	0.0223	*
Sit_laboralEmpleado Público	-0.6399	0.0912	-7.02	0.0000	***
Sit_laboralEmpleado Privado	-0.5553	0.0971	-5.72	0.0000	***
Act_EconomSector 2	0.6861	0.1285	5.34	0.0000	***
Act_EconomSector 3	0.4278	0.1291	3.31	0.0009	***
Act_EconomSector 4	0.4369	0.1325	3.30	0.0010	***
Miemb_hogDe 4 a 5 Miembros	0.1938	0.0604	3.21	0.0013	**
Miemb_hogMás de 5 Miembros	0.4212	0.0770	5.47	0.0000	***
Reg_tenencPropio	-0.2179	0.0564	-3.86	0.0001	***
GeneroEstCivilHombre otro	-0.0176	0.0830	-0.21	0.8323	
GeneroEstCivilMujer casada	-0.2483	0.0873	-2.84	0.0045	**
GeneroEstCivilMujer otro	-0.1438	0.0812	-1.77	0.0764	.

[1]Nivel de significancia: ' *** ' 0.000 , ' ** ' 0.001, ' * ' 0.01, ' . ' 0.05, ' ' 0.1

[2]Las variables categóricas tienen como característica de referencia a los trabajadores que no poseen ahorros, no tienen deudas, sus ingresos durante la cuarentena se han aumentado, sus gastos en la cuarentena se han disminuido o se han mantenido, con edades entre los 18 y 24 años, desempleados, en el sector económico 1, con 1 a 3 miembros en el hogar, vivienda arrendada y hombres solteros.

Tabla 3.9: Regresión probit para el modelo de demanda de crédito. **Elaboración:** Autor.

En la tabla 3.9 se muestra la estimación de los coeficientes, errores estándar, valor del estadístico de Wald y el *p*-valor asociado; se puede observar que, exceptuando el coeficiente de la variable `GeneroEstCivil` para la categoría *Hombre otro*, todos los coeficientes son significativos al 95 % de confianza.

En este caso la interpretación de los coeficientes no es tan sencilla como en el modelo logit, sin embargo, existen formas limitadas en las que se puede interpretar

los coeficientes de la regresión probit [66]. Como se afirma en [67], un coeficiente positivo significa que un aumento en el predictor conduce a un aumento en la probabilidad predicha. Un coeficiente negativo significa que un aumento en el predictor conduce a una disminución en la probabilidad predicha.

En este sentido, los datos sugieren que la probabilidad de necesitar un crédito es mayor para los trabajadores que poseen obligaciones de deuda en los sectores regulados y no regulados por la banca. Asimismo, dicha probabilidad aumenta para los trabajadores que han dejado de percibir ingresos y para aquellos que han visto reducida su renta durante la crisis sanitaria, como también para los trabajadores cuyos gastos durante la cuarentena han sido mayores.

Los trabajadores con hogares formados por 5 o más miembros tienen mayor probabilidad de necesitar un crédito que los hogares con menos integrantes, al igual que aquellos trabajadores cuya actividad económica se encuentra ubicada en el sector 2. Cabe destacar que la probabilidad de necesitar un crédito aumenta mientras la edad del trabajador incrementa, pero cada vez en menor medida, hasta la edad de 60 años donde la probabilidad decrece.

Por otro lado, tener ahorros, contar con vivienda propia y estar empleado en el sector público, son características de los trabajadores que disminuyen la probabilidad de necesitar un crédito. De manera análoga, cuando la persona es mujer y se encuentra casada, es menos propensa a necesitar de financiamiento.

3.3.2. Pruebas de bondad de ajuste

Los estadísticos en la tabla 3.10 muestran que a un nivel de significancia del 5 %, las tres pruebas de bondad de ajuste resultan significativas, como resultado, el modelo probit estimado es adecuado para los datos.

Prueba	Estadístico	p-valor
Devianza	3187.863	0.0000
Pearson	2989.329	0.0000
C_g	2946.000	0.0000

Tabla 3.10: Estadísticos de la Devianza, Pearson y Hosmer Lemeshow para el modelo probit. **Elaboración:** Autor.

Al mismo tiempo, el coeficiente de determinación de McFadden, igual a 0.2128, indica un ajuste robusto y estadísticamente significativo.

Respecto al poder discriminativo del modelo, la prueba de Kolmogorov-Smirnov presenta un estadístico igual a 0.455; por consiguiente, la diferencia entre las distribuciones de acumulación empírica de la tasa de verdaderos positivos y falsos positivos es significativa (Figura 3.14).

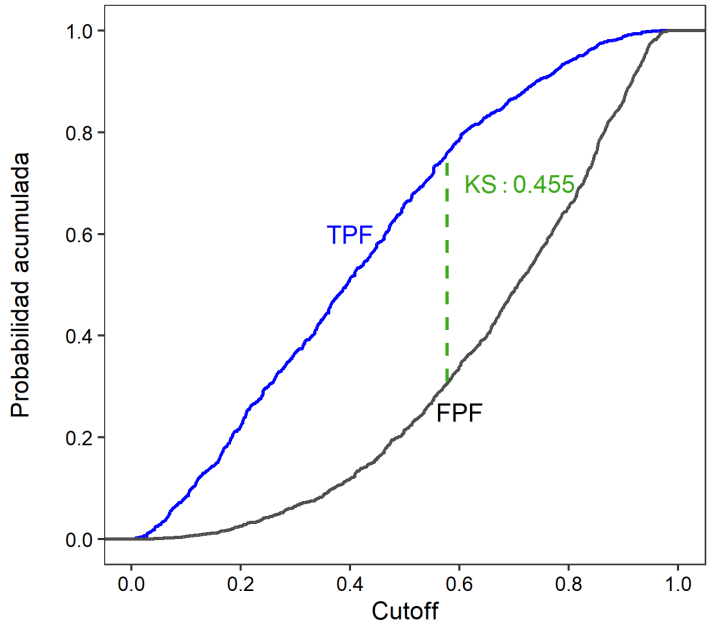


Figura 3.14: Distribución acumulada y estadístico *K-S* de la regresión probit. **Elaboración:** Autor.

Sumando a esto, el gráfico de la curva ROC (Figura 3.15) muestra un comportamiento alejado de la recta de clasificación aleatoria, con un índice de AUROC igual a 0.798; por ende, se puede considerar que el modelo tiene un alta capacidad predictiva. Además, el coeficiente de GINI igual a 0.595, indica una desigualdad significativa.

Por otra parte, tomando de referencia el punto de corte 0.578, correspondiente al valor donde se máxima el estadístico *K-S*, se elabora la tabla de clasificación (Tabla 3.11), la cual indica que el modelo tiene una asertividad global del 72.34 %.

Por último, al analizar el criterio de información de Akaike, el modelo presenta un valor de $AIC = 3237.9$.

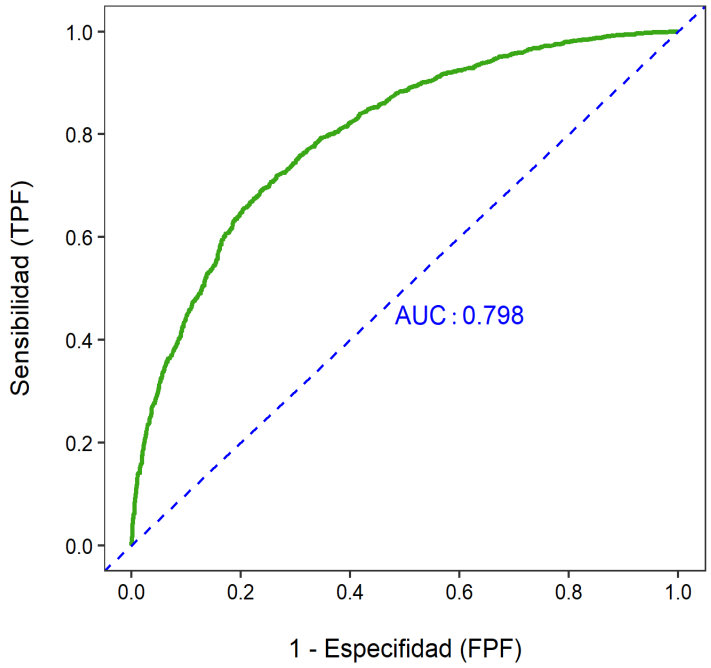


Figura 3.15: Curva ROC para el modelo probit. **Elaboración:** Autor.

Estimado	Observado		Total	Ajuste
	0	1		
0	1000	501	1501	76.10 %
1	314	1131	1445	69.30 %
Total	1314	1632	2946	72.34 %

Tabla 3.11: Tabla de clasificación del modelo probit. **Elaboración:** Autor.

3.3.3. Multicolinealidad

En la tabla 3.12 se presentan los factores generalizados de la inflación de la varianza para el modelo probit estimado; si se analizan los GVIF más altos, se observa que ningún valor es superior a la cota establecida para considerar problemas de multicolinealidad. Por lo tanto, se puede concluir que no existe relación de dependencia entre las variables explicativas.

	GVIF	Df	GVIF^(1/(2*Df))
Tien_ahorros	1.10	1.00	1.05
Tien_deudas	1.60	4.00	1.06
Ing_durCOVID	1.71	2.00	1.14
Efect_GastoCOVID	1.04	1.00	1.02
Rango_edad	2.38	4.00	1.11
Sit_laboral	3.55	3.00	1.24
Act_Econom	2.08	3.00	1.13
Miemb_hog	1.16	2.00	1.04
Reg_tenenc	1.09	1.00	1.04
GeneroEstcivil	1.48	3.00	1.07

Tabla 3.12: Factor GVIF para los parámetros estimados del modelo probit. **Elaboración:** Autor.

3.3.4. Adecuación del modelo

En cuanto a los supuestos para los residuos del modelo probit, se debe verificar que siguen una distribución normal con media cero y varianza finita [68]. Para dicho propósito, se empiezan revisando los gráficos exploratorios de normalidad.

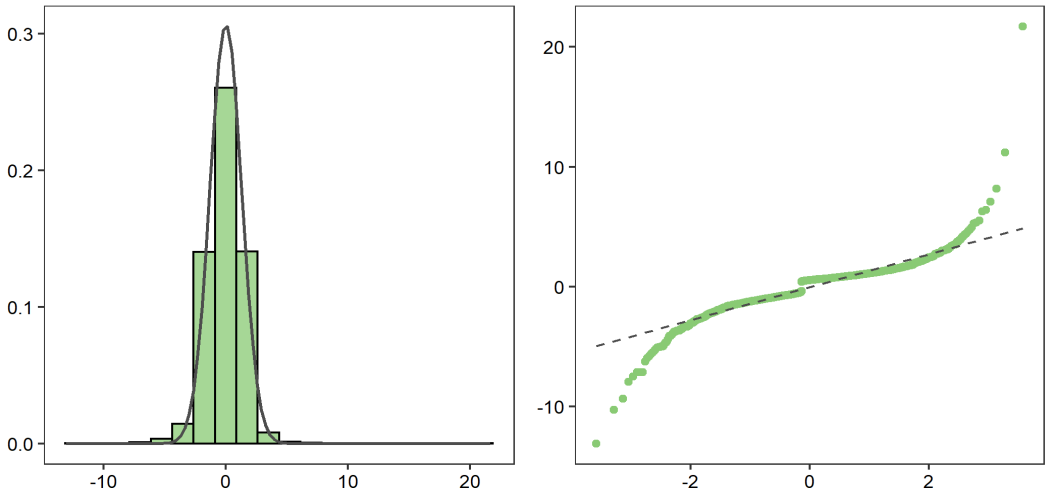


Figura 3.16: Gráficos exploratorios de normalidad. **Elaboración:** Autor.

En la figura 3.16 se muestran el histograma y el gráfico cuantil cuantil para los residuos del modelo probit. Como se puede observar, el histograma presenta simetría y similitud a una distribución normal; por el contrario, la gráfica cuantil cuantil parece tener valores extremos con un patrón de una distribución de colas pesadas.

De manera analítica, para contrastar el supuesto de normalidad de los residuos se realizan las pruebas de Kolmogorov-Smirnov, Anderson-Darling y Jarque Bera[7]. En la tabla 3.13 se muestran los valores del estadístico y el p-valor asociado de los test.

Prueba	Estadístico	p-valor
Kolmogorov-Smirnov	0.1674	0.0000
Anderson-Darling	72.074	0.0000
Jarque Bera	49538	0.0000

Tabla 3.13: Pruebas de normalidad para los residuos del modelo probit. **Elaboración:** Autor.

En todos los casos, los valores de probabilidad asociados son estadísticamente nulos. Por lo tanto, se puede concluir que los residuos del modelo siguen una distribución normal.

Por otro lado, en la figura 3.17 se presentan los residuos de Pearson para el modelo probit estimado. Si se analizan aquellos cuyo valor absoluto es mayor a 2, el 4.11 % de los errores se consideran significativos.

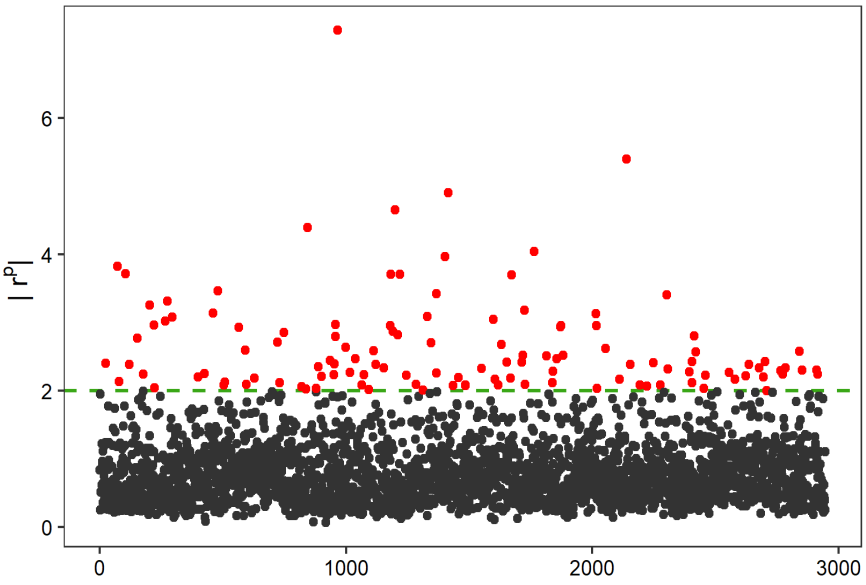


Figura 3.17: Residuos de Pearson del modelo probit estimado. **Elaboración:** Autor.

[7]Las pruebas de Anderson-Darling y Jarque Bera son test no paramétricos para comprobar si los datos de una muestra provienen de una distribución específica [69]

Asimismo, la figura 3.18 muestra los residuos de la Devianza. Si se toman en cuenta aquellos cuyo valor absoluto es superior a 2, se obtiene que el 1.56 % corresponden a errores significativos.

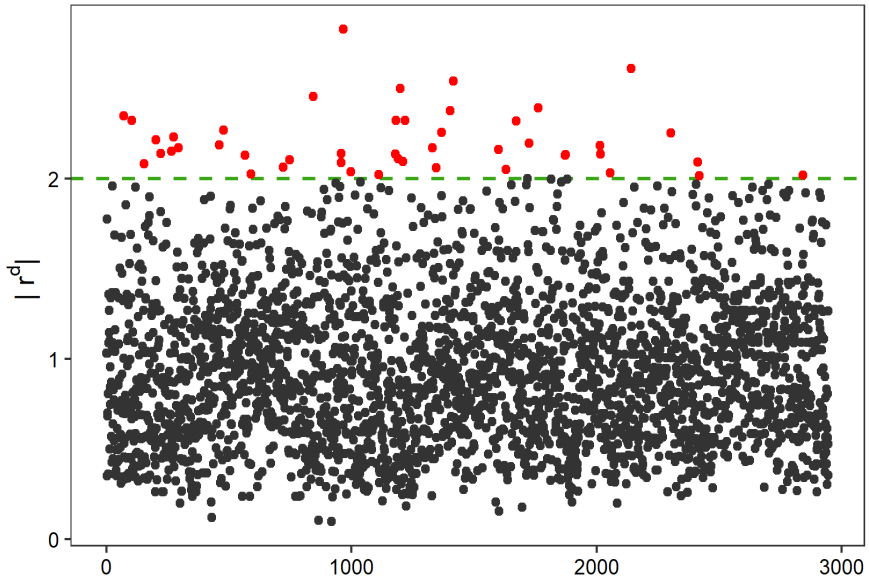


Figura 3.18: Residuos de la Devianza del modelo probit estimado. **Elaboración:** Autor.

Finalmente, la prueba de Cook muestra que la regresión probit estimada no posee ningún valor atípico considerable.

Capítulo 4

Resultados

En este capítulo se proceden a realizar las pruebas de bondad de ajuste de los modelos estimados anteriormente utilizando la muestra de validación. Además, se escogerá cuál modelo (logit o probit) se retiene para explicar la demanda de crédito en este trabajo.

4.1. Backtesting Modelo Logit

En la tabla 4.1 se presentan las medidas de McFadden, AUROC y coeficiente de GINI en las bases de modelamiento y validación.

Medida	Muestra de modelamiento	Muestra de validación
McFadden	0.2132	0.2062
AUROC	0.7980	0.7905
GINI	0.5960	0.5810

Tabla 4.1: Medidas de bondad de ajuste para el modelo logit en las bases de modelamiento y validación. **Elaboración:** Autor.

Se puede observar que, con respecto a las medidas obtenidas con la muestra de modelamiento, las pruebas de bondad de ajuste para la regresión logística en la muestra de validación no difieren de manera significativa.

De igual forma, como se muestra en la tabla 4.2, el ajuste global del modelo obtenido con la base de validación disminuye en 0.73 % con respecto al ajuste global obtenido con la base de modelamiento.

Estimado	Observado		Total	Ajuste
	0	1		
0	213	135	348	74.22 %
1	74	315	389	70.00 %
Total	287	450	737	71.64 %

Tabla 4.2: Tabla de clasificación del modelo logit para la muestra de validación. **Elaboración:** Autor.

Como resultado, el modelo logit es adecuado para explicar la demanda de crédito en Ecuador durante la crisis por la enfermedad COVID-19.

4.2. Backtesting Modelo Probit

En el caso del modelo probit, se empieza revisando los supuestos de normalidad para los residuos. Como se muestra en la figura 4.1, el histograma presenta un ajuste adecuado a una distribución normal pero el gráfico cuantil cuantil tiende a alejarse de la recta de referencia y parece tener valores extremos.

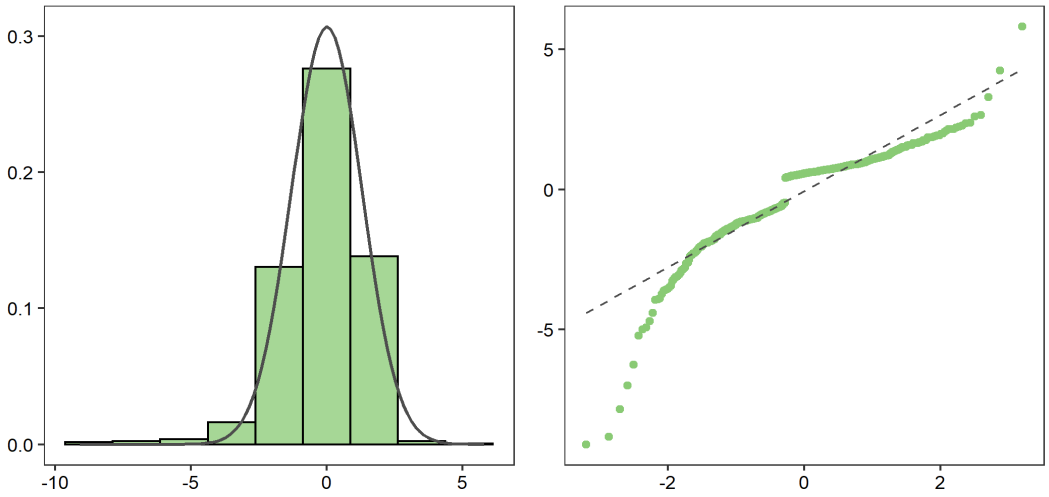


Figura 4.1: Gráficos exploratorios de normalidad para los residuos del modelo probit en la muestra de validación. **Elaboración:** Autor.

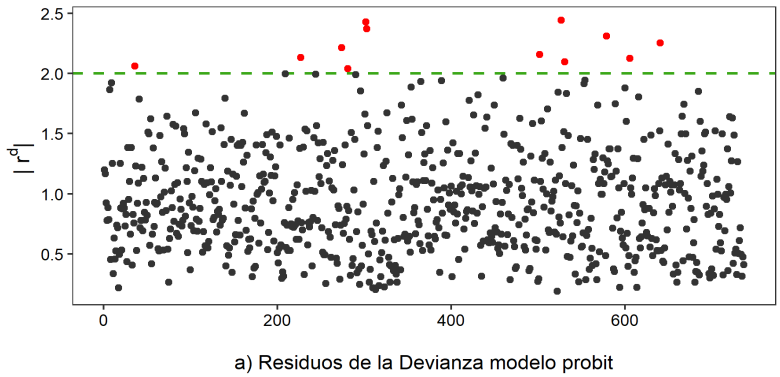
De manera analítica, para contrastar la hipótesis de normalidad de los residuos se realizan las pruebas de Kolmogorov-Smirnov, Anderson-Darling y Jarque Bera.

Prueba	Estadístico	p-valor
Kolmogorov-Smirnov	0.2247	0.0000
Anderson-Darling	27.113	0.0000
Jarque Bera	1327.1	0.0000

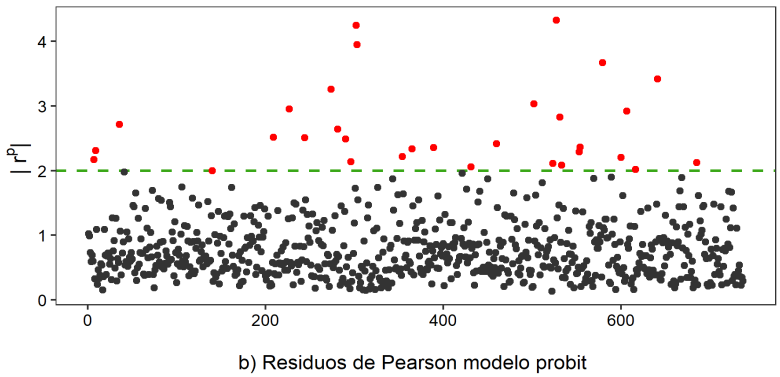
Tabla 4.3: Pruebas de normalidad para los residuos del modelo probit en la muestra de validación. **Elaboración:** Autor.

Como se muestra en la tabla 4.3, los valores de probabilidad asociados son estadísticamente nulos; por lo tanto, se puede afirmar que los residuos de la regresión probit siguen una distribución normal.

Por otro lado, si analizamos lo residuos de la Devianza en la figura 4.2(a) y consideramos aquellos cuyo valor absoluto es superior a 2, el 1.62 % de los errores corresponden a valores significativos. De la misma manera, si examinamos los residuos de Pearson en la figura 4.2(b) y consideramos aquellos cuyo valor absoluto es mayor a 2, el 4.2 % de los errores corresponden a valores significativos.



a) Residuos de la Devianza modelo probit



b) Residuos de Pearson modelo probit

Figura 4.2: Residuos de la Devianza y de Pearson para el modelo probit en la muestra de validación. **Elaboración:** Autor.

Además, la prueba de Distancia de Cook para los residuos de la regresión probit en la muestra de validación indica que ningún valor es superior a 1. En consecuencia, el modelo no presenta valores extremos significativos.

En cuanto a las medidas de bondad de ajuste, la tabla 4.4 indica que las pruebas aplicadas a la muestra de validación no difieren en gran medida de los valores obtenidos con la muestra de modelamiento.

Medida	Muestra de modelamiento	Muestra de validación
McFadden	0.2128	0.2063
AUROC	0.7980	0.7903
GINI	0.5950	0.5806

Tabla 4.4: Medidas de bondad de ajuste para el modelo probit en las bases de modelamiento y validación. **Elaboración:** Autor.

Asimismo, la tabla de clasificación (Tabla 4.5) para la regresión probit en la muestra de validación, indica que el ajuste global del modelo disminuye en 0.56 % con respecto al ajuste global del modelo obtenido con la muestra de modelamiento.

Estimado	Observado		Total	Ajuste
	0	1		
0	216	137	353	75.26 %
1	71	313	384	69.56 %
Total	287	450	737	71.78 %

Tabla 4.5: Tabla de clasificación del modelo probit para la muestra de validación. **Elaboración:** Autor.

En consecuencia, no existe diferencia en la asertividad de las dos muestras y por lo tanto el modelo probit es adecuado para explicar la demanda de crédito en Ecuador durante la crisis por la enfermedad COVID-19.

4.3. Elección del modelo

Para demostrar empíricamente cuál modelo tiene una capacidad predictiva mayor, se han tomado diferentes medidas de bondad de ajuste fuera de la muestra de modelamiento.

En base al criterio de la curva ROC, el modelo cuya gráfica se acerque más a la esquina superior izquierda tendrá mejor precisión.

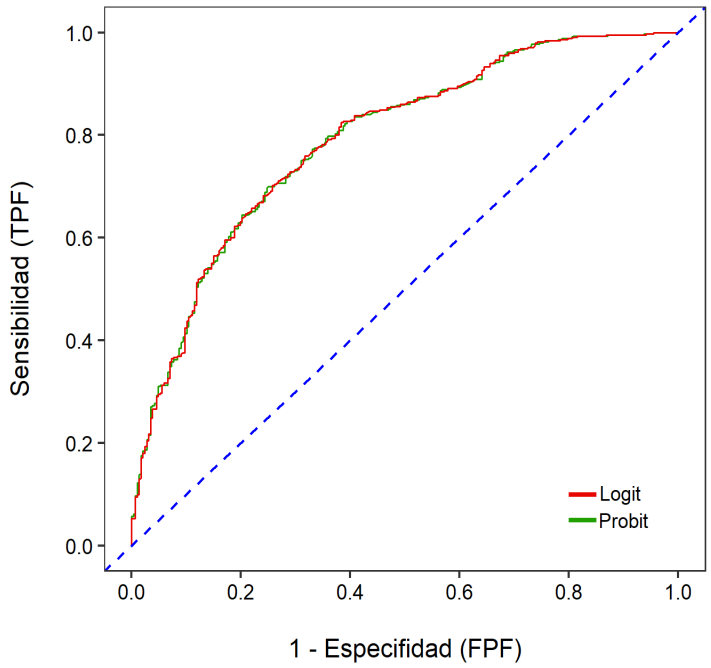


Figura 4.3: Comparaciones Curvas ROC de los modelos logit y probit en la muestra de validación. **Elaboración:** Autor.

Como se muestra en la figura 4.3, existe cierta complicación a la hora de determinar cuál se aproxima más a dicha esquina. Es por ello que se analizan los índices AUROC y GINI de cada modelo.

La tabla 4.6 indica que el modelo logit cuenta con un AUROC de 0.7905 y el modelo probit con un AUROC de 0.7903. Asimismo, el modelo logit cuenta con un GINI de 0.5810 y el modelo probit con un GINI de 0.5806. A pesar de que el modelo logit cuenta con los índices más elevado, la diferencia es mínima.

Medida	Modelo Logit	Modelo Probit
AUROC	0.7905	0.7903
GINI	0.5810	0.5806

Tabla 4.6: Comparación de los índices de AUROC y GINI de los modelos logit y probit en la muestra de validación. **Elaboración:** Autor.

De acuerdo con la información de la tabla 4.7, el modelo logit es el de menor tasa de falsos positivos, pero el de mayor tasa de falsos negativos. Además, tiene el

mayor indicador de sensibilidad, es decir, es el mejor para predecir los trabajadores que necesitarán de un crédito. Por último, en cuanto a la tasa de errores, el modelo probit es el de menor tasa.

Medida	Modelo Logit	Modelo Probit
Tasa de aciertos	71.64 %	71.78 %
Tasa de errores	28.36 %	28.22 %
Especificidad	74.22 %	75.26 %
Sensibilidad	70.00 %	69.56 %
Tasa de falsos negativos	25.78 %	24.74 %
Tasa de falsos positivos	30.00 %	30.44 %

Tabla 4.7: Medidas de predicción de los modelos en la base de validación. **Elaboración:** Autor.

En la figura 4.4 se muestran las curvas estimadas de las funciones de distribución de los modelos logit y probit.

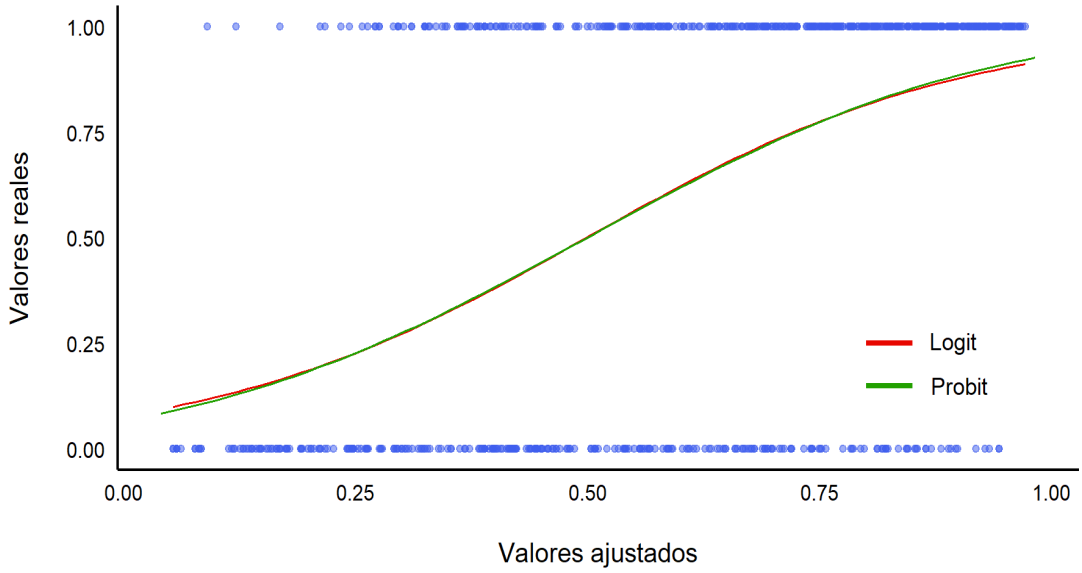


Figura 4.4: Curvas estimadas de las funciones de distribución acumuladas para los modelos logit y probit. **Elaboración:** Autor.

Como se puede observar, existe una gran concentración de observaciones con valor 1 para la variable dependiente (o trabajadores que necesitan crédito). Por tanto, el uso del modelo logit sería el más adecuado, debido a que su función de distribución tiene más masa en las colas que la distribución normal, y observaciones a las que se les asignarían una baja probabilidad con la regresión probit serían más comunes bajo la regresión logit.

Por otro lado, el modelo logit ha sido preferido para estudiar los determinantes de la necesidad de acceso a crédito en trabajos como Vizhñay y Samaniego [13], Carballo *et al.* [12], Botello [70], y Díaz [14]. La importancia de utilizar la regresión logit radica en la naturaleza cualitativa de las variables explicativas. Como menciona Enchautegui [71], si todas las variables exógenas son discretas, el modelo logit es recomendable debido a que permite una interpretación más clara de los coeficientes.

En consecuencia, la estimación logit es preferida para explicar la demanda de crédito en Ecuador durante la crisis por COVID-19.

Capítulo 5

Conclusiones y Recomendaciones

En la presente investigación se especificó un modelo de elección binaria para identificar los determinantes en la probabilidad de requerir un crédito por parte de los trabajadores públicos, privados, autónomos y desempleados en Ecuador durante la crisis por COVID-19. Se trabajó con la información recogida por la encuesta para evaluar los efectos de la crisis sanitaria sobre los trabajadores y las organizaciones del sector EPS, elaborada por los profesores del Departamento de Economía Cuantitativa y del Departamento de Matemática de la Escuela Politécnica Nacional.

Con el fin de determinar qué función de distribución de probabilidad asociada se escogería para el modelo de elección binaria, se desarrolló una propuesta metodológica con una serie de pasos importantes, que van desde la depuración de la base de datos, análisis descriptivo de la muestra, selección de variables, particionamiento de la muestra en una base de modelamiento y otra de validación, estimación simultánea de dos modelos (logit y probit) y finalmente la elaboración de varias pruebas estadísticas de bondad de ajuste sobre las bases antes mencionadas.

Se comprobó que los desempeños de los modelos logit y probit son muy similares, hallando una mínima diferencia en los resultados obtenidos con los distintos indicadores. En base a que, tanto los valores ligeramente superiores de AUROC, GINI y Sensibilidad del modelo logit sobre el probit en la muestra de validación, como la gráfica de las curvas estimadas de distribución 4.4 donde el modelo logit es más adecuado debido a la gran concentración de observaciones en las colas, y sobre todo a la múltiple literatura relacionada que utiliza en su mayoría un modelo logit sobre el probit, se estableció la función logística como la función de distribución de probabilidad asociada para el modelo de elección binaria.

En relación a las características que influyen en la probabilidad de necesitar un

crédito durante la crisis por COVID-19 en Ecuador, la evidencia empírica señala como principal determinante el no poseer ahorros. Otro factor importante es el tener deudas pendientes, siendo mayor la probabilidad cuando dicha deuda se establece con alguna entidad perteneciente al sector no regulado por la banca o cuando la misma se efectúa con diferentes entidades dentro y fuera del sistema financiero.

De la misma manera, la disminución o total pérdida de ingresos durante la pandemia se relaciona positivamente con la probabilidad de necesitar financiamiento, así como también el incremento de los gastos y el número de miembros en el hogar.

Adicionalmente, la edad guarda una relación positiva con la necesidad de acceder a un crédito aunque de modo decreciente cuando avanza el ciclo de vida. A su vez, trabajadores que se dedican a actividades económicas tales como: alojamiento y comida, agricultura, ganadería, selvicultura y pesca, comercio al por mayor y al por menor, construcción, transporte y almacenamiento, servicios diversos e industrias manufactureras son más propensos a demandar un crédito.

Por otro lado, las variables que influyen de manera negativa sobre la probabilidad de necesitar un financiamiento son: el tipo de vivienda (si el trabajador tiene vivienda propia), la situación laboral (si es trabajador autónomo o empleado en el sector público o privado), el género y estado civil (si es mujer y está casada).

Para futuras investigaciones se podría extender el estudio considerando un comportamiento no lineal de las variables explicativas a través de modelos GAM (Modelo Aditivo Generalizado). Además, se podría realizar una comparación entre los distintos indicadores de bondad de ajuste obtenidos, permitiendo determinar cuál tiene mejores resultados.

Finalmente, el modelo desarrollado en este trabajo puede ser aplicado como un instrumento para una adecuada planificación económica y financiera, que permita identificar las características de la población que más necesita a través de la creación de líneas especiales de crédito.

Bibliografía

[1] Organización Mundial de la Salud (OMS), "Nuevo coronavirus - China", 2020. [En línea]. Disponible: https://tinyurl.com/yyep5wwr/.

[2] ——, "Alocución de apertura del Director General de la OMS en la rueda de prensa sobre la COVID-19 celebrada el 11 de marzo de 2020", 2020. [En línea]. Disponible: https://tinyurl.com/y2wvhrx3.

[3] BBC News Mundo, "Estamos frente a una crisis generalizada del capitalismo democrático mundial y del no democrático, como el de China", 2020. [En línea]. Disponible: https://www.bbc.com/mundo/noticias-52055657.

[4] N. Fernandes, "Economic Effects of Coronavirus Outbreak (COVID-19) on the World Economy," *SSRN Electronic Journal*, 2020. DOI: 10.2139/ssrn.3557504.

[5] BBC News Mundo, "Coronavirus: el colapso en la economía china por el coronavirus (y por qué es una 'gran amenaza' para el mundo)", 2020. [En línea]. Disponible: https://www.bbc.com/mundo/noticias-internacional-51916056.

[6] ——, "Coronavirus en Estados Unidos. La pandemia dispara el desempleo: 10 millones en dos semanas, nuevo récord histórico", 2020. [En línea]. Disponible: https://www.bbc.com/mundo/noticias-52142353.

[7] CEPAL, NU, "Informe sobre el impacto económico en América Latina y el Caribe de la enfermedad por coronavirus (COVID-19)," 2020.

[8] A. Durojaiye, S. A. Yusuf y O. Balogun, "Determinants of Demand for Microcredit among Grain Traders in Southwestern States, Nigeria," *IOSR Journal of Agriculture and Veterinary Science*, vol. 7, n.º 11, págs. 01-09, 2014. DOI: 10.9790/2380-071130109.

[9] N. Loayza, "Causas y consecuencias de la informalidad en el Perú," *Revista Estudios Económicos*, n.º 15, págs. 43-64, 2008.

[10] M. Dini y G. Stumpo, *Mipymes en América Latina: un frágil desempeño y nuevos desafíos para las políticas de fomento*. CEPAL, 2018.

[11] F. Pucutay, "Los modelos logit y probit en la investigación social," *Centro de Investigación y desarrollo (CIDE), Perú*, 2002.

[12] I. E. Carballo, M. K. Grandes y L. V. Molouny, "Determinantes de la demanda potencial de microcrédito en Argentina," *Cuadernos de Administración*, vol. 29, n.º 52, pág. 199, 2016. DOI: 10.11144/javeriana.cao29-52.cddp.

[13] A. Vizhñay y A. Samaniego, "Determinantes del acceso al crédito en el Ecuador," *Revista Espacios, 40 (13)*, págs. 25-36, 2019.

[14] O. Díaz, "Determinantes del acceso al microcrédito para emprendedores bolivianos," Banco Central de Bolivia, inf. téc., 2008.

[15] Ministerio de Salud Pública, "Acuerdo *Nº* 00126-2020," inf. téc., 2020, Disponible: https://tinyurl.com/yyzjoa8o.

[16] Ministerio de Defensa Nacional, "Decreto Ejecutivo *Nº* 1017," inf. téc., 2020, Disponible: https://tinyurl.com/y5qvnl4q.

[17] M. Ruiz, F. Castellani y col., "El impacto del COVID-19 en las economías de la región (Centroamérica)," inf. téc., 2020. DOI: 10.18235/0002279.

[18] Ministerio del Trabajo, "Velamos por la estabilidad laboral", 2020. [En línea]. Disponible: https://tinyurl.com/y6sb5949.

[19] D. Gómez Cabrera, E. F. Sánchez Trujillo y col., "Modelos de Eleccion Discreta: Revisión y aplicación mediante cuadratura Gaussiana," Tesis de mtría., Universidad EAFIT, 2008.

[20] B. Alegre y G. Cahuana, "Modelos de elección binaria y su aplicación en el riesgo crediticio en la Caja Municipal Cusco," Tesis de pregrado, Universidad Nacional de San Antonio Abad del Cusco, 2020.

[21] S. Sarkar y H. Midi, "Importance of assessing the model adequacy of binary logistic regression," *Journal of Applied Sciences*, vol. 10, n.º 6, págs. 479-486, 2010.

[22] B. Granda y M. Elena, "Factores determinantes del acceso y racionamiento del crédito en las MIPYMES ecuatorianas," Tesis de mtría., FLACSO, Quito, Ecuador, 2011.

[23] D. N. Gujarati y D. C. Porter, *Econometría*, 5.ª ed. México, D. F.: McGraw-Hill, 2009.

[24] W. H. Greene, *Econometric Analysis*, 7.ª ed. United States: Pearson, 2012.

[25] K. G. King-keé, "Métodos de Mínimos Cuadrados Ponderados para la estimación de los Modelos Lineales Generalizados," Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Lima, Perú, 2001.

[26] Y. Cabrero y A. García, *Análisis Estadístico de Datos Espaciales con QGIS Y R*. Editorial UNED, 2015.

[27] J. Nelder y R. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society A.*, n.º 135, págs. 370-384, 1972.

[28] K. Knight, *Mathematical Statistics*. United States: Chapman & Hall/CRC, 2000.

[29] A. Agresti, *An Introduction to Categorical Data Analysis*, 2.ª ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2007.

[30] J. Aldrich, "R.A. Fisher and the making of maximum likelihood 1912-1922," *Statistical Science*, vol. 12, págs. 162-176, 1997. DOI: 10.1214/ss/1030037906.

[31] A. Novales, *Econometría*, 2.ª ed. Madrid: McGraw-Hill, 1993.

[32] R. Vélez y A. García, *Principios de inferencia estadística*. UNED, Universidad Nacional de Educación a Distancia, 2012.

[33] P. McCullagh y J. Nelder, *Generalized Linear Models*, 2.ª ed. Chapman & Hall, 1983.

[34] T. Figueroa, "La fecundidad y su relación con variables socioeconómicas, demográficas y educativas aplicando el Modelo de Regresión Poisson," Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Lima, Perú, 2005.

[35] A. J. Dobson y A. G. Barnett, *An Introduction to Generalized Linear Models*, 3.ª ed. Chapman & Hall/CRC, 2008.

[36] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 5.ª ed. South-Western CENGAGE Learning, 2012.

[37] R. D. Luce, *A Theory of Individual Choice Behavior*. John Wiley y Sons, 1954.

[38] D. McFadden y col., "Conditional logit analysis of qualitative choice behavior," 1973.

[39] K. Train, *Métodos de elección discreta con simulación*, 2.ª ed. 2009.

[40] J. H. Stock y M. W. Watson, *Introducción a la Econometría*, 3.ª ed. Madrid: Pearson Education, S. A., 2012.

[41] C. Hernández y col., "Mínimos cuadrados versus verosimilitud," *Ciencia y Mar, IX*, vol. 27, págs. 41-45, 2005.

[42] L. Thurstone, "A Law of Comparative Judgement," *Psychological Review 34*, págs. 273-286, 1927.

[43] E. D. Hahn y R. Soyer, "Probit and logit models: Differences in the multivariate realm," *The Journal of the Royal Statistical Society, Series B*, págs. 1-12, 2005.

[44] E. Medina, "El uso de los modelos de elección discreta para la predicción de crisis cambiarias: El caso latinoamericano," Tesis doct., Universidad Autónoma de Madrid, 2003.

[45] L. Cayuela, "Modelos Lineales Generalizados (GLM)," *Materiales de un curso del R del IREC*, 2009.

[46] T. Iglesias, "Métodos de Bondad de Ajuste en Regresión Logística," Tesis de mtría., Universidad de Granada, España, 2013.

[47] D. Collett, *Modelling Binary Data*. Chapman y Hall/CRC, 2002. DOI: 10.1201/b16654.

[48] D. Hosmer y S. Lemeshow, *Applied Logistic regression*, 2.ª ed. Wiley, 1989.

[49] M. Ruiz y col., "Creación de un modelo de predicción de default para microempresas y emprendedores," Tesis de mtría., Colegio Universitario de Estudios Financieros.

[50] A. Pérez, "Modelo de activación de tarjetas de crédito en el mercado crediticio ecuatoriano a través de una metodología analítica y automatizada en R," Tesis de pregrado, Escuela Politécnica Nacional, 2014.

[51] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, n.º 8, págs. 861-874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

[52] R. Anderson, *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. 2007.

[53] T. Costa y col., "Bondad de ajuste y elección del punto de corte en regresión logística basada en distancias. Aplicación al problema de Credit Scoring," en *Anales del Instituto de Actuarios Españoles*, Instituto de Actuarios Españoles, 2012, págs. 19-40.

[54] T. B. Arnold y J. W. Emerson, "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions," *The R Journal*, vol. 3, n.º 2, pág. 34, 2011. DOI: 10.32614/rj-2011-016.

[55] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, n.º 253, págs. 68-78, 1951.

[56] A. Castro, "Regresión Lineal," en *Monografías de Matemática y Estadística*, Escuela Politécnica Nacional, 2008.

[57] C. Davis y col., "An example of dependencies among variables in a conditional logistic regression," *Modern statistical methods in chronic disease epidemiology*, págs. 140-147, 1986.

[58] S. Moreno, "El Modelo Logit Mixto para la construcción de un Scoring de Crédito," Tesis de mtría., Universidad Nacional de Colombia, 2013.

[59] J. Neter, C. Nachtsheim y col., *Applied linear statistical models*. McGraw-Hill Irwin New York, 2005.

[60] J. Fox y S. Weisberg, *An R companion to applied regression*, 3.ª ed. Sage publications, 2018.

[61] V. Almeida, *Censo de Población y Vivienda (CPV) 2010*, Instituto Nacional de Estadística y Censos (INEC), 2010. [En línea]. Disponible en: `https://tinyurl.com/y57psvlj`.

[62] Real Academia de Ingeniería, *"coerción"*, Diccionario Español de Ingeniería, 2020. [En línea]. Disponible en: `https://tinyurl.com/y3ugaub4`.

[63] D. McFadden, *Quantitative methods for analyzing travel behavior of individuals: some recent developments*. Institute of Transportation Studies, University of California Berkeley, 1977.

[64] A. Mondal, "Classifications in R: Response Modeling, Credit Scoring and Credit Rating using Machine Learning Techniques", 2016. [En línea]. Disponible: `https://tinyurl.com/yynkbt3y`.

[65] J. F. Gonzáles, "Distribución del ingreso y pobreza en América Latina," Tesis doct., Universidad Nacional Autónoma de México, 2013.

[66] N. Garrido, "Construcción de un Modelo de Scoring de aprobación para cartera comercial de una institución financiera pública mediante Modelos Aditivos Generalizados," Tesis de mtría., Escuela Politécnica Nacional, 2018.

[67] J. Bruin, "Probit Regression a Stata Annotated Output", UCLA: Statistical Consulting Group, 2011. [En línea]. Disponible: `https://stats.idre.ucla.edu/stata/ado/analysis/`.

[68] G. Serrano, "Observaciones Anómalas en Modelos de Variable Dependiente Cualitativa," Tesis doct., Universidad Complutense de Madrid, 1993.

[69] I. Pedrosa y col., "Pruebas de bondad de ajuste en distribuciones simétricas, ¿qué estadístico utilizar?" *Universitas psychologica*, vol. 14, n.º 1, págs. 245-254, 2015.

[70] H. Botello, "Determinantes del acceso al crédito: Evidencia a nivel de la firma en Bolivia," *Libre Empresa*, vol. 12, n.º 1, págs. 45-62, 2015.

[71] M. E. Enchautegui, "Módulo de estudio sobre modelos Probit y Logit," *Puerto Rico: Universidad de Puerto Rico*, 2000.

[72] J. Rojo, *Árboles de clasificación y regresión*. Madrid: Laboratorio de Estadística, 2005.

Anexos

Anexo A

Datos y Variables

A.1. Lista de variables

Variable	Tipo	Descripción
ID_respuesta	Numérico	Identificador de respuesta.
Pais	Carácter	País de origen.
Extranj_Ec_anios	Numérico	Corresponde al número de años viviendo en Ecuador, dado el caso que no sea ecuatoriano.
Visa_Cmigrat	Carácter	Tipo de visa o condición migratoria.
Provincia	Carácter	Provincia en la que reside.
Rango_edad	Factor	Edad.
Genero	Carácter	Género.
N_Instruccion	Factor	Nivel de instrucción alcanzado.
Est_civil	Carácter	Estado civil.
Miemb_hog	Numérico	Número de personas en el hogar.
Mpercp_ing	Numérico	Corresponde al número de personas en el hogar que son perceptoras de ingreso.
Reg_tenenc	Carácter	Tipo de vivienda.
Sit_laboral	Carácter	Situación actual en el mercado laboral.
Act_Econom	Carácter	Actividad económica a la que se dedica.
Act_EconomOtro	Carácter	Corresponde a una actividad económica en particular.
Seg_Mercado	Carácter	Segmento de mercado en el que se encuentra.
Seg_MercadoOtro	Carácter	Corresponde a un segmento de mercado en particular.
Sit_laboralCOVID	Carácter	Situación laboral a causa de la cuarentena.

Tabla A.1: Descripción de las variables en la base de datos. **Elaboración:** Autor.

Variable	Tipo	Descripción
Segur_med	Carácter	Tipo de seguro médico.
Ing_antCOVID	Carácter	Ingresos antes de la cuarentena.
Ing_durCOVID	Carácter	Variación de los ingresos durante la cuarentena.
Aumt_IngCOVID	Carácter	Corresponde al incremento de los ingresos durante la cuarentena.
Dismy_IngCOVID	Carácter	Corresponde a la disminución de los ingresos durante la cuarentena.
Gast_antCOVID	Carácter	Gastos antes de la cuarentena.
Efect_GastoCOVID	Carácter	Variación de los gastos durante la cuarentena.
Aumt_GastCOVID	Carácter	Corresponde al incremento de los gastos durante la cuarentena.
Dismy_GastCOVID	Carácter	Corresponde a la disminución de los gastos durante la cuarentena.
Tien_ahorros	Factor	Si tiene ahorros = 1 caso contrario 0.
Tien_deudas	Factor	Si tiene deudas = 1 caso contrario 0.
Neces_prestam	Factor	Si necesita un préstamo = 1 caso contrario 0.
Calif_neCredit	Factor	Necesidad por un crédito: 5 muy urgente y 1 nada urgente.
Monto_Dcredt	Carácter	Monto de crédito, dado que necesite un préstamo.
Periodic_Dconv	Carácter	Corresponde al período de pago que le convendría, dado que necesite un préstamo.
Couta_DPag	Carácter	Cuota de pago que le convendría pagar, dado que necesite un préstamo.
Recibe Bono	Factor	Recibe bono = 1 caso contrario 0.
Deud_Banco	Factor	Tiene deuda en banco = 1 caso contrario 0.
Deud_Cooper	Factor	Tiene deuda en cooperativa = 1 caso contrario 0.
Deud_cnfamil	Factor	Tiene deuda con algún familiar = 1 caso contrario 0.
Deud_persona	Factor	Tiene deuda con otra persona = 1 caso contrario 0.
Deud_fundac	Factor	Tiene deuda en fundación = 1 caso contrario 0.
Deud_casacomerc	Factor	Tiene deuda en casa comercial = 1 caso contrario 0.
Deud_tarjetacred	Factor	Tiene deuda con algún emisor de tarjeta de crédito = 1 caso contrario 0.

Tabla A.2: Descripción de las variables en la base de datos (Continuación). **Elaboración:** Autor.

A.2. Método CHAID

El método CHAID (Chi-squared Automatic Interaction Detection) es uno de los algoritmos más utilizados en la creación de árboles, puede trabajar con variables de tipo numérico, factor y carácter. Dada una variable predictora, el algoritmo agrupa aquellas categorías consideradas estadísticamente homogéneas y deja las categorías heterogéneas inalteradas. Luego, de entre todas las variables predictoras potenciales elige la que tenga el mayor coeficiente de χ^2 para formar la primera rama del árbol.

El coeficiente χ^2 mide la asociación entre dos variables nominales u ordinales y se define como [72]:

$$\chi^2 = \sum\sum \frac{(n_{i,j} - n'_{i,j})^2}{n'_{i,j}} \tag{A.1}$$

donde,

$n_{i,j}$ es la frecuencia observada de la celda (i, j),

$n'_{i,j}$ es la frecuencia esperada de la celda (i, j).

Valores cercanos a cero de este coeficiente indicaran que no hay asociación entre las variables. Por otro lado, valores grandes de este coeficiente indicaran la existencia de asociación entre las variables.

La salida del software AnswerTree para el árbol de decisión se muestra en la figura A.1. El programa muestra el coeficiente χ^2 (`Chi-square`) y el p-valor (`P-value`) de la prueba, al 95 % de confianza valores de `P-value` menores a 0.05 muestran asociación entre las variables.

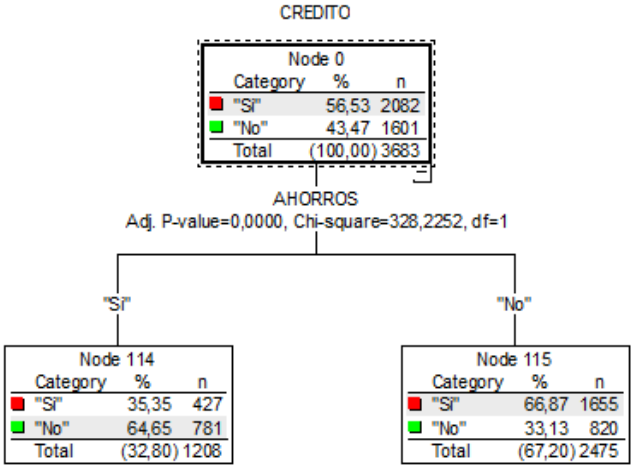


Figura A.1: Árbol de clasificación con el método CHAID. **Elaboración:** Autor.

En este caso, existe una clara asociación entre las variables `CREDITO` y `AHORROS`.

A.3. Agrupación de categorías usando árboles

Efect_GastoCOVID

Se forman 2 categorías:

- Se ha mantenido, Disminuido, y
- Aumentado

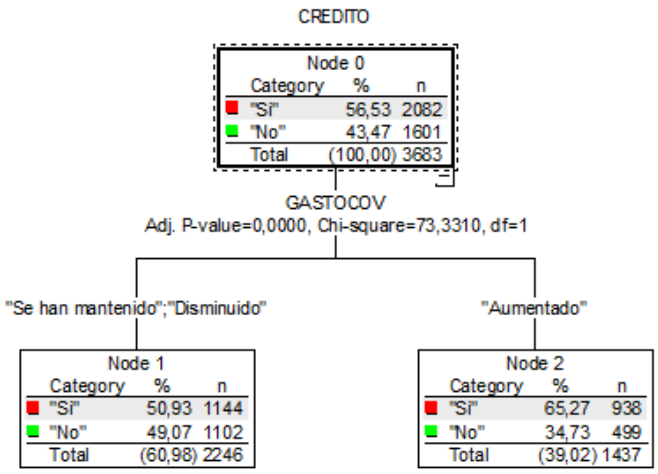


Figura A.2: Agrupación por variación de gastos durante la cuarentena. **Elaboración:** Autor.

Ing_durCOVID

Se forman 3 categorías:

- Aumentado, Se ha mantenido,
- Disminuido, y
- No estoy percibiendo ingresos durante la cuarentena.

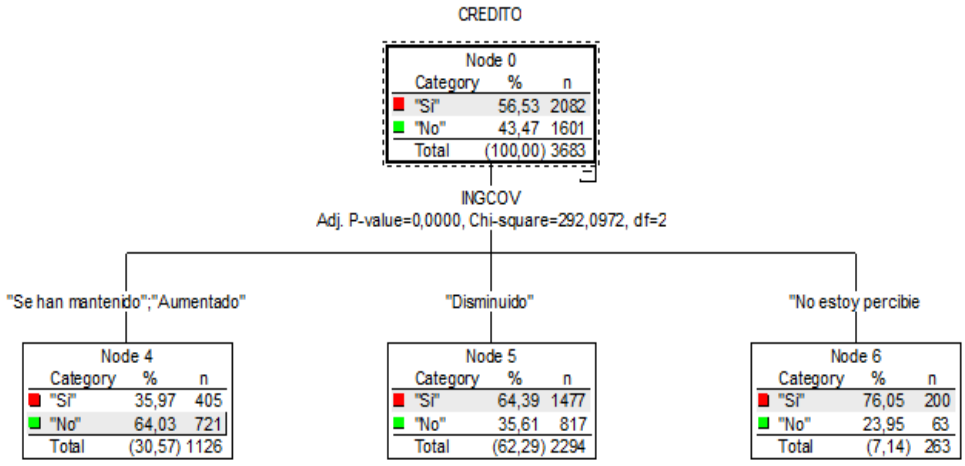


Figura A.3: Agrupación por variación de la renta durante la cuarentena. **Elaboración:** Autor.

Miemb_hog

Se forman 3 grupos:

- De 1 a 3 miembros,
- De 4 a 5 miembros, y
- Más de 5 miembros.

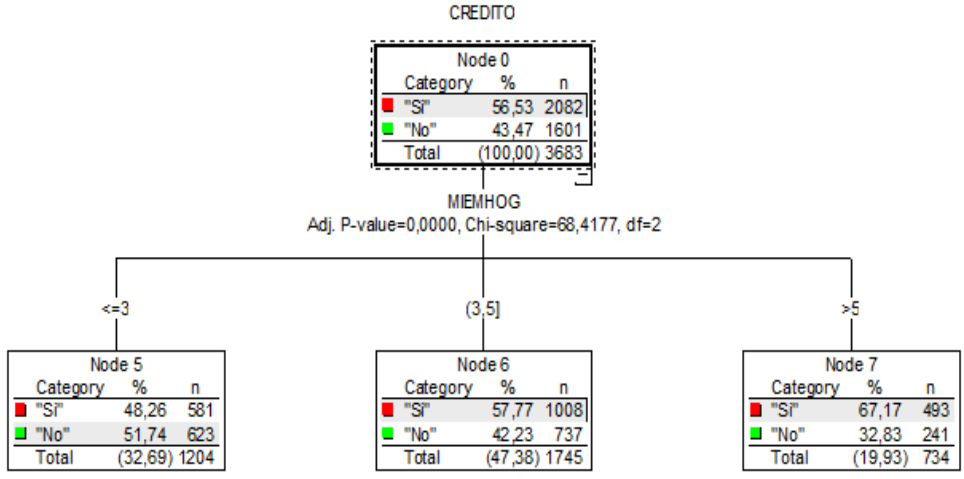


Figura A.4: Agrupación por número de personas en el hogar. **Elaboración:** Autor.

Act_Econom

Se forman 4 sectores para la agrupación de actividades económicas (Tabla A.3).

Sector	Actividad económica
Sector 1	Actividades inmobiliarias Actividades financieras y de seguros Información y comunicación
Sector 2	Actividades de alojamiento y de servicio de comida Agricultura, ganadería, silvicultura y pesca Comercio al por mayor Comercio al por menor Construcción Industrias manufactureras Otras actividades de servicios Servicio doméstico Transporte y almacenamiento
Sector 3	Artes, entretenimiento y recreación Enseñanza Otro Sin actividad por desempleo
Sector 4	Actividades de atención de la salud humana y de asistencia social Actividades profesionales, científicas y técnicas

Tabla A.3: Agrupación por actividad económica. **Elaboración:** Autor.

Tien_deudas

Se categoriza la deuda en 5 sectores (Tabla A.4).

Sector	Tipo de deuda
Sector regulado banca	Bancos Tarjetas de crédito
Sector regulado otro	Cooperativas Casa comercial Fundación
Sector no regulado	Otra persona Familiar
Ambos sectores	Sector regulado y no regulado
Ninguno	No tiene deudas

Tabla A.4: Categorización de la deuda por sectores. **Elaboración:** Autor.

Genero/Est_civil

Se utiliza un árbol de clasificación para estudiar la interacción entre las variables `Genero` y `Est_civil`. De esta manera, se forman 4 grupos: `Hombre casado`, `Hombre otro`, `Mujer casada`, y `Mujer otro`.

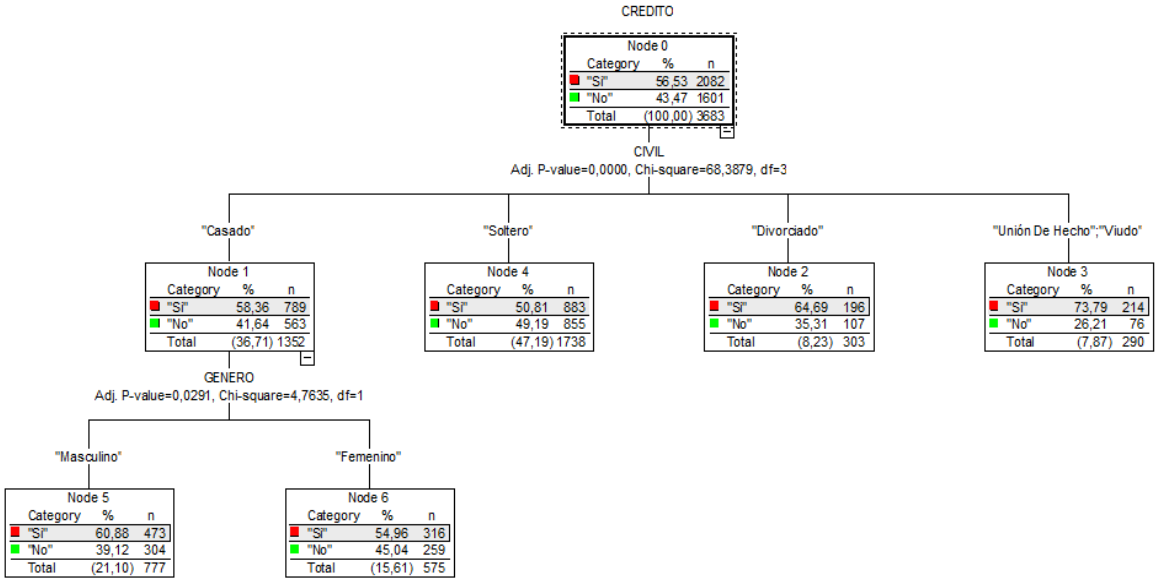


Figura A.5: Interacción de las variables: `Genero` y `Est_civil`. **Elaboración:** Autor.

Anexo B

Códigos en R

B.1. Estimación y validación de los modelos

```
#Carga de librerías
library(tidyverse) #paquete para manipulación de datos
library(stats) #paquete de modelos lineales generalizados
library(caTools) #paquete función (particionar)
library(xlsx) #paquete para leer y exportar archivos
library(xtable) #paquete para escribir en codigo LateX
library(InformationValue) #paquete para punto corte
library(grid) #paquete para manipulación de gráficas
library(vcd) #paquete para manipulación de gráficas
library(DescTools) #pruebas de bondad de ajuste
library(data.table) #paquete manipulación de datos
library(plotROC) #paquete para curva ROC
library(scales) #paquete de colores
library(ROCit) #paquete calculo curva ROC
library(vcdExtra) #paquete para matriz de confusión
library(ResourceSelection) #prueba Holsmer Lemeshow
library(MASS) #paquete para usar la funcion stepAIC
library(car) #prueba de multicolinealidad
library(nortest) #pruebas de normalidad
library(tseries) #Pruebas de normalidad

#Seteo de los directorios de trabajo
getwd()
setwd("C:\\Users\\pmoli\\Documents\\Base Poli COVID19")
dir()

#Lectura de la base de datos
```

```r
load("DataCOV19.RData")

#Revision de las variables
names(BDDCOV19)
str(BDDCOV19)

#Preparación de la data con las variables de la encuesta
data<-BDDCOV19%>%dplyr::select(ID_respuesta,Provincia,
                          Rango_edad, Genero, N_Instruccion,
                          Est_civil,Miemb_hog,Mpercp_ing,
                          Reg_tenenc, Sit_laboral,
                          Act_Econom,Sit_laboralCOVID,
                          Segur_med,Ing_antCOVID,
                          Ing_durCOVID,Gast_antCOVID,
                          Efect_GastosCOVID,Tien_ahorros,
                          Neces_prestam,`Recibe Bono`,
                          Deud_Banco,Deud_Cooper,Deud_cnfamil,
                          Deud_persona,Deud_fundac,
                          Deud_casacomerc,Deud_tarjetacred,
                          Tien_deudas)
rm(BDDCOV19)
names(data)

#Revision de la variable respuesta
tab1<-with(data, table(Neces_prestam))
tab1
#Porcentajes
round(prop.table(tab1)*100,2)
#No hay problemas de sesgo

#Agrupación de variables usando un arbol de
#decisión método CHAID

#Variable gastos durante cuarentena
table(data$Efect_GastosCOVID)
data<-data%>%mutate(Efect_GastosCOVID=ifelse(
      Efect_GastosCOVID=="Se han mantenido" |
      Efect_GastosCOVID=="Disminuido",
      "Se han mantenido o se han disminuido",
      "Aumentado"))
table(data$Efect_GastosCOVID)

#Variable ingresos durante la cuarentena
table(data$Ing_durCOVID)
```

```
data<-data%>%mutate(Ing_durCOVID=ifelse(
      Ing_durCOVID=="Se han mantenido" |
      Ing_durCOVID=="Aumentado",
      "Se ha mantenido o aumentado",
      Ing_durCOVID))
table(data$Ing_durCOVID)

#Variable miembros en el hogar
data<-data%>%mutate(Miemb_hog=ifelse(
      Miemb_hog<=3,"De 1 a 3 Miembros",
      ifelse(Miemb_hog>3 & Miemb_hog<=5,
      "De 4 a 5 Miembros",
      "Más de 5 Miembros")))
#Coercion a Factor
data<-data%>%mutate(Miemb_hog=factor(Miemb_hog))
with(data,table(Miemb_hog))

#Variable Actividad Económica
sect1<-c("Actividades de atención de la salud humana
          y de asistencia social",
         "Actividades profesionales, científicas y
          técnicas")
sect2<-c("Actividades de alojamiento
          y de servicio de comida",
         "Agricultura, ganadería, silvicultura
          y pesca",
         "Comercio al por mayor",
         "Comercio al por menor","Construcción",
         "Industrias manufactureras",
         "Otras actividades de servicios",
         "Servicio doméstico",
         "Transporte y almacenamiento")
sect3<-c("Artes, entretenimiento y recreación","Otro",
         "Sin actividad por desempleo","Enseñanza")
sect4<-c("Actividades inmobiliarias",
         "Actividades financieras y de seguros",
         "Información y comunicación")
data<-data%>%mutate(Act_Economica=ifelse(
      Act_Econom %in% sect4, "Sector 1",
      ifelse(Act_Econom %in% sect2, "Sector 2",
      ifelse(Act_Econom %in% sect3, "Sector 3",
      "Sector 4"))))
table(data$Act_Economica)

```

```
#Variable deuda
data<-data%>%mutate(Tien_deudas=ifelse(
      (Deud_Banco=="Sí"| Deud_tarjetacred=="Sí") &
      (Deud_cnfamil=="No"&Deud_persona=="No"),
      "SecRegulado_Banca",
      ifelse((Deud_Cooper=="Sí"| Deud_fundac=="Sí"|
      Deud_casacomerc=="Sí") &
      (Deud_cnfamil=="No"&Deud_persona=="No"),
      "SecRegulado_Otro",
      ifelse((Deud_cnfamil=="Sí"| Deud_persona=="Sí") &
      (Deud_Banco=="No" & Deud_Cooper=="No" &
      Deud_fundac=="No"& Deud_casacomerc=="No" &
      Deud_tarjetacred=="No"),"Sec_NoRegulado",
      ifelse(Deud_Banco=="No"& Deud_Cooper=="No" &
      Deud_cnfamil=="No"& Deud_persona=="No" &
      Deud_fundac=="No" & Deud_casacomerc=="No" &
      Deud_tarjetacred=="No","No","AmbosSectores")))))
table(data$Tien_deudas)
#Se ordena las categorías para la estimación
data<-data%>%mutate(Tien_deudas=factor(Tien_deudas,
      levels=c("No","SecRegulado_Banca",
      "SecRegulado_Otro","Sec_NoRegulado",
      "AmbosSectores")))

#Interacción variable genero con estado civil
table(data$Est_civil)
table(data$Genero)
data<-data%>%mutate(GeneroEstcivil=ifelse(
      Genero=="Masculino" & Est_civil=="Casado",
      "Hombre casado",
      ifelse(Genero=="Masculino"&Est_civil=="Soltero",
      "Hombre otro",
      ifelse(Genero=="Masculino"&Est_civil=="Divorciado",
      "Hombre otro",
      ifelse(Genero=="Masculino"&
      Est_civil=="Unión De Hecho", "Hombre otro",
      ifelse(Genero=="Masculino" & Est_civil=="Viudo",
      "Hombre otro",
      ifelse(Genero=="Femenino"&Est_civil=="Casado",
      "Mujer casada",
      ifelse(Genero=="Femenino" & Est_civil=="Soltero",
      "Mujer otro","Mujer otro"))))))))
table(data$GeneroEstcivil)

```

```r
#NOTA:
#Para correr el modelo se debe fijar la categoria de
#referencia en algunas variables ya que
#el paquete R no permite especificar se debe realizar
#manualmente
#Especificar el nivel de referencia para el modelo

#Variable Sit Laboral Categoria de Referencia "Desempleado"
data<-data%>%mutate(Sit_laboral=factor(Sit_laboral,
             levels = c("Desempleado",
             "Autónomo u Organizaciones de la EPS",
             "Empleado Público","Empleado Privado")))

#Variable Ingreso Durante Cuarentena Categoria de referencia
#"Ingresos se han mantenido o aumentado"
data<-data%>%mutate(Ing_durCOVID=factor(Ing_durCOVID,
             levels = c("Se ha mantenido o aumentado",
             "No estoy percibiendo ingresos durante
             la cuarentena","Disminuido")))

#Variable Gasto durante la cuarentena categoría de referencia
#"gastos se han mantenido o se han disminuido"
data<-data%>%mutate(Efect_GastosCOVID=factor(
      Efect_GastosCOVID,
      levels = c("Se han mantenido o se
      han disminuido", "Aumentado")))

#Marca modelamiento / validación (80% / 20%)
set.seed(12345)
sample<-sample.split(data$ID_respuesta, SplitRatio = 0.8)
mod<-setDT(subset(setDF(data), sample == TRUE))
val<-setDT(subset(setDF(data), sample == FALSE))

################Estimación Modelo Logit#########################

modLOGit<-glm(Neces_prestam ~ Tien_ahorros+Tien_deudas+
                              Ing_durCOVID+
                              Efect_GastosCOVID+
                              Rango_edad+Sit_laboral+
                              Act_Economica+
                              Miemb_hog+
                              Reg_tenenc+
                              GeneroEstcivil,
              family=binomial(link = "logit"),data=mod)
```

```
204
205 #Resumen del modelo
206 summary(modLOGit)
207
208 #Calculo de los Odds
209 odds<-round(exp(coefficients(modLOGit)),4)
210 odds
211
212 #Pruebas de bondad de ajuste
213
214 #Estadístico de la Devianza
215 devLoG<-sum(residuals(modLOGit,type="deviance")^2)
216 pvalue_devLog<-1-pchisq(devLoG,
217                  modLOGit$df.null-modLOGit$df.residual)
218 pvalue_devLog
219 #Estadístico de la Devianza individual
220 #dev_ind<-anova(modLOGit,test="Chisq")
221 #dev_ind
222
223 #Estadístico de Pearson
224 pearLog<-sum(residuals(modLOGit,type="pearson")^2)
225 pvalue_pearLog<-1-pchisq(pearLog,
226                   modLOGit$df.null-modLOGit$df.residual)
227 pvalue_pearLog
228
229 #Prueba de Holsmer Lemoshow Cg
230 hoslem.test(mod$Neces_prestam, fitted(modLOGit))
231
232 #Calculo del Pseudo R2 de McFadden
233 rMcLoG<-PseudoR2(modLOGit, which="McFadden")
234 rMcLoG
235 #Alternativa de calculo
236 1-modLOGit$deviance/modLOGit$null.deviance
237
238 #Grafico Curva ROC
239 mod<-mod%>%mutate(Neces_prestam=
240            ifelse(Neces_prestam=="Si",1,0))
241 mod<-mod%>%mutate(Predichos=predict(
242            modLOGit, mod, type="response"))
243
244 #Elementos de la grafica para calcular
245 #el área bajo la curva AUC
246 p1<-ggplot(mod,aes(d=Neces_prestam,m=Predichos))
247     + geom_roc()
```

```
p1
#Curva ROC
logroc<-ggplot(mod,aes(d=Neces_prestam,
        m=Predichos)) + theme_bw()+
        geom_roc(n.cuts=0, colour="#f18914") +
        theme(axis.text=element_text(colour =
        "black"),
        plot.title = element_text(hjust = 0.5))+
        scale_x_continuous("\n1 - Especifidad (FPF)",
        breaks=seq(0, 1, by = .2))+
        scale_y_continuous("Sensibilidad (TPF)\n",
        breaks = seq(0, 1, by = .2)) +
        geom_abline(intercept=0, slope=1,
        colour="blue", linetype="dashed") +
        annotate("text", x=0.6, y=0.45, parse=TRUE,
        label=paste0("AUC: ",round(calc_auc(p1)$AUC,3)),
        colour="blue")+
        ggExtra::removeGridX()+
        ggExtra::removeGridY()
logroc
#Indice de Gini
#2*AUROC-1
gini<-2*round(calc_auc(p1)$AUC,4)-1
gini<-gini*100
gini

#Estadistico Kolmogorov-Smirnov
dres<-data.frame(pred=predict(modLOGit,
      mod, type="response"),
      var=mod$Neces_prestam)
ROC<-rocit(score=dres$pred,class=dres$var)
ksplot<-ksplot(ROC)
#Calculo del punto optimo de corte
cutoff<-ksplot$`KS Cutoff`
#Calculo del estadístico KS
kstat<-as.numeric(ksplot$`KS stat`)

#Grafica K-S
ksploti<-ggplot(mod, aes(x=Predichos,
         group=Neces_prestam,
         color=Neces_prestam))+
         stat_ecdf(size=.7) +
         scale_x_continuous("Cutoff",
         breaks = seq(0,1,.2), limits = c(0,1))+
```

```
292         scale_y_continuous("Probabilidad acumulada\n",
293         breaks = seq(0,1,.2))+
294         annotate("segment", x=cutoff, xend=cutoff,
295         y=.31, yend=.77, colour="#f18914", size=.7,
296         linetype="dashed")+
297         theme_bw()+
298         scale_colour_gradient(low="blue",high="gray30")+
299         ggExtra::removeGridX()+
300         ggExtra::removeGridY()+
301         theme(legend.position="none",
302         axis.text = element_text(colour = "black"))+
303         annotate("text", x=.71, y=.7, parse=TRUE,
304         label=paste0(" ",'KS: ',round(kstat,3)),
305         colour="#f18914") +
306         annotate("text", x=0.4, y=0.6, parse=TRUE, size=4,
307         label="TPF", colour="blue")+
308         annotate("text", x=0.6, y=0.25, parse=TRUE, size=4,
309         label="FPF", colour="black")
310 ksploti
311
312 #Tabla de clasificación
313 res<-predict(modLOGit,mod,type="response")
314 res<-ifelse(res>cutoff,1,0)
315 mc<-table(res,mod$Neces_prestam)
316 names(mc)<-c("Si","No")
317 mc[1,1] # Verdaderos positivos
318 mc[2,2] # Verdaderos negativos
319 mc[1,2] # Falsos positivos
320 mc[2,1] # Falsos negativos
321 prop.table(mc)*100
322 #Porcentaje global
323 round((mc[1,1]+mc[2,2])*100/sum(mc),2)
324
325 #Residuos de pearson
326 respearson<-residuals(modLOGit,type="pearson")
327 respearson
328
329 #Gráfico de residuos
330 pearres<-ggplot(mod,
331         aes(x=seq(1,2946,1),y=abs(respearson)))+
332         geom_hline(yintercept = 2, color = "#f18914",
333         linetype = "dashed", size=0.7) +
334         geom_point(aes(color=ifelse(abs(respearson)>=2,
335         'red', 'gray20')), size=1.5) +
```

```
          scale_color_identity() + theme_bw() +
          theme(axis.text.x=element_text(colour="black"),
          axis.title.x=element_blank(),
          axis.text.y=element_text(colour="black"),
          axis.title.y=element_text(size = 12, angle=90,
          vjust = 0.5, hjust = 0.5)) +
          ggExtra::removeGridX()+
          ggExtra::removeGridY()+
          ylab(expression(paste("| ",r^{p},"| ")))
pearres
#Porcentaje errores significativos
table(abs(respearson)>2)

#Residuos de la devianza
resdeviance<-residuals(modLOGit3,type="deviance")
resdeviance

#Gráfico de los residuos
desvres<-ggplot(mod,aes(x=seq(1,2946,1),
          y = abs(resdeviance))) +
          geom_hline(yintercept = 2, color = "#f18914",
          linetype = "dashed", size=0.7) +
          geom_point(aes(color=ifelse(abs(resdeviance)>=2,
          'red', 'gray20')), size=1.5) +
          scale_color_identity() + theme_bw() +
          theme(axis.text.x = element_text(colour = "black"),
          axis.title.x = element_blank(),
          axis.text.y = element_text(colour = "black"),
          axis.title.y = element_text(size = 12, angle=90,
          vjust = 0.5, hjust = 0.5)) +
          ggExtra::removeGridX()+
          ggExtra::removeGridY()+
          ylab(expression(paste("| ",r^{d},"| ")))
desvres
#Valores significativos
table(abs(resdeviance)>2)

#Multicolinealidad de los predictores
viflog<-car::vif(modLOGit)
viflog

#Distancias de Cook
cook<-cooks.distance(modLOGit)
#valores influyentes
```

```
significativas<-cook>1
table(significativas)


################Estimación Modelo Probit#########################
modProb<-glm(Neces_prestam ~ Tien_ahorros+Tien_deudas+
                             Ing_durCOVID+
                             Efect_GastosCOVID+
                             Rango_edad+Sit_laboral+
                             Act_Econom+
                             Miemb_hog+
                             Reg_tenenc+
                             GeneroEstcivil,
          family=binomial(link = "probit"),data=mod)
summary(modProb)

#Estadistico de la Devianza
desvprob<-sum(residuals(modProb,type="deviance")^2)
pvalue<-1-pchisq(desvprob,
          modProb$df.null-modProb$df.residual)

#Estadístico de Pearson
pearprob<-sum(residuals(modProb,type="pearson")^2)
pvalue<-1-pchisq(pearprob,
          modProb$df.null-modProb$df.residual)

#Prueba de Holsmer Lemoshow Cg
hoslem.test(mod$Neces_prestam, fitted(modProb))

#Calculo del Pseudo R2
PseudoR2(modProb, which="McFadden")

#Grafico Curva ROC
mod<-mod%>%mutate(Neces_prestam=ifelse(
     Neces_prestam=="Si",1,0))
mod<-mod%>%mutate(Predichos=predict(
     modProb, mod, type="response"))

#Calculo el área bajo la curva AUC
p1<-ggplot(mod, aes(d=Neces_prestam,m=Predichos))
    + geom_roc()
p1
#Curva ROC
logroc<-ggplot(mod,aes(d=Neces_prestam,m=Predichos))+
```

```
        theme_bw()+
        geom_roc(n.cuts = 0, colour="#3AA717") +
        theme(axis.text = element_text(colour = "black"),
        plot.title = element_text(hjust = 0.5))+
        scale_x_continuous("\n1 - Especifidad (FPF)",
        breaks = seq(0, 1, by = .2))+
        scale_y_continuous("Sensibilidad (TPF)\n",
        breaks = seq(0, 1, by = .2)) +
        geom_abline(intercept=0, slope=1, colour="blue",
        linetype="dashed") +
        annotate("text", x=0.6, y=0.45, parse=TRUE,
        label=paste0("AUC: ",round(calc_auc(p1)$AUC,3)),
        colour="blue")+
        ggExtra::removeGridX()+
        ggExtra::removeGridY()
logroc

#Indice de Gini
#2*AUROC-1
gini<-2*round(calc_auc(p1)$AUC,4)-1
gini<-gini*100
gini

#Estadistico Kolmogorov-Smirnov
dres<-data.frame(pred=predict(modProb, mod,
      type="response"), var=mod$Neces_prestam)
ROC<-rocit(score=dres$pred, class=dres$var)
ksplot<-ksplot(ROC)
#Calculo del punto optimo de corte
cutoff<-ksplot$`KS Cutoff`
#Calculo del estadístico KS
kstat<-as.numeric(ksplot$`KS stat`)

#Grafica K-S
ksploti<-ggplot(mod,aes(x=Predichos,
         group=Neces_prestam, color=Neces_prestam))+
         stat_ecdf(size=.7) +
         scale_x_continuous("Cutoff",
         breaks = seq(0,1,.2), limits = c(0,1))+
         scale_y_continuous("Probabilidad acumulada\n",
         breaks = seq(0,1,.2))+
         annotate("segment", x=cutoff, xend=cutoff,
         y=.31, yend=.76, colour="#3AA717", size=.7,
         linetype="dashed")+
```

```
          theme_bw()+
          scale_colour_gradient(low="blue",high="gray30")+
          ggExtra::removeGridX()+
          ggExtra::removeGridY()+
          theme(legend.position="none",
          axis.text = element_text(colour = "black"))+
          annotate("text", x=.71, y=.7, parse=TRUE,
          label=paste0(" ",'KS: ',round(kstat,3)),
          colour="#3AA717") +
          annotate("text", x=0.4, y=0.6, parse=TRUE,
          size=4,label="TPF", colour="blue")+
          annotate("text", x=0.6, y=0.25, parse=TRUE,
          size=4,label="FPF", colour="black")
ksploti

#Matriz de clasificación
res<-predict(modProb, mod, type="response")
res<-ifelse(res > cutoff, 1, 0)
mc <- table(res,mod$Neces_prestam)
names(mc)<-c("Si","No")
mc[1,1] # Verdaderos positivos
mc[2,2] # Verdaderos negativos
mc[1,2] # Falsos positivos
mc[2,1] # Falsos negativos
prop.table(mc)*100
#Porcentaje global
round((mc[1,1]+mc[2,2])*100/sum(mc),2)

#Valores influyentes
#Pruebas de normalidad
residuos<-modProb$residuals
#Prueba de Kolmogorov-Smirnov
lillie.test(residuos)
#Prueba de Jarque Bera
jarque.bera.test(residuos)
#Prueba de Anderson-Darling
ad.test(resprob$resprob)

#Residuos de pearson
respearson<-residuals(modProb,type="pearson")

#Gráfico
pearres<-ggplot(mod, aes(x=seq(1,2946,1),
          y = abs(respearson))) +
```

```
          geom_hline(yintercept=2, color="#3AA717",
          linetype = "dashed", size=0.7) +
          geom_point(aes(color=ifelse(abs(respearson)>=2,
          'red', 'gray20')), size=1.5) +
          scale_color_identity() +
          theme_bw() +
          theme(axis.text.x=element_text(colour="black"),
          axis.title.x=element_blank(),
          axis.text.y=element_text(colour = "black"),
          axis.title.y=element_text(size = 12, angle=90,
          vjust = 0.5, hjust = 0.5)) +
          ggExtra::removeGridX()+
          ggExtra::removeGridY()+
          ylab(expression(paste("| ",r^{p},"| ")))
pearres
#Residuos significativos
table(abs(respearson)>2)

#Residuos de la devianza
resdeviance<-residuals(modProb,type="deviance")
#Gráfico
desvres<-ggplot(mod, aes(x=seq(1,2946,1),
          y = abs(resdeviance))) +
          geom_hline(yintercept=2, color="#3AA717",
          linetype = "dashed", size=0.7) +
          geom_point(aes(color=ifelse(abs(resdeviance)>=2,
          'red', 'gray20')), size=1.5) +
          scale_color_identity() +
          theme_bw() +
          theme(axis.text.x=element_text(colour="black"),
          axis.title.x=element_blank(),
          axis.text.y=element_text(colour="black"),
          axis.title.y=element_text(size = 12, angle=90,
          vjust = 0.5, hjust = 0.5)) +
          ggExtra::removeGridX()+
          ggExtra::removeGridY()+
          ylab(expression(paste("| ",r^{d},"| ")))
desvres
#valores significativos
table(abs(resdeviance)>2)
#Distancias de cook
cook<-cooks.distance(modProb)
#valores influyentes
influyentes<-cook>1
```

```
table(influyentes)

#Grafica Comparación Distribuciones
val<-val%>%mutate(PredichosLog=modLOGit3$fitted.values)
val<-val%>%mutate(Neces_prestam=ifelse(
                  Neces_prestam=="Si",1,0))
val<-val%>%mutate(PredichosProb=modProb$fitted.values)

logprob<-ggplot(val,aes(x=PredichosLog,y=Neces_prestam))
       + geom_point(color="#3F5EEE",fill="#69b3a2",
         alpha=0.5,size=1) +
       stat_smooth(data=val, method="glm",
       method.args=list(family=binomial(link="logit")),
       se=FALSE, colour="#FE8D06", size=0.35) +
       stat_smooth(data=val, aes(x=PredichosProb,
                                 y=Neces_prestam),
       method="glm", method.args=list(
       family=binomial(link="probit")), se=FALSE,
       colour="gray30", size=0.35)+
       theme(axis.line = element_line(colour="black"),
       axis.ticks.x = element_blank(),
       axis.text.x = element_text(size = rel(0.9),
       colour = "black"), axis.ticks.y = element_blank(),
       panel.background = element_blank(),
       plot.title = element_text(size = rel(1.5)),
       axis.text.y = element_text(colour = "black",
       size = rel(0.9)),
       axis.title.x = element_text(size = rel(0.9)),
       axis.title.y = element_text(size = rel(0.9)))+
       scale_y_continuous(breaks = seq(0,1,0.25)) +
       annotate("text", x=0.894, y=0.25, parse=TRUE,
       size = rel(3), label="- Logit", colour="#FE8D06") +
       annotate("text", x=0.9, y=0.15, parse=TRUE,
       size = rel(3), label="- Probit", colour="gray30")+
       xlab("\nValores ajustados")+
       ylab("Valores reales\n")
```

B.2. Pruebas de Distancia de Cook

```
#Distancia de Cook

#Muestra de modelamiento
#Modelo logit
```

```
>table(cooks.distance(modLOGit3)>1)
FALSE
2946
#Modelo probit
>table(cooks.distance(modProb)>1)
FALSE
2946

#Muestra de validación
#Modelo logit
>table(cooks.distance(modLOGit3_val)>1)
FALSE
737
#Modelo probit
>table(cooks.distance(modProb_val)>1)
FALSE
737
```