UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE MATEMÁTICA CON MENCIÓN EN ESTADÍSTICA

"MODELOS DE ELECCIÓN BINARIA Y SU APLICACIÓN EN EL RIESGO CREDITICIO EN LA CAJA MUNICIPAL CUSCO"

TESIS PARA OPTAR AL TÍTULO DE LICENCIADO EN MATEMÁTICA MENCIÓN ESTADÍSTICA

TESISTAS : **Br. BETTY ALEGRE RAMOS.**

Br. GABRIELA ROCIO CAHUANA HUAYLLAPUMA.

ASESORA : **Mgt. RINA MARICELA ZAMALLOA CORNEJO.**

CUSCO – PERU

2020

DEDICATORIA Y AGRADECIMIENTOS

La presente tesis esta dedicado:

A mis hijos Camila, Estefano y Hana, mi principal motivo de existencia y que día día me dan fuerzas para seguir adelante.

A mi esposo Gonzalo Suarez por su amor , comprensión y su apoyo constante en todos mis triunfos.

A mis padres Luis Alegre y Gladys Ramos porque ellos siempre estuvieron a mi lado brindándome su apoyo y sus consejos para ser de mí una mejor persona.

A mis hemanas Milagros y Rosmery por su gran apoyo y todos mis familiares.

Agradesco a Dios y a la Virgencita Inmaculada Concepción por darme la vida y darme muchas bendiciones y principalmente haberme otorgado una familia maravillosa , quienes han creido siempre en mí.

Un eterno agradecimiento para mi asesora Mgt. Rina Zamalloa Cornejo, quien con su conocimiento y experiencia nos ayudo a sacar adelante el presente trabajo de tesis y por los años de enseñanza en nuestra formación Universitaria.

…..Un agradecimiento especial a mi amiga y compañera Gabriela por todos los momentos que compartimos y por todos esos sacrificios que se ven reflejados en nuestra presente tesis.

Betty Alegre Ramos

Agradezco a Dios por darme la vida, por estar presente en mi día a día, por ponerme pruebas y guiarme en el camino para superarlos.

A mis padres Alberto Cahuana y Fortunata Huayllapuma, por el apoyo incondicional que vienen brindándome a lo largo de mi vida, por su amor, paciencia, comprensión y ánimo para seguir adelante, a mis hermanos y familia por su apoyo.

A Ronald Mendoza mi compañero de vida, por estar a mi lado apoyándome para continuar, por su atención, amor y confianza que en todo momento ha demostrado. A Axel Ronald Mendoza Cahuana, por ser mi motor y motivo, por darme las fuerzas para lograr lo que me propongo. Gracias amada familia.

Agradezco también mi asesora Mgt. Rina Zamalloa por su guía, paciencia, orientación y enseñanzas para la realización del presente trabajo de tesis. Y en general a todos los docentes de la Escuela Profesional de Matemática y Estadística por sus enseñanzas. Gracias queridos docentes.

A mis amigas Betty (Compañera de tesis), Blanca, Arelí, Ninoska, Paola y Milagros, por su amistad y los momentos que compartimos en nuestra etapa de estudiantes. Gracias amigas mías.

Gabriela Rocio Cahuana Huayllapuma.

Índice

Lista de tablas

Lista de Figuras

PRESENTACIÓN

SEÑORES:

DECANO DE LA FACULTAD DE CIENCIAS QUÍMICAS, FÍSICAS Y MATEMÁTICAS.

DOCENTE COORDINADOR DE LA CARRERA PROFESIONAL DE MATEMÁTICA Y ESTADÍSTICA.

DOCENTES MIEMBROS DEL JURADO.

De acuerdo al Reglamento de Grados y Títulos de la Carrera Profesional de Matemática y Estadística de la Facultad de Ciencias Químicas, Físicas y Matemáticas de la Universidad Nacional de San Antonio Abad del Cusco y con el fin de optar el Título Profesional de Licenciado en Matemática y Estadística mención Estadística, presentamos el trabajo de Tesis titulado "**MODELOS DE ELECCIÓN BINARIA Y SU APLICACIÓN EN EL RIESGO CREDITICIO EN LA CAJA MUNICIPAL CUSCO**", del año 2014.

Debemos precisar que el logro de este trabajo, es el resultado del esfuerzo y dedicación de varios años, con el único anhelo de cumplir nuestras metas; por lo que en este trabajo se vierten los conocimientos teóricos y prácticos adquiridos a lo largo de nuestros años de estudio.

Por lo expuesto, señores miembros del jurado, dejamos a vuestra consideración el presente trabajo para su revisión esperando sus apreciaciones y sugerencias para el enriquecimiento del mismo.

Atentamente,

Betty Alegre Ramos

Gabriela Rocio Cahuana Huayllapuma

UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO - UNSAAC

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE MATEMÁTICA Y ESTADÍSTICA

"MODELOS DE ELECCIÓN BINARIA Y SU APLICACIÓN EN EL RIESGO CREDITICIO DE LA CAJA MUNICIPAL CUSCO"

Autores: Gabriela Rocio Cahuana Huayllapuma
Betty Alegre Ramos

Asesora: Mgt. Rina Maricela Zamalloa Cornejo

Fecha: Cusco, Febrero del 2020.

RESUMEN

En este trabajo se analiza y predice a través del modelo de Regresión Logística estimado, los Factores de Riesgo asociados a la ocurrencia del Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco; se presenta un enfoque cuantitativo, utilizando el método descriptivo, correlacional, explicativo y transversal; se trabajó con una muestra seleccionada por muestreo probabilístico aleatorio simple en una cantidad de 330, escogida de la cantidad de clientes de la Caja Municipal de Ahorro y Crédito Cusco, a través de la estimación del modelo de Regresión Logística Binaria Múltiple, validación de supuestos, evaluación de la bondad de ajuste y la capacidad predictiva del modelo propuesto. De los resultados se concluye que existe un modelo estimado para la evaluación del Riesgo Crediticio, así como factores asociados a la ocurrencia del Riesgo Crediticio.

Palabras clave: Riesgo Crediticio, Caja Municipal de Ahorro y Crédito Cusco, Riesgo de Crédito, Modelo de Regresión Logística Binaria Múltiple.

NATIONAL UNIVERSITY OF SAN ANTONIO ABAD DEL CUSCO

SCIENCE FACULTY

PROFESSIONAL SCHOOL OF MATHEMATICS AND STATISTICS

"MODELS OF BINARY ELECTION AND ITS APPLICATION IN THE CREDIT RISK OF THE CAJA MUNICIPAL CUSCO"

Authors: Gabriela Rocio Cahuana Huayllapuma
Betty Alegre Ramos

Tutor: Mgt. Rina Maricela Zamalloa Cornejo

Fecha: Cusco, Febrero del 2020.

ABSTRACT

This paper analyzes and predicts through the estimated Logistic Regression model, the Risk Factors associated with the occurrence of Credit Risk in the Caja Municipal de Ahorro y Crédito Cusco; a quantitative approach is presented, using the descriptive, correlational, explanatory and transversal method; we worked with a sample selected by simple random probabilistic sampling in an amount of 330, chosen from the number of clients of the Municipal Savings and Credit Fund Cusco, through the estimation of the Multiple Binary Logistic Regression model, validation of assumptions, evaluation of the goodness of fit and the predictive capacity of the proposed model. From the results it is concluded that there is an estimated model for the evaluation of Credit Risk, as well as factors associated with the occurrence of Credit Risk.

Keywords: Credit Risk, Credit Risk, Multiple Binary Logistic Regression Model

INTRODUCCIÓN

Los modelos de regresión han llegado a ser un componente integral de algunos análisis de datos referentes con la descripción de la relación entre una variable respuesta y una o más variables explicativas. En general, los modelos de elección binaria, específicamente la Regresión Logística es adecuada cuando la variable de respuesta Y es politómica (admite varias categoría de respuesta, tales como empeora mucho, empeora poco, se mantiene, mejora poco, mejora mucho), pero es especialmente útil cuando solo hay dos posibles respuestas (cuando la variable de respuesta es dicotómica), que es el caso más común.

La Regresión Logística es muy utilizada en las diferentes ciencias como son: médicas, sociales, económicas y otras; en la presente investigación será aplicada al estudio del Riesgo Crediticio en la Caja Municipal Cusco, debido a que es un problema constante el hecho que una o varias personas que obtienen un préstamo o varios, éstos no son devueltos en su oportunidad y ello es lo que genera el Riesgo Crediticio, los datos a utilizarse están comprendidos en el periodo de un año, de enero a diciembre del 2014, para estudiar la ocurrencia o no ocurrencia del riesgo crediticio en la población de estudio, 950 prestamistas atendidos en la Caja en mención. El propósito de este trabajo es construir un modelo que tenga la capacidad de predecir la ocurrencia de riesgo crediticio, así como conocer los factores influyentes en la ocurrencia del mismo.

El presente trabajo de investigación consta de la siguiente estructura:

Primeramente, se efectúa el Planteamiento del Problema: Situación Problemática, Formulación del Problema, Justificación, Objetivos y Limitaciones de la Investigación.

A continuación, se plantea el Marco Teórico Conceptual: Antecedentes y Marco Conceptual.

Seguidamente se desarrolla la Hipótesis y Variables: Hipótesis, Identificación de Variables e Indicadores.

Para luego desarrollar la Metodología: Delimitación Geográfica, Metodología de la Investigación; Unidad de Análisis, además de establecer la Población y Muestra, para lo cual se utilizan las Técnicas de Muestreo y se procede a la recolección de Información.

Luego se presentan los Resultados

Finalmente se presentan las conclusiones y recomendaciones a los que se llegó con el trabajo de tesis, seguidas de la bibliografía y los Anexos A, B y C.

I. PLANTEAMIENTO DEL PROBLEMA

1.1. Situación Problemática

En el estudio del modelo de Regresión Lineal se observa que es una técnica versátil que permite predecir el comportamiento de una variable dependiente en función de una o más variables independientes siempre y cuando las variables sean cuantitativas, asimismo, la variable dependiente o variable respuesta debe estar relacionada linealmente con cada una de las variables independientes o variables regresoras, del mismo modo los residuos siguen una distribución normal con media cero y varianza constante para los distintos valores de las variables independientes, sin embargo cuando las variables a estudiar son cualitativas o categóricas; la Regresión Lineal deja de ser útil, por ejemplo en el caso de que se realicen pruebas médicas a una persona para determinar si tiene o no una enfermedad o, que un cliente devuelva o no un crédito bancario, es en estos casos que se recurren a otras técnicas multivariadas y específicamente modelos de Elección Binaria como por ejemplo la Regresión Logística, ya que ésta es especialmente útil por contener una variable respuesta dicotómica, es decir cuando estamos interesados en pronosticar la probabilidad de que ocurra o no un suceso o evento determinado; esta variable puede tomar el valor de uno cuando el evento es de éxito, o tomar el valor cero cuando el evento es de fracaso. En este caso la predicción nos permitirá determinar la variable dependiente en función de las variables independientes.

Es necesario mencionar, que en los últimos años el uso de modelos de elección binaria ha ido en aumento en las diferentes áreas de investigación, llegando a ser un instrumento idóneo, debido a su facilidad de predicción, tal como lo muestran algunos estudios al trabajar con probabilidades de éxito y fracaso, y hoy en día con el avance de los sistemas informáticos, la recolección, análisis e interpretación de datos (llegando a ser incluso miles) son realizados de manera rápida; asimismo, ayuda en el cálculo de decenas de modelos de regresión en un

tiempo muy corto, para luego quedarse con el más apropiado de ellos y finalmente tomar decisiones respecto a un problema.

En el presente estudio de investigación haremos uso de los modelos de elección Binaria, específicamente la Regresión Logística para aplicarla en el análisis del riesgo crediticio, considerando al cliente que no paga a tiempo las cuotas de su deuda, como moroso con lo que se medirá el nivel de riesgo crediticio, riesgo crediticio que puede ser evitado si se conocen los factores que influyen en el mismo y así poder evitar problemas en las entidades financieras.

1.2. Formulación del Problema

a. Problema General

¿Cuál es el modelo de elección binaria que mejor se ajusta al Riesgo Crediticio de la Caja Municipal de Ahorro y Crédito Cusco?

b. Problema Específico

¿Qué factores están asociados a la ocurrencia del Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco?

1.3. Justificación

El presente estudio de investigación se desarrolla con el propósito de dar a conocer otras formas de predicción que existe y que son diferentes a la Regresión Lineal, en este caso específico para el estudio de variables categóricas binarias, que ayudará a identificar el modelo estimado para la variable en estudio y las variables que inciden en la ocurrencia del riesgo crediticio en la Caja Municipal de Ahorro y Crédito Cusco.

La investigación ayuda a generar un modelo para entender uno de los aspectos importantes, conocer los factores de riesgo que tienen mayor influencia en la presencia de riesgo crediticio que es una preocupación principalmente de las entidades bancarias especialmente en las Cajas Municipales.

1.4. Objetivos

a. Objetivo General

Identificar y predecir a través del modelo de Regresión Logística estimado, los Factores de Riesgo asociados a la ocurrencia del Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco.

b. Objetivos Específicos

- Elaborar el marco teórico apropiado para la elección del modelo de Regresión Logística que se ajusta al Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco.
- Determinar los factores asociados y factores no asociados al Riesgo Crediticio de la Caja Municipal de Ahorro y Crédito Cusco, utilizando el modelo de elección binaria que mejor se ajusta.

1.5. Limitaciones de la Investigación

Una de las limitaciones que causó mayor dificultad fue la recolección de los datos, ya que por política de la Caja Municipal de Ahorro y Crédito Cusco, no bridan información sobre los préstamos otorgados.

Sin embargo, esto fue posible por tratarse de información que se utilizaría en una investigación de tesis y que la misma no requería de los datos personales de los usuarios (nombres, dirección, DNI, número de celular o teléfono).

II. MARCO TEÓRICO CONCEPTUAL

2.1. Antecedentes

a. Internacional

- Arcelia Mirna Cabrera Cruz (2014), en su trabajo de tesis “Diseño de Credit Scoring para Evaluar el Riesgo Crediticio en una Entidad de Ahorro y Crédito Popular” (México),

propuso el siguiente objetivo: Diseñar un modelo de credit scoring para evaluar el riesgo crediticio de los solicitantes de préstamo de la Entidad Oxaqueña. La investigación es de tipo exploratorio, porque no se tienen registros de estudios previos, es decir, se trata de un tema poco explorado, en el contexto en el que se presenta. De acuerdo a la clasificación que realizan Hernández, Fernández y Baptista (2006) sobre los diseños de investigación, este estudio se puede ubicar dentro de un diseño de investigación no experimental transeccional o transversal que tiene como propósito describir las variables y analizar su incidencia e interrelación en un momento dado, de tipo correlacional dado que se pretende determinar la relación que existe entre el perfil del cliente y su comportamiento crediticio. En dicho trabajo de tesis se llegaron a las siguientes conclusiones: la hipótesis planteada en la presente investigación se cumple puesto que se comprobó que el perfil del cliente tiene que ver con su nivel de cumplimiento, así se obtuvo que las variables que mejor pueden explicar la morosidad de la institución, de acuerdo al diseño del modelo de credit scoring obtenido, son: Oficina, Buró, Producto y Estado civil. Por lo anterior se puede determinar que dentro del credit scoring el perfil del cliente tiene que ver con su nivel de cumplimiento, puesto que el credit scoring logra diferenciar un buen y un mal cliente.

- Lisbeth Vaneza Paredes Medina (2014), en su trabajo de tesis "Análisis de Riesgo Crediticio y su Incidencia en la Liquidez de la Cooperativa de Ahorro y Crédito Frandesc Ltda., de la Ciudad de Riobamba, Provincia de Chimborazo" (Ecuador), propuso el siguiente objetivo: Realizar el análisis del riesgo crediticio y determinar su incidencia en la liquidez de la Cooperativa de Ahorro y Crédito FRANDESC Ltda., de la ciudad de Riobamba, provincia de Chimborazo. La metodología utilizada en la investigación es cualitativa, ya que descifra el análisis de la problemática de la disminución de la rentabilidad en la Cooperativa de Ahorro y Crédito "FRANDESC Ltda.", de la ciudad de

Riobamba. Se llegó a las siguientes conclusiones: Se ha detectado deficiencias en la unidad de gestión financiera que presenta la cooperativa "FRANDESC Ltda., y se ha identificado las falencias en el nivel del riesgo crediticio y la liquidez, y por ende en el proceso final de créditos otorgados a los socios y es evidente que el personal del área administrativa desconozca sobre el análisis de riesgo crediticio y la liquidación para aplicar una evaluación y aumentar el nivel de rentabilidad dicha cooperativa."

- Nubia Velandia Velandia (2013), en su trabajo de tesis "Establecimiento de un Modelo Logit para la Medición del Riesgo de Incumplimiento en Créditos para una Entidad Financiera del Municipio de Arauca, Departamento de Arauca" (Colombia), propuso el siguiente objetivo: Definir un modelo de regresión Logit por medio del cual se logre estimar la probabilidad de incumplimiento en los créditos otorgados a las pymes del Municipio de Arauca. La metodología utilizada tiene un enfoque cuantitativo y es de tipo explicativo, enfoque cuantitativo porque en este estudio se recogen datos cuantitativos los cuales posteriormente se demuestran y comprueban. Se llegó a las siguientes conclusiones: El riesgo en las entidades financieras es un pilar fundamental pues a través de él se realiza un importante filtro para la colocación de cartera, llámese riesgo de crédito, de mercado, de liquidez, operativo y de financiación del terrorismo y lavado de activos; ésta situación es hoy día un verdadero reto que hace parte del buen nombre de las entidades y que conjunto con el código de ética y buen gobierno hacen que las entidades busquen día por día posicionarse con mejor imagen, la cual está orientada a estos pilares. Así como, para lograr un balance del riesgo de crédito es importante tener en cuenta la responsabilidad social, las personas , el profesionalismo con que se atiende, el buen servicio, que junto con las grandes transformaciones tecnológicas, van a permitir dar soluciones oportunas para cada caso y de esta manera contribuir a un mejor país, con soluciones reales y trascendentales que permitan que las entidades financieras tengan

además de la opción del apalancamiento para realizar obras, cumplir sueños, la ventaja de ser realmente un aliado de las personas y empresas para beneficio de la comunidad.

- Horacio Fernández Castaño y Fredy Ocaris Pérez Ramírez, (2005) en su investigación "El Modelo Logístico: Una Herramienta Estadística para Evaluar el Riesgo de Crédito" (Colombia), y en el cual se utilizó la metodología de medición de riesgo de crédito, basados en modelos Logit y Probit que permitan mejorar el control, la toma de decisiones de la administración financiera y la gestión de los riesgos. Se llegó a la siguiente conclusión En los modelos de Scoring generalmente la variable respuesta o dependiente es de carácter dicotómico, mientras que las demás variables explicativas pueden ser continuas, categóricas (convertidas a dummys) y/o dicotómicas. Para ajustar este tipo de modelos, los principales métodos estadísticos usados son las regresiones lineales, Logit o Probit, y que la complejidad de los sistemas de medición de riesgo, aún para el caso del enfoque estandarizado, exigirá un esfuerzo notable de los entes supervisores, por mejorar sus capacidades técnicas. Estos esfuerzos serán considerablemente superiores cuando se trate de estudiar la consistencia de los sistemas de rating internos de los bancos.

b. Nacional

- Maribel Giovana Sarco Yampasi (2017) en su trabajo de tesis "Factores que determinan el otorgamiento de crédito de la financiera Credinka en la ciudad de Ayaviri, 2015" (Puno); propuso el siguiente objetivo: "Identificar los factores que determinan significativamente en el momento de aprobar el otorgamiento de crédito de la Financiera Credinka en la ciudad de Ayaviri". La metodología utilizada es un modelo Logístico para estimar la probabilidad de conseguir el otorgamiento de crédito en función de distintas variables. Esta metodología permite evaluar cuáles son los determinantes que influyen en el otorgamiento de crédito. Se llegó a la siguiente conclusión: Los factores influyentes

para la determinación de préstamos financieros en la financiera Credinka son: el Ingreso económico, Gasto, número de Hijos, Edad del cliente, Seguro de salud, tipo de Vivienda y Material de construcción de la vivienda. Los factores con mayor influencia son la edad con 0.037 y el ingreso con 0.062 de significancia en vista de que ambos se encuentran en el rango entre 0 y 1 más cercanos al cero.

- Rosa Silvia Calderon Espinola (2014), en su trabajo de tesis "La Gestión del Riesgo Crediticio y su Influencia en el Nivel de Morosidad de la Caja Municipal de Ahorro y Crédito de Trujillo - Agencia Sede Institucional – Periodo 2013" (Trujillo). Propuso el siguiente objetivo: Determinar cómo influye la Gestión del Riesgo Crediticio en el nivel de morosidad de la Caja Municipal de Ahorro y Crédito de Trujillo S.A. – Agencia Sede Institucional - periodo 2013. Se utilizó la metodología inductiva y Deductiva teniendo en cuenta como base la investigación bibliográfica, información documentaria y la recolección de información a través de la aplicación de una encuesta. Se llegó a las siguientes conclusiones: la gestión del riesgo crediticio realizado en la Agencia Sede Institucional de Caja Municipal de Ahorro y Credito de Trujillo S.A. en el periodo 2013 influyó disminuyendo los niveles de morosidad. En cuanto a la morosidad, Caja Trujillo inició el periodo 2013 con ratio de mora elevado de 7.36% y terminó el 2013 con un ratio de 5.51% producto de los castigos realizados, de la venta de adjudicados, de la venta de cartera a la FOCMAC, y de la gestion realizada por el personal de creditos para reducir este indicador. Es de precisas que la venta de cartera y la realizacion de adjudicados considera los malos creditos destinados al sector inmobiliario que se dio en su momento producto de la inadecuada metodologia de evaluacion aplicada, toda vez que la Caja Trujillo utilizó el mismo analisis y tecnologia crediticia que se emplea para la aprobacion de un credito microempresarial.

- Patricia Mirella Pantoja Vilchez (2016), en su trabajo de tesis "Propuesta de un Modelo Logit para evaluar el Riesgo Crediticio en las Cajas Municipales de Ahorro y Crédito: Caso de la Caja Municipal de Huancayo, periodo 2011-2015" (Lima), propuso la siguiente hipótesis: Contribuir a la innovación y por ende a la reducción de los niveles de riesgo de crédito de la CMAC Huancayo, una IMF representativa en el sistema de las cajas municipales, dada su actual metodología crediticia. Utilizó el tipo de investigación experimental, por motivos que la variable dependiente del modelo se define como la probabilidad de que un cliente incumpla en el reembolso de su deuda (Riesgo Creditico) en función del número de días retraso en el pago que suponga un coste para la CMAC, en ese caso se consideró, retraso a cerca de los 30 días de atraso. Llegando a la siguiente conclusión: En la investigación, se diseñó un modelo de calificación estadística para una cartera minorista de créditos de la CMAC Huancayo, aplicando la técnica de regresión logística binaria para datos individuales, en la que se obtuvo un poder moderado de calificación, capaz de predecir correctamente como máximo un 77.7% de los créditos de la cartera, mejorando la capacidad de predicción obtenido en modelos de regresión logística para IMF ya publicados en Latinoamérica y corroborado por un porcentaje similar de capacidad de predicción por estudios anteriores en la diversas instituciones de microfinanzas en el ámbito internacional. A este respecto, las medidas de valoración del modelo globalmente indican un ajuste aceptable en la regresión logística y las variables explicativas que incrementan o reducen la probabilidad de impago del cliente han sido agrupadas en tres: variables sociodemográficas, variables de comportamiento (variables cualitativas) y variables financieras.
- Pablo Valdivia Fernández (2016), en su trabajo de Tesis "El Riesgo de Crédito y su Influencia en la Liquidez de la Caja Municipal de Ahorro y Crédito Cusco S.A. en el periodo 2009-2013" (Cusco), propuso el objetivo: Analizar el incremento del riesgo

crediticio y su relación con la liquidez de la Caja Cusco, en la Región Cusco durante el periodo 2009-2013, dicho trabajo se aplicó la investigación exploratoria lo cual permitió ampliar el tema seleccionado y analizar las posibles soluciones habiendo planteado una hipótesis previa. Del mismo se utilizó la investigación de nivel descriptivo, correlacional y propositivo por medio de la relación de variables y la determinación de la relación de los objetos de la investigación de campo con el problema, el nivel descriptivo se refiere a describir el comportamiento de las variables o su representación en frecuencias lo que indica que se aplicará la estadística descriptiva para estimar la presencia de las variables en el problema y la herramienta utilizada para ello fueron las encuestas aplicadas a la población de estudio. Se llegaron a las siguientes conclusiones: El riesgo crediticio de la Caja Cusco es elevado y se puede considerar preocupante debido al comportamiento de la cartera vencida, esta obedece a que al momento de otorgar el crédito no se realiza un estudio exhaustivo de la situación del cliente de su patrón de comportamiento y fidelidad con la Caja Cusco; la liquidez de la Caja Cusco se puede considerar poco adecuada ya que para otorgar créditos el tiempo que utilizan los analistas de créditos en su mayoría oscilan entre 15 y 30 días perdiendo clientes potenciales y perjudicando a la rotación y reinversión de recursos.

2.2. Marco Conceptual:

Tomando en cuenta que los Modelos Lineales Generalizados son una extensión del Modelo Lineal General; para el presente caso en el marco conceptual comenzaremos desarrollando algunos principales aspectos del Modelo Lineal General dado que es un buen punto de partida para el estudio de los Modelos Lineales Generalizados y posteriormente desarrollar los modelos de elección binaria y la Regresión Logística como caso particular de la distribución Binomial y éste de los Modelos Lineales Generalizados.

2.2.1. Modelo Lineal General:

El Modelo Lineal General surge por la necesidad de expresar en forma cuantitativa relaciones entre un conjunto de variables, en la que una de ellas es llamada variable respuesta o dependiente siendo las otras variables explicativas o independientes.

Sea Y una variable aleatoria cuya función de distribución de probabilidad pertenece a una familia de distribuciones de probabilidades $\boldsymbol{H}$ y es explicada por un conjunto de variables $X_1, X_2, X_3, \dots, X_k$ las cuales son fijadas antes de conocer Y. La esperanza condicional de Y está dada por:

$$E(Y/X_1, X_2, \dots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad (\mathbf{1})$$

Al extraer una muestra aleatoria de tamaño n $\{(y_i, x_{i1}, \dots, x_{ik}): i = 1,2, \dots, n\}$, de una población en la cual la variable respuesta Y, y las variables independientes $X_1, X_2, \dots, X_k$, se relacionan linealmente, cada observación de la muestra puede ser expresada como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1,2, \dots, n \qquad (\mathbf{2})$$

En la ecuación (2), el término ε_i es una perturbación aleatoria no observable denominada error aleatorio, la cual tiene esperanza cero, varianza σ^2 (constante); y dos errores cualesquiera ε_i y $\varepsilon_{i'}$, $\forall i \neq i'$ son incorrelacionados entre sí.

Utilizando la notación matricial, la ecuación (2) se puede expresar como:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (\mathbf{3})$$

Donde, $\boldsymbol{Y}' = (Y_1, Y_2, \dots, Y_n)$ es un vector de variables aleatorias observables, denominado vector respuesta de orden n; $\boldsymbol{X}$ es la matriz de variables independientes de orden $n \times (k+1)$, $\boldsymbol{\beta}$ el vector de parámetros desconocidos de orden $(k+1)$ y $\boldsymbol{\varepsilon}$ el vector de errores aleatorios de orden n.

El vector de respuestas Y de la expresión (3) está formado por dos componentes, una sistemática y otra aleatoria. La primera componente constituida por la combinación lineal $\boldsymbol{X\beta}$, predictor lineal, el cual está representado como:

$$\boldsymbol{\eta = X\beta} \qquad (\mathbf{4})$$

La segunda componente, formada por el vector aleatorio Y, con elementos independientes entre sí, caracterizada por una distribución $h \in H$ con vector de esperanzas $\boldsymbol{\mu}$ y matriz de covarianza $\sigma^2 I$.

Por otro lado, calculando la esperanza de Y en (3) se tiene que:

$$\boldsymbol{E(Y) = X\beta = \mu}$$

Una característica distintiva del modelo lineal general, es que la variable respuesta Y está medida en escala numérica, mientras que las variables independientes pueden ser numéricas o categóricas y además son independientes entre sí. (Gonzales King-Keé, 2001)

2.2.1.1. Variables Independientes Numéricas:

Cuando todas las variables independientes $X_1, X_2, \dots, X_k$ son continuas, el modelo (2) es denominado modelo de regresión lineal múltiple. Los parámetros $\beta_1, \beta_2, \dots, \beta_k$, son denominados coeficientes de regresión cada β_j representa el cambio esperado en la variable respuesta, Y, por cada unidad de cambio en X_j, considerando a las demás variables independientes constantes. Siempre que el recorrido de las variables independientes incluya el cero. El coeficiente β_0 puede ser interpretado como la media de la distribución de Y.

Dos o más variable independientes pueden tener un efecto sobre la variable dependiente cuando interactúan; en este caso, y siempre que la interacción sea interpretable, estas componentes deben ser consideradas en el modelo para lograr un mejor ajuste y una interpretación óptima de los resultados. (Gonzales King-Keé, 2001)

2.2.1.2. Variables Independientes Cualitativas o Categóricas:

Cuando el predictor lineal η está formado únicamente por variables cualitativas, estas son denominadas factores y los valores que toman corresponden a los niveles de factor. Estos niveles pueden no tener un orden asociado a ellos, como en el caso del color de pelaje, la raza de un animal, etc., (es el caso de las variables independientes de tipo nominal); también

pueden tener un orden que no signifique magnitud, como una escala de preferencias (variables independientes ordinales); o tener asociada una escala de medición numérica, como por ejemplo las cantidades de fertilizante utilizada en un experimento agrícola.

Cuando las observaciones son clasificadas por dos o más factores, hablamos de un análisis multifactorial y los tratamientos son las combinaciones entre los niveles de los factores considerados. Por ejemplo, al considerar un modelo con 2 factores A y B, será necesario incluir términos de la forma $\alpha_i + \beta_j$, mientras que si existe interacciones se incluirán términos de la forma $(\alpha\beta)_{ij}$; siendo la representación de un modelo de 2 factores:

$$y_{ijm} = \delta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijm} \qquad i = 1,2,\dots,k;\ j = 1,2,\dots,b;\ m = 1,2,\dots,n$$

Donde δ es la media general; α_i es el efecto del i-ésimo nivel del factor A; β_j el efecto del j-ésimo nivel del factor B; $(\alpha\beta)_{ij}$ el efecto de la interacción entre A y B; ε_{ijm} la componente aleatoria con características dadas en la ecuación (2); y_{ijm} es la respuesta del m-ésimo sujeto correspondiente al i-ésimo nivel del factor A y el j-ésimo nivel del factor B.

Para tres o más factores solo necesitaremos una extensión del modelo anterior.

Para que los modelos de rango incompleto puedan ser representados de la forma (3), es necesario utilizar **variables artificiales o contrastes**, con valores numéricos que representen a las categorías originales.

Para ilustrar claramente esta situación consideraremos un experimento con una variable respuesta y un único factor de k niveles y n repeticiones por cada nivel, la disposición de los datos será entonces:

Tabla 1:
Experimento con una Variable Respuesta y un Factor con k Niveles

Factores Niveles	**Observaciones**				
$\mathbf{1}$	y_{11}	…	y_{1j}	…	y_{1n}
⋮	⋮	…	⋮	…	⋮
$\boldsymbol{i}$	y_{i1}	…	y_{ij}	…	y_{in}
⋮	⋮	…	⋮	…	⋮
$\boldsymbol{k}$	y_{k1}	…	y_{kj}	…	y_{kn}

Donde y_{ij} es la j-ésima observación correspondiente al i-ésimo nivel del tratamiento, con $i = 1,2, \dots, k$ y $j = 1,2, \dots, n$, el modelo será:

$$y_{ij} = \delta + \alpha_i + \varepsilon_{ij} \qquad i = 1,2, \dots, k; \;\; j = 1,2, \dots, n$$

La representación matricial del modelo será:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad (\mathbf{5})$$

La matriz X se define de acuerdo a los objetivos del estudio, pues la interpretación de los parámetros dependerá de la manera como se defina esta matriz:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{ij} \\ \vdots \\ Y_{kn} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \dots 0 \dots \\ 1 & 1 & 0 \dots 0 \dots \\ \vdots & \vdots & \vdots \vdots \vdots \cdots \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_i \\ \vdots \\ \alpha_k \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1n} \\ e_{21} \\ \vdots \\ e_{ij} \\ \vdots \\ e_{kn} \end{bmatrix}$$

Para garantizar que $X'X$ sea invertible, las columnas de la matriz X deben ser linealmente independientes; para ello, si el factor A tiene k niveles, se definirá una variable artificial con $k - 1$ niveles.

El tipo de reparametrización que utilizaremos, es llamado "reparametrización del punto central". Así por ejemplo, si una variable A tiene k niveles, lo primero que hay que hacer es determinar la categoría que será usada como referencia. Suponiendo que elegimos la última categoría, tendríamos que la i-ésima columna de la matriz X, contiene a 1 en la i-ésima fila. -1 en la última fila y cero en las restantes. Si α_i denota el parámetro que corresponde al i-ésimo nivel del factor A las $k - 1$ columnas producen estimadores de los parámetros independientes α_1, α_2, α_{k-1}.

Por ejemplo

Si tenemos una variable A con 2 categorías, la reparametrización del punto central estará dada por:

$$X_i = \begin{cases} 1\ si\ la\ observación\ pertenece\ al\ i-ésimo\ nivel\ del\ factor\ A. \\ -1\ caso\ contrario \end{cases}$$

Para una variable con más de 2 niveles:

$$X_i = \begin{cases} 1\ si\ la\ observación\ pertenece\ al\ i-ésimo\ nivel\ del\ factor\ A. \\ 0\ caso\ contrario \\ -1\ nivel\ de\ referencia \end{cases}$$

Con esta reparametrización comparamos el efecto de cada una de las categorías de las variables independientes con el efecto de la categoría usada como referencia.

Además de las variables dependientes continuas, existen muchos otros tipos de variables dependientes, principalmente las **variables categóricas**, ya sean binarias, multinomiales y ordinales; así como las variables discretas. (Gonzales King-Keé, 2001)

2.2.2. Modelos Lineales Generalizados (MLG)

Para dar inicio al desarrollo de los Modelos Lineales Generalizados MLG comenzaremos mencionando que los primeros que estudiaron dichos modelos fueron Nelder y Wedderburn, extendiendo la teoría de los modelos lineales, incorporando de esta manera la posibilidad de modelar variables respuestas continuas o categóricas con distribuciones del error no necesariamente homocedásticos. (Gonzales King-Keé, 2001)

Los MLG son usados cuando se tiene una única variable aleatoria respuesta Y asociada a un conjunto de variables explicativas $X_1, X_2, .., X_k$. Éstos son una familia de modelos estadísticos que permiten relacionar variables dependientes con una combinación lineal de variables independientes.

Los MLG, fueron creados para establecer modelos útiles para una mayor variedad de tipos de variables de respuesta.

En un MLG, es posible estimar una función llamada "función de enlace" del valor medio de la respuesta, como una función lineal de valores de las variables explicativas.

En este caso, los datos pueden pertenecer de forma más general a alguna de las distribuciones de probabilidad de la familia exponencial.

El modelo se representa por:

$$g\big(E(Y/X)\big) = g(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \eta(x) \qquad (\mathbf{6})$$

Donde g es la función de enlace.

La función lineal de las variables explicativas, $\eta(x)$, comúnmente se le llama predictor lineal.

A continuación se muestran los principales Modelos Lineales Generalizados en la siguiente tabla:

Tabla 2:
Principales Modelos Lineales Generalizados

Naturaleza de la Variable Respuesta	Componente Sistemático	Componente Aleatorio	Función de Enlace	Modelo Lineal	
▪ Numérica cuantitativa	▪ Numérico ▪ Categórico ▪ Mixto	▪ Normal ▪ Normal ▪ Normal	▪ Identidad ▪ Identidad ▪ Identidad	▪ Regresión lineal ▪ ANOVA o de diseño experimental ▪ ANCOVA o de diseño experimental con variables concomitantes	**ML**
▪ Categórica binaria - No agrupada - Agrupada (frecuencias)	▪ Mixto ▪ Categórico	▪ Binomial (1) ▪ Bernoulli ▪ Binomial (n)	▪ Logit ▪ Logit ▪ Probit	▪ Regresión Logística ▪ Análisis Logit ▪ Análisis Probit	
▪ Categórica Politómica - No agrupada - Agrupada (frecuencias)	▪ Mixto ▪ Categórico	▪ Multinomial ▪ Multinomial	▪ Logit generalizado ▪ Logit generalizado	▪ Regresión logística multinomial ▪ Análisis Logit multinomial	**MLG**
▪ Recuento ▪ Frecuencia	▪ Mixto ▪ Categórico	▪ Poisson ▪ Poisson	▪ Logarítmica ▪ Logarítmica	▪ Regresión de Poisson ▪ Análisis log lineal	

En la Tabla 2 se puede observar cómo el Modelo Lineal es el caso más elemental del Modelo Lineal Generalizado. Las coincidencias y las diferencias entre uno y otro hacen

posible, en el caso del MLG, un tratamiento matemático y estadístico adecuado a los niveles de medida de las variables que contiene. (Lopez Gonzales & Ruiz Soler, 2011)

Los Modelos Lineales Generalizados tienen 2 componentes: Componente Aleatorio, Componente Sistematico y; una Función de Enlace.

2.2.1.1. Componente Aleatorio:

El componente aleatorio identifica a la variable respuesta y su distribución de probabilidad. Sea el vector aleatorio $Y = (y_1, y_2, \dots, y_n)'$ donde $y_i, i = 1,2, \dots, n$ son independientes e idénticamente distribuidos. La distribución de probabilidad de este vector pertenece a la familia exponencial (normal, log-normal, Poisson, gamma,…) que tiene la siguiente forma:

$$f_y(y, \theta, \emptyset) = exp\left\{\frac{1}{a(\emptyset)}\big(y\,\theta - b(\theta) + c(y, \emptyset)\big)\right\};\ \emptyset > 0 \qquad (7)$$

θ	:	Es un parámetro canónico (natural), como puede ser alguna función de la media.
$\emptyset$	:	Es el parámetro de dispersión y,
$a(\cdot), b(\cdot) y\ c(\cdot)$	:	Son funciones conocidas y determinan la función de probabilidad como la binomial o la normal.

Si $\emptyset$ es conocido, este es un modelo de la familia exponencial lineal.

Si $\emptyset$ es desconocido, es un modelo de dispersión exponencial.

Por ejemplo la función de densidad de la distribución normal está dada por:

$$f(y;\ \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

Puede ser reescrita como:

$$f(y;\ \mu, \sigma^2) = exp\left[\frac{1}{\sigma^2}\left(y\mu - \frac{\mu^2}{2}\right) - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)\right)\right]$$

Donde;

$$E[y] = \mu = \theta$$

$$a(\emptyset) = \sigma^2 = \emptyset;$$

$$b(\theta) = \sigma^2 = \frac{\mu^2}{2};$$

$$c(y, \emptyset) = -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + ln(2\pi\sigma^2)\right)$$

En este caso vemos que la variable dependiente sigue una distribución normal, que por definición es continua. Pero existen ocasiones en que la variable dependiente sigue una distribución no continua y, por tanto, los valores estimados por el modelo seguirán el mismo tipo de distribución que los datos de partida. Para verificar si los datos son o no normales se recomienda en primer lugar conocer el tipo de variable respuesta y su naturaleza; y en segundo lugar analizar los residuos del modelo una vez ajustado el modelo. Por ejemplo una variable con distribución Binomial se usa para proporciones y datos de presencia/ausencia (la tasa de mortalidad, tasa de infección, porcentaje de éxito reproductivo, presencia o ausencia de una determinada enfermedad).

Para mayor detalle a continuación se aprecian algunas distribuciones de la familia exponencial $f(y, \theta, \phi)$.

Tabla 3:
Algunas Distribuciones de la Familia Exponencial

Distribución	$\boldsymbol{a(\emptyset)}$	$\boldsymbol{\theta}$	$\boldsymbol{b(\theta)}$	$\boldsymbol{c(y, \emptyset)}$	$\boldsymbol{\mu(\theta)}$	$\boldsymbol{v(\theta)}$
Normal $N(\mu, \sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$	θ	1
Poisson $P(\mu)$	1	$ln(\mu)$	e^θ	$-ln(y!)$	e^θ	μ
Binomial $B(m, \pi)$	1	$ln\left(\frac{\mu}{m-\mu}\right)$	$mln(1+e^\theta)$	$ln\binom{m}{y}$	$\frac{me^\theta}{1+e^\theta}$	$\frac{\mu}{m}(m-\mu)$
Binomial Negativa $BN(\mu, k)$	1	$ln\left(\frac{\mu}{\mu-k}\right)$	$-kln(1-e^\theta)$	$ln\left(\frac{\Gamma(k+y)}{\Gamma(k)y!}\right)$	$\frac{ke^\theta}{1-e^\theta}$	$\mu\left(\frac{\mu}{k}+1\right)$
Gamma $G(\mu, v)$	v^{-1}	$-\frac{1}{\mu}$	$\ln(-\theta)$	$vln(vy) - \ln(y) - ln\Gamma(v)$	$-\frac{1}{\theta}$	μ^2
Normal Inversa $IG(\mu, \sigma^2)$	σ^2	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left[ln(2\pi\sigma^2) + \frac{1}{\sigma^2 y}\right]$	$(-2\theta)^{-1/2}$	μ^3

Para calcular la media de y_i en los MLG se usa la función de verosimilitud, que en el caso de una distribución normal es:

$$L(\theta, \phi, y_i) = lnf_y(y_i, \theta, \phi)$$

$$L(y_i, \theta) = \frac{1}{a(\emptyset)}(y_i\theta - b(\theta) + c(y, \emptyset)$$

$$E\left(\frac{dL}{d\theta}\right) = 0; \frac{dL}{d\theta} = \frac{\{y_i - b'(\theta)\}}{a(\emptyset)}; \text{entonces } E\left(\frac{dL}{d\theta}\right) = E\left(\frac{\{y_i - b'(\theta)\}}{a(\emptyset)}\right) = 0$$

$$E\left(\frac{dL}{d\theta}\right) = \frac{\{\mu_i - b'(\theta)\}}{a(\emptyset)} = 0$$

$$E(y_i) = \mu_i = b'(\theta) \qquad (\mathbf{8})$$

De forma similar procedemos para calcular la varianza de la distribución normal:

$$E\left(\frac{d^2L}{d\theta^2}\right) - E\left(\frac{dL}{d\theta}\right)^2 = 0$$

$$\frac{d^2L}{d\theta^2} = \frac{b''(\theta)}{a(\emptyset)}$$

$$E\left(\frac{b''(\theta)}{a(\emptyset)}\right) = \frac{b''(\theta)}{a(\emptyset)}$$

$$E\left(\frac{d^2L}{d\theta^2}\right) - E\left(\frac{dL}{d\theta}\right)^2 = \frac{b''(\theta)}{a(\emptyset)} - \left(\frac{\{\mu_i - b'(\theta)\}}{a(\emptyset)}\right)^2 = \frac{b''(\theta)}{a(\emptyset)} - \frac{Var(Y)}{\left(a(\emptyset)\right)^2} = 0$$

$$Var(y_i) = a(\emptyset)b''(\theta) \qquad (\mathbf{9})$$

2.2.1.2. Componente Sistemático:

El componente sistemático de un Modelo Lineal Generalizado especifica las variables explicativas. Estos entran linealmente como predictores en la parte derecha de la ecuación del modelo. Es decir, el componente sistemático especifica las variables que son el $\{x_i\}$ en la fórmula:

$$\eta(x) = \sum_{i=1}^{p} \beta_i x_i \qquad (\mathbf{10})$$

Donde;

x_i : Son variables explicativas del modelo que no pueden estar altamente correlacionadas.

β_i : Son parámetros cuyos valores son desconocidos y necesitan ser estimados.

Esta combinación lineal de las variables explicativas se denomina predictor lineal.

Algunos $\{x_i\}$ se pueden basar en otros, del modelo. Por ejemplo, tal vez $x_3 = x_1 x_2$, para permitir la interacción entre x_1 y x_2 en sus efectos en Y, o tal vez $x_3 = x_1^2$, para permitir un efecto curvilíneo de x_1. (Los Modelos Lineales Generalizados usan minúsculas para cada x para enfatizar que valores x se tratan como fijos en lugar una variable aleatoria). (Agresti, 2007)

2.2.1.3. Función de Enlace:

Esta función es útil en los Modelos Lineales Generalizados porque permite restringir los valores de μ, porque tiene en cuenta otras distribuciones a parte de la normal. La función de enlace $g(\mu)$, relaciona el componente aleatorio y el componente sistemático vinculando al predictor lineal η con el valor esperado de la variable respuesta, $E(Y/X) = \mu$, a través de la función.

$$g(\mu) = X\beta$$
$$\eta = g(\mu);\ \mu = g^{-1}(\eta) \qquad (\mathbf{11})$$

Donde g es función diferenciable, monótona e invertible.

En los modelos lineales generalizados, se utilizan diversas familias de distribuciones exponenciales como la Binomial, Poisson, Gamma, Normal y Normal Inversa entre otras.

Cada distribución elegida tiene una función de enlace asociada y una estructura del error correspondiente para la estimación (función de varianza).

En estos modelos, la varianza de Y puede ser una función de la media μ:

$$Var(Y) = \phi(\mu)$$

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

En la tabla siguiente se presentan algunas funciones de enlace y funciones de varianza para algunas distribuciones utilizadas en modelos lineales generalizados:

Tabla 4:
Algunas Funciones de Enlace y Funciones de Varianza de la Familia Exponencial

Familia de Distribuciones	Función de Enlace	Función de Varianza
Normal/Gauseana	Identidad μ	l
Binomial	Logarítmica $Log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\mu(1-\mu)}{n}$
Poisson	Logística $Log(\mu)$	μ
Gamma	Reciproca $\frac{1}{\mu}$	μ^2
Normal inversa	Reciproca^2 $\frac{1}{\mu^2}$	μ^3

Además otras funciones de enlace son:

- Función Logit : $\eta = ln\left\{\frac{\mu}{1-\mu}\right\}$
- Función Probit : $\eta = \phi^{-1}(\mu)$
- Función Log-log complementaria : $\eta = ln\{-\ln(1-\mu)\}$.

(Agresti, 2007)

2.2.3. Estimación de los Modelos Lineales Generalizados:

Los dos métodos clásicos para estimar los parámetros desconocidos de un modelo lineal general, son de Máxima Verosimilitud (MV) y el método de Mínimos Cuadrados Generalizados (MCG); siendo el método de Mínimos Cuadrados Ponderados (MCP) un caso particular de este último.

A continuación, estudiaremos las condiciones bajo las cuales se realiza la estimación de Máxima Verisimilitud por ser éste el método más óptimo ya que tiene propiedades de consistencia, eficiencia y asintótica. Los resultados de los estimadores, serán válidos en el contexto de los modelos lineales generalizados.

2.2.3.1. Estimación de Máxima Verosimilitud:

La función de verosimilitud de la función de la familia exponencial está dada por:

$$L = f(y_1, \dots, y_n; \theta_1, \dots, \theta_n) = \prod_{i=1}^{n} exp\left\{\frac{1}{a(\phi)}\left(y_i\theta_i - b(\theta_i) + c(y_i, \phi)\right)\right\} \qquad (\mathbf{12})$$

Aplicando logaritmo natural a ambos miembros de la ecuación anterior:

$$lnL = ln\prod_{i=1}^{n} exp\left\{\frac{1}{a(\phi)}\left(y_i\theta_i - b(\theta_i) + c(y_i, \phi)\right)\right\}$$

Aplicando la propiedad de producto de logaritmos, y renombrando a lnL por l se tendrá la siguiente ecuación:

$$l = ln\sum_{i=1}^{n} exp\left\{\frac{1}{a(\phi)}\left(y_i\theta_i - b(\theta_i) + c(y_i, \phi)\right)\right\}$$

Luego;

$$l = lnexp\sum_{i=1}^{n}\left\{\frac{1}{a(\phi)}\left(y_i\theta_i - b(\theta_i) + c(y_i, \phi)\right)\right\}$$

Por propiedad de logaritmos:

$$l = \sum_{i=1}^{n}\left\{\frac{1}{a(\phi)}\left(y_i\theta_i - b(\theta_i) + c(y_i, \phi)\right)\right\};\ \ lnexp = 1$$

Distribuyendo la sumatoria:

$$l = \sum_{i=1}^{n}\frac{1}{a(\phi)}y_i\theta_i - \sum_{i=1}^{n}\frac{1}{a(\phi)}b(\theta_i) + \sum_{i=1}^{n}\frac{1}{a(\phi)}c(y_i, \phi)$$

$$l = \frac{1}{a(\phi)}\left\{\sum_{i=1}^{n}[y_i\theta_i - b(\theta_i)] + \sum_{i=1}^{n}c(y_i, \phi)\right\}$$

Derivando respecto a los parámetros β del modelo:

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)}\sum_{i=1}^{n}\left[y_i\frac{d\theta_i}{d\beta} - \frac{db(\theta_i)d\theta_i}{d\theta_i d\beta}\right]$$

Según la ecuación se tiene que:

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \frac{d\theta_i}{d\beta}$$

Reemplazando $\frac{d\theta_i}{d\beta} = \frac{d\theta_i d\mu_i}{d\mu_i d\beta}$ se tiene:

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \frac{d\theta_i d\mu_i}{d\mu_i d\beta}$$

También:

$$\frac{d\mu_i}{d\beta} = \frac{d\mu_i d\big(g(\mu_i)\big)}{d\big(g(\mu_i)\big) d\beta}$$

Invirtiendo $\frac{d\mu_i}{d(g(\mu_i))}$ y por definición de función de enlace $g(\mu_i) = \boldsymbol{x'}_i\beta$

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \frac{d\theta_i}{d\beta} \left[\frac{d\big(g(\mu_i)\big)}{d\mu_i}\right]^{-1} \frac{d\boldsymbol{x'}_i\beta}{d\beta}$$

Invirtiendo $\frac{d\theta_i}{d\mu_i}$ y derivando $\frac{dx\prime_i\beta}{d\beta}$

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \left[\frac{d\big(g(\mu_i)\big)}{d\mu_i}\right]^{-1} \boldsymbol{x'}_i$$

Por la ecuación tenemos que $\mu_i = db(\theta)$ entonces:

$$\frac{d\mu_i}{d\theta} = \frac{d^2 b(\theta_i)}{d\theta^2} = b''(\theta)$$

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \left[\frac{d^2 b(\theta_i)}{d\theta_i^2}\right]^{-1} \left[\frac{d\big(g(\mu_i)\big)}{d\mu_i}\right]^{-1} \boldsymbol{x'}_i$$

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} \sum_{i=1}^{n} [y_i - \mu_i] \, [b''(\theta_i)]^{-1} \left[\frac{d\big(g(\mu_i)\big)}{d\mu_i}\right]^{-1} \boldsymbol{x'}_i$$

Para maximizar, igualamos a cero la derivada encontrada, de modo que queda:

$$\frac{dl}{d\beta} = \frac{1}{a(\phi)} [\boldsymbol{Y} - \boldsymbol{\mu}] \boldsymbol{W} \boldsymbol{\Delta} \boldsymbol{X'} = \boldsymbol{0} \qquad (\boldsymbol{13})$$

Donde;

$$\boldsymbol{W} = \left[b''(\theta_i) \left(\frac{d\big(g(\mu_i)\big)}{d\mu_i} \right)^2 \right]^{-1} ; \boldsymbol{\Delta} = \frac{d\big(g(\mu_i)\big)}{d\mu_i}$$

$$\boldsymbol{YW\Delta X'} = \boldsymbol{\mu W \Delta X'} \qquad (\mathbf{14})$$

Los parámetros $\boldsymbol{\mu}$, $\boldsymbol{W}$ y $\boldsymbol{\Delta}$ son funciones no lineales, por lo que es necesario usar procesos numéricos iterativos para estimar β.

El método numérico de Newton – Raphson da solución a la ecuación $f(x)$, que está basada en una aproximación de Taylor para una función $f(x)$ en la vecindad del punto x_0; es decir,

$$f(x) = f(x_0) + (x - x_0) f'(x_0) = 0$$

Obteniéndose:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

De forma general:

$$x^{(m+1)} = x^{(m)} - \frac{f(x^m)}{f'(x^m)} \qquad (\mathbf{15})$$

Para resolver $\frac{dl}{d\beta} = 0$, usando el método de Newton – Raphson en su forma multivariada, cuya ecuación es:

$$\beta^{(m+1)} = \beta^{(m)} + \left[I\big(\beta^{(m)}\big) \right]^{-1} \frac{\partial l}{\partial \beta}\bigg|_{\beta = \beta^{(m)}} \qquad (\mathbf{16})$$

Donde

$I\big(\beta^{(m)}\big)$; es la inversa de las derivadas de segundo orden del (β).

El método de Newton – Raphson es útil para hallar las derivadas parciales de segundo orden. Otro método muy usado es el método de Fisher, en general, más sencillo (que coincide con el método de Newton – Raphson en el caso de las funciones de enlace canónico). Este método implica una sustitución de la matriz de derivadas parciales, es decir, la matriz de

información observada por la matriz de información esperada de Fisher. (Contreras Vilca, 2012).

La información de Fisher mide, que tanta información da una variable aleatoria $X = (x_1, \dots, x_n)$ sobre un parámetro desconocido θ. Si el logaritmo de $f(X, \theta)$ es doblemente diferenciable respecto a θ, la información de Fisher puede ser escrita como:

$$I(\theta) = -E\left[\frac{\partial^2 l(Y,\theta)}{\partial\theta\partial\theta'}\right] \qquad (\mathbf{17})$$

En base a la ecuación anterior, calcularemos la información de Fisher de β.

Así:

$$\frac{\partial^2 l(Y,\theta)}{\partial\theta\partial\theta'} = \frac{\partial}{\partial\beta'}\left(\frac{\partial l}{\partial\beta}\right)$$

Por la ecuación se tiene:

$$\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'} = \frac{\partial}{\partial\beta'}\left(\frac{1}{a(\phi)}[\boldsymbol{Y}-\boldsymbol{\mu}]\boldsymbol{W\Delta X}'\right)$$

$$\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'} = \frac{1}{a(\phi)}\frac{\partial}{\partial\beta'}([\boldsymbol{Y}-\boldsymbol{\mu}]\boldsymbol{W\Delta X}')$$

$$\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'} = \frac{1}{a(\phi)}\left\{\frac{\partial}{\partial\beta'}([\boldsymbol{Y}-\boldsymbol{\mu}]\boldsymbol{W\Delta X}')\right\}$$

Aplicando multiplicación de derivadas:

$$\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'} = \frac{1}{a(\phi)}\left\{-\boldsymbol{W\Delta X}'\frac{\partial\mu}{\partial\beta'} + [\boldsymbol{Y}-\boldsymbol{\mu}]\frac{\partial}{\partial\beta'}\boldsymbol{W\Delta X}'\right\}$$

Luego;

$$-E\left[\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'}\right] = -E\left[\frac{1}{a(\phi)}\left\{-\boldsymbol{W\Delta X}'\frac{\partial\mu}{\partial\beta'} + [\boldsymbol{Y}-\boldsymbol{\mu}]\frac{\partial}{\partial\beta'}\boldsymbol{W\Delta X}'\right\}\right]$$

Por la ecuación (13) sabemos que $[\boldsymbol{Y}-\boldsymbol{\mu}]\frac{\partial}{\partial\beta'}\boldsymbol{W\Delta X}' = \mathbf{0}$ y por

$$-E\left[\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'}\right] = \frac{1}{a(\phi)}\boldsymbol{W\Delta X}'\frac{\partial\mu}{\partial\beta'} + 0 = \frac{1}{a(\phi)}\boldsymbol{W\Delta X}'\frac{\partial\mu}{\partial g(\mu)}\frac{\partial g(\mu)}{\partial\beta'}$$

$$-E\left[\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'}\right]=\frac{1}{a(\phi)}\boldsymbol{W\Delta X'}\left(\frac{\partial g(\mu)}{\partial\beta'}\right)^{-1}\frac{\partial \boldsymbol{X'\beta}}{\partial\beta'}=\frac{1}{a(\phi)}\boldsymbol{X'W\Delta}\left(\frac{\partial g(\mu)}{\partial\mu}\right)^{-1}\boldsymbol{X}$$

$$-E\left[\frac{\partial^2 l(Y,\beta)}{\partial\beta\partial\beta'}\right]=\frac{1}{a(\phi)}\boldsymbol{X'W\Delta\Delta^{-1}X}=\frac{1}{a(\phi)}\boldsymbol{X'WX}$$

Luego la matriz de información de Fisher de β es estimada por:

$$I(\beta)=\frac{1}{a(\phi)}\boldsymbol{X'WX} \qquad (\mathbf{18})$$

Reemplazando las ecuaciones (13) y (18) en la ecuación (16)

$$\beta^{(m+1)}=\beta^{(m)}+\left[I\left(\beta^{(m)}\right)\right]^{-1}\frac{\partial l}{\partial\beta}\Big|_{\beta=\beta^{(m)}}$$

$$\beta^{(m+1)}=\beta^{(m)}+\left[\frac{1}{a(\phi)}\boldsymbol{X'WX}\right]^{-1}\frac{1}{a(\phi)}[\boldsymbol{Y}-\boldsymbol{\mu}]\frac{\partial}{\partial\beta'}\boldsymbol{W\Delta X'}$$

$$\beta^{(m+1)}=\beta^{(m)}+[\boldsymbol{XWX'}]^{-1}[\boldsymbol{Y}-\boldsymbol{\mu}]\boldsymbol{W}\frac{\partial g(\mu_i)}{\partial\mu_i}\boldsymbol{X'}$$

$$\beta^{(m+1)}=\left[\boldsymbol{X'W^{(m)}X}\right]^{-1}\boldsymbol{W^{(m)}X'z^{(m)}} \qquad (\mathbf{19})$$

Donde:

$$\boldsymbol{W^{(m)}}=\left[b''(\theta_i)\left(\frac{d\left(g(\mu_i)\right)}{\partial\mu_i}\right)^2\right]^{-1};$$

$$n_i^{(m)}=\sum_{r=1}^{p}x_{ir}\beta_r^{(\beta)}$$

$$z^{(m)}=n_i^{(m)}+\left[y_i-\mu_i^{(m)}\right]g'\left(\mu_i^{(m)}\right)$$

(Tomaiconza Ataulluco & Pari Sallo, 2014)

2.2.4. Evaluación del Modelo Lineal Generalizado

El objetivo es evaluar la validez externa del modelo ajustado, para lograr esto es necesario verificar la bondad del ajuste y la adecuación (cumplimiento de supuestos) del modelo ajustado.

2.2.4.1. Bondad de Ajuste del Modelo

Una vez estimados los parámetros, se debe valorar la magnitud de la discrepancia entre los datos observados y esperados. Según Nelder y McCullagh (1991), el ajuste de un modelo a partir del conjunto de datos observados puede ser encarado como una manera de sustituir las observaciones por un conjunto de valores estimados $\hat{\mu}$ para un modelo con un número de parámetros relativamente pequeño.

Una discrepancia pequeña entre los datos observados y $\hat{\mu}$ puede ser tolerable, en cuanto que una discrepancia grande no. De esta manera, si se admite una combinación satisfactoria de la distribución de la variable respuesta y de la función de enlace, el objetivo es determinar cuántos términos son necesarios en la estructura lineal para una descripción razonable de los datos. Un número grande de variables independientes (explicativas) puede 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 independientes puede llevar a un modelo de fácil interpretación pero que se ajuste pobremente a los datos. Lo que se desea en realidad es un modelo intermedio.

En el proceso del ajuste del modelo se evalúan generalmente un conjunto de modelos que constituyen aproximaciones a los datos observados. Dos modelos que pueden intervenir en las comparaciones a menudo son:

a. Modelo Saturado:

En este modelo el número de parámetros estimados es igual al número de observaciones. En datos individuales, utilizar este modelo implicaría estimar un número de parámetros igual al tamaño muestral.

El modelo saturado es un modelo de forma similar al modelo propuesto que describe de modo perfecto los datos. Por tanto, tiene poca utilidad desde el punto de vista de ajuste de un

modelo. Sin embargo, es útil para medir como un ajuste concreto se parece a un ajuste "perfecto". El modelo saturado asociado a un modelo propuesto viene caracterizado por:

Utilizar la misma distribución para la respuesta (no necesariamente con los mismos parámetros).

Utilizar el mismo enlace.

El número de parámetros es igual al número de datos ($p = n$), y por lo tanto no quedan grados de libertad para los residuos.

b. Modelo Nulo:

Este es un modelo muy simple, el cual se utiliza como modelo de referencia. Contiene como único parámetro al valor esperado μ, para todas las observaciones. Habitualmente es incapaz de representar adecuadamente la estructura de los datos, asume un efecto nulo de las variables independientes. (Figueroa Arboccó, 2005)

2.2.4.2. Función de Devianza

La Devianza del modelo fue propuesta por Nelder y Weddernum (1972) sirve para evaluar el ajuste del modelo, así como comparar modelos. La Devianza residual o simplemente Devianza compara el logaritmo de la verosimilitud del modelo saturado con el logaritmo de la verosimilitud de un modelo ajustado. El valor de la función de logaritmo de verosimilitud, para el modelo ajustado, nunca puede ser mayor que el valor del modelo saturado, porque el modelo ajustado contiene menos parámetros, entonces la Devianza del modelo es siempre mayor o igual a cero y se define como:

$$D_p(\mu) = 2[lnL(modelo\ saturado) - lnL(\hat{\mu})]$$

$$D_p(\mu) = 2[l(modelo\ saturado) - l(\hat{\mu})] \qquad (\mathbf{20})$$

Que es la distancia entre el logaritmo de la función verosimilitud del modelo saturado (con n parámetros) y el modelo en investigación (con p parámetros). Un valor pequeño del desvío indica que para un número menor de parámetros, se obtiene un ajuste tan bueno como cuando

se ajuste un modelo saturado. Por tanto, la Devianza es una medida de distancia de los valores ajustados $\hat{\mu}'s$ en relación con los datos observados.

La Devianza es siempre mayor o igual a cero. Para probar la adecuación de un Modelo Lineal Generalizado, el valor de la Devianza debe ser comparado con el percentil de alguna distribución de probabilidad referente. En la práctica, la función Devianza se compara con los percentiles de una distribución χ^2_{n-p}. (McCullagh y Nelder, 1991).

Las pruebas de hipótesis son:

H_0 : El modelo ajustado es adecuado. (se acepta si $S_p(\mu) \leq \chi^2_{(\alpha,n-p)}$)

H_1 : El modelo ajustado no es adecuado. (se acepta si $S_p(\mu) > \chi^2_{(\alpha,n-p)}$)

Donde: $S_p(\mu) = \emptyset^{-1} D_p(\mu)$

Aunque la varianza residual es una medida de ajuste muy usada, McCullagh y Nelder (1989) argumentan que (al menos en el MLG Binomial) valores grandes de Devianza no siempre son evidencia de un ajuste pobre.

Para mayor detalle, a continuación se muestran las funciones de Devianza para algunas distribuciones:

Tabla 5

Funciones de Devianza de Algunas Distribuciones

Distribuciones	Devianza
Normal	$D_p = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$
Binomial	$D_p = 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right]$
Poisson	$D_p = 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (y_i - \hat{\mu}_i)\right]$
Binomial Negativo	$D_p = 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (y_i - k) ln\left(\frac{\hat{\mu}_i - k}{y_i - \hat{\mu}_i}\right)\right]$
Gamma	$D_p = 2\sum_{i=1}^{n} \left[y_i \ln\left(\frac{\hat{\mu}_i}{y_i}\right) + (y_i - k) ln\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right)\right]$
Normal Inverso	$D_p = 2\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{y_i \hat{\mu}_i^2}$

(Tomaiconza Ataulluco & Pari Sallo, 2014)

2.2.4.3. Coeficiente de Determinación (R^2)

La medida R^2 es definida como la reducción proporcional en la incertidumbre, debido a la inclusión de las variables independientes. Bajo ciertas condiciones, también puede ser interpretada como la varianza explicada por el modelo ajustado.

Se han propuesto varios R^2 basados en las definiciones de residuales, sin embargo una medida R^2 preferida es aquella basada en el desvío, que tiene la siguiente forma:

$$R^2 = 1 - \frac{D_p(\hat{\mu})}{D_p(\hat{\mu}_0)}$$

Donde $D_p(\hat{\mu})$ y $D_p(\hat{\mu}_0)$ son las funciones devianza de los modelos ajustado y nulo, respectivamente.

Esta medida satisface las siguientes propiedades:

1) $0 \leq R^2 \leq 1$
2) No decrece a medida que se añaden las variables independientes.
3) Tiene una interpretación en términos del contenido de información de los datos.

2.2.4.4. Estadística Chi-Cuadrado de Pearson

El uso de la χ^2 también es frecuentemente recomendada, se calcula como la suma de los residuales de Pearson al cuadrado; la cual toma la siguiente forma:

$$\chi^2\ pearson = \sum_{i=1}^{n} \left(\frac{y_i - \mu_i}{\sqrt{V(\mu_i)}} \right)^2$$

La estadística χ^2 se aproxima asintóticamente a una distribución χ^2_{n-p}. Las pruebas de hipótesis son las mismas que para la función de Devianza, y el modelo ajustado es adecuado sí:

$$\chi^2\ pearson \leq \chi^2_{(\alpha, n-p)}$$

2.2.5. Análisis de Variables Independientes: Pruebas de Significancia del Modelo Ajustado

Los métodos de inferencia en los Modelos Lineales Generalizados se basan fundamentalmente en la teoría de Máxima Verosimilitud. Según ésta, existen tres estadísticas para probar hipótesis relativas a los parámetros β que son deducidas de las distribuciones asintóticas de las funciones adecuadas de las estimaciones de los β .

Estas son:

1) Test de Razón de Verosimilitud.
2) Test de Wald: también conocida como de Máxima Verosimilitud por algunos autores, se basa en la distribución normal asintótica del vector $\hat{\beta}$.
3) Test de Score: obtenida a partir de la función score.

Las hipótesis a probar se definen de la siguiente manera:

$$H_0 \quad : \quad \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 \quad : \quad al\ menos\ un\ \beta_i \neq 0$$

2.2.5.1. Test de Razón de Verosimilitud (LRT)

Conocido también como prueba "G" compara el logaritmo de la verosimilitud del modelo nulo con el logaritmo de verosimilitud del modelo ajustado. Está definido por:

$$LRT = -2\ ln\left(\frac{L(modelo\ nulo)}{L(modelo\ ajustado)}\right)$$

O equivalentemente:

$$LRT = -2[lnL(modelo\ nulo) - lnL(modelo\ ajustado)] \sim \chi_p^2$$

Se rechaza H_0 si $LRT \geq \chi^2_{(\alpha,p)}$

En el caso de las distribuciones Binomial y Poisson, el valor de LRT es igual que el de las Devianzas respectivas, pues $\emptyset = 1$.

2.2.5.2. Test de Wald (W)

El test de Wald es un método alternativo para probar la significancia del modelo, y se define por:

$$W = [\hat{\beta}_{ajustado} - \beta_{nulo}]' \hat{Var}^{-1}(\hat{\beta})[\hat{\beta}_{ajustado} - \beta_{nulo}]$$

Donde $\hat{Var}^{-1}(\hat{\beta})$ es la estimación de la matriz de varianza – covarianza asintótica de $\hat{\beta}$, entonces:

$$W = \frac{1}{\emptyset}[\hat{\beta}_{ajustado} - \beta_{nulo}]'(X^T \widehat{W} X)[\hat{\beta}_{ajustado} - \beta_{nulo}] \quad (\mathbf{21})$$

Se aproxima a una distribución χ^2 con grados de libertad igual a la dimensión de $\hat{\beta}_{ajustado}$ y se rechaza H_0 si $W \geq \chi^2_{(\alpha,p)}$.

Para muestras grandes LRT y Wald dan resultados similares, aunque no los mismos, sin embargo para muestras pequeñas, pueden diferir, varias investigaciones han demostrado que el uso de LRT da aproximaciones más confiables para muestras pequeñas que con la prueba de Wald (McCullagh et al., 2001; Agresti., 1996) debido a que para muestras grandes la varianza se infla.

2.2.5.3. Test de Score (S)

También conocido como el Test de Rao, se define como:

$$S = U(\hat{\beta}_{nulo})^T \hat{Var}_0(\hat{\beta}_{ajustado})^{-1} U(\hat{\beta}_{nulo}) \quad (\mathbf{22})$$

Donde $\hat{V}ar_0(\hat{\beta}_{ajustado})$ es la varianza asintótica de $\hat{\beta}_{ajustado}$ bajo $H_0: \beta = \hat{\beta}_{nulo}$. Esta estadística se define también de la siguiente forma:

$$S = \emptyset(\hat{\beta}_{nulo})^T (X^T \widehat{W}_0 X)^{-1} U(\hat{\beta}_{nulo}) \quad (\mathbf{23})$$

Donde $\widehat{W}_0$ es estimado bajo H_0.

Asintóticamente y bajo la hipótesis nula, las tres estadísticas definidas LRT, W y S se distribuyen como χ^2_p. Para las hipótesis relativas a un único coeficiente β, la estadística de

Wald es la más usada. Para hipótesis relativas a varios coeficientes, la razón de Máxima Verosimilitud es preferida por ser un test uniformemente más poderoso. (Figueroa Arboccó, 2005)

2.2.6. Modelos Lineales Generalizados para datos Binarios

2.2.6.1. Distribución Binomial y su Modelo

Sea $Y \sim B(n,p)$, la función de densidad de Y esta dada por:

$$f(y;\theta,\emptyset) = \binom{n}{y} p^y (1-p)^{n-y}$$

$$f(y;\theta,\emptyset) = exp\left(yln(p) - (y-n)ln\,(1-p) + \binom{n}{y}\right)$$

$$f(y;\theta,\emptyset) = exp\left(yln\left(\frac{p}{1-p}\right) - nln\left(\frac{p}{1-p}\right) + ln\binom{n}{y}\right)$$

$$f(y;\theta,\emptyset) = exp\left(yln\theta - nln\left(1+e^{\theta}\right) + ln\binom{n}{y}\right)$$

$$a(\emptyset) = 1$$

$$b(\theta) = nln\left(1+e^{\theta}\right)$$

$$c(y,\emptyset) = ln\binom{n}{y}$$

$$\mu(\theta) = np = n\frac{e^{\theta}}{1+e^{\theta}}$$

$$\theta = logit(p) = ln\left(\frac{p}{1-p}\right) \Leftrightarrow e^{\theta} = \frac{p}{1-p} \Leftrightarrow e^{\theta} = \left(1+e^{\theta}\right)p \Leftrightarrow p = \frac{e^{\theta}}{1+e^{\theta}}$$

Supongamos que tenemos n variables de respuestas independientes $Y_i \sim B(1,p_i)$.

$$f(y_i,p_i) = p_i^{y_i}(1-p_i)^{1-y_i}, y_i = 0{,}1$$

Cada individuo i está asociado a un vector específico $X_i, i = 1{,}2,\dots,n$ resultando del vector de covarianzas.

Como $E(y_i) = p_i$, para este modelo se tiene, $\theta_i = \eta_i = \boldsymbol{X_i}'\boldsymbol{\beta}$

Concluimos que la función de enlace canónico es una función $logit$. Asumiendo una probabilidad de suceso, $P(y_i = 1) = p_i$ está relacionada con el vector X_i a través de,

$$p_i = \frac{exp(\boldsymbol{X_i'\beta})}{1 + exp(\boldsymbol{X_i'\beta})}$$

La función $F: \mathbb{R} \to [0,1]$, definida por, $F(x) = \frac{exp(x)}{1+exp(x)}$ es una función de distribución. El MLG definido por el modelo binomial con función de enlace canónico $logit$ es conocido por Modelo de Regresión Logística. (Tomaiconza Ataulluco & Pari Sallo, 2014)

2.2.7. Modelos de Elección Binaria

A continuación comenzaremos desarrollando la distribución binomial, seguidamente desarrollaremos los modelos de elección binaria, como son el Modelo de Regresión Logística (caso particular del modelo de Distribución Binomial) y Modelo de Regresión Probit, por ser parte de los modelos lineales generalizados; asimismo, tomando en cuenta que nuestro interés es predecir los valores de una variable dicotómica (binaria); es decir una variable que solo puede tomar dos valores, los cuales son complementarios y no comparables, profundizaremos el desarrollo de la Regresión Logística cuya variable dependiente es binaria o dicotómica.

A continuación se muestra la clasificación de los modelos de elección binaria:

Tabla 6:
Clasificación de los Modelos de Elección Binaria

N° de alternativas	Tipo de alternativas	Tipo de función	El regresor se refiere a	
			Características (de los individuos)	Atributos (de las alternativas)
Modelos de respuesta dicotómica (2 alternativas)	Complementarias	Lineal	Modelo de Probabilidad Lineal Tipificado	
		Logística	Modelo Logit	
		Normal tipificada	Modelo Probit	
Modelos de respuesta múltiple (más de 2 alternativas)	No ordenadas	Logística	Logit Multinomial • Logit Anidado • Logit Mixto	Logit Condicional • Logit Anidado • Logit Mixto
		Normal tipificada	Probit Multinomial Probit Multivariante	Probit Condicional Probit Multivariante
	Ordenadas	Logística	Logit Ordenado	
		Normal tipificada	Probit Ordenado	

2.2.7.1. Distribución Binomial

La distribución binomial fue desarrollada por Jakob Bernoulli (Suiza, 1654-1705) y es la principal distribución de probabilidad discreta para variables dicotómicas, es decir, que solo pueden tomar dos posibles resultados. Bernoulli definió el proceso conocido por su nombre. Dicho proceso, consiste en realizar un experimento aleatorio una sola vez y observar si cierto suceso ocurre o no, siendo p la probabilidad de que ocurra (éxito) y $q = 1\text{-}p$ de que no ocurra (fracaso), por lo que la variable solo puede tomar dos posibles valores, el 1 si ocurre y el 0 sino sucede.

La distribucion binomial es una generalización de la distribución de Bernoulli, cuando en lugar de realizar el experimento aleatorio una sola vez, se realiza n, siendo cada ensayo independiente del anterior. (Martínez Gómez & Marí Benlloch, 2002).

Consideremos una variable aleatoria X que da el número de éxitos que aparecen al repetir n veces de forma independiente un experimento en idénticas condiciones. En esta situación diremos que X sigue una distribución Binomial.

Las características principales de este modelo de distribución son:

- Repetir n pruebas independientes unas de otras.
- Para cada una de las pruebas solo pueden darse dos resultados: éxito o fracaso
- La probabilidad de éxito en cada prueba es de p.

En tales condiciones, diremos que la variable aleatoria $Y =$"N° de éxitos en las n pruebas" sigue una distribución Binomial de parámetros n y p, y lo escribiremos como $Y \sim B(n, p)$.

Observamos que la variable aleatoria Y solo puede tomar los valores $0, 1, 2, 3, \dots, n$ siendo por tanto una variable aleatoria discreta.

Así pues, la funcion de probabilidad binomial es la siguiente:

$$f(x) = P(Y = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad para\ x = 0,1,2,3, \dots, n \qquad (\mathbf{24})$$

Donde:

$\binom{n}{x} = \frac{n!}{x!(n-x)!}$: Es la combinación lineal de n tomadas de x en x.

Y : Es la variable dependiente que sigue una distribución Binomial.

n y p : Son los parámetros de la distribución Binomial.

Propiedades de la Distribución Binomial

Dentro de las características de la distribución Binomial podemos mencionar:

- La esperanza o media de la variable aleatoria binomial es:

$$\mu = E(Y) = np \quad (\mathbf{25})$$

- La varianza es:

$$\sigma^2 = np(1-p) \quad (\mathbf{26})$$

2.2.7.2. Modelo de Regresión Logística:

Los modelos de regresión logística son herramientas que permiten explicar el comportamiento de una variable respuesta discreta (binaria o con más de dos categorías) a través de una o varias variables independientes explicativas de naturaleza cuantitativa y/o cualitativa. Según el tipo de variable respuesta estaremos hablando de regresión logística binaria (variable dependiente con 2 categorías), o de regresión logística multinomial (variable dependiente con más de 2 categorías), pudiendo ser esta última de respuesta nominal u ordinal. Los modelos de respuesta discreta son un caso particular de los modelos lineales generalizados formulados por Nelder yWedderburn en 1972, al igual que los modelos de regresión lineal o el análisis de la varianza. (Iglesias Cabo, 2013)

2.2.7.2.1. Función Logística

La Función Logística expresa una relación entre dos o más variables de forma que a cada elemento x de las categorías de la variable independiente, X, le corresponde un único elemento $\pi(x)$ de la variable dependiente y está representada por:

$$f(x) = \frac{1}{1+e^{-x}} \quad (\mathbf{27})$$

Su gráfica es una curva S o Sigmoidea, tiene un único punto de inflexión en el que cambia la concavidad y la rapidez del crecimiento. (Meza Saldaña, Reyes Cervantes, Pérez Salvador, & Tajonar Sanabria, 2006).

Además la función Logística verifica que: $0 \leq f(x) \leq 1$ y puede ser interpretada en términos de probabilidad.

Por otra parte la función logística verifica las siguientes propiedades:

- $\lim_{x\to-\infty} \frac{1}{1+e^{-x}} = 0$
- $\lim_{x\to+\infty} \frac{1}{1+e^{-x}} = 1$
- $f(0) = \frac{1}{1+e^{-0}} = \frac{1}{2}$

Asimismo; multiplicando al numerador y denominador de la función logística por el factor e^x, obtenemos la función de distribución Logística, de la siguiente forma:

$$F(x) = \frac{1.e^x}{1.e^x + e^{-x}.e^x}$$

$$F(x) = \frac{e^x}{1+e^x} \qquad (\mathbf{28})$$

Es necesario realizar una transformación para lo cual partimos de un modelo de regresión que permite explicar el comportamiento de la variable dependiente Y en función de una serie de k variables independientes, y un término de perturbación u.

$$Y = f(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u) \qquad (\mathbf{29})$$

Si consideramos a Y (variable respuesta), como una variable dicotómica de forma que; si $Y = 1$, indica que el individuo presenta el evento de interés, caso contrario será $Y = 0$, esta variable así definida sigue una distribución binomial, cuyos parámetros son 1 y p, donde p es la probabilidad de que un individuo presenta el evento de interés, o que $Y = 1$, para este caso la media de la variable binomial será, $E(Y) = 1.p = p$.

Considerando que este modelo de regresión es lineal, con el supuesto habitual de que $E(u) = 0$, se tendría la expresion anterior de la siguiente forma:

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = p$$

Y no existiendo restricción sobre los parámetros del modelo, las estimaciones de los parámetros están dadas por la suma $\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$.

Pero existe un problema si, la expresión anterior es superior a la unidad o inferior a cero, debido a que no tiene sentido tener estimaciones de probabilidades fuera del intervalo $0 < p \leq 1$, esto no ocurre en regresión lineal ya que al ser Y normal condicionada a los valores de las X_k, puede tomar cualquier valor; pero no así cuando Y es binomial, además las hipótesis de varianza constante y que la perturbación siga una distribución normal tampoco se cumple. Es por esto que se toman otras medidas como son los Odds, que es el cociente $\frac{p}{1-p}$ (probabilidad de presentar el evento entre la probabilidad de no presentarla); aquí los valores del cociente serán números positivos, debido a que $0 < p \leq 1$; además si p es poco probable o es próximo a cero el cociente también lo será, y en caso de ser cercano a 1, el cociente también.

Si consideramos la transformación mediante el logaritmo natural del parámetro p; donde los posibles valores serán cualquiera de los números reales, positivo o negativo, con lo que se soluciona el problema mencionado. A esta transformación de p se denomina transformación Logística o transformación Logit de la probabilidad de p.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) \qquad (\mathbf{30})$$

Entonces el modelo que permite resolver el problema, queda representado por:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad (\mathbf{31})$$

Este modelo fue planteado en estudios de cohortes y porsteriormente para casos y controles, y su atractivo está en que sus parámetros son interpretables como una medida de riesgo asociado a las variables predictoras. (Rodas Guizado, 2011)

El modelo de regresión logística es un modelo estadístico en el que se desea conocer la relación entre una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos valores (regresión logística múltiple o multinomial).

Una o más variables explicativas independientes, o covariables, ya sean cualitativas o cuantitativas, siendo la ecuación inicial del modelo de tipo exponencial, si bien su transformación logarítmica (logit) permite su uso como una función lineal.

Como se ve, las covariables pueden ser cuantitativas o cualitativas. Las covariables cualitativas deben ser dicotómicas, tomando valores 0 para su ausencia y 1 para su presencia (esta codificación es importante, ya que cualquier otra codificación provocaría modificaciones en la interpretación del modelo). Pero si la covariable cualitativa tuviera más de dos categorías, para su inclusión en el modelo debería realizarse una transformación de la misma en varias covariables cualitativas dicotómicas ficticias o de diseño (las llamadas variables dummy), de forma que una de las categorías se tomaría como categoría de referencia. Con ello cada categoría entraría en el modelo de forma individual. En general, si la covariable cualitativa posee n categorías, habrá que realizar $n - 1$ covariables ficticias.

Por sus características, los modelos de regresión logística permiten dos finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente, lo que lleva implícito también clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, conocer la odds ratio para cada covariable).

- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente, según la probabilidad que tenga de pertenecer a una de ellas dada la presencia de determinadas covariables.

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico). (SEOC)

2.2.7.2.2. Definición del Modelo de Regresión Logística

Sea Y una variable dependiente binaria (con dos posibles valores: 0 y 1). Sea un conjunto de k variables independientes, $(X_1, X_2, \dots X_k)$, observadas con el fin de predecir y/o explicar el valor de Y.

El objetivo consiste en determinar:

$$P[Y = 1/X_1, X_2, \dots, X_k] \rightarrow P[Y = 0/X_1, X_2, \dots, X_k] = 1 - P[Y = 1/X_1, X_2, \dots, X_k]$$

Para ello, se construye el modelo

$$P[Y = 1/X_1, X_2, \dots X_k] = p(X_1, X_2, \dots X_k; \beta) = p \qquad (\mathbf{32})$$

Donde: $p(X_1, X_2, \dots, X_k; \boldsymbol{\beta})$: $R^k \rightarrow [0,1]$ es una función que recibe el nombre de función de enlace (Función de probabilidad) cuyo valor depende de un vector de parámetros $\boldsymbol{\beta}' = (\beta_1, \beta_2, \dots, \beta_k)$. (De la Fuente Fernández, 2011).

2.2.7.2.3. Forma Muestral del Modelo de Regresión Logística

Sea $Y' = (y_1, y_2, \dots, y_k)$ el vector de componentes de una variable dependiente binaria donde cada observación y_i puede tomar dos valores, 0 (Fracaso) y 1 (Exito); además están distribuidas independientemente y con una distribución de Bernoulli; esto es, $y_i \sim B(1, p_i)$, sean también un conjunto de k variables independientes $X_1, X_2, \dots X_k$ que explican el valor de Y. Entonces la función de probabilidad estará dada por:

$$P(Y = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \qquad i = 1,2, \dots, n \tag{33}$$

Y para una muestra de tamaño n la función de distribución conjunta será:

$$P(Y = y_1, Y = y_2, \dots, Y = y_n) = \prod_{i=1}^{n} P(Y = y_i)$$

$$P(Y = y_1, Y = y_2, \dots, Y = y_n) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{34}$$

De aquí que la esperanza y la varianza, serán:

$$E(y_i) = p_i \tag{35}$$

$$Var(y_i) = p_i(1 - p_i) \tag{36}$$

2.2.7.2.4. Función de Verosimilitud

Con el fin de estimar $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)$ y analizar el comportamiento del modelo estimado se toma una muestra aleatoria de tamaño n dada por $(x_i, y_i), i = 1,2, \dots, n$; donde el valor de las variables independientes es $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$ e $y_i \in [0,1]$ es el valor observado de Y en el i-ésimo elemento de la muestra.

Como $(Y/X_1, X_2, \dots, X_k) \in B[1, p(X_1, X_2, \dots X_k; \beta)]$ esto debido a que la variable dependiente toma solo dos resultados (éxito y fracaso), cuando el número de éxitos en n repeticiones tiene una distribución Binomial $B(n, p)$; entonces la función de probabilidad de una observación de cualquier y es:

$$P(Y/x) = P(Y = y) = p^y(1 - p)^{1-y}$$

La función de verosimilitud viene dada por:

$$L[\beta/(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)] = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \tag{37}$$

Donde $p_i = p(x_1; \beta) = p[(x_{i1}, x_{i2}, \dots, x_{ik}); \beta], i = 1,2, \dots, n$

La Regresión Logística se clasifica en: Regresión Logística Binaria (simple y múltiple) y Regresión Logística Multinomial.

2.2.7.3. Modelo de Regresión Logística Binaria

Los modelos de regresión logística binaria resultan con mayor interés ya que la mayor parte de las circunstancias analizadas en cualquier rama responden a este modelo (presencia o no, éxito o fracaso, etc.). Como se ha visto, la variable dependiente será dicotómica que se codificará como 0 o 1 (respectivamente, "ausencia" o presencia"). Este aspecto de la codificación de las variables no es banal (influye en la forma en que se realizan los cálculos matemáticos), y habrá que tenerlo muy en cuenta si se emplean paquetes estadísticos que no recodifican automáticamente las variables cuando éstas se encuentran codificadas de forma diferente (por ejemplo, el uso frecuente de 1 para la presencia y – 1 o 2 para la ausencia).

Sea

$$p(X_1, X_2, \dots X_k; \boldsymbol{\beta}) = F(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

donde:

$F(x) = \frac{e^x}{1+e^x}$; es la distribución o función de densidad acumulada.

Por la transformación logística, de la ecuación (32). Partimos de la hipótesis de que los datos siguen el modelo:

$$\ln\left(\frac{p(X_1, X_2, \dots X_k; \beta)}{1 - p(X_1, X_2, \dots X_k; \beta)}\right) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad (\mathbf{38})$$

Este es el modelo Logit. La expresión (38), por definición de logaritmos es equivalente a la siguiente ecuación

$$\frac{P(Y = 1/X_1, X_2, \dots X_k)}{P(Y = 0/X_1, X_2, \dots X_k)} = \frac{p(X_1, X_2, \dots X_k; \beta)}{1 - p(X_1, X_2, \dots X_k; \beta)} = e^{\beta_0} \times e^{\beta_1 X_1} \times e^{\beta_2 X_2} \times \dots \times e^{\beta_k X_k} \qquad (\mathbf{39})$$

Que es el coeficiente de probabilidades, conocido como factor de riesgo, donde la variable Y toma el valor de 1 cuando existe la presencia de la característica y en ausencia de la característica toma el valor de 0.

Al despejar la probabilidad p, se tendrá:

$$p(X_1, X_2, \dots X_k; \boldsymbol{\beta}) = \frac{\exp\left(\beta_o + \sum_{i=1}^{k} \beta_i X_i\right)}{1 + \exp\left(\beta_o + \sum_{i=1}^{k} \beta_i X_i\right)} \quad (\mathbf{40})$$

Entonces la probabilidad p sigue una distribución logística.

Donde:

$p(X_1, X_2, \dots X_k; \boldsymbol{\beta})$: Es la probabilidad en presencia de las covariables $X_1, X_2, \dots X_k$.

$X_1, X_2, \dots X_k$: Es un conjunto de k covariables que forman parte del modelo;

β_0 : Es la constante del modelo o término independiente;

β_i : Los coeficientes de las covariables.

(De la Fuente Fernández, 2011)

2.2.7.3.1. Función de Verosimilitud para la Regresión Logística Binaria

Tomando en cuenta la forma matricial del predictor lineal $(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$, de la ecuación (40) se tiene:

$$p(X_1, X_2, \dots X_k; \boldsymbol{\beta}) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) = \frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}} \quad (\mathbf{41})$$

Entonces la función de verosimilitud según la ecuación (37) viene dada por:

$$L(\boldsymbol{\beta}) = L\{[(X_1', y_1), \dots, (X_n', y_n)]; \boldsymbol{\beta}\} = \prod_{i=1}^{n} \left(\frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}}\right)^{y_i} \left(\frac{1}{1 + e^{X_i' \beta}}\right)^{1 - y_i} \quad (\mathbf{42})$$

2.2.7.3.2. Estimación de Parámetros del Modelo de Regresión Logística Binaria

Una vez que tenemos el modelo pasamos a estimar el vector de parámetros $\boldsymbol{\beta}$ mediante el método de máxima verosimilitud, que consiste en elegir el valor de $\widehat{\boldsymbol{\beta}}$**,** como estimador para $\boldsymbol{\beta}$, nuestro objetivo es que la función de verosimilitud $L(\boldsymbol{\beta})$ sea máxima y para esto aplicamos logaritmo a la función en la ecuación de la siguiente manera:

$$lnL(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = ln\left[\prod_{i=1}^{n} \left(\frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}}\right)^{y_i} \left(\frac{1}{1 + e^{X_i' \beta}}\right)^{1 - y_i}\right]$$

De la ecuación (41) reemplazando por p_i, se tiene:

$$lnL(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) = ln\left[\prod_{i=1}^{n} {p_i}^{y_i} (1 - p_i)^{1 - y_i}\right] \quad (\mathbf{43})$$

Resolviendo el logaritmo, se tiene:

$$l(\boldsymbol{\beta}) = lnL(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i lnp_i + (1 - y_i)\ln(1 - p_i)] \quad (\mathbf{44})$$

Luego derivando la ecuación (44) respecto a $\boldsymbol{\beta}$, e igualando a cero, se obtiene:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - p_i] = 0 \quad (\mathbf{45})$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} X_i[y_i - p_i] = 0 \quad (\mathbf{46})$$

Donde:

$p_i = p(X_i; \boldsymbol{\beta}), i = 1,2, \dots, n$

Se llega a las ecuaciones (45) y (46) mediante métodos iterativos de derivación (Método de Newton Raphson).

2.2.7.3.3. Evaluación del Modelo

La evaluación del modelo estimado es un aspecto muy importante en el análisis de la regresión, porque comprende:

- La validación de supuestos de independencia de las observaciones, la no multicolinealidad de las variables explicativas.
- El análisis de los diversos tipos de residuos nos permite detectar observaciones atípicas.
- El análisis de influencia, permite detectar a conjuntos de observaciones que influyen en diversos aspectos del análisis de regresión, como el ajuste del modelo o la estimación de los parámetros del modelo.

Si el modelo cumple todos estos análisis estará en condiciones de ser utilizada, y por tanto ayudará al cumplimiento de objetivos.

2.2.7.4. Regresión Logística Binaria Simple

El modelo de Regresión Logística Binaria Simple tiene como objetivo, estudiar la relación existente entre una variable respuesta (con dos posibles valores 1 y 0) y una sola variable independiente o explicativa.

Asimismo, se sabe que en cualquier problema de regresión la cantidad clave es el valor promedio de la variable respuesta, dado el valor de la variable independiente. Esta cantidad se denomina media condicional y se expresa como "$E(Y/x)$" donde Y denota a la variable respuesta y x denota un valor de la variable independiente.

La cantidad $E(Y/x)$ se lee "El valor esperado de Y, dado el valor de x". En la regresión lineal, asumimos que esta media puede ser expresada como una ecuación lineal en x (o alguna transformación de x o Y), como:

$$E(Y/x) = \beta_0 + \beta_1 x$$

Esta expresión implica que es posible para $E(Y/x)$ tomar un valor como x que varíe entre $-\infty$ y $+\infty$ y por consiguiente los valores verdaderos de $E\ (Y/x)$ para proporcionar una evaluación razonable de la relación entre la variable respuesta y la variable independiente con los datos dicotómicos la media condicional debe ser mayor o igual que 0 y menor o igual a 1 es decir, $[0 \leq E(Y/x) \leq 1]$.

El cambio en la $E(Y/x)$ por unidad de cambio en x se vuelve progresivamente más pequeñas según la media condicional se aproxima a 0 o 1. Se dice que la curva tiene forma de S. Se asemeja a un diagrama de una distribución acumulada de la variable aleatoria. No parece sorprendente que se hayan utilizado algunas distribuciones acumuladas bien conocidas para proporcionar un modelo para $E\ (Y/x)$ en el caso cuando y es dicotómica.

2.2.7.4.1. Formulación del Modelo

A fin de simplificar la notación, utilizamos la cantidad $\pi(x) = E(Y/x)$ para representar la media condicional de Y dado x cuando se utiliza la distribución logística. La forma específica del modelo de regresión logística que utilizamos es:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \qquad (\mathbf{47})$$

Una transformación de $\pi(x)$ que es fundamental para nuestro estudio de regresión logística es la transformación logit. Esta transformación se define en términos de $\pi(x)$, como:

$$g(x) = ln\left[\frac{\pi(x)}{1 + \pi(x)}\right]$$

$$= \beta_0+\beta_1 x \qquad (\mathbf{48})$$

La importancia de esta transformación es que $g(x)$ tiene muchas de las propiedades deseables de un modelo de regresión lineal. El logit, $g(x)$, es lineal en sus parámetros, puede ser continuo, y puede variar entre $-\infty$ y $+\infty$, dependiendo del rango de x.

La segunda diferencia importante entre los modelos lineales y la regresión logística se refiere a la distribución condicional de la variable respuesta. En el modelo de regresión lineal, suponemos que una observación de la variable respuesta puede estar expresada como $y = E(Y/x) + \varepsilon$. La cantidad ε se denomina el error y expresa una desviación de la media condicional. La hipótesis más común es que ε sigue una distribución normal con media cero y alguna varianza que es constante en los niveles de la variable independiente. De ello se deduce que la distribución condicional de la variable respuesta dado x será normal con media $E(Y/x)$ y varianza que es constante. Este no es el caso de una variable respuesta dicotómica. En esta situación podemos expresar el valor de la variable respuesta dado x como $y = \pi(x) + \varepsilon$. Aquí la cantidad ε podrá asumir uno de dos posibles valores. Si $y = 1$, entonces $\varepsilon = 1 - \pi(x)$ con probabilidad $\pi(x)$, y si $y = 0$ el valor de $\varepsilon = -\pi(x)$ con probabilidad $1 -$

$\pi(x)$. Por lo tanto, ε tiene una distribución con media cero y varianza igual a $\pi(x)[1 - \pi(x)\ ]$. Es decir, la distribución condicional de la variable respuesta sigue una distribución binomial con una probabilidad dada por la media condicional, $\pi(x)$.

En resumen, se ha observado que en un análisis de regresión cuando la variable respuesta es dicotómica:

(1) La ecuación de la media condicional de la regresión debe formularse para estar delimitado entre 0 y 1. Hemos señalado que el modelo de regresión logística, $\pi(x)$ dada en la ecuación (47), satisface esta restricción.

(2) La distribución binomial, no la normal, describe la distribución de los errores y será la distribución estadística sobre la cual está basada el análisis.

(3) Los principios que guían un análisis mediante regresión lineal también nos guían en una regresión logística.

2.2.7.4.2. Ajuste del Modelo de Regresión Logística

Supongamos que tenemos una muestra de n observaciones independientes del par $(x_i, y_i), i = 1,2, \dots, n$, donde y_i denota el valor de una variable respuesta dicotómica y x_i es el valor de la variable independiente para el i-ésimo sujeto. Además, supongamos que la variable respuesta ha sido codificada como 0 o 1, que representa la ausencia o la presencia de la característica, respectivamente. Para ajustar el modelo de regresión logística en la ecuación (47) a un conjunto de datos requiere que estimemos los valores de β_0 y β_1, los parámetros desconocidos.

En la regresión lineal, el método utilizado con mayor frecuencia para estimar los parámetros desconocidos es mínimos cuadrados. En ese método elegimos los valores de β_0 y β_1 que minimizan la suma de las desviaciones cuadradas de los valores observados de Y a partir de los valores pronosticados basados en el modelo. Bajo las hipótesis usuales para la regresión lineal, el método de mínimos cuadrados produce estimadores con una serie de

propiedades estadísticas deseables. Desafortunadamente, cuando el método de mínimos cuadrados es aplicado a un modelo con una respuesta dicotómica, los estimadores ya no tienen estas mismas propiedades.

El método general de estimación que conlleva a los mínimos cuadrados para funcionar bajo el modelo de regresión lineal (cuando los términos del error son normalmente distribuidos) se denomina máxima verosimilitud. Este método proporcionará la base para nuestro enfoque de estimación con el modelo de regresión logística. En un sentido muy general, el método de máxima verosimilitud produce valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observados. Para aplicar este método debemos primero construir una función, llamada función de verosimilitud. Esta función expresa la probabilidad de los datos observados en función de los parámetros desconocidos. Los estimadores de máxima verosimilitud de estos parámetros son escogidos para ser los valores que maximizan esta función. Así, los estimadores resultantes son los que están más estrechamente con los datos observados. Ahora describimos cómo encontrar estos valores a partir del modelo de regresión logística.

2.2.7.4.3. Estimación de los Parámetros del Modelo

Si Y se codifica como 0 o 1, entonces la expresión para $\pi(x)$ dada en la ecuación (47) proporciona (para un valor arbitrario de $\boldsymbol{\beta} = (\beta_0, \beta_1)$, el vector de parámetros) la probabilidad condicional de Y igual a 1 dado x. Esto se denota como $P(Y = 1/x)$. De ello se deduce que la cantidad $1 - \pi(x)$ da la probabilidad condicional de Y igual a 0, dado x, $P(Y = 0/x)$. Por tanto, para los pares (x_i, y_i), donde $y_i = 1$, el aporte de la función de máxima verosimilitud es $\pi(x_i)$, y para aquellos pares donde $y_i = 0$ el aporte de la función de máxima verosimilitud es $1 - \pi(x_i)$, donde la cantidad $\pi(x_i)$ denota el valor de $\pi(x)$ calculado en x_i. Una manera conveniente de expresar la contribución a la función de verosimilitud para el par (x_i, y_i) es a través de la expresión:

$$\pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (\mathbf{49})$$

Dado que se supone que las observaciones son independientes, la función de verosimilitud se obtiene como el producto de los términos dados en la expresión (49) de la siguiente manera:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i} \quad (\mathbf{50})$$

El principio de máxima verosimilitud indica que usamos como estimación de $\boldsymbol{\beta}$ el valor que maximiza la expresión en la ecuación (50). Sin embargo, es más fácil matemáticamente trabajar con el log de la ecuación (50). Esta expresión, logaritmo de verosimilitud, se define como:

$$L(\beta) = ln[l(\beta)] = ln\left[\prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}\right]$$

$$L(\beta) = \sum_{i=1}^{n} \{y_i ln[\pi(x_i)] + (1-y_i) ln[1-\pi(x_i)]\} \quad (\mathbf{51})$$

Para encontrar el valor de β que maximiza $L(\beta)$ derivamos $L(\beta)$ con respecto a β_0 y β_1 e igualamos la expresión resultante a cero. Estas ecuaciones, son conocidas como las ecuaciones de verosimilitud, y son:

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \quad (\mathbf{52})$$

Y

$$\frac{\partial L(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_i[y_i - \pi(x_i)] = 0 \quad (\mathbf{53})$$

En la regresión lineal, las ecuaciones de verosimilitud, obtenidas por la diferencia de las sumatorias de la función de las desviaciones cuadradas con respecto a β son lineales en los parámetros desconocidos y por lo tanto se resuelven fácilmente. Para la regresión logística las

expresiones en las ecuaciones (52) y (53) son no lineales en β_0 y β_1, y así requieren los métodos especiales para su solución. El valor $\boldsymbol{\beta}$ dado por la solución de las ecuaciones (52) y (53) se llama la estimador de máxima verosimilitud y se denota como $\hat{\beta}$. En general, el uso del símbolo $\hat{}$ denota la estimación de máxima verosimilitud de la cantidad respectiva. Por ejemplo, $\hat{\pi}(x_i)$. Esta cantidad proporciona una estimación de la probabilidad condicional de Y igual a 1, dado x igual a x_i.

Una interesante consecuencia de la ecuación (52) es:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i)$$

Es decir, la suma de los valores observados de y es igual a la suma de los valores predictivos (esperados). Esta propiedad será especialmente útil cuando hablemos de evaluar el ajuste del modelo. (Hosmer & Lemeshow, 2000)

2.2.7.4.4. Prueba para la Significancia de los Coeficientes

Después de estimar los coeficientes, nuestra primera visión al modelo ajustado se refiere a la evaluación de la significancia de las variables en el modelo. Esto normalmente implica la formulación y prueba de una hipótesis estadística para determinar si las variables independientes en el modelo se relacionan significativamente con la variable respuesta. El método para realizar esta prueba es bastante general y difiere de un tipo de modelo próximo solo en los detalles específicos. Comenzamos discutiendo el enfoque general para una única variable independiente.

Un método para probar la significancia del coeficiente de una variable en cualquier modelo se refiere a la siguiente pregunta. ¿El modelo que incluye la variable independiente nos dice más acerca de la variable respuesta que un modelo que no incluya dicha variable? Esta pregunta se responde comparando los valores observados de la variable respuesta a las predictoras por cada uno de los dos modelos; el primero y el segundo sin la variable

independiente. La función matemática utilizada para comparar los valores observados y pronosticados depende del problema particular. Si los valores pronosticados con la variable en el modelo son mejores, o más exacto en cierto sentido, que cuando la variable no está en el modelo, se considera que la variable independiente es significativa. Es importante señalar que no estamos considerando la pregunta, si los valores previstos son una representación exacta de los valores observados en un sentido absoluto (esto se denominaría bondad de ajuste). El método general para evaluar la significancia de las variables se ilustra fácilmente en el modelo de regresión lineal, y su uso allí motivará el enfoque utilizado para la regresión logística. Una comparación de los dos enfoques resaltará las diferencias entre el modelamiento de variables respuesta dicotómicas y continuas.

En regresión lineal, se aborda la evaluación de la significancia del coeficiente de la pendiente formando lo que se denomina tabla de análisis de varianza. Esta tabla particiona la suma total de desviaciones cuadradas de observaciones sobre su media en dos partes:

1) La suma de las desviaciones cuadradas de las observaciones sobre la línea de regresión SSE, (o suma de cuadrados residuales), y
2) La suma de cuadrados de valores predictores, basados en el modelo de regresión, sobre la media de la variable dependiente SSR, (o por la suma de cuadrados de regresión).

Esto es solo una manera conveniente de mostrar la comparación de los valores observados a los predictores bajo dos modelos. En la regresión lineal, la comparación de los valores observados y predictores se basa en el cuadrado de la distancia entre los dos. Si y_i denota el valor observado y $\hat{y}_i$ denota el valor estimado para el individuo i-ésimo bajo el modelo, entonces la estadística utilizada para evaluar esta comparación es:

$$SSE = \sum_{I=1}^{n} (y_i - \hat{y}_i)^2$$

Bajo el modelo que no contiene la variable independiente en cuestión el único parámetro es β_0, y $\hat{\beta}_0 = \bar{y}$ la media de la variable respuesta. En este caso, $\hat{y}_i = \bar{y}$ y SSE es igual a la varianza total. Cuando incluimos la variable independiente en el modelo, cualquier disminución en SSE se debe al hecho de que el coeficiente de la pendiente para la variable independiente no es cero. El cambio en el valor de SSE es debido a la fuente de la variabilidad de la regresión, denotada SSR. Es decir:

$$SSR = \left[\sum_{I=1}^{n}(y_i - \bar{y}_i)^2\right] - \left[\sum_{I=1}^{n}(y_i - \hat{y}_i)^2\right]$$

En regresión lineal, el interés se centra en el tamaño de SSR. Un valor elevado indica que la variable independiente es importante, mientras que un valor pequeño indica que la variable independiente no ayuda a predecir la respuesta.

El principio que rige la regresión logística es el mismo: Comparar los valores observados de la variable respuesta a valores pronosticados obtenidos a partir de modelos con y sin la variable independiente. En regresión logística, la comparación de los valores observados y predictivos se basa en el logaritmo de la función de verosimilitud definida en la ecuación (51). Para comprender mejor esta comparación, es útil conceptualmente pensar en el valor observado de la variable respuesta como un valor predictivo resultante de un modelo saturado. Un modelo saturado es uno que contiene tantos parámetros como sus puntos de datos. (Un ejemplo sencillo de un modelo saturado es un modelo de regresión lineal la función ajustada cuando solo hay dos puntos de datos, $n = 2$).

La comparación de los valores observados predictivos utilizando la función de verosimilitud se basa en la siguiente expresión.

$$D = -2ln\left[\frac{(máxima\ verosimilitud\ del\ modelo\ ajustado)}{(máxima\ verosimilitud\ del\ modelo\ saturado)}\right] \qquad (\mathbf{54})$$

La cantidad dentro de los corchetes en la expresión anterior se llama razón de verosimilitud. Es necesario usar el logaritmo y multiplicar este valor por -2 para obtener una cantidad cuya distribución es conocida y por lo tanto puede ser utilizado para propósitos de prueba de hipótesis. Esta prueba se llama prueba de la razón de verosimilitud. Entonces las ecuaciones (51) y (54) se convierte en:

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \qquad (\mathbf{55})$$

Donde $\hat{\pi}_i = \hat{\pi}(x_i)$.

La estadística, D, en la ecuación (55) se llama Devianza y desempeña un papel central en algunos enfoques para evaluar la bondad de ajuste. La Devianza mostrada en la ecuación (55), cuando se calcula para la regresión lineal, es idénticamente igual a la SSE.

Además, en un escenario donde los valores de la variable respuesta son 0 o 1, la probabilidad del modelo saturado es 1. Específicamente, se desprende de la definición de un modelo saturado $\hat{\pi}_i = y_i$ y lo más probable es que:

$$l(modelo\ saturado) = \prod_{i=1}^{n} {y_i}^{y_i} \times (1-y_i)^{(1-y_i)} = 1 \qquad (\mathbf{56})$$

Así, de la ecuación (54) se desprende que la Devianza es:

$$D = -2\ln(máxima\ verosimilitud\ del\ modelo\ ajustado) \qquad (\mathbf{57})$$

Algunos paquetes estadísticos, como el SAS, muestran el valor de la devianza como en la ecuación (57) en lugar del logaritmo de verosimilitud para el modelo ajustado. Se quiere destacar que consideramos la devianza en los mismos términos que consideramos la suma de cuadrados residuales en la regresión lineal en el contexto de las pruebas de la importancia de un modelo ajustado.

Para evaluar el significado de una variable independiente comparamos el valor de D con y sin la variable independiente en la ecuación. El cambio en D debido a la inclusión de la variable independiente en el modelo se obtiene como:

$$G = D(modelo \sin la\ variable) - D(modelo\ con\ la\ variable)$$

Esta estadística desempeña el mismo papel en la regresión logística como el numerador de la prueba F parcial en la regresión lineal. Porque la probabilidad del modelo saturado es común a ambos valores de D que se diferencia para calcular G, esto puede ser expresado como:

$$G = -2 \ln \left[\frac{(verosimilitud\ sin\ la\ variable)}{(verosimilitud\ con\ la\ variable)} \right] \quad (\mathbf{58})$$

Para el caso específico de una sola variable independiente, es fácil demostrar que cuando la variable no está en el modelo, la estimación de máxima verosimilitud β_0 es $ln\left({}^{n_1}/_{n_0}\right)$ donde $n_1 = \sum y_i$ y $n_0 = \sum(1 - y_i)$ y el valor pronosticado es constante, n_1/n. En este caso, el valor de G es:

$$G = -2 \ln \left[\frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n} \hat{\pi}_i{}^{y_i} (1 - \hat{\pi}_i)^{(1-y_i)}} \right] \quad (\mathbf{59})$$

O

$$G = 2 \left\{ \sum_{i=1}^{n} [y_i ln(\hat{\pi}_i) + (1 - y_i) ln(1 - \hat{\pi}_i)] - [n_1 ln(n_1) + n_0 ln(n_0) - n ln(n)] \right\} \quad (\mathbf{60})$$

Bajo la hipótesis de que β_1 es igual a cero, la estadística G sigue una distribución Chi-Cuadrado con 0 grados de libertad. También se necesitan supuestos matemáticos adicionales; sin embargo, para el caso anterior son más bien no restrictivos e implican tener un tamaño de muestra n suficientemente grande.

El cálculo del logaritmo de verosimilitud y la prueba de razón de verosimilitud son características estándar de todo el programa de regresión logística. Esto hace que sea fácil

comprobar el significado de la adición de nuevos términos al modelo. En el caso simple de una sola variable independiente, primero ajustamos un modelo que contenga solamente el término constante. A continuación, ajustamos un modelo que contenga la variable independiente junto con la constante. Esto da lugar a un nuevo registro de probabilidad. La prueba de razón de verosimilitud se obtiene multiplicando la diferencia entre estos dos valores por -2. Este resultado, junto con el p-valor asociado a la distribución Chi-Cuadrado, puede obtenerse a partir de muchos paquetes estadísticos. (Hosmer & Lemeshow, 2000)

Estadístico de Wald:

La prueba de Wald es obtenida comparando el estimador de máxima verosimilitud de la pendiente del parámetro $\hat{\beta}_1$, con una estimación de su error estándar. La relación resultante, bajo la hipótesis de que $\hat{\beta}_1 = 0$ sigue una distribución normal estándar. Aunque todavía no hemos discutido formalmente cómo se obtienen las estimaciones de los errores estándar de los parámetros estimados. La prueba de Wald para el modelo de regresión logística es:

- Formulación de Hipótesis

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

- Estadístico de prueba

$$W = \frac{\hat{\beta}_1}{\widehat{SE}(\hat{\beta}_1)} \sim \chi^2_{\alpha,1} \qquad (61)$$

Donde:

$\hat{\beta}_1$: Es el coeficiente de regresión logística muestral.

$\widehat{SE}(\hat{\beta}_1)$: Es el error estándar del coeficiente de regresión logística muestral.

$\chi^2_{\alpha,1}$: Es la distribución Chi-Cuadrada con un nivel de significación α y un grado de libertad

- Regla de Decisión*:*

Si $W > \chi^2_{\alpha,1}$ rechazamos H_0 con un nivel de significancia α.

- Conclusión:

La variable independiente influye en la probabilidad de las características de la variable dependiente, si la variable independiente es cualitativa los grados de libertad son igual al número de categorías menos uno. (Flores Manrique, 2002)

Hauck y Donner (1977) examinaron los resultados de la prueba de Wald y encontraron que se comporta de una manera absurda, y a menudo no rechazaban la hipótesis nula cuando el coeficiente era significativo. Recomendaron que se utilizara la prueba de razón de verosimilitud.

Jennings (1986) también ha analizado la adecuación de inferencias en la regresión logística basada en estadísticas de Wald. Sus conclusiones son similares a las de Hauck y Donner. Tanto la prueba de la razón de verosimilitud, G y la prueba de Wald, W, requieren el cálculo de la estimación de máxima verosimilitud para β_1.

En resumen, el método para probar la significancia del coeficiente de una variable en la regresión logística es similar a la utilizada en la regresión lineal; sin embargo, se utiliza la función de probabilidad de una variable respuesta dicotómica. (Hosmer & Lemeshow, 2000)

Estimación del Intervalo de Confianza

Un complemento importante a la prueba para la significación del modelo, es el cálculo e interpretación de los intervalos de confianza para los parámetros de interés. En algunos ajustes puede ser de interés proporcionar estimaciones del intervalo para los valores supuestos.

La base para la construcción de los estimadores del intervalo, es la misma teoría estadística que utilizamos para formular las pruebas para la significancia del modelo. En particular, los estimadores del intervalo de confianza para la pendiente y la intersección se basan en sus respectivas pruebas de Wald. Los extremos de un intervalo de confianza de $100\left(1-\frac{\alpha}{2}\right)\%$ para el coeficiente de la pendiente son:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1) \quad (\mathbf{62})$$

Y para la intersección son:

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_0) \quad (\mathbf{63})$$

Donde $1-\alpha/2$ es el punto superior $100\left(1-\frac{\alpha}{2}\right)\%$ de la distribución normal estándar y $\widehat{SE}(.)$ denota un estimador basado en el modelo del error estándar del estimador de parámetros respectivo. Podemos ampliar la discusión de la fórmula utilizada para el cálculo de los estimadores de los errores estándar. Por el momento utilizamos el hecho de que los valores estimados se proporcionan tras el ajuste de un modelo y, además, muchos paquetes proporcionan los extremos del intervalo de estimaciones.

El logit es la parte lineal del modelo de regresión logística y, como tal, es la más parecida a la línea ajustada en un modelo de regresión lineal. El estimador del logit es:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad (\mathbf{64})$$

El estimador de la varianza del estimador del logit requiere de la obtención de la varianza de una suma. En este caso es:

$$\hat{Var}[\hat{g}(x)] = \hat{Var}(\hat{\beta}_0) + x^2\hat{Var}(\hat{\beta}_1) + 2x\hat{Cov}(\hat{\beta}_0, \hat{\beta}_1) \quad (\mathbf{65})$$

En general, la varianza de una suma es igual a la suma de la varianza de cada término y el doble de la covarianza de cada par de términos posibles formados a partir de los componentes de la suma. Los puntos finales de un intervalo de confianza basado en Wald de $100\ (1-\alpha)\%$ para el logit son:

$$\hat{g}(x) \pm z_{1-\alpha/_2}\widehat{SE}[\hat{g}(x)] \quad (\mathbf{66})$$

Donde $\widehat{SE}[\hat{g}(x)]$ es la raíz cuadrada positiva de la varianza del estimador en (65).

Y los puntos finales de un intervalo de confianza del 95% se obtienen de los puntos finales respectivos del intervalo de confianza para el logit. Los puntos finales del intervalo de confianza basado en Wald de $100\ (1-\alpha)\%$ para el valor ajustado son:

$$\frac{e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]}}{1+e^{\hat{g}(x)\pm z_{1-\alpha/2}\widehat{SE}[\hat{g}(x)]}} \quad (67)$$

(Hosmer & Lemeshow, 2000)

2.2.7.5. Modelo de Regresión Logística Binaria Múltiple

En la sección anterior desarrollamos el Modelo de Regresión Logística en el contexto univariante, como en el caso de la regresión lineal, la fuerza de una técnica de modelado radica en su capacidad para modelar muchas variables, algunas de las cuales pueden estar en diferentes escalas de medición. Ahora se generalizará el modelo logístico para el caso de más de una variable independiente, el cual se denominará "caso multivariable". El centro de la consideración de múltiples modelos logísticos será la estimación de los coeficientes en el modelo y la prueba de significancia. En este modelo se introducirá el uso de variables ficticias para modelar variables independientes de escala nominal. En todos los casos se asumirá que hay una colección predeterminada de variables a examinar. (Flores Manrique, 2002)

Consideremos una colección de p variables independientes denotadas por el vector $x' = (x_1,\ x_2, \dots, x_p)$. Por el momento supondremos que al menos una de estas variables están en escala cuantitativa continua. La probabilidad condicional de que la respuesta esté presente se denotará por $P(Y = 1/x) = \pi(x)$. El logit del modelo de regresión logística múltiple está dado por la ecuación:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (68)$$

Y el modelo de regresión logística múltiple es:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}} \quad (69)$$

Si alguna variable independiente categórica es de escala nominal como raza, sexo, grupo de tratamiento, etc., es inapropiado incluirlas en el modelo como si fueran variables de escala

cuantitativa. Los números utilizados para representar los distintos niveles de estas variables de escala nominal son meramente identificadores y no tienen significancia numérica. En esta situación el método es utilizar una colección de variables dummy (o variables ficticias). Supongamos, por ejemplo, que una de las variables independientes es raza, que ha sido codificada como "blanco", "negro" y "otro". En este caso, se necesitarán dos variables ficticias. Una posible estrategia de codificación es que cuando el resultado es "blanca", las dos variables ficticias, D_1 y D_2, serían igual a cero; cuando el resultado es "negro", D_1 sería igual a 1 mientras D_2 todavía sería igual a 0; cuando la raza resultante es "otro", usaríamos $D_1 = 0$ y $D_2 = 1$. (Salcedo Poma, 2002)

Las diferentes estrategias de creación e interpretación de las variables ficticias se discutirán detalladamente más adelante.

El Modelo de Regresión Logística Binaria Múltiple se estima a través de los métodos de máxima verosimilitud, los mismos que se encuentran en los softwares estadísticos que permiten analizar datos mediante este método.

Asumiremos que disponemos de una muestra de n observaciones independientes y p variables independientes definidas por $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i), i = 1, \dots, n$ donde y_i toma valores de 0 o 1, para estimar $\beta' = (\beta_0, \dots, \beta_p)$ que es el vector de parámetros desconocidos.

Para el Modelo de Regresión Lineal Múltiple se usa el método de Mínimos Cuadrados para estimar β, el cual minimiza la suma de cuadrados del error, pero cuando la variable respuesta es binaria aplicar este método no provee las mismas propiedades cuando es usado en variables respuestas continuas.

Por ello se usará el método de Máxima Verosimilitud, ya que obtendremos parámetros estimados que maximizan la probabilidad de obtener un conjunto de datos observados.

La función de verosimilitud expresa la probabilidad de los datos observados como función de parámetros desconocidos. Los estimadores de Máxima Verosimilitud de esos parámetros son aquellos que están en concordancia con los datos observados. (Salcedo Poma, 2002)

La función de probabilidad es casi idéntica a la que figura en la ecuación (50), el único cambio es que $\pi(x)$ ahora se define como en la ecuación (69). Habrá $p+1$ ecuaciones maximoverosímiles que se obtienen por medio de la diferenciación del logaritmo de la función de máxima verosimilitud con respecto a los $p+1$ coeficientes. Las ecuaciones maximoverosímiles que resultan pueden ser expresados así:

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0$$

Y

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0$$

Para $i = 1,2,\dots,n$, $j = 1,2,\dots,p$.

Como en el modelo univariante, la solución de las ecuaciones de verosimilitud requiere un programa especial disponible en la mayoría de los paquetes estadísticos, si no es en todos. Por lo que $\widehat{\boldsymbol{\beta}}$ denota la solución a estas ecuaciones. Así, los valores ajustados para el modelo de regresión logística múltiple son $\hat{\pi}(x_i)$, el valor de la expresión en la ecuación (69) calculado con $\widehat{\boldsymbol{\beta}}$, y x_i.

El método de estimación de las varianzas y covarianzas de los coeficientes estimados se deriva de la teoría bien desarrollada de la estimación de máxima verosimilitud. Esta teoría establece que los estimadores se obtienen de la matriz de segundas derivadas parciales de la función de máxima verosimilitud. Estas derivadas parciales tienen la siguiente forma general:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi_i(1-\pi_i) \qquad (\mathbf{70})$$

Y

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \qquad (\mathbf{71})$$

Para $j, l = 0, 1, 2, \ldots, p$ donde π_i indica $\pi(x_i)$. La matriz de orden $(p+1) \times (p+1)$ que contiene el negativo de los términos dados en las ecuaciones (70) y (71) se denota como $\boldsymbol{I(\beta)}$**.** Esta matriz se denomina matriz de información observada. Las varianzas y covarianzas de los coeficientes estimados son obtenidos a partir de la inversa de la matriz de información como $Var(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})$. Excepto en casos muy especiales no es posible escribir una expresión explícita para los elementos de esta matriz. Por lo tanto, usaremos la notación $Var(\beta_j)$ para denotar el j-ésimo elemento diagonal de esta matriz, que es la varianza de $\hat{\beta}_j$, y $Cov(\beta_j, \beta_i)$ para denotar un elemento arbitrario fuera de la diagonal, que es la covarianza de $\hat{\beta}_j$ y $\hat{\beta}_i$. Los estimadores de varianzas y covarianzas, que se denota por $\hat{Var}(\widehat{\boldsymbol{\beta}})$, se obtienen mediante la evaluación $Var(\boldsymbol{\beta})$ en $\widehat{\boldsymbol{\beta}}$**.** Nosotros usaremos $\hat{Var}(\widehat{\boldsymbol{\beta}_j})$ y $\hat{Cov}(\hat{\beta}_j, \hat{\beta}_l), j, l = 0,1,2, \ldots, p$ para indicar los valores de la matriz.

En la mayoría de los casos, tendremos ocasión de utilizar únicamente la estimación de errores estándar de los coeficientes estimados, que se designan como:

$$\widehat{SE}(\hat{\beta}_j) = \left[\hat{Var}(\hat{\beta}_j)\right]^{1/2} \qquad (\mathbf{72})$$

Para $j = 0,1,2, \ldots, p$. Vamos a utilizar esta notación en la elaboración de métodos para las pruebas de coeficiente y estimación de intervalo de confianza.

Una formulación de la matriz de información que será útil cuando se discute el ajuste del modelo y la evaluación del ajuste es $\hat{\boldsymbol{I}}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{X'VX}$ donde $\boldsymbol{X}$ es una matriz $n \times (p+1)$, matriz que contiene los datos de los sujetos o elementos, y $\boldsymbol{V}_{n \times n}$ es una matriz diagonal que tiene por elementos $\hat{\pi}_i(1 - \hat{\pi}_i)$. Es decir, la matriz $\boldsymbol{X}$ es:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \dots x_{1p} \\ 1 & x_{21} & x_{22} \dots x_{2p} \\ \vdots & \vdots & \vdots \cdots \vdots \\ 1 & x_{n1} & x_{n2} \dots x_{np} \end{bmatrix}_{n\times(p+1)}$$

Y la matriz $\boldsymbol{V}$ es:

$$\boldsymbol{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \dots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \dots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \dots & 0 & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}_{n\times n}$$

(Hosmer & Lemeshow, 2000)

2.2.7.5.1. Prueba de la Significancia del Modelo

Una vez que hemos estimado el modelo de Regresión Logística Binaria Múltiple, comenzamos el proceso de evaluación del modelo. Como en el caso univariante presentado en la sección anterior, el primer paso en este proceso es, por lo general, evaluar la significancia de las variables en el modelo. La prueba de razón de verosimilitud para la significación global de los coeficientes p para las variables independientes en el modelo se realiza exactamente de la misma manera que en el caso univariado. La prueba se basa en la estadística G dada en la ecuación (58). La única diferencia es que los valores ajustados, $\hat{\pi}$, bajo el modelo se basan en el vector que contiene $p + 1$ parámetros $\widehat{\boldsymbol{\beta}}$. Bajo la hipótesis nula de que los p coeficientes de la "pendiente" para las variables independientes son iguales a cero, G será la distribución de Chi-Cuadrado con p grados de libertad.

(Hosmer & Lemeshow, 2000)

Usualmente en la estimación del Modelo de Regresión Logística, como en el Modelo de Regresión Lineal Múltiple se efectúan pruebas con objetivos diferentes, siendo estos:

- Determinar si una variable explicativa tiene coeficientes igual a cero.
- Determinar si un conjunto de variables explicativas tienen coeficientes igual a cero.
- Determinar la calidad del ajuste global del modelo.

Seguidamente se presentan las siguientes pruebas para evaluar la significancia de los coeficientes:

Estadístico de Wald (W)

Como ya vimos este estadístico evalúa la significancia de los coeficientes, y se define como el vector matriz de los coeficientes estimados del modo siguiente:

- Formulación del hipótesis

$$H_0 \quad : \beta_i = 0$$
$$H_1 \quad : \ \beta_i \neq 0, \ Para\ algún\ i = 1,2, \dots, p$$

- Estadístico de Prueba

$$W = \hat{\beta}'\left[V\hat{a}r\left(\hat{\beta}\right)\right]^{-1}\hat{\beta} = \hat{\beta}'\left[\hat{I}\left(\hat{\beta}\right)\right]^{-1}\hat{\beta} = \hat{\beta}'(X'VX)\hat{\beta} \sim \chi^2_{\alpha,p+1} \qquad (\mathbf{73})$$

Donde:

$X_{n\times(p+1)}$ y $V_{n\times n}$: son las matrices.

- Regla de Decisión

Si $W > \chi^2_{\alpha,p} + 1$ rechazamos H_0 con un nivel de significancia fijado α.

- Conclusión

La variable independiente influye en la probabilidad del suceso.

(Flores Manrique, 2002)

Puntuación Eficiente de Rao

La Puntuación eficiente de Rao, cumple un papel similar que el estadístico t de Student para variables que no se incluyen en el modelo. Supongamos que β_i es un parámetro asociado a la variable X_i, bajo el supuesto que se incluirá en la ecuación en el paso siguiente. La puntuación eficiente de Rao permite contrastar la hipótesis nula.

$$H_0 \quad : \beta_i = 0$$
$$H_1 \quad : \beta_i \neq 0, para\ al\ menos\ un\ i = 1,2, \dots, p$$

(Tomaiconza Ataulluco & Pari Sallo, 2014)

Al principio, todas las variables están fuera del modelo. Para que ingresen en el modelo nos fijamos en la que tiene menor p – valor asociado a la puntuación eficiente de Rao y es menor

que 0.05. De estas se tomará la variable que tenga mayor valor en la puntuación eficiente de Rao. (Martín Martín, Cabero Morán, & De Paz Santana, 2008)

2.2.7.5.2. Bondad de Ajuste del Modelo

En regresión logística existen varias medidas de ajuste global para comparar la diferencia entre valores esperados y valores observados. Dos de las más populares, dada su disponibilidad en los distintos softwares, son el test basado en la devianza D y el estadístico χ^2 de Pearson.

Estadístico de Devianza (D)

El Estadístico de Devianza desempeña, en la Regresión Logística, el mismo papel que la suma de cuadrados residuales en la regresión lineal. Este estadístico realiza una comparación de las variables independientes del modelo de regresión logística mediante la prueba de verosimilitud con la significancia de los $p + 1$ parámetros, bajo la hipótesis para determinar si las variables independientes influyen significativamente en el modelo.

En Regresión Logística la comparación de los valores observados Y y los valores esperados $\hat{Y}$ obtenidos del modelo "con" y "sin" la variable en cuestión. Para ello se emplea la función de verosimilitud que está dada por:

$$D = -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right]$$

- Formulación de la hipótesis

  $H_0 \quad : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

  $H_1 \quad : \beta_i \neq 0, \quad Para\ algún\ i = 1,2,\ldots,p$

- Estadístico de Prueba

  $D \sim \chi^2_{\alpha,p+1}$

- Regla de Decisión

  Si $D > \chi^2_{\alpha,p+1}$, rechazamos H_0

- Conclusión

Al menos uno de los coeficientes es diferente de cero y la variable correspondiente influye en la probabilidad del suceso en estudio.

Una vez encontrado el mejor conjunto de variables explicativas que predicen a la variable Y, seguidamente se debe evaluar mediante el estadístico de Wald cada coeficiente para determinar cuál o cuáles ingresan al modelo. (Ballón Beltran & Bernabé Ponte, 2015)

Prueba Chi-Cuadrado de Pearson

Para medir la bondad del ajuste también se utilizan las medidas del error que cuantifican la diferencia entre el valor observado y el estimado. Para esta prueba la hipótesis planteada es la siguiente:

- Formulación del hipótesis

$$H_0 \quad : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$
$$H_1 \quad : \ \beta_i \neq 0, \ \ Para\ algún\ i = 1,2, \dots, p$$

Que es lo mismo que:

H_0 : No existe diferencia entre el modelo saturado y el ajustado.

H_1 : Existe diferencia entre el modelo saturado y el ajustado.

Se construye un estadístico que recoge los residuos estandarizados o de Pearson del modelo Logit, que se definen como la diferencia entre el valor observado de la variable respuesta y el estimado, dividido por la estimación de la desviación típica, ya que la esperanza es nula. A través del contraste de multiplicadores de Lagrange, se puede calcular el estadístico conocido con el nombre de χ^2 de Pearson, que se define como:

$$\chi_c^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{p}_i)^2}{\hat{p}_i(1 - \hat{p}_i)} \qquad (\mathbf{74})$$

Que se distribuye con una $\chi^2_{(\alpha, n-k-1)}$;

La conclusión se da para un nivel de significancia α, se rechaza la hipótesis nula, si: $\chi_c^2 < \chi^2_{(\alpha, n-p-1)}$ o que el p-valor es menor que el valor de α.

Este estadístico es similar a la suma de cuadrados de los residuos del modelo de regresión convencional. El ajuste del modelo será mejor cuanto más cerca esté el valor del estadístico de cero. Para saber a partir de que valor puede considerarse el ajuste como aceptable es necesario conocer la distribución del estadístico. Este estadístico, bajo la hipótesis nula, se distribuye como una Chi-Cuadrado con $(n-k-1)$ grados de libertad, por lo que su valor se compara con el valor teórico de las tablas de la Chi-Cuadrado para contrastar la hipótesis nula. Si el valor calculado es superior al valor teórico se rechaza la hipótesis nula lo que equivale a decir que el error cometido es significativamente distinto de cero, es decir, se trataría de un mal ajuste. (Medina Moral, 2003)

Estadístico de Hosmer-Lemeshow (C_g)

El estadístico C_g se basa en la agrupación de las probabilidades estimadas bajo el modelo de regresión $\hat{p}_1, \hat{p}_2, \dots, \hat{p}_N$. La idea básica es que el primer grupo estará formado aproximadamente por las $\frac{N}{G}$ observaciones cuyas probabilidades esperadas sean más pequeñas, el segundo estará formado por los siguientes $\frac{N}{G}$ más pequeños y así sucesivamente. Los puntos de corte así generados se denominan deciles de riesgo. La siguiente tabla muestra las frecuencias esperadas y observadas, en cada uno de los grupos, utilizados en el cálculo del estadístico C_g, denotando como $d_i, i = 1,2, \dots,10$, los deciles de riesgo de las probabilidades estimadas.

Tabla 7:
Frecuencias Esperadas y Observadas para C_g

	Respuesta			
	$Y = 1$		$Y = 0$	
Grupos	**Observado**	**Esperado**	**Observado**	**Esperado**
$\hat{p}_j < d_1$	O_{11}	e_{11}	O_{01}	e_{01}
$d_1 \leq \hat{p}_j < d_2$	O_{12}	e_{12}	O_{02}	e_{02}
…	…	…	…	…
$d_9 \leq \hat{p}_j < d_{10}$	O_{1G}	e_{1G}	O_{0G}	e_{0G}
Total	O_1	e_1	O_0	e_0

El número de individuos observados para los que ocurrió el suceso y para los que no ocurrió, en cada uso de los grupos es (frecuencias observadas):

$$O_{1G} = \sum_{k=1}^{n_g} y_k$$

$$O_{0g} = \sum_{k=1}^{n_g} (1 - y_k)$$

Siendo n_g el número de observaciones en el grupo g.

Analógicamente, el número esperado de individuos para los que ocurrirá el suceso y para los que no, se denotan por (frecuencias esperadas):

$$e_{1g} = \sum_{k=1}^{n_g} \hat{p}(x_k)$$

$$e_{0g} = \sum_{k=1}^{n_g} \left(1 - \hat{p}(x_k)\right)$$

El Estadístico C_g se obtiene entonces comparando estos valores observados y esperados de la siguiente forma:

$$C_g = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{\left(O_{kg} - e_{kg}\right)^2}{e_{kg}} \qquad (\mathbf{75})$$

A través de estudios de simulación se demostró que cuando $J = N; R + 1 < G$ (el número de covariables más 1 es menor que el número de grupos), bajo la hipótesis del modelo logístico, C_g tiene distribución asintótica χ^2_{G-2} (Hosmer and Lemeshow [1989]).

El inconveniente en el uso de este estadístico radica su dependencia en la elección de los puntos de corte, dando lugar a estimaciones diferentes para los mismos datos por distintos programas, pudiendo llegar incluso a darse la situación extrema de aceptar la hipótesis nula de un ajuste adecuado por parte de algún programa y de rechazo por otro. Es por esta razón,

por la que el estadístico C_g se considera algo inestable, aunque la mayor parte de los softwares siguen apostando por la implementación de este test.

Estadístico de Hosmer-Lemeshow (H_g)

El siguiente test propuesto por Hosmer y Lemeshow se basa en la formación de los grupos de acuerdo a unos puntos de corte fijos y preestablecidos.

El número de grupos a utilizar puede ser arbitrario, aunque los autores recomendaron el uso de 10 (Hosmer and Lemeshow [1980]). La tabla muestra las frecuencias esperadas y observadas para cada uno de estos grupos.

Tabla 8:
Frecuencias Esperadas y Obervadas para H_g

	Respuesta			
	$Y = 1$		$Y = 0$	
Grupos	**Observado**	**Esperado**	**Observado**	**Esperado**
$0 \leq \hat{p}_j < 0.1$	O_{11}	e_{11}	O_{01}	e_{01}
$0.1 \leq \hat{p}_j < 0.2$	O_{12}	e_{12}	O_{02}	e_{02}
...	...	...	...	...
$0.9 \leq \hat{p}_j < 1.0$	O_{110}	e_{110}	O_{010}	e_{010}
Total	O_1	e_1	O_0	e_0

La formulación del estadístico y su distribución asintótica es la misma que para el anterior C_g:

$$H_g = \sum_{k=0}^{1} \sum_{g=1}^{G} \frac{\left(O'_{kg} - e'_{kg}\right)^2}{e'_{kg}} \qquad (76)$$

En el trabajo realizado por Hosmer y Lemeshow pusieron de maniesto que entre todos los test basados en las probabilidades estimadas que habían desarrollado, los más razonables eran estos dos, C_g y H_g. Aunque no queda claro cuál será más adecuado entre ellos: en un primer momento, tras varias simulaciones, vieron como H_g parecía más potente que C_g, pero más tarde señalaron como más adecuado a C_g pensando que se ajustaba mejor a la distribución Chi-Cuadrado (Hosmer and Lemeshow [1989]). En su último trabajo sobre este tema

(Hosmer et al. [1997]) los autores recomendaron usar estos estadísticos para confirmar la falta de ajuste señalada tras la utilización de otros métodos.

Nótese finalmente que estos estadísticos no son más que estadísticos Chi-Cuadrado de bondad de ajuste por lo que para que su distribución asintótica sea Chi-Cuadrado, dicha aproximación será válida siempre que al menos el 80% de las frecuencias estimadas bajo el modelo sean mayores que 5 y todas mayores que 1. (Iglesias Cabo, 2013)

2.2.7.5.3. Tasas de Clasificación Correcta

La Tabla 9, permite evaluar la eficacia del modelo para clasificar nuevos individuos, ya sea en el primer o segundo grupo, se elige un punto de corte de 0,5, si el valor de la probabilidad es $\geq 0{,}5$ se acepta a "1" como respuesta y para la probabilidad $< 0{,}5$, se considera que la variable de respuesta toma el valor de "0".

Los valores de la tabla de clasificación se definen como:

Tabla 9:
Tabla de Clasificación

		Predictivo		
		Si	**No**	**Total**
Observado	**Si**	n_{11}	n_{12}	$n_{11}+n_{12}$
	No	n_{21}	n_{22}	$n_{21}+n_{22}$
Total		$n_{11}+n_{21}$	$n_{12}+n_{22}$	n

Donde los valores de n_{11} y n_{22} son los correctamente clasificados, que es lo mismo a decir $\frac{n_{11}+n_{22}}{n} \times 100\%$, es el porcentaje de observaciones bien clasificadas por el modelo de Regresión Logística estimado, y se espera que este porcentaje sea el más alto posible a fin de concluir que el modelo obtenido clasifica bien las observaciones en estudio y; n_{12} y n_{21}, son los incorrectamente clasificados; es decir $\frac{n_{12}+n_{21}}{n} \times 100\%$, es el porcentaje de observaciones mal clasificadas, mediante el modelo de Regresión Logística estimado.

2.2.7.5.4. Sensibilidad y Especificidad

Teniendo en cuenta esta tabla se definen la sensibilidad y la especificidad que, son pruebas diagnósticas.

La sensibilidad y la especificidad son las medidas tradicionales y básicas del valor diagnóstico de una prueba. Miden la discriminación diagnóstica de una prueba en relación a un criterio de referencia, que se considera la verdad.

Estos indicadores en principio permiten comparar directamente la eficacia de una prueba con el de otras y esperar resultados similares cuando son aplicadas en diferentes países, regiones o ámbitos.

a. La Sensibilidad (S)

La sensibilidad del modelo se refiere a la capacidad que tiene éste para detectar como positivos los casos que poseen la característica. En términos coloquiales, si al modelo le presentamos solo casos positivos, la sensibilidad determina la capacidad que tiene el modelo de no equivocarse. La sensibilidad queda definida como:

$$Sensibilidad = \frac{VerdaderosP}{VerdaderosP + FalsosN} \quad (\mathbf{77})$$

Y a través de las frecuencias de las casillas puede indicarse de forma operativa como:

$$\frac{n_{11}}{n_{11} + n_{21}} \times 100\%$$

b. La Especificidad (E)

La especificidad del modelo se refiere a la capacidad que tiene éste para discriminar correctamente los casos que no poseen la característica. Es decir, sobre un conjunto de casos que no poseen la característica, determina en qué grado no va a confundirlos con casos que poseen la característica. La especificidad la definimos como:

$$Especificidad = \frac{VerdaderosN}{VerdaderosN + FalsosP} \quad (\mathbf{78})$$

Que podemos indicar, también, como:

$$\frac{n_{22}}{n_{12}+n_{22}} \times 100\%$$

Un modelo con buena capacidad predictiva debería tener valores altos tanto de sensibilidad como de especificidad.

c. Estimación de Sensibilidad y Especificidad

Para ilustrar el significado de estos conceptos a través de sus estimaciones, supóngase que se tienen N sujetos de los que se conoce su estatus verdadero (enfermo o no) y se les ha practicado el test o prueba que se está evaluando y cuyo resultado puede ser inequívocamente positivo o negativo.

Estas características pueden entonces estimarse fácilmente a partir de una tabla de 2×2 como se muestra en la siguiente tabla.

Tabla 10:
Resultados de la Prueba y la Existencia de la Enfermedad

		Criterio de Verdad		Total
		Enfermos	No enfermos	
Prueba diagnóstico	Positivos	a	b	$a+b$
	Negativos	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d$

Donde:

- a : Pacientes con la enfermedad diagnosticados como “positivos” por la prueba.
- b : Pacientes sin la enfermedad diagnosticados como “positivos” por la prueba.
- c : Pacientes con la enfermedad diagnosticados como “negativos” por la prueba.
- d : Pacientes sin la enfermedad diagnosticados como “negativos” por la prueba.

Puede apreciarse que cada celda de la Tabla 10 refleja una característica que también suele calificarse de la manera siguiente:

$$Sensibilidad = \frac{Verdaderos\ Positivos}{Total\ de\ Enfermos} = \frac{a}{a+c}$$

$$Especificidad = \frac{Verdaderos\ Negativos}{Total\ de\ No\ Enfermos} = \frac{d}{b+d} \quad (\mathbf{79})$$

(Camarero Rioja, Almazán Llorente, & Mañaz Ramirez, 2013)

2.2.7.6. Calidad de Ajuste

2.2.7.6.1. Coeficiente Pseudo-R^2 de Mc-Fadden

Si identificamos al estadístico $(-2lnL)$, conocido también como estadístico de variación; identificaremos por $(-2lnL)_0$ al valor inicial de $(-2lnL)$, que es el mínimo valor bajo el modelo nulo, dado solo por el término de la constante y denotaremos por $(-2lnL)_k$ al mínimo valor de $(-2lnL)$ bajo el modelo ajustado con todos los parámetros, se obtiene la siguiente expresión del Coeficiente Pseudo-R^2 de Mc-Fadden:

$$R^2_{MF} = 1 - \frac{(-2lnL)_k}{(-2lnL)_0} \quad (\mathbf{80})$$

Siendo su rango teórico de valores $0 \leq R^2_{MF} \leq 1$, aunque muy raramente su valor se aproxima a uno. Suele considerarse una buena calidad de ajuste cuando está entre los valores $0{,}2 \leq R^2_{MF} \leq 0{,}4$ y excelente cuando es superior.

2.2.7.6.2. Coeficiente Pseudo-R^2 de Cox y Snell

En este caso usamos directamente la función de verosimilitud, y no el estadístico $(-2lnL)$. Por lo que denotamos por $L_0 = exp\left(\frac{-(-2lnL)_0}{2}\right)$, el máximo de verosimilitud bajo el modelo nulo dado solo para el término de la constante y denotamos por $L_k = exp\left(\frac{-(-2lnL)_k}{2}\right)$, el máximo de verosimilitud bajo el modelo ajustado con todos los parámetros. Entonces definimos el coeficiente Pseudo-R^2 de Cox y Snell de la siguiente manera:

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_k}\right)^{\left(\frac{2}{N}\right)} = 1 - exp\left(\frac{(-2lnL)_k - (-2lnL)_0}{N}\right) \quad (\mathbf{81})$$

Su rango teórico de valores $0 \leq R_{CS}^2 \leq 1$, lo que hace poco interpretable al depender de L_0, pero también se considera una buena calidad de ajuste cuando está entre los valores $0{,}2 \leq R_{CS}^2 \leq 0{,}4$, y excelente cuando es superior.

2.2.7.6.3. Coeficiente Pseudo-R^2-de Nagelkerke

Nagelkerke define R^2 así:

$$R_N^2 = \left(\frac{R_{CS}^2}{1 - (L_0)^{\left(\frac{2}{N}\right)}} \right) = \frac{1 - exp\left(\frac{(-2lnL)_k - (-2lnL)_0}{N} \right)}{1 - exp\left(\frac{-(-2lnL)_0}{2} \right)} \qquad (\mathbf{82})$$

En este caso su rango de valores es $0 < R_N^2 < 1$, por lo que se puede interpretar del mismo modo que el coeficiente de determinación de la regresión lineal clásica, aunque es más difícil que alcance valores cercanos a 1.

2.2.7.7. Estimación de Intervalos de Confianza

El método utilizado para estimar los intervalos de confianza para un modelo de variable múltiple es esencialmente el mismo que para el modelo logístico simple.

Los límites al $100(1-\alpha)\%$ para un intervalo de confianza para los coeficientes obtenidos de $\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_1)$ para la pendiente y coeficientes de $\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\hat{\beta}_0)$ para el término constante.

La idea básica es la misma solo que ahora hay más términos que intervienen en la sumatoria. Se deduce de $L(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ que es una expresión general para el estimador del Logit para un modelo que contiene p variables independientes es:

$$\hat{L}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \qquad (\mathbf{83})$$

Una forma alternativa de expresar el estimador de logit en la ecuación (82) mediante el uso de la notación de vectores como $L(x) = X'\hat{\beta}$, donde, el vector $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)$ denota el estimador de los coeficientes $p+1$ y el vector $X' = \left(x_0, x_1, x_2, \ldots, x_p\right)$ representa la constante y un conjunto de valores de las p covariables en el modelo, donde $x_0 = 1$.

De ello deducimos a partir de $V\hat{a}r\left[\hat{L}(x)\right] = V\hat{a}r\left(\hat{\beta}_0\right) + x^2 V\hat{a}r\left(\hat{\beta}_1\right) + 2xC\hat{o}v\left(\hat{\beta}_0, \hat{\beta}_1\right)$.

una expresión para el estimador de la varianza del estimador de logística en $\hat{L}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$ es:

$$V\hat{a}r\left[\hat{L}(x)\right] = \sum_{j=1}^{p} x_j^2 \, V\hat{a}r\left(\hat{\beta}_j\right) + \sum_{j=0}^{p} \sum_{l=j+1}^{p} 2x_j x_i C\hat{o}v\left(\hat{\beta}_j, \hat{\beta}_i\right) \qquad (\mathbf{84})$$

Podemos expresar este resultado mucho más conciso mediante el uso de la expresión de la matriz para el estimador de la varianza del estimador de los coeficientes. A partir de la expresión de la matriz de información observada, tenemos que:

$$V\hat{a}r\left(\hat{\beta}\right) = (X'VX)^{-1}$$

Deduciendo esta fórmula se llega a la siguiente expresión equivalente para el estimador en la ecuación (83) es:

$$V\hat{a}r[\hat{g}(X)] = X'V\hat{a}r\left(\hat{\beta}\right)X = X'^{\left(X'VX\right)^{-1}}X$$

Afortunadamente, todos los buenos paquetes de software de Regresión Logística proporcionan la opción para el usuario para crear una nueva variable que contiene los valores estimados de la ecuación anterior o el error estándar para todos los sujetos en el conjunto de datos. Esta característica elimina la carga computacional asociada con los cálculos de la matriz en la ecuación anterior y permite al usuario calcular rutinariamente valores ajustados y las estimaciones del intervalo de confianza. (Hosmer & Lemeshow, 2000)

2.2.7.8. Interpretación de los parámetros

Para la interpretación de los coeficientes del modelo basta con tener en cuenta, si el modelo ajustado es bueno, entonces se dice que el modelo es significativo, pero además se debe analizar el grado de asociación estadística existente en sus parámetros, a partir de la ecuación (38) se tiene:

$$ln\left(\frac{p\left(X_1, \ldots, X_p; \beta\right)}{1 - p\left(X_1, \ldots, X_p; \beta\right)}\right) = ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Donde el "Odds Ratio" que es el factor de riesgo está dado por la razón de esta expresión:

$$\frac{p(X_1, \dots, X_p; \beta)}{1 - p(X_1, \dots, X_p; \beta)} = e^{\beta_0} \times e^{\beta_1 x_1} \times \dots \times e^{\beta_p x_p}$$

Entonces para:

$$\frac{\frac{p(X_1 + 1, \dots, X_p; \beta)}{1 - p(X_1 + 1, \dots, X_p; \beta)}}{\frac{p(X_1, \dots, X_p; \beta)}{1 - p(X_1, \dots, X_p; \beta)}} = e^{\beta_1}$$

Por tanto, e^{β_1} es el factor de cambio en el "Odds Ratio" (OR) de riesgo si el valor de la variable X_1 cambia en una unidad.

Así, si $\beta_1 > 0$ o $(\beta_1 < 0)$ el factor será mayor que 1 y $p(X_1, \dots, X_p; \beta)$ aumentará (o disminuirá).

Si $\beta_1 = 0$ la variable X_1 no ejerce ningún efecto en p_i, β_0 es un ajuste de escala. Su mejor interpretación se obtiene calculando el valor de $p(X_1, \dots, X_p; \beta)$ en los valores medios de $X_1, \dots, X_p$ y usar como variables explicativas sus valores estandarizados.

En Regresión Logística la medida de asociación más empleada, es el OR debido a que el número e es la base de los logaritmos naturales y elevados a un coeficiente de regresión logística del factor, si es mayor que 1 supone un aumento unitario, indica que el factor de riesgo es mayor.

Si el modelo de Regresión Logística es significativo y una de las variables independientes es dicotómica con valores de 0 y 1, el número e elevado al coeficiente de Regresión Logística es el OR, denominado factor de riesgo o protección que implica un aumento unitario de la variable independiente. En el caso de una variable cuantitativa, e elevado a β_1 es el número de veces que aumenta la probabilidad de padecer una enfermedad por cada unidad de aumento de la variable independiente, o dicha de otra manera, cuantas veces es

más probable que padezca la enfermedad una persona que presenta síntomas relacionadas a ella.

2.2.7.9. Construcción del Modelo de Regresión

Para construir un modelo de regresión, nos centraremos en el tipo de variables que deseamos introducir (Categóricas y continuas) y posteriormente, veremos los métodos que los paquetes estadísticos nos ofrecen actualmente para obtener el modelo de regresión más fiable.

2.2.7.9.1. Selección de las Variables del Modelo

Una vez conocido el procedimiento de ajuste de modelos de regresión logística binaria o binomial, el siguiente paso es seguir estrategias para la selección de las variables que mejor explica a la variable respuesta. Para ello se adoptará el principio de parsimonía, que consiste en seleccionar el modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla en términos de cocientes de ventajas.

Se debe tener cuidado con las variables independientes cualitativas que se transforman en varias variables dummies, siempre que se incluya o excluya una de estas variables, todas las demás categorías deben ser incluidas o excluidas en bloque, ya que de lo contrario implicaría que se habría recodificado la variable, y por lo tanto la interpretación de la misma no sería la correcta. Además, se debe tener en cuenta la significancia que pudiera tener cada variable dummy; no siempre todas las categorías de una variable independiente son significativas, o todas no significativas; por lo que, cuando ocurra esta situación es recomendable contrastar el modelo completo frente al modelo sin la covariable mediante la prueba de razón de verosimilitud, decidiendo incluir o excluir la variable independiente dependiendo del resultado de la prueba y del interés de la variable independiente. Si se obtiene significación en este contraste, la variable permanecería en el modelo, si no se obtiene significación y la covariable es de interés, su inclusión en el modelo es a criterio del investigador. (Moral Pelaez, 2006)

El modelo de regresión se puede construir utilizando los siguientes métodos:

a. Método Hacia Adelante "Forward"

Consiste en ir introduciendo las variables en el modelo únicamente si cumplen una serie de condiciones hasta que no se pueda introducir ninguna más, hasta que ninguna cumpla la condición impuesta, tiene los siguientes pasos:

1) Se inicia con un modelo vacío de variables independientes (solo β_0).
2) Se ajusta un modelo y se calcula el p-valor del contraste de razón de verosimilitud que resulta de incluir cada variable por separado.
3) Se selecciona el modelo con el p-valor más significativo.
4) Se ajusta de nuevo un modelo con la(s) variable(s) seleccionada(s) y se calcula el p-valor de añadir cada variable no seleccionada anteriormente por separado.
5) Se selecciona el modelo con el p-valor más significativo.
6) Se repite los pasos 4 y 5 hasta que no queden variables significativas para incluir.

b. Método Hacia Atrás "Backward"

Se introducen en el modelo todas las variables y se van suprimiendo si cumplen una serie de condiciones definidas a priori hasta que no se puedan eliminar más, es decir ninguna variable cumple la condición impuesta, tiene los siguientes pasos.

1) Se inicia con un modelo general que incluyen todas las variables independientes candidatas.
2) Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.
3) Se selecciona para eliminar a la variable menos significativa.
4) Se repiten los pasos 2 y 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda el ajuste.

c. Método Paso a Paso "Stepwise"

Este método combina los dos pasos anteriores, adelante y atrás introduciendo o eliminando variables del modelo si cumplen una serie de condiciones definidas a priori hasta que ninguna variable satisfaga ninguna de las condiciones expuestas de entrada o salida del modelo. Pero no todos los métodos llegan a la misma solución necesariamente. Este método está basado en contrastes condicionales de razón de verosimilitud:

1) Si partimos del modelo vacío, solo con la constante, este método consiste en partir del modelo inicial, y en cada paso se ajustarán todos aquellos modelos que resultan de incluir cada una de las variables explicativas que no están en el modelo seleccionado en el paso anterior.
2) Entonces se llevan a cabo contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el modelo seleccionado en el paso anterior y en la hipótesis alterna el modelo resultante de la inclusión de cada variable. De este modo se seleccionarán las variables para las que el contraste sea significativo, y se incluiría en el modelo aquella variable asociada al mínimo p-valor de entre todos los menores o iguales que α_1.
3) La inclusión de variables mediante este método continúa hasta que ninguno de estos contrastes condicionales sea significativo.
4) Por otra parte, a la misma vez, se considera en cada paso la posibilidad de eliminar alguno de los parámetros del modelo seleccionado en el paso anterior (método hacia atrás). Pero no se puede eliminar en un paso la variable que acaba de entrar en el paso anterior, por lo que se fijará para la eliminación de variables un nivel de significancia α_2 mayor que α_1.
5) Al igual que antes, para la eliminación de variables se realizarán contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el modelo que resulta de la

eliminación de cada variable y en la hipótesis alterna el modelo seleccionado en el paso anterior. Así, las variables candidatas a eliminar serán aquellas cuyo p-valor sea mayor de α_2 y se eliminará la variable con mayor p-valor de estos. La eliminación de variables continúa hasta que todos estos contrastes condicionales resulten significativos.

6) Así finalmente, se llegará a un paso en el que ninguno de los contrastes condicionales de introducción de variables sean significativos y todos los de eliminación de variables sean significativos.

d. Método de Introducir todas las variables Obligatoriamente "Enter"

Este último método de selección de variables para construir el modelo de regresión, produce que el proceso de selección de variables sea manual, partiendo de un modelo inicial, en el que se obliga a que entren todas las variables seleccionadas, se va evaluando que variable es la que menos participa en él y se elimina, volviendo a construir un nuevo modelo de regresión aplicando el mismo método, pero excluyendo la variable seleccionada y aplicando el mismo proceso de selección. Este proceso se repite reiteradamente hasta que se considere que el modelo obtenido es el que mejor se ajusta a las condiciones impuestas y que no se puede eliminar ninguna variable más de las que los componen. *(Moral Pelaez, 2006)*

2.2.7.10. Medidas de Asociación

A continuación presentamos algunas medidas de asociación que permiten medir si dos variables están asociadas o no, en este caso la variable de estudio o dependiente con cada una de las variables independientes.

2.2.7.10.1. Determinación de Factores de Riesgo

Un factor de riesgo es cualquier característica o circunstancia detectable de una persona o grupo de personas que se asocian con un aumento en la probabilidad de padecer, desarrollar o estar especialmente expuestos a un proceso mórbido. Estos factores de riesgo, sean biológicos, ambientales, de comportamiento, socio-culturales o económicos, pueden

(sumándose unos a otros), aumentar el efecto aislado de cada uno de ellos produciendo un fenómeno de interacción. Entonces podemos decir que el término riesgo, implica la presencia de una característica o factor (o de varios) que incrementa la probabilidad de consecuencias adversas. En este sentido el riesgo constituye una medida de probabilidad estadística de que en un futuro se produzca un acontecimiento por lo general no deseado.

2.2.7.10.2. Cuantificación del Riesgo

La cuantificación del riesgo constituye un elemento fundamental en la formulación de prioridades que no deben dejarse a la casualidad. Existen diferentes maneras de cuantificar ese riesgo.

2.2.7.10.2.1. Riesgo Absoluto

Mide la incidencia del daño en la población total.

2.2.7.10.2.2. Riesgo Relativo

Compara la frecuencia con que ocurre el daño entre los que tienen el factor de riesgo y los que no lo tienen, mide la fuerza de la asociación entre la exposición y la enfermedad, para esto veremos la siguiente tabla:

Tabla 11:
Tabla 2×2 para el Cálculo de Medidas de Asociación en Estudios de Seguimiento

	Enfermos	*No Enfermos*	*Total*
Expuestos	a	b	$a+b$
No Expuestos	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

El riesgo relativo indica la probabilidad de que se desarrolle la enfermedad en los expuestos a un factor de riesgo en relación al grupo de los no expuestos. Su cálculo se estima dividiendo la incidencia de la enfermedad en los expuestos (I_e) entre la incidencia de la enfermedad en los no expuestos $(I_\emptyset)$.

$$Riesgo\ Relativo = \frac{Incidencia\ en\ Expuestos}{Incidencia\ en\ No\ Expuestos} = \frac{I_e}{I_\emptyset} = \frac{a/(a+b)}{c/(c+d)}$$

En los estudios de casos y controles, dado que la incidencia es desconocida, el método de estimación del riesgo relativo es diferente y se estima calculando el Odds Ratio. (Camarero Rioja, Almazán Llorente, & Mañaz Ramirez, 2013)

2.2.7.10.2.3. Razón, Odd y Odd Ratio

a. Razón:

Una razón o ratio es el cociente entre dos cantidades y señala cuantas veces una cantidad es mayor o menor respecto a la otra.

De forma habitual esta razón o ratio suele denominarse con el término Odd.

$$Odd = \frac{p}{q} = \frac{p}{(1-p)}$$

El término Odd en inglés se refiere a la razón que se establece entre la ocurrencia (o su probabilidad) de un suceso respecto a su no ocurrencia. Se interpreta como ventaja comparativa. Es muy usual, por ejemplo, en el mundo de las apuestas.

Podemos interpretar el Odd en términos de probabilidad.

b. Odd Ratio:

El Odd Ratio es una razón de Odds, abreviadamente OR y puede interpretarse como ventaja comparativa o, como razón de probabilidades.

Cuando el Odd Ratio alcanza el valor 1 quiere decir que no hay diferencias.

$$Odd\ Ratio = OR = \frac{Odd_A}{Odd_B} = \frac{\frac{p_A}{(1-p_A)}}{\frac{p_B}{(1-p_B)}}$$

Para el estudio de casos y controles, Odd Ratio, es el cociente de Odds de exposición observada en casos y la Odd de exposición en el grupo control.

Tabla 12:
Tabla 2×2 en Estudios de Casos y Controles

	Casos	**Controles**
Expuestos	a	b
No Expuestos	c	d

$$OR = \frac{Odds\ de\ Exposición\ en\ Casos}{Odds\ de\ Exposición\ en\ Controles}$$

Donde:

$$Odds\ de\ Exposición\ en\ Casos = \frac{Casos\ \ Expuestos}{Casos\ No\ Expuestos} = \frac{a}{c}$$

$$Odds\ de\ Exposición\ en\ Controles\ = \frac{Controles\ Expuestos}{Controles\ No\ Expuestos} = \frac{b}{d}$$

Por lo tanto el Odds Ratio es:

$$OR = \frac{a/c}{b/d} = \frac{a \times d}{b \times c} \quad (\mathbf{84})$$

(Camarero Rioja, Almazán Llorente, & Mañaz Ramirez, 2013)

b.1. Intervalo de Confianza del Odd Ratio

Como el Odd Ratio es la estimación de la asociación de un determinado factor con una incidencia, por lo que resulta necesario calcular la medida de variabilidad de esta estimación. El intervalo de confianza es el rango en el que se encuentra el verdadero valor del Odd Ratio; esto permite tener una buena estimación cuando el Odd Ratio se aproxima a 1, pero se hace menos estable para Odds Ratios mayores. La fórmula para el intervalo de Odd Ratio es:

$$IC = OR\left(1 \mp \frac{z}{Xhm}\right)$$

Donde:

OR	:	Odds Ratio calculado
z	:	Valor en la tabla normal para un nivel de confianza del 95%.
Xhm	:	Chi Cuadrado de hm, cuya fórmula es: $Xhm = \sqrt{\frac{(n-1)(a\times d-b\times c)^2}{(a+b)(c+d)(a+c)(b+d)}}$

b.2. Interpretación

- Si el resultado de $OR > 1$, la asociación es positiva; es decir, la presencia del factor se asocia a la mayor ocurrencia del evento, y se le considera Factor de Riesgo.

- Si el resultado de $OR < 1$, la asociación es negativa; es decir, la presencia del factor no se asocia con la mayor ocurrencia del evento, y contrario al caso anterior se le considera Factor de Protección.
- Si el resultado de $OR = 1$, no existe asociación entre las variables; es decir que la cantidad de veces que el evento ocurra será igual con o sin la presencia del factor, la relación es 1 a 1.

Los resultados del Intervalo de Confianza, permiten establecer si una asociación es estadísticamente significativa; si este resultado incluye el 1, se dice que la asociación no es estadísticamente significativa y; si no incluye el 1, la asociación es estadísticamente significativa, en resumen se considera la siguiente tabla:

Tabla 13:
Tabla Resumen de Interpretación de OR

Valor Odds Ratio	Intervalo de Confianza		Tipo de Asociación
	Inferior	Superior	
Igual a 1			No existe asociación
Mayor de 1	> 1	> 1	Significativa, Factor de Riesgo
Menor de 1	< 1	< 1	Significativa, Factor de Protección
Mayor de 1	< 1	> 1	No significativa
Menor de 1	< 1	> 1	No significativa

2.2.7.10.2.4. Coeficiente Q de Yule

Este coeficiente es muy útil para el caso de medir la asociación entre dos variables categóricas y se define como:

Tabla 14*:*
Tabla 2×2

		Y		Total
		y_1	y_2	
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d$

$$Q = \frac{(a \times d) - (b \times c)}{(a \times d) + (b \times c)} \quad (85)$$

Cuya propiedad es que, está acotada entre $-1 \leq Q \leq 1$.

Su **interpretación** se da de la siguiente manera:

Tabla 15*:*
Resumen de Interpretación de Coeficiente Q de Yule

Valor Q		**Conclusión**
Si $Q < 0$	:	Se verifica una asociación negativa.
Si $Q = 0$	:	Se dice que las variables son independientes.
Si $Q > 0$	:	Se verifica una asociación positiva.
Si $Q = 1$	:	Se verifica que $b \times c = 0$ y existe asociación completa entre las variables.

2.2.7.10.2.5. Prueba Chi-Cuadrado (χ^2) de Pearson

En este caso el objetivo de esta prueba es contrastar la hipótesis de independencia entre dos factores o variables que se basa en la información proporcionada por las frecuencias observadas contenidas en la tabla de contingencia:

Tabla 16:
Tabla de Contingencia $N \times M$

		Y						**Total**
		y_1	y_2	…	y_j	…	y_m	
X	x_1	n_{11} (e_{11})	n_{12} (e_{12})	…	n_{1j} (e_{1j})	…	n_{1m} (e_{1m})	$N_{1\circ}$
	x_2	n_{21} (e_{21})	n_{22} (e_{22})	…	n_{2j} (e_{2j})	…	n_{2m} (e_{2m})	$N_{2\circ}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	x_i	n_{i1} (e_{i1})	n_{i2} (e_{i2})	…	n_{ij} (e_{ij})	…	n_{im} (e_{im})	$N_{i\circ}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	x_p	n_{p1} (e_{p1})	n_{p2} (e_{p2})	…	n_{pj} (e_{pj})	…	n_{pm} (e_{pm})	$N_{p\circ}$
Total		$N_{\circ 1}$	$N_{\circ 2}$	…	$N_{\circ j}$	…	$N_{\circ m}$	N

Donde:

$$N_{i\circ} = \sum_{j=1}^{m} n_{ij} \qquad N_{\circ j} = \sum_{j=1}^{p} n_{ij} \qquad N = \sum_{i} N_{i\circ} = \sum_{j} N_{\circ j}$$

Se analizan dos variables (que admiten distintas modalidades) mediante una tabla de contingencia, donde una ocupa las filas y otra las columnas, la intersección entre fila y columna da lugar a una celda o casilla, cuya frecuencia observada es n_{ij}. Luego se contrasta la hipótesis nula que presupone la independencia de ambas variables, mediante el estadístico χ^2 de Pearson. Los pasos a seguir son:

- Formulación de Hipótesis

  H_0 : Ambas variables son Independientes

  H_1 : Existe una relación de Dependencia

- Cálculo del Estadístico Observado

  Definimos el estadístico observado o calculado.

$$\chi_c^2 = \sum_{i=1}^{p}\sum_{j=1}^{m}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \sim \chi^2_{(p-1).(m-1)}$$

- Estadístico Teórico

  Que sigue una distribución χ^2 con $(p-1).(m-1)$ grados de libertad si es cierta la hipótesis nula H_0, con $e_{ij} > 5;\ 1 \leq i \leq p;\ 1 \leq j \leq m$.

- Determinación de la Región Crítica

  La región crítica para el contraste de independencia se determina:

$$\chi^2_{\alpha,gl} = P\left[\chi^2_{(p-1).(m-1)} \geq p/H_0\right] = \alpha$$

- Regla de Decisión

  La decisión que se toma, para un nivel de significancia α será:

  $\chi_c^2 < \chi^2_{\alpha,gl} \Rightarrow$ Se acepta H_0 (no existe diferencia significativa al nivel α).

  $\chi_c^2 > \chi^2_{\alpha,gl} \Rightarrow$ Se rechaza H_0 (Existe diferencia significativa al nivel α).

  (Tomaiconza Ataulluco & Pari Sallo, 2014)

2.2.7.11. Variables Ficticias (Dummy)

Las variables ficticias, también llamadas falsas, surgen de la necesidad de involucrar variables cualitativas (atributos o de categorías) en el análisis de Regresión Logística. Por ejemplo, se tienen las siguientes variables:

- Estado Civil (soltero, casado, viudo, divorciado).
- Zona de residencia (urbana, rural, sub-urbano).
- Religión (católico, testigo, musulmán, evangélico)
- Ocupación (empleador, cuenta propia, asalariado, trabajador sin remuneración).

Como se observa cada una de estas variables tiene más de dos categorías, en estos casos se hace necesario la inclusión de variables de las que sospeche su importancia, y puedan proporcionar estimaciones más precisas. Para ser incorporadas en el modelo de regresión deben ser codificadas convenientemente, y la regla es introducir en el modelo tantas variables imaginarias como categorías menos uno tenga la variable cualitativa (si la variable cualitativa tiene m categorías, se introducirán en el modelo de regresión $m-1$ variables falsas). Una variable ficticia puede tomar solo los valores uno o cero, para identificar las diferentes categorías de la variable cualitativa, y solo es aplicable en los casos en los que la ecuación de regresión tiene una constante o intercepto. (Rodas Guizado, 2011)

2.2.11.2. Codificación

Mostraremos la codificación mediante un ejemplo, con la variable *ocupación.*

Tabla 17:
Codificación de Variables Ficticias

Categoría Ocupacional	**Variables Ficticias**		
	Empleador	**Cuenta Propia**	**Asalariado**
Empleador	1	0	0
Cuenta Propia	0	1	0
Asalariado	0	0	1
Trabajador sin Remuneración	0	0	0

Se crean tres variables dicotómicas la primera de ellas sería "Empleador", quien lo sea tendrá el valor 1 en esa variable y el valor cero en las variables "Cuenta Propia" y "Asalariado". Los "Por Cuenta Propia" tendrán valor 1 en la segunda variable, en las otras, de forma similar para "Asalariado". No se necesita crear, en cambio, una variable llamada "Trabajador sin Remuneración", lo será quien tenga cero en las tres anteriores. Esta última es la categoría "Base o de referencia" de las variables ficticias. Una vez realizada esta transformación, estas variables pueden ser incorporadas en una ecuación de regresión. (Chitarroni, 2002).

2.2.8. Caja Municipal de Ahorro y Crédito Cusco

La Caja Municipal de Ahorro y Crédito Cusco S.A., es una empresa pública con personería jurídica de derecho privado, creada bajo el ámbito del Decreto Ley N° 23039 del 14 de mayo de 1980, con autorización de funcionamiento mediante Resolución S.B.S. N° 218-88 del 22 de marzo de 1988, se inició con los servicios de crédito prendario, contando con la Asesoría Técnica de la GTZ en el marco del convenio Perú-Alemania. Al segundo año de funcionamiento tuvo la autorización para la captación de ahorros del público, al tercer año para el otorgamiento de créditos a la micro y pequeña empresa y, posteriormente otras modalidades de créditos. (Caja Municipal Cusco, 2017, pág. 19)

2.2.9. Riesgo

El riesgo es la valorización de una situación negativa, probable y futura que ocasiona un daño (pérdida del valor económico) y debido a ello, sus características básicas sobre la incertidumbre en cuanto a la posibilidad de que se materialice en una pérdida efectiva, sobre la cuantía de la pérdida y sobre el momento en que se materializa la pérdida. (Vásconez E., 2010)

2.2.10. Crédito

El crédito, entendido como un proceso, consiste en que una persona natural o persona jurídica otorgue un financiamiento a otra persona natural o jurídica, a cambio de que esta última, en un período posterior devuelva el financiamiento recibido conjuntamente con una retribución, conocida como tasa de interés compensatoria o tasa de interés activa, la cual expresa el valor del dinero en el tiempo, más otros gastos asociados al financiamiento, si los hubiese.

En ese sentido, la persona natural o jurídica que otorga el financiamiento, se le conoce como agente prestamista o acreedor. Mientras que, la persona natural o jurídica que recibe el financiamiento se le llama agente prestatario o deudor.

Cuando el agente prestamista es una institución jurídica, por ejemplo una empresa financiera (Banco, Caja Municipal, Rural u otras) y el agente prestatario es una persona natural o persona jurídica (empresa), el financiamiento a otorgar debe ser evaluado y analizado por la empresa financiera, determinando a priori el riesgo de crédito que dicha operación involucra, determinar si el prestatario devolverá el financiamiento recibido, ya que de no hacerlo implicaría una pérdida para la empresa financiera.

2.2.11. Cartera Crediticia

Si la solicitud de crédito ha sido aprobada por la empresa financiera, entonces se habrá otorgado un crédito, de hacerlo extensivo a otras solicitudes de créditos y pasando por el proceso antes descrito, se aprobarán otro créditos, pasando a formar un portafolio crediticio, llamado también cartera de créditos, el cual expresa el conjunto de créditos que se hayan aprobado y/o desembolsado y que está bajo la administración de la empresa financiera. (Vela Zavala & Caro Anchay, 2010)

2.2.12. Riesgo de Crédito

El riesgo de crédito es la posibilidad de sufrir una pérdida como consecuencia de un impago o morosidad por parte de la contrapartida en una operación financiera, es decir, el riesgo de que no se llegue a pagar la deuda.

El riesgo de crédito supone una variación en los resultados financieros de un activo financiero una cartera de inversión tras la quiebra o impago de una empresa. Por tanto, es una forma de medir la probabilidad que tiene un deudor (derecho de pago) frente a un acreedor (derecho de cobro) de cumplir con sus obligaciones de pago, ya sea durante la vida del activo financiero o a vencimiento.

Este tipo de riesgo está relacionado directamente con los problemas que pueda presentar la compañía, de una forma individual. En cambio, el riesgo de mercado (en el que se incluye riesgo de divisa, de precio, de volatilidad, etc.) tiene un componente de riesgo sistemático (es aquel que se deriva de la incertidumbre global del mercado que afecta en mayor o menor grado a todos los activos existentes en la economía).

Una característica a tener en cuenta es la forma de la distribución del riesgo de crédito. Mientras que el riesgo de mercado toma una distribución normal, lo que quiere decir que es simétrica dando las mismas probabilidades a ambos lados de la distribución, el riesgo de crédito es asimétrica negativa. Con una asimetría negativa, hay más valores a la izquierda de la distribución, es decir, de la media. Además, la media de la distribución es menor a la media de la distribución normal. (Vásconez E., 2010)

2.2.13. Créditos a pequeñas empresas

Financiamiento crediticio de manera directa e indirecta, destinados a financiar actividades de producción, comercialización o prestación de servicios, otorgados a personas jurídicas o personas naturales que poseen las siguientes características:

- Registrar un nivel de endeudamiento total en el Sistema Financiero (sin incluir los créditos hipotecarios para vivienda) superior a S/. 20 000 pero no mayor a S/. 300 000 en los últimos seis (6) meses.

Otros criterios para considerar un financiamiento como crédito a pequeñas empresas:

- Si posteriormente, el endeudamiento total del deudor en el sistema financiero (sin incluir los créditos hipotecarios para vivienda) excediese los S/. 300 000 por seis (6) meses consecutivos, los créditos que inicialmente se clasificaron como a pequeñas empresas deberán ser reclasificados como créditos a medianas empresas.
- Si posteriormente, el endeudamiento total del deudor en el sistema financiero (sin incluir los créditos hipotecarios para vivienda) disminuyese posteriormente a un nivel no mayor a S/. 20 000 por seis (6) meses consecutivos, los créditos que inicialmente se clasificaron como a pequeñas empresas deberán ser reclasificados como créditos a microempresas. (Vela Zavala & Caro Anchay, 2010)

2.2.14. Créditos a Microempresas

Financiamiento crediticio de manera directa e indirecta, destinados a financiar actividades de producción, comercialización o prestación de servicios, otorgados a personas jurídicas o personas naturales que poseen las siguientes características:

- Registrar un nivel de endeudamiento total en el Sistema Financiero (sin incluir los créditos hipotecarios para vivienda) no mayor a S/. 20 000 en los últimos seis (6) meses.

Otros criterios para considerar un financiamiento como crédito a microempresas:

- Si posteriormente, el endeudamiento total del deudor en el sistema financiero (sin incluir los créditos hipotecarios para vivienda) excediese a un nivel mayor a S/. 20 000 por seis (6) meses consecutivos, los créditos que inicialmente se clasificaron como de microempresa deberán ser reclasificados según el nivel de endeudamiento que corresponda. (Vela Zavala & Caro Anchay, 2010)

2.2.15. Morosidad

La morosidad de un crédito se define como una situación en la que el deudor se ha retrasado 60 días a más en el pago de los intereses y/o el principal de su deuda. Se trata de una situación de alto riesgo pero que aún no ha caído en la categoría de crédito fallido (irrecuperable).

Cuando un crédito se considera irrecuperable se eliminan de la contabilidad de morosos. (Bedregal & Barba, 2017)

III. HIPÓTESIS Y VARIABLES

3.1. Hipótesis

a. Hipótesis general

Existe un modelo de regresión binaria que se ajusta al riesgo crediticio de la Caja Municipal de Ahorro y Crédito Cusco.

b. Hipótesis Específica

Existen factores asociados al riesgo crediticio en la Caja Municipal de Ahorro y Crédito Cusco.

3.2. Identificación de Variables e Indicadores

Dentro de las variables consideradas en el presente trabajo de investigación identificamos las siguientes:

a. Variable Dependiente:

Para el presente estudio, la variable dependiente es la "**Riesgo Crediticio" (Morosidad**).

Tabla 18*:*
Tabla Descriptiva de la Variable Dependiente

Variable	Descripción	Tipo	Escala de Medición	Criterios de Medición	Indicador de Calificación
Riesgo Crediticio (Morosidad)	Préstamo con un retraso de pago de 60 días a más.	Cualitativa	Nominal	Si: 1 No: 0	Presencia Si

b. Variables independientes:

Estas variables describen características cualitativas y cuantitativas del sujeto de crédito, que interviene en el estudio como factores de riesgo en el Riesgo Crediticio (morosidad) y son Edad del Cliente, Sexo del Cliente, Actividad del Cliente, Antigüedad del Negocio, Destino del Préstamo, Antecedentes en el Clearing, Pasivo Financiero, Capital del Préstamo y Número de Cuotas, las cuales se detallan en la Tabla 19:

Tabla 19:
Matriz de Operacionalización de Variables Independientes

VARIABLE	DEFINICIÓN	INDICADOR	DIMENSIÓN	ESCALA
Edad del cliente	Tiempo que ha vivido el cliente contando desde su nacimiento.	Fecha de nacimiento	Numérico	Cuantitativo
Sexo del cliente	Es el conjunto de las peculiaridades que caracterizan a los individuos de una especie dividiéndolos en masculino y femenino.	Caracteres sexuales secundarios	1. Masculino 0. Femenino	Nominal dicotómico
Actividad del cliente	Actividad económica que realiza el cliente.	Comprobantes de pago. Ficha RUC.	1. Indep. sin Est. Sup. 2. Indep. con Est. Sup. 3. Empl. Profesional 4. Empl. Técnico 5. Empl. con Oficio	Nominal politómico
Antigüedad del negocio	Antigüedad de la actividad económica que realiza el cliente.	Datos de la SUNAT. Comprobantes de pago.	Numérico	Cuantitativo
Destino del préstamo	Descripción de la razón por la cual el cliente adquiere el préstamo.	Solicitud de préstamo del cliente.	1. Implem. de negocio 2. Adquis. de bien inmueble 3. Adquis. de bien mueble. 4. Construc. de vivienda	Nominal politómico
Antecedentes en el clearing	Historial crediticio del cliente.	Sistema de entidades bancarias.	1. Si 0. No	Nominal dicotómico
Pasivo financiero	Registro crediticio de deudas del cliente en otras entidades financieras en el momento de solicitar el préstamo.	Sistema de entidades bancarias.	1. Si 0. No	Nominal dicotómico
Capital del préstamo	Monto del préstamo otorgado sin considerar los intereses.	Solicitud de préstamo del cliente.	Numérico	Cuantitativo
Número de cuotas	Cantidad de cuotas en las que el cliente debe abonar a la entidad financiera para la cancelación de su préstamo.	Programación emitida por la entidad financiera.	Numérico	Cuantitativo

IV. METODOLOGÍA

4.1. Delimitación Geográfica

La investigación se realiza en la Caja Municipal de Ahorro y Crédito Cusco, del distrito de Wanchaq, provincia del Cusco y departamento del Cusco, con número de RUC 20114839176.

4.2. Metodología de la Investigación

4.2.1.Tipo de Investigación

El presente trabajo de investigación se desarrolla de forma cuantitativa, descriptiva, correlacional, explicativa y transversal.

- El enfoque cuantitativo (que representa, un conjunto de procesos) es secuencial y probatorio. Cada etapa precede a la siguiente y no podemos "brincar" o eludir pasos. El orden es riguroso, aunque desde luego, podemos redefinir alguna fase. Parte de una idea que va acotándose y, una vez delimitada, se derivan objetivos y preguntas de investigación, se revisa la literatura y se construye un marco o una perspectiva teórica. De las preguntas se establecen hipótesis y determinan variables; se traza un plan para probarlas (diseño); se miden las variables en un determinado contexto; se analizan las mediciones obtenidas utilizando métodos estadísticos, y se extrae una serie de conclusiones respecto de la o las hipótesis. (Hernandez Sampieri, Fernandez Collado , & Baptista Lucio, 2014)
- Con los estudios descriptivos se busca especificar las propiedades, las características y los perfiles de personas, grupos, comunidades, procesos, objetos o cualquier otro fenómeno que se someta a un análisis. Es decir, únicamente pretenden medir o recoger información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren, esto es, su objetivo no es indicar

cómo se relacionan éstas. (Hernandez Sampieri, Fernandez Collado , & Baptista Lucio, 2014)

- El estudio correlacional tiene como finalidad conocer la relación o grado de asociación que exista entre dos o más conceptos, categorías o variables en una muestra o contexto en particular. En ocasiones sólo se analiza la relación entre dos variables, pero con frecuencia se ubican en el estudio vínculos entre tres, cuatro o más variables. Para evaluar el grado de asociación entre dos o más variables, en los estudios correlacionales primero se mide cada una de éstas, y después se cuantifican, analizan y establecen las vinculaciones. Tales correlaciones se sustentan en hipótesis sometidas a prueba. (Hernandez Sampieri, Fernandez Collado , & Baptista Lucio, 2014)
- Los estudios explicativos van más allá de la descripción de conceptos o fenómenos o del establecimiento de relaciones entre conceptos; es decir, están dirigidos a responder por las causas de los eventos y fenómenos físicos o sociales. Como su nombre lo indica, su interés se centra en explicar por qué ocurre un fenómeno y en qué condiciones se manifiesta o por qué se relacionan dos o más variables. (Hernandez Sampieri, Fernandez Collado , & Baptista Lucio, 2014)
- El estudio transversal se define como un tipo de investigación observacional que analiza datos de variables recopiladas en un periodo de tiempo sobre una población muestra o subconjunto predefinido. Este tipo de estudio también se conoce como estudio de corte transversal, estudio transversal y estudio de prevalencia. (Hernandez Sampieri, Fernandez Collado , & Baptista Lucio, 2014)

4.3. Unidad de Análisis

La Unidad de Análisis del presente trabajo de investigación de la Caja Municipal de Ahorro y Crédito Cusco, son los clientes de la Caja Municipal de Ahorro y Crédito Cusco, a los que se les otorgó créditos en el periodo de enero a diciembre del 2014.

4.4. Población y Muestra

4.4.1 Población de Estudio

Durante el período del 01 de enero al 31 de diciembre del año 2014 la Caja Municipal de Ahorro y Crédito Cusco tuvo en su cartera de clientes a un total de 950 clientes (pequeñas empresas y microempresas), a los que se les otorgó crédito según su base de datos.

4.4.2. Muestra

4.4.2.1. Tamaño de Muestra

Para obtener el tamaño de muestra, usamos la fórmula para poblaciones finitas, que se muestra a continuación:

$$n = \frac{NZ_0^2 p(1-p)}{e^2(N-1) + Z_0^2 p(1-p)}$$

Donde:

$N = 950$	:	Población (Total de clientes año 2014).
$Z_0 = 1.96$	:	Valor en la tabla normal estándar para un nivel de confianza del 95%.
$p = 0.1$	:	Probabilidad de ocurrencia del evento (Morosidad).
$1 - p = 0.9$	:	Probabilidad de no ocurrencia de riesgo crediticio (morosidad).
$e = 0.026$	:	Error de muestreo.

El valor de $p = 0.1$, éxito de que ocurra el evento, fue considerado de un estudio realizado por la propia Caja Municipal, en el que concluyen que, "El 10% de los clientes que reciben préstamos financieros de la Caja Municipal, presenta Morosidad", y por coplemento $1 - p = 0.9$ (fracaso), y el error de muestreo de 2.6%, fue facilitado por el personal de la Caja Municipal de Ahorro y Crédito Cusco y se eligió para obtener una estimación más precisa,

ya que los datos obtenidos se recolectaron de las fichas de crédito llenadas por el personal de la Caja Municipal, al momento de otorgar y culminar el periodo de préstamo (pero no durante el período de préstamo).

Reemplazando los datos en la fórmula se tiene:

$$n = \frac{950 \times 1.96^2 \times 0.1 \times 0.9}{0.026^2 \times (950 - 1) + 1.96^2 \times 0.1 \times 0.9}$$

$$n = 330.04 \approx 330$$

Por lo tanto el tamaño muestral con el que se trabajó fue 330 clientes.

4.5. Técnicas de Selección de Muestra

Para la selección de los clientes que entran en la muestra, utilizamos el muestreo probabilístico aleatorio simple sin reposición. Los números aleatorios fueron generado en la hoja de cálculo Excel con la siguiente función:

$$f_x = ALEATORIO.ENTRE(1; 950)$$

Luego de obtener los números aleatorios de acuerdo al **Cuadro A.1. del Anexo A**, se procedió a recolectar la información de los clientes seleccionados en la muestra de la base de datos de la Caja Municipal de Ahorro y Crédito Cusco.

Para continuar con el procesamiento de la información en el Programa Estadístico SPSS versión 22**.**

V. RESULTADOS Y DISCUSIÓN

5.1. Procesamiento, Análisis e Interpretación de Resultados

5.1.1. Analisis de Datos

Consideramos conveniente analizar los datos obtenidos de la muestra, para verificar la correlación entre las variables independientes de forma individual; es decir, la asociación de cada variable, esto con el propósito de llegar a una mejor interpretación de los datos.

Entre las nueve variables independientes consideradas distinguimos que, cuatro son cuantitativas y el resto son categóricas, dentro de estas últimas, 3 son dicotómicas y 2 son politómicas. Identificadas las variables, pasamos a realizar el análisis descriptivo de las mismas, de acuerdo al tipo de variable.

5.1.2. Análisis Descriptivo

En el análisis inicial de los datos obtenidos, empezamos con la totalidad de la muestra que son 330 clientes atendidos durante el año 2014, de estos datos podemos afirmar que:

Tabla 20:
Morosidad del Cliente

Morosidad	Frecuencia	Porcentaje
No	231	70,0
Si	99	30,0
Total	**330**	**100,0**

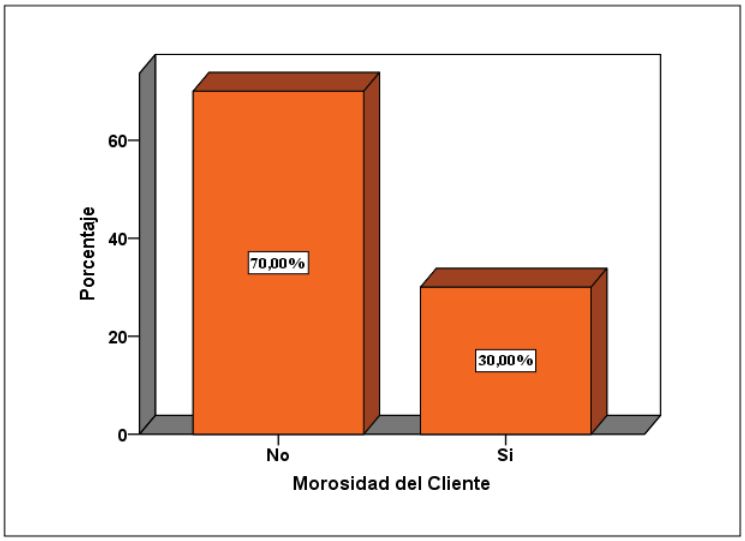

Figura 1. Morosidad del Cliente

La Tabla 20, indica que de la muestra bajo estudio, 330 clientes atendidos en la Caja Municipal de Ahorro y Crédito Cusco durante el periodo del 01 de enero al 31 de diciembre del 2014, el 30,0% de clientes presentaron morosidad, mientras que el 70,0% de clientes no presentaron morosidad durante el lapso de préstamo.

5.1.2.1. Tiempo de Retraso en el período de préstamo dentro de la Morosidad

Seguidamente dentro de los préstamos con morosidad durante el periodo indicado anteriormente, se desea saber la cantidad de días de retraso más frecuente que se haya presentado, así se tiene:

Tabla 21*:*
Frecuencias de Morosidad

Días de retraso	Frecuencia	Porcentaje	Porcentaje Acum.
<=67	13	13,1	12,1
[68-75]	26	26,3	39,4
[76-83]	31	31,3	70,7
[84-91]	25	25,3	96,0
>=92	4	4,0	100,0
Total	**99**	**100,0**	

La Tabla 21 indica que la morosidad de los préstamos ocurridos según el número de días de retraso: El 13,1% de los clientes tienen de 60 a 67 días de retraso; 26,3% de clientes tienen un retraso de 68 a 75 días; 31,3% personas se retrasaron pagando su préstamo de 76 a 83 días; 25,3% de clientes se retrasaron pagando de 84 a 91 días y 4,0% de clientes tienen de 92 a más días de retraso en el pago de su préstamo.

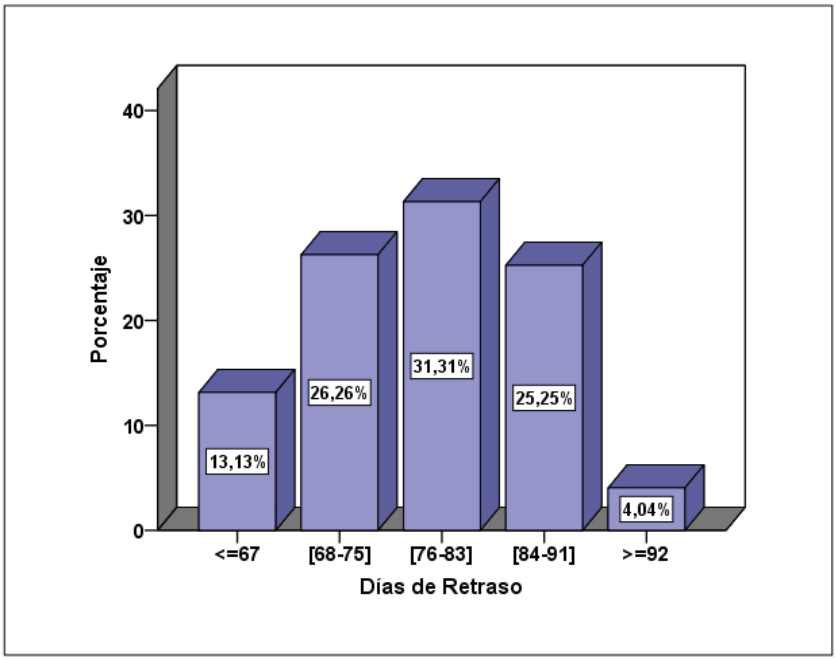

Figura 2. Frecuencias de Morosidad

Como se observa en la Figura 2. podemos decir que el 31,3% de los clientes se retrasaron en el pago del préstamo entre 76 a 83 días, siendo este tiempo el de mayor frecuencia.

5.1.2.2. Variables Independientes Cuantitativas

a. Tabla Descriptiva

Tabla 22:
Estadísticos Descriptivos

	N	Mínimo	Máximo	Media	Desv. Estánd.	Varianza
Edad del Cliente	330	26	73	47,53	9,57	91,54
Antigüedad del Negocio	330	2	30	17,91	8,35	69,64
Capital del Préstamo	330	5000	40000	18256,06	7986,43	63783017,87
N° de Cuotas en Meses	330	12	48	25,70	11,19	125,11

En la Tabla 22 se observa los estadísticos descriptivos como la media, desviación estándar y la varianza de las variables cuantitativas consideradas en el estudio.

b. Tablas de Frecuencia

A continuación se tiene la base de datos utilizada para la aplicación, se presentan mediante tablas de frecuencias y diagramas de barras, tanto para datos cuantitativos como categóricos.

Tabla 23:
Frecuencias de la Variable Edad del cliente

Edad	Frecuencia	Porcentaje
<=33	37	11,2
[34-41]	54	16,4
[42-49]	56	17,0
[50-57]	162	49,1
[58-65]	19	5,8
>=66	2	,6
Total	**330**	**100,0**

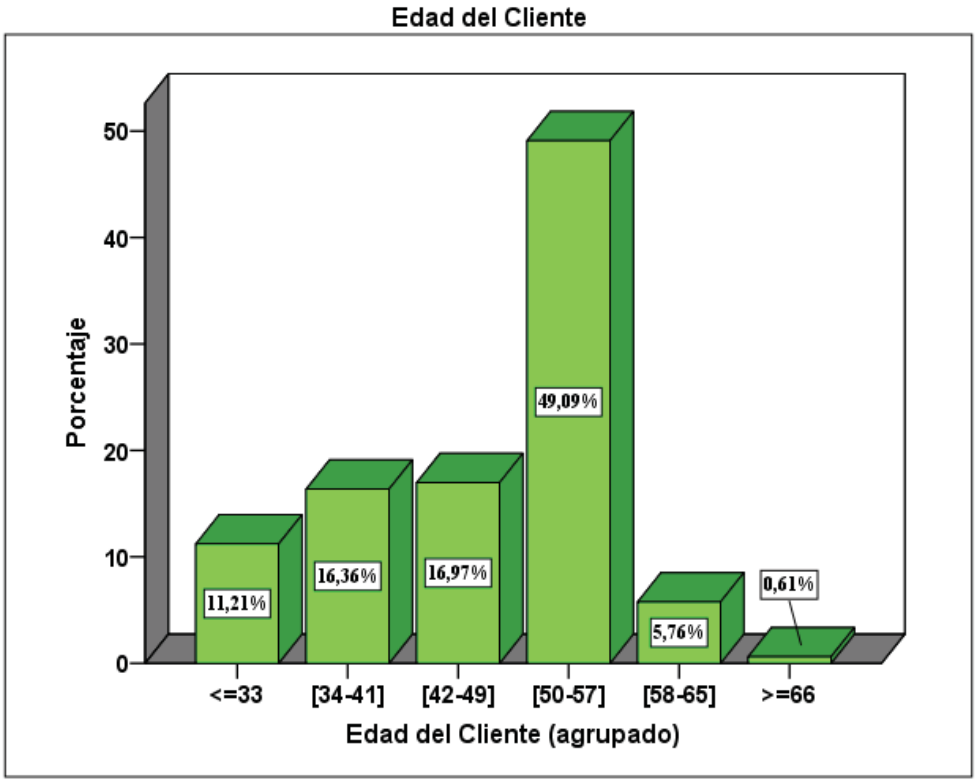

Figura 3. Frecuencias de la Variable Edad del cliente

Tabla 24:
Frecuencias de la Variable Antigüedad del Negocio

Antigüedad del Negocio	Frecuencia	Porcentaje
<=7	67	20,3
[8-11]	13	3,9
[12-15]	33	10,0
[16-19]	31	9,4
[20-23]	65	19,7
[24-27]	95	28,8
>=28	26	7,9
Total	330	100,0

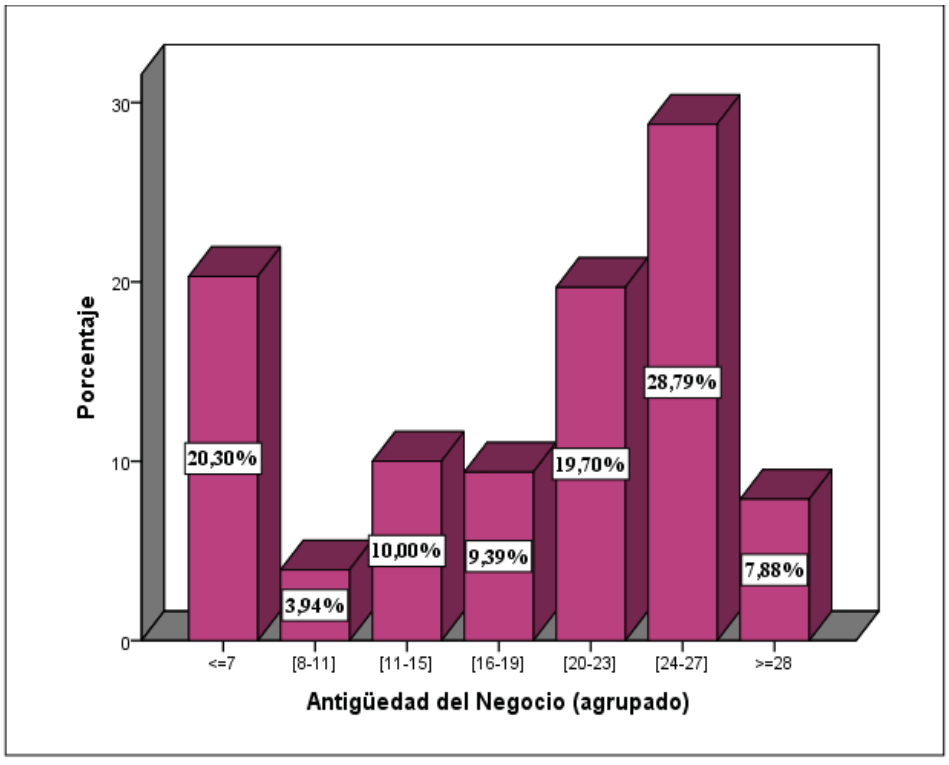

Figura 4. Frecuencias de la Variable Antigüedad del Negocio

Tabla 25:
Frecuencias de la Variable Capital de Préstamo

Capital del Préstamo	Frecuencia	Porcentaje
<=9375	28	8,5
]9375-13750]	67	20,3
]13750-18125]	97	29,4
]18125-22500]	51	15,5
]22500-26875]	39	11,8
]26875-31250]	30	9,1
]31250-35625]	5	1,5
>35625	13	3,9
Total	**330**	**100,0**

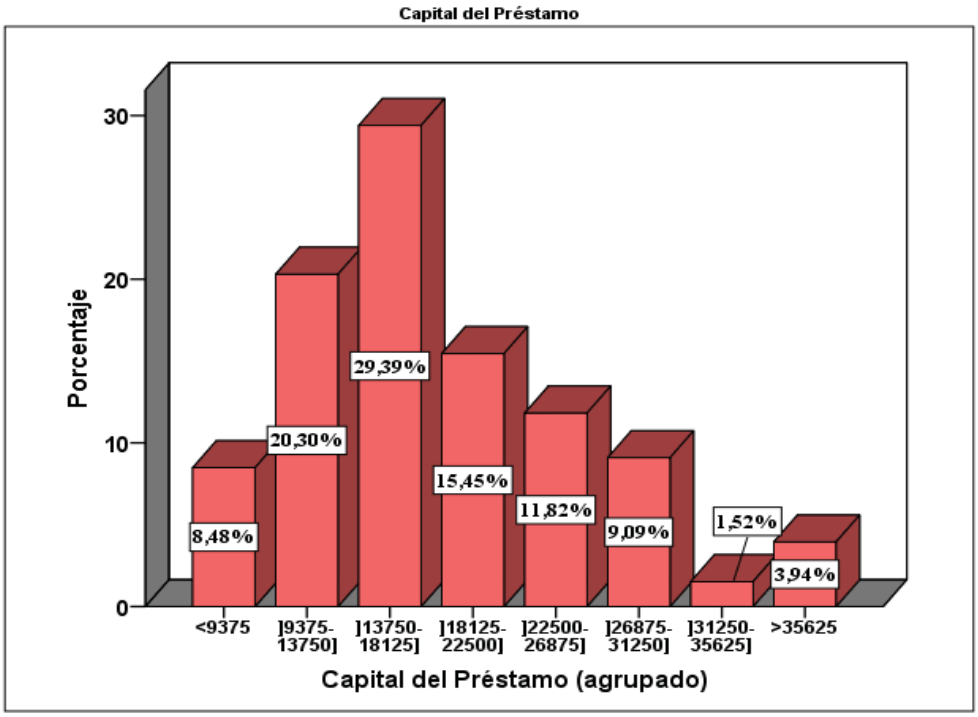

Figura 5. Frecuencias de la Variable Capital de Préstamo

Tabla 26:
Frecuencias de la Variable Número de Cuotas

Número de Cuotas	Frecuencia	Porcentaje
<=19	123	37,3
[20-26]	96	29,1
[27-33]	21	6,4
[34-40]	51	15,5
>=41	39	11,8
Total	330	100,0

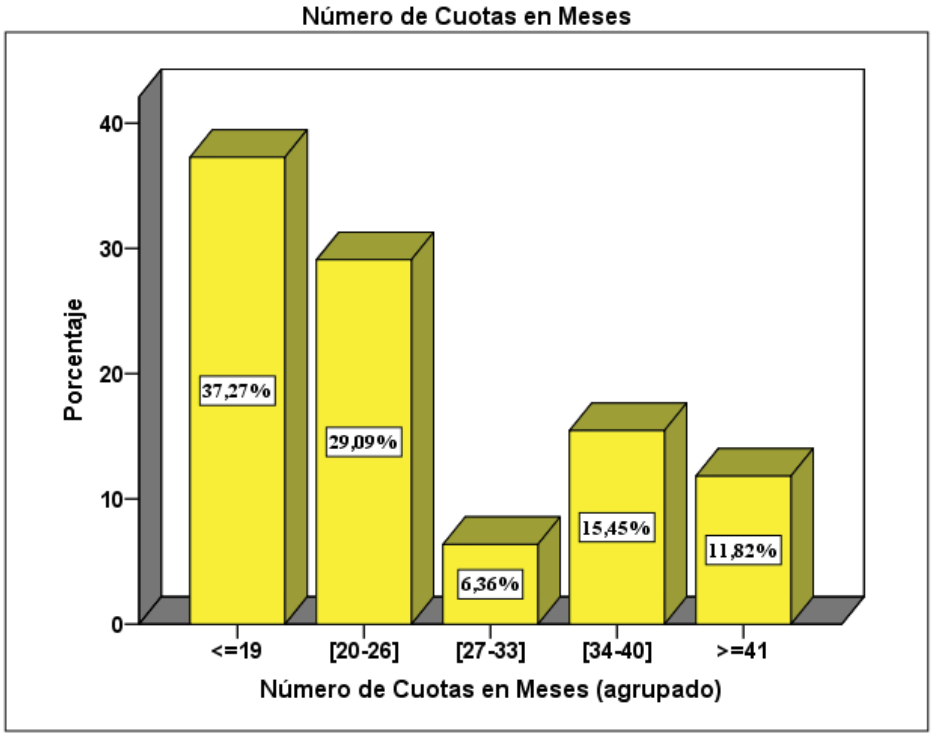

Figura 6. Frecuencias de la Variable Número de Cuotas

5.1.2.3. Variables Independientes Categóricas

a. Tabla de Frecuencias de Datos Categóricos

Tabla 27*:*
Frecuencias de la Variable Sexo del cliente

Sexo	Frecuencia	Porcentaje
Femenino	166	50,3
Masculino	164	49,7
Total	**330**	**100,0**

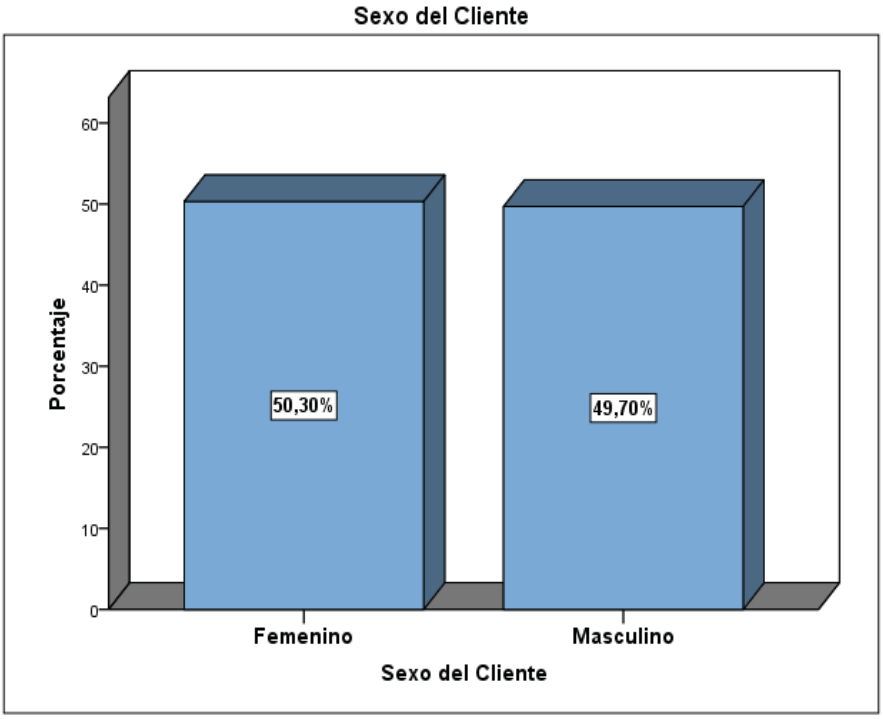

Figura 7. Frecuencias de la Variable Sexo del cliente

Tabla 28:
Frecuencias de la Variable Antecedentes en Clearing

Antecedentes en el clearing	Frecuencia	Porcentaje
Si	130	39,4
No	200	60,6
Total	**330**	**100,0**

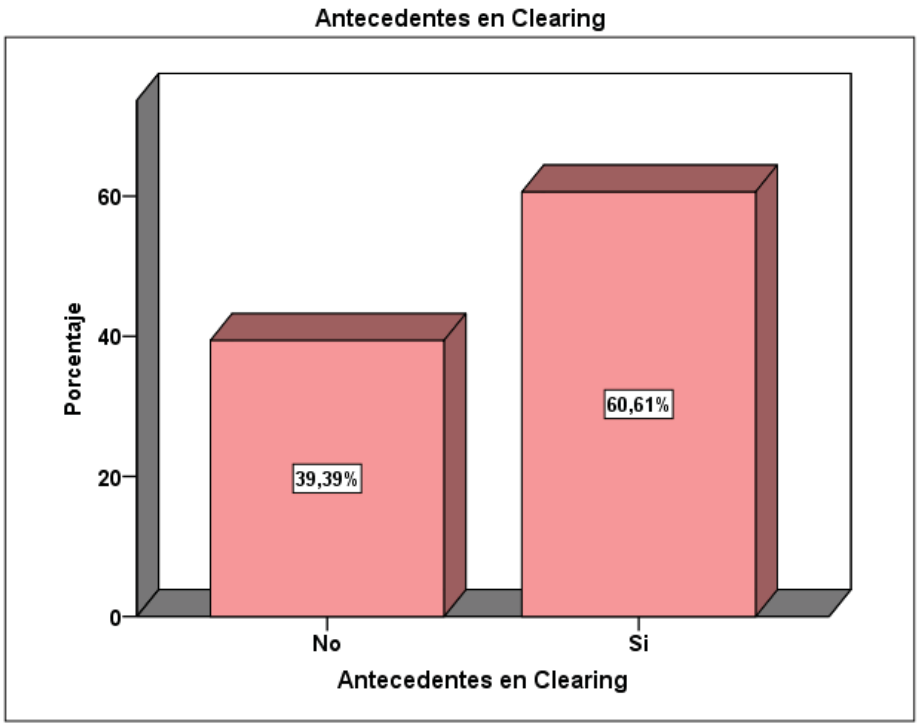

Figura 8. Frecuencias de la Variable Antecedentes en Clearing

Tabla 29:
Frecuencias de la Variable Pasivo Financiero

Pasivo Financiero	Frecuencia	Porcentaje
Si	193	58,5
No	137	41,5
Total	**330**	**100,0**

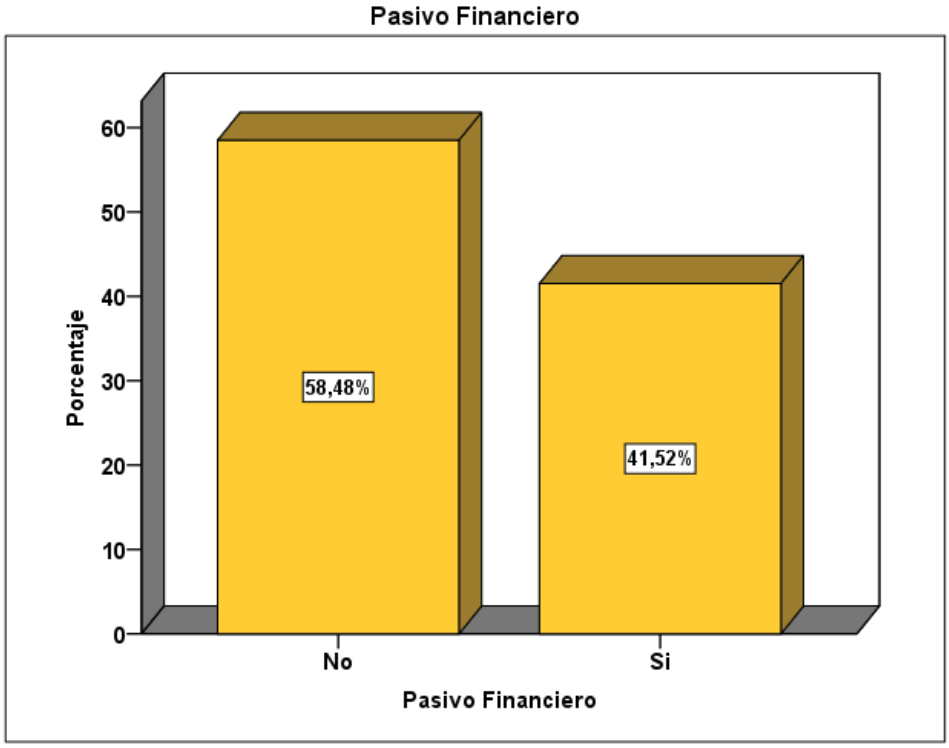

Figura 9. Frecuencias de la Variable Pasivo Financiero

Tabla 30:
Frecuencias de la Variable Actividad del cliente

Actividad	Frecuencia	Porcentaje
Independiente sin Estudios Superiores	100	30,30
Independiente con Estudios Superiores	71	21,52
Empleado Profesional	66	20,00
Empleado Técnico	27	8,18
Empleado con Oficio	66	20,00
Total	**330**	**100,00**

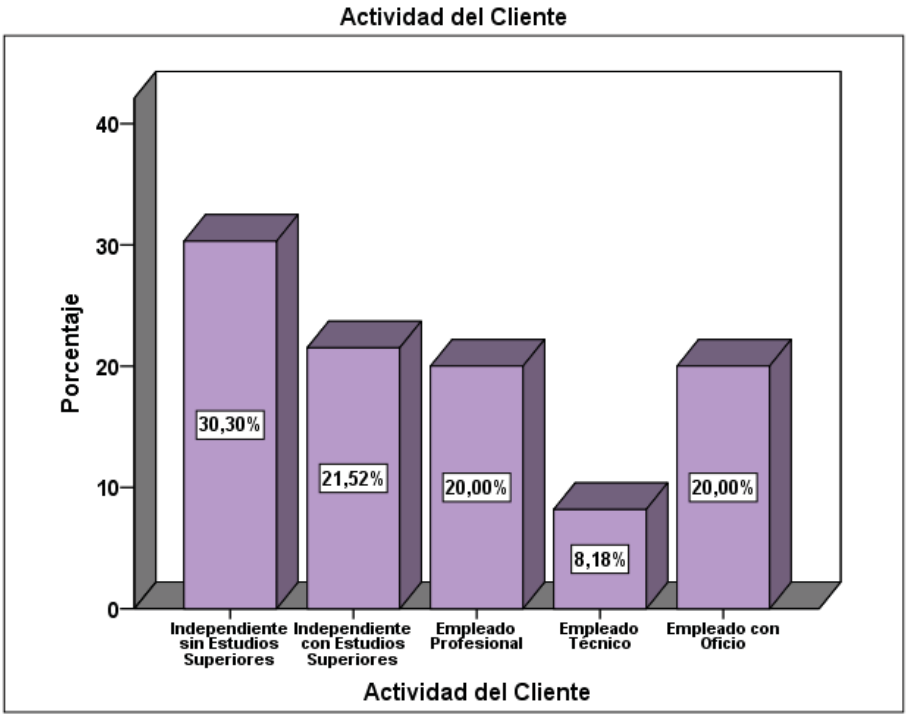

Figura 10. Frecuencias de la Variable Actividad del cliente

Tabla 31:
Frecuencias de la Variable Destino del Préstamo

Actividad	Frecuencia	*Porcentaje*
Implementación de Negocio	106	32,12
Adquisición de Bien Inmueble	83	25,15
Adquisición de Bien para el Hogar	68	20,61
Construcción de Vivienda	73	22,12
Total	**330**	**100,0**

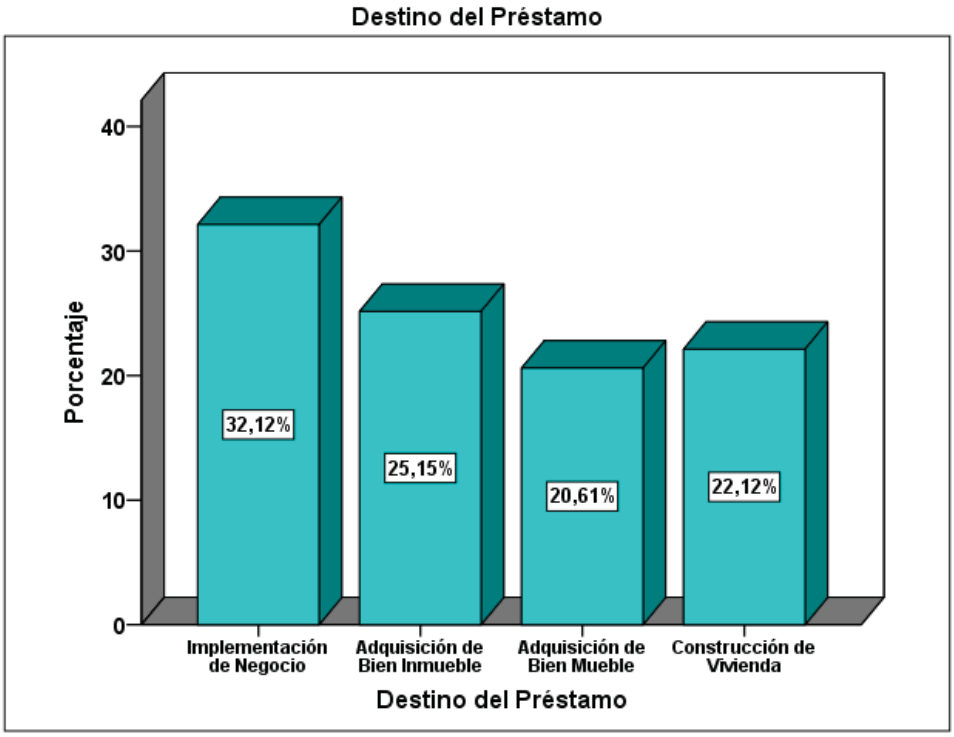

Figura 11. Frecuencias de la Variable Destino del Préstamo

5.1.3. Análisis e Interpretación de Resultados

5.1.3.1. Formulación del Modelo de Elección Binaria Propuesta

$$p = \frac{e^L}{1+e^L} = \frac{1}{1+e^{-L}}$$

donde:

$$\hat{L} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \cdots + \hat{\beta}_9 X_9$$

$$p = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X_1+\hat{\beta}_2 X_2+\hat{\beta}_3 X_3+\cdots+\hat{\beta}_9 X_9)}}$$

donde:

X_1	:	Antigüedad del Negocio
X_2	:	Pasivo Financiero
X_3	:	Capital del Préstamo
X_4	:	Destino del Préstamo
X_5	:	Número de Cuotas
X_6	:	Sexo del Cliente
X_7	:	Actividad del Cliente
X_8	:	Antecedentes en Clearing
X_9	:	Edad del Cliente

Hipótesis

H_0 : $\hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = \cdots = \hat{\beta}_9$

H_1 : $\hat{\beta}_i \neq 0$, para por lo menos un $i = 1,2,3,\ldots,9$

Que es lo mismo, decir:

H_0 : Las variables $(X_1, X_2, \ldots, X_9)$ no influyen en el Riesgo Crediticio.

H_1 : Las variables $(X_1, X_2, \ldots, X_9)$ influyen en el Riesgo Crediticio.

5.1.3.2. Modelo de Elección Binaria Ajustada utilizando el Método Forward

Para la selección de variables, consideramos las 9 variables independientes, que anteriormente fueron analizadas individualmente, en este punto lo realizamos en conjunto

para elegir las variables que explican la presencia del Riesgo Crediticio y su influencia. En esta selección usamos el método "Forward", este método comienza con el bloque inicial, que en su selección solo considera a la constante $\hat{\beta}_0 = -0{,}847$, como se aprecia a continuación en la siguiente tabla:

Tabla 32:
Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-0,847	0,120	49,751	1	0,000	0,429

En la Tabla 32. se puede observar a las variables que no estaban en el modelo, pero que ingresan una a una, debido a que el valor de la Puntuación Eficiente de Rao es representativo y tiene un nivel de significancia inferior a 0.05.

Tabla 33:
Variables con Puntuación Eficiente de Rao representativo y Valor Significativo en cada Paso

Pasos	Variables	Puntuación Eficiente de Rao	gl	Sig.
Paso 1	Antigüedad del Negocio	145,607	1	0,000
Paso 2	Pasivo Financiero	49,985	1	0,000
Paso 3	Capital del Préstamo	38,415	1	0,000
Paso 4	Destino del Préstamo	34,590	3	0,000
Paso 5	Número de Cuotas	18,693	1	0,000
Paso 6	Sexo	13,217	1	0,000
Paso 7	Actividad del Cliente	9,533	4	0,049
Paso 8	Antecedentes en Clearing	0,571	1	0,450

a. La Variable Antecedentes en Clearing no ingresa en el modelo por ser, no significativa.

Por lo tanto, se puede indicar que en el paso inicial la función estimada es:

$$\hat{L} = -0{,}847$$

Donde $-0{,}847$ es el valor de la constante.

En el Paso 1, vemos que la variable, Antigüedad del Negocio ingresa porque presenta una Puntuación Eficiente de Rao de $145{,}607$ y es significativa con un nivel del 5% entonces la nueva función estimada es:

$$\hat{L} = 2{,}685 - 0{,}222X_1$$

Donde X_1 es Antigüedad del Negocio.

A partir del ingreso de la siguiente variable, analizamos desde el Bloque 1.

En el Paso 2, vemos que la variable, Pasivo Financiero ingresa porque presenta una Puntuación Eficiente de Rao de 49,985 y es significativa con un nivel del 5% entonces la nueva función estimada es:

$$\hat{L} = 3{,}647 - 0{,}246X_1 - 2{,}356X_2$$

Donde X_2 es la variable Pasivo Financiero.

En el Paso 3, la variable que ingresa es Capital del Préstamo, con una Puntuación Eficiente de Rao de 38,415 por ser significativa, entonces la nueva función estimada es:

$$\hat{L} = 3{,}647 - 0{,}246X_1 - 3{,}759X_2 + 0{,}000X_3$$

Donde X_3 es la variable Capital de Préstamo.

En el Paso 4, la variable que ingresa es Destino del Préstamo, con una Puntuación Eficiente de Rao de 34,590 por ser significativa, entonces la nueva función estimada es:

$$\hat{L} = 0{,}689 - 0{,}210X_1 - 3{,}888X_2 + 0{,}000X_3 + 2{,}870X_4$$

Donde X_4 es la variable Destino del Préstamo.

En el Paso 5, la variable que ingresa es Número de Cuotas, con una Puntuación Eficiente de Rao de 18,693 por ser significativa, entonces la nueva función estimada es:

$$\hat{L} = 2{,}137 - 0{,}182X_1 - 3{,}247X_2 + 0{,}000X_3 + 2{,}620X_4 - 0{,}172X_5$$

Donde X_5 es la variable Número de Cuotas.

En el Paso 6, ingresa la variable Sexo del Cliente, con una Puntuación Eficiente de Rao de 13,217 por ser significativa, entonces la nueva función estimada es:

$$\hat{L} = 3{,}527 - 0{,}187X_1 - 3{,}745X_2 + 0{,}000X_3 + 2{,}412X_4 - 0{,}189X_5 - 2{,}084X_6$$

Donde X_6 es la variable Sexo del Cliente.

En el Paso 7, ingresa la variable Actividad del Cliente, con una Puntuación Eficiente de Rao de 9,533 por ser significativa, entonces la nueva función estimada es:

$\hat{L} = 2{,}869 - 0{,}210X_1 - 3{,}641X_2 + 0{,}000X_3 + 3{,}036X_4 - 0{,}229X_5 - 2{,}259X_6 + 3{,}737X_7$

Donde X_7 es la variable Actividad del Cliente.

Tabla 34*:*
Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 7[g]	Sexo(1)	-2,259	,665	11,542	1	,001	,104	,028	,385
	Actividad			8,499	4	,075			
	Actividad(1)	1,367	,882	2,404	1	,121	3,923	,697	22,082
	Actividad(2)	,234	1,192	,039	1	,844	1,264	,122	13,058
	Actividad(3)	,309	1,067	,084	1	,772	1,363	,168	11,026
	Actividad(4)	3,737	1,370	7,437	1	,006	41,973	2,861	615,713
	Antigüedad	-,210	,042	24,839	1	,000	,811	,746	,880
	Destino préstamo			21,226	3	,000			
	Destino préstamo(1)	-1,062	,925	1,320	1	,251	,346	,056	2,117
	Destino préstamo(2)	,013	,884	,000	1	,988	1,013	,179	5,735
	Destino préstamo(3)	3,036	,929	10,674	1	,001	20,831	3,370	128,767
	Pasivo finan(1)	-3,641	,794	21,040	1	,000	,026	,006	,124
	Capital	,000	,000	26,392	1	,000	1,000	1,000	1,000
	Número cuotas	-,229	,057	16,284	1	,000	,796	,712	,889
	Constante	2,869	1,182	5,896	1	,015	17,620		

g. Variables especificadas en el paso 7: Actividad.

En el siguiente paso, la posible variable a ingresar es: Antecedentes en el Clearing, mas no es significativa, por lo que el ingreso de las variables concluye en el Paso 7. Como se puede observar en la Tabla 33. Los coeficientes de las variables que ingresan en la ecuación, se observan en la Tabla 34. y se detallan en la **Tabla 73 del Anexo B**.

Bondad de Ajuste del Modelo Seleccionado:

Realizaremos esta prueba para ver los supuestos del modelo de Regresión Logística y la Bondad de Ajuste del modelo.

Contraste de Bondad de Ajuste del Modelo:

Para realizar la prueba de Bondad de Ajuste del modelo seleccionado, los resultados se muestran en el Anexo B, con el Método "Forward", procesado en el programa SPSS, así observamos la disminución del estadístico $(-2lnL)$ de la siguiente manera:

Tabla 35*:*
Resumen del Modelo

Variables en el Modelo	-2 log de la Verosimilitud
Solo la Constante	403,170
Con 7 Variables[a]	84,149

a. Antigüedad, Pasivo, Capital, Destino, Num. Cuota, Sexo, Actividad del Cliente

Inicialmente en el Bloque 0, el estadístico $(-2lnL)$ con solo la constante es de 403,170. (**Ver Tabla 63 del Anexo B),** y en el Bloque 1, en el ingreso de variables al modelo. Al ingresar la primera variable, Antigüedad del Negocio, el Estadistico $(-2lnL)$ es de 245,206, luego con el ingreso de la variable Pasivo Financiero disminuye, al ingresar la variable Capital del Préstamo el estadístico $(-2lnL)$ es de 163,202, con el ingreso de la variable Destino del Préstamo disminuye a 129,956, con el ingreso de la variable Número de Cuotas nuevamente disminuye a 107,409, con el ingreso de la variable Sexo vuelve a disminuir y llega a 93,798, finalmente con el ingreso de la variable Actividad del Cliente disminuye a 84,149 **(Ver Tabla 69 del Anexo B).**

a. Prueba de Devianza:

Como vimos en la Tabla 34. el valor del estadístico el estadístico $(-2lnL)$ para el modelo que contiene solo la constante, es igual a 403,170 y con las 7 variables explicativas ingresadas en el modelo el valor del estadístico $(-2lnL)$ disminuye a 84,149.

Por lo indicado podemos verificar lo siguiente:

Hipótesis:

H_0 : $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1 : $\beta_i \neq 0$, Para algún $i = 1,2, \dots, k$

Que es igual a decir:

H_0 : Los datos se ajustan al modelo.

H_1 : Los datos no se ajustan al modelo.

El estadístico es:

$$D = (-2lnL_{Modelo\ 0}) - (-2lnL_{Modelo\ 7})$$

$$D = (403{,}170) - (84{,}149)$$

$$D \approx 319{,}021$$

Como $319{,}021 > \chi^2_{0.05;(8)} = 15{,}507$, rechazamos H_0 y concluimos que los datos se ajustan al modelo, o que las 7 variables Antigüedad del Negocio, Pasivo Financiero Capital del Préstamo, Destino del Préstamo, Número de Cuotas, Sexo del Cliente, Actividad del Cliente influyen en el modelo.

Tasa de Clasificación Correcta:

Una vez comprobada la Bondad de Ajuste del modelo, nos interesa conocer su capacidad predictiva, es decir su eficacia para predecir si un cliente va a incurrir a morosidad o no, para ello tenemos la siguiente Tabla:

Tabla 36:
Tabla de Clasificación[a,b]

			Pronosticado		
			Morosidad del Cliente		Corrección de porcentaje
	Observado		No	Si	
Paso 0	Morosidad del	No	231	0	100,0
	Cliente	Si	99	0	0,0
	Porcentaje global				70,0

a. La constante se incluye en el modelo.
b. El valor de corte es 0,500

La Tabla 36. de clasificación, permite evaluar el ajuste del modelo de regresión que hasta este momento tiene un solo parámetro en la ecuación, esto lo hace comparando los valores pronosticados con los valores observados. Por defecto se emplea un punto de corte,

de 0,5 de probabilidad de Y para clasificar a los clientes, es decir, que los clientes con una probabilidad $< 0{,}5$ se clasifican como cero (No presentan morosidad), mientras que si la probabilidad resulta $\geq 0{,}5$ se clasifican como uno (Si presentan morosidad).

En esta tabla se muestran los casos bien clasificados en la diagonal principal, y los casos mal clasificados en la segunda diagonal, por esto es que se dice que: La capacidad predictiva es en este caso del 70,0% (Ajuste global), es la probabilidad de clasificar correctamente clientes que presentan morosidad. Esto es:

$$\frac{231+0}{231+0+99+0} = \frac{231}{330} = 70{,}0\%$$

Aquí podemos observar el indicador de ***Especificidad***, es decir, la probabilidad de que identifique como cliente que no presenta morosidad a quien efectivamente no lo presente; esto es, de los $224+7$ pacientes que **NO** presentan morosidad, 224 fueron pronosticados como que no tienen morosidad, es decir hay un porcentaje de aciertos del $\frac{224}{224+7} = 97{,}0\%$.

Y el indicador de ***Sensibilidad*** medirá la probabilidad de que la prueba identifique como clientes que presenta morosidad a quien efectivamente presenta morosidad; esto es, de los $91+8$ clientes que presentan morosidad, 91 fueron pronosticados como que presentan morosidad; es decir, hay un porcentaje de aciertos del $\frac{91}{91+8} = 91{,}9\%$.

Calidad de Ajuste:

Para analizar la calidad de ajuste de nuestro modelo, observamos la siguiente tabla:

Tabla 37:
Resumen del Modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
0	403,170		
1	245,206[a]	0,380	0,539
2	200,415[b]	0,459	0,651
3	163,272[b]	0,517	0,733
4	129,956[c]	0,563	0,798
5	107,409[c]	0,592	0,839
6	93,798[d]	0,608	0,863
7	84,149[d]	0,620	0,879

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.
b. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.
c. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de ,001.
d. La estimación ha terminado en el número de iteración 8 porque las estimaciones de parámetro han cambiado en menos de ,001.

En este punto, el análisis se realiza con los coeficientes

Pseudo R cuadrado de Mc Fadden, Cox y Snell y *R cuadrado de Nagelkerke*,

calculando los $Pseudo - R^2$, tenemos:

a. Coeficiente Pseudo-R^2 de Mc Fadden:

$$R^2_{MF} = 1 - \frac{(-2lnL)_k}{(-2lnL)_0}$$

$$R^2_{MF} = 1 - \frac{84{,}149}{403{,}170}$$

$$R^2_{MF} = 0{,}791$$

Como se observa el $0 \leq R^2_{MF} = 0791 \leq 1$ se concluye que la calidad de ajuste es excelente.

b. Coeficiente Pseudo-R^2 de Cox y Snell:

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_k}\right)^{\left(\frac{2}{N}\right)} = 1 - exp\left(\frac{(-2lnL)_k - (-2lnL)_0}{N}\right)$$

$$R^2_{CS} = 1 - exp\left(\frac{84{,}149 - 403{,}170}{330}\right)$$

$$R^2_{CS} = 1 - 0{,}380$$

$$R^2_{CS} = 0{,}620$$

Se aprecia que $0 \leq R^2_{CS} = 0{,}620 \leq 1$, entonces se concluye que la calidad de ajuste es excelente.

c. Coeficiente Pseudo-R^2 de Nagelkerke:

$$R^2_N = \left(\frac{R^2_{CS}}{1 - (L_0)^{\left(\frac{2}{N}\right)}}\right) = \frac{1 - \exp\left(\frac{(-2lnL)_k - (-2lnL)_0}{N}\right)}{1 - \exp\left(\frac{-(-2lnL)_0}{N}\right)}$$

$$R_N^2 = \frac{1 - \exp\left(\frac{84{,}149 - 403{,}170}{330}\right)}{1 - \exp\left(\frac{-403{,}170}{330}\right)}$$

$$R_N^2 = \frac{0{,}620}{0{,}705}$$

$$R_N^2 = 0{,}879$$

El valor de $0 \leq R_N^2 = 0{,}879 \leq 1$ y se aproxima a más a la unidad, por lo que se concluye que la calidad de ajuste es buena.

5.1.3.3. Modelo de Elección Binaria Ajustada utilizando el Método Backward

Una vez que las variables fueron seleccionadas según el método "Forward" (Hacia adelante). Ahora analizamos con todas las variables seleccionadas para verificar si la calidad del ajuste mejora. Para ello tenemos la probabilidad con el modelo ajustado:

$$p = \frac{1}{1 + e^{L}}$$

Donde:

$$\hat{L} = 2{,}869 - 0{,}210\boldsymbol{AntNeg} - 3{,}641\boldsymbol{PasFin} + 0{,}000\boldsymbol{CapPres} + 3{,}036\boldsymbol{DestCred_3} - 0{,}229\boldsymbol{NumCuo} - 2{,}259\boldsymbol{Sexo} + 3{,}737\boldsymbol{ActClie_4}$$

$$p = \frac{1}{1 + e^{-(2{,}869 - 0{,}210\boldsymbol{AntNeg} - 3{,}641\boldsymbol{PasFin} + 0{,}000\boldsymbol{CapPres} + 3{,}036\boldsymbol{DestCred_3} - 0{,}229\boldsymbol{NumCuo} - 2{,}259\boldsymbol{Sexo} + 3{,}737\boldsymbol{ActClie_4})}}$$

Tabla 38:
Variables en la Ecuación Modelo

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 3[g]	Sexo(1)	-2,259	,665	11,542	1	,001	,104	,028	,385
	Actividad(4)	3,737	1,370	7,437	1	,006	41,973	2,861	615,713
	Antigüedad	-,210	,042	24,839	1	,000	,811	,746	,880
	Destino préstamo(3)	3,036	,929	10,674	1	,001	20,831	3,370	128,767
	Pasivo finan(1)	-3,641	,794	21,040	1	,000	,026	,006	,124
	Capital	,000	,000	26,392	1	,000	1,000	1,000	1,000
	Número cuotas	-,229	,057	16,284	1	,000	,796	,712	,889
	Constante	2,869	1,182	5,896	1	,015	17,620		

En la primera columna se muestra a las variables en estudio, y la segunda columna muestra al valor de los coeficientes estimados (B) de las variables independientes, que se interpretan de la siguiente manera:

- $\hat{\beta}_0 = 2{,}869$; es la constante del modelo, y significa que cuando todas las variables regresoras tengan valor cero, la razón de probabilidades de incidencia y no incidencia tomará este valor.
- $\hat{\beta}_1 = -0{,}210$*;* es el coeficiente de la variable Antigüedad del Negocio y, significa que la variable Antigüedad del Negocio es un factor que disminuye la probabilidad de ocurrencia de la morosidad porque es menor que 1.
- $\hat{\beta}_2 = -3{,}641$; es el coeficiente de la variable Pasivo Financiero y, significa que la variable Pasivo Financiero es un factor que disminuye la probabilidad de ocurrencia de morosidad por que es menor que 1.
- $\hat{\beta}_3 = 0{,}000$; es el coeficiente de la variable Capital del Préstamo y, significa que dicha variable no aumenta ni disminuye la probabilidad de ocurrencia de morosidad dado que es igual a 0.
- $\hat{\beta}_4 = 3{,}036$; es el coeficiente de la variable Destino del Préstamo (Adquisición de bien mueble) y, significa que la referida variable incrementa la probabilidad de ocurrencia de morosidad porque es mayor que 1.
- $\hat{\beta}_5 = -0{,}229$; es el coeficiente de la variable Número de Cuotas y, significa que la variable Número de Cuotas disminuye la probabilidad de ocurrencia de la morosidad porque es menor que 1.
- $\hat{\beta}_6 = -2{,}259$; es el coeficiente de la variable Sexo del Cliente y significa que la variable Sexo (Femenino) disminuye la probabilidad de ocurrencia de la morosidad porque es menor que 1.

- $\hat{\beta}_7 = 3{,}737$; es el coeficiente de la variable Actividad del Cliente y, significa que la variable Actividad del Cliente (Empleado técnico) incrementa la probabilidad de ocurrencia de la morosidad porque es mayor que 1.

Odds Ratio (Exp(B))

Los valores de Exp(B), son consideradas en la evaluación dentro del modelo. Las variables Destino del Préstamo y Actividad del Cliente están relacionadas de manera positiva con la ocurrencia de la morosidad; mientras que las variables Antigüedad del Negocio, Pasivo Financiero, Número de Cuotas y Sexo del Cliente tienen una relación negativa con la morosidad y por último la variable Capital del Préstamo no tiene relación negativa ni positiva con la ocurrencia.

- El $OR = 0{,}811 < 1$ de la Antigüedad del Negocio, representa la disminución de riesgo por unidad de variación, un cliente cuyo negocio tiene 5 años de antigüedad tiene más riesgo de presentar morosidad que una persona que cuyo negocio tiene 20 años de antigüedad.
- El $OR = 0{,}026 < 1$ del Pasivo Financiero, representa la disminución de riesgo por unidad de variación, un cliente que tiene pasivo financiero tiene más riesgo de presentar morosidad que una persona no tiene pasivo financiero.
- El $OR = 1{,}000$ del Capital del Préstamo, significa que dicha variable no influye en la ocurrencia de la morosidad.
- El $OR = 20{,}831 > 1$ indica que los clientes cuyo Destino del Préstamo es la adquisición de un bien mueble, tiene 20 veces más de riesgo de presentar morosidad que los clientes que presenten otro tipo de destino del préstamo.
- El $OR = 0{,}796 < 1$ del Número de Cuotas, representa la disminución de riesgo por unidad de variación, un cliente que tiene una cantidad de cuotas menor tiene más riesgo de presentar morosidad que un cliente cuyo número de cuotas es mayor.

- El $OR = 0{,}104 < 1$ indica la disminución del riesgo por unidad de variación, un cliente es de sexo masculino tiene mayor probabilidad de presentar morosidad que un cliente de sexo femenino.
- El $OR = 41{,}973 > 1$ indica que los clientes cuya Actividad económica es Empleado Técnico, tiene 41 veces más de riesgo de presentar morosidad que los clientes que presenten otro tipo de Actividad económica.

Interpretación de Tablas para el Modelo Seleccionado:

En este ítem, con el método "Backward" confirmamos los datos seleccionados mediante el método "Forward".

Tabla 39:
Resumen de Procesamiento de Casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	330	100,0
	Casos perdidos	0	0,0
	Total	330	100,0
Casos no seleccionados		0	0,0
	Total	**330**	**100,0**

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos.

En esta Tabla podemos observar la cantidad de clientes que entran en el estudio de análisis, tenemos 330 datos incluidos, ningún caso perdido.

Tabla 40:
Codificación de Variable Dependiente

Valor original	Valor interno
No	0
Si	1

En la Tabla 40. podemos observar la codificación de la variable dependiente Riesgo Crediticio (Morosidad).

Tabla 41:
Codificaciones de Variables Categóricas

			Codificación de parámetro			
		Frecuencia	**(1)**	**(2)**	**(3)**	**(4)**
Actividad del	1	100	1,000	,000	,000	,000
Cliente	2	71	,000	1,000	,000	,000
	3	66	,000	,000	1,000	,000
	4	27	,000	,000	,000	1,000
	5	66	,000	,000	,000	,000
Destino del	1	106	1,000	,000	,000	
Préstamo	2	83	,000	1,000	,000	
	3	68	,000	,000	1,000	
	4	73	,000	,000	,000	
Pasivo	No	193	1,000			
Financiero	Si	137	,000			
Antecedentes en	No	130	1,000			
Clearing	Si	200	,000			
Sexo del Cliente	Femenino	166	1,000			
	Masculino	164	,000			

En la Tabla 41. se observa la codificación de las categorías de las variables independientes (Categóricas); que codifica con 1 para la presencia y con 0 para la ausencia de la variable predictora.

Bloque 0, Bloque inicial:

Tabla 42:
Historial de Iteraciones[a,b,c]

Iteración		**Logaritmo de la verosimilitud -2**	**Coeficientes Constante**
Paso 0	1	403,326	-0,800
	2	403,170	-0,847
	3	403,170	-0,847

a. La constante se incluye en el modelo.
b. Logaritmo de la verosimilitud -2 inicial: 403.170
c. La estimación ha terminado en el número de iteración 3 porque las estimaciones de parámetro han cambiado en menos de .001.

En este Bloque inicial se calcula la verosimilited de un modelo que solo tiene el término constante β_0.

El estadístico $(-2lnL)$ mide hasta que punto un modelo se ajusta bien a los datos. Cuanto más pequeña sea el valor, mejor será el ajuste enconces podemos decir que la constante se incluye en el modelo.

Para el proceso iterativo de estimación del primer parámetro β_0, se ha necesitado tres ciclos para estimar correctamente el término de la constante $\beta_0 = -0{,}847$, porque la variación de $(-2lnL)$ entre los tres bucles ha cambiado en menos del criterio fijado por el programa en $(0{,}001)$.

Tabla 43:
Tabla de Clasificacióna,b

			Pronosticado		
			Morosidad del Cliente		Corrección de porcentaje
	Observado		No	Si	
Paso 0	Morosidad del Cliente	No	231	0	100,0
		Si	99	0	0,0
	Porcentaje global				70,0

a. La constante se incluye en el modelo.
b. El valor de corte es 0,500

En este primer paso el modelo ha clasificado correctamente a un 70,0% de los casos y ningún cliente con morosidad ha sido clasificado correctamente.

Tabla 44:
Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-0,847	0,120	49,751	1	0,000	0,429

La Tabla 44. muestra en este primer bloque, que en la ecuación de regresión solo aparece el parámetro estimado $\beta_0 = -0{,}847$ el error estándar $E.T. = 0{,}120$ y la significancia estadística con la Prueba de Wald que es un estadístico que sigue una distribución Chi – Cuadrada con 1 grado de libertad y la estimación de la $OR = e^{\beta_0} = e^{-0{,}847} = 0{,}429$.

Tabla 45:
Las Variables que No están en la Ecuación

		Puntuación	gl	Sig.
Variables	Edad	97,237	1	,000
	Sexo(1)	30,006	1	,000
	Actividad	20,028	4	,000
	Actividad(1)	2,460	1	,117
	Actividad(2)	,931	1	,335
	Actividad(3)	2,078	1	,149
	Actividad(4)	15,215	1	,000
	Antiguedad	145,607	1	,000
	Destinoprestamo	126,760	3	,000
	Destinoprestamo(1)	34,404	1	,000
	Destinoprestamo(2)	12,756	1	,000
	Destinoprestamo(3)	118,154	1	,000
	Antecedente(1)	1,511	1	,219
	Pasivofinan(1)	85,368	1	,000
	Capital	2,299	1	,129
	Numcuota	58,980	1	,000

a. Los chi-cuadrados residuales no se calculan debido a redundancias.

En la Tabla 45. podemos observar la Puntuación Eficiente de Rao y la significación estadística que está asociada a la Prueba de Wald.

Bloque 1: Método = Retroceder por paso (razón de verosimilitud)

Tabla 46:
Historial de Iteraciones[a,b,c,d]

			Coeficientes														
	Iteración	Logaritmo de la verosimilitud -2	Constante	Edad	Sexo(1)	Actividad(1)	Actividad(2)	Actividad(3)	Actividad(4)	Antigüedad	Destinoprestamo(1)	Destinoprestamo(2)	Destinoprestamo(3)	Antecedente(1)	Pasivofinan(1)	Capital	Numcuota
Paso 1	1	169,730	,483	,004	-,448	,254	,120	,109	,693	-,077	-,226	-,141	1,155	-,089	-1,144	,000	-,033
	2	114,378	,729	,017	-,861	,492	,176	,263	1,306	-,130	-,434	-,188	1,686	-,185	-1,895	,000	-,072
	3	92,635	1,152	,028	-1,330	,741	,202	,385	2,054	-,179	-,664	-,161	2,119	-,283	-2,584	,000	-,124
	4	84,896	1,729	,035	-1,816	,953	,151	,374	2,841	-,220	-,867	-,096	2,532	-,373	-3,186	,000	-,179
	5	83,221	2,155	,040	-2,162	1,084	,035	,292	3,387	-,249	-,971	-,028	2,873	-,462	-3,594	,000	-,219
	6	83,107	2,300	,042	-2,279	1,124	-,033	,254	3,565	-,259	-,995	-,002	3,008	-,506	-3,736	,000	-,233
	7	83,106	2,313	,042	-2,289	1,128	-,042	,251	3,580	-,260	-,996	,000	3,021	-,511	-3,749	,000	-,234
	8	83,106	2,313	,042	-2,289	1,128	-,042	,251	3,580	-,260	-,996	,000	3,021	-,511	-3,749	,000	-,234
Paso 2	1	169,810	,602		-,451	,256	,123	,111	,697	-,072	-,229	-,140	1,156	-,091	-1,146	,000	-,033
	2	114,685	1,180		-,882	,504	,189	,266	1,329	-,111	-,450	-,186	1,703	-,186	-1,904	,000	-,072
	3	93,017	1,889		-1,369	,775	,232	,389	2,107	-,147	-,693	-,156	2,173	-,275	-2,596	,000	-,125
	4	85,326	2,629		-1,854	1,019	,203	,393	2,915	-,179	-,892	-,074	2,625	-,358	-3,191	,000	-,181
	5	83,683	3,155		-2,194	1,174	,112	,334	3,471	-,202	-,991	,015	2,991	-,442	-3,588	,000	-,222
	6	83,573	3,335		-2,307	1,225	,058	,307	3,653	-,209	-1,014	,051	3,132	-,483	-3,724	,000	-,235
	7	83,573	3,350		-2,316	1,229	,052	,305	3,669	-,210	-1,015	,054	3,145	-,488	-3,735	,000	-,237
	8	83,573	3,350		-2,316	1,229	,052	,305	3,669	-,210	-1,015	,054	3,145	-,488	-3,735	,000	-,237
Paso 3	1	170,012	,545		-,448	,268	,144	,115	,702	-,072	-,241	-,154	1,154		-1,139	,000	-,032
	2	115,010	1,052		-,873	,532	,231	,271	1,348	-,112	-,473	-,211	1,685		-1,892	,000	-,070
	3	93,397	1,683		-1,352	,831	,297	,390	2,152	-,147	-,726	-,186	2,128		-2,573	,000	-,122
	4	85,799	2,329		-1,827	1,108	,303	,391	2,978	-,180	-,932	-,105	2,559		-3,146	,000	-,178
	5	84,242	2,744		-2,152	1,297	,260	,334	3,544	-,202	-1,036	-,022	2,903		-3,516	,000	-,216
	6	84,149	2,861		-2,252	1,362	,236	,311	3,723	-,209	-1,061	,011	3,027		-3,633	,000	-,228
	7	84,149	2,869		-2,259	1,367	,234	,309	3,737	-,210	-1,062	,013	3,036		-3,641	,000	-,229
	8	84,149	2,869		-2,259	1,367	,234	,309	3,737	-,210	-1,062	,013	3,036		-3,641	,000	-,229

a. Método: Retroceder por paso (Razón de verosimilitud)
b. La constante se incluye en el modelo.
c. Logaritmo de la verosimilitud -2 inicial: 403,170
d. La estimación ha terminado en el número de iteración 8 porque las estimaciones de parámetro han cambiado en menos de ,001.

En la Tabla 46. se muestra el proceso de iteración, que se realiza para siete coeficientes, la constante (ya incluida en el paso anterior), las variables Edad, Sexo, Actividad del Cliente, Antigüedad del Negocio, Destino del Préstamo, Antecedentes en el Clearing, Pasivo Financiero, Capital de Préstamo y Número de Cuotas

Se observa como disminuye el estadístico $(-2lnL)$ respecto al paso anterior el modelo solo con la constante tenía un valor de $(-2lnL) = 403{,}170$, mientras que ahora se reduce a 84,149 y el proceso termina en 3 bubles o ciclos.

Tabla 47:
Pruebas Ómnibus de Coeficientes de Modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	320,065	14	,000
	Bloque	320,065	14	,000
	Modelo	320,065	14	,000
Paso 2[a]	Escalón	-,467	1	,494
	Bloque	319,598	13	,000
	Modelo	319,598	13	,000
Paso 3[a]	Escalón	-,576	1	,448
	Bloque	319,022	12	,000
	Modelo	319,022	12	,000

a. Un valor negativo de chi-cuadrados indica que el valor de chi-cuadrados ha disminuido del paso anterior.

La Tabla 47. de prueba global o prueba de ómnibus de coeficientes del modelo, muestra el estadístico Chi – Cuadrado que evalúa la hipótesis nula de que los coeficientes β_k de todos los términos (excepto la constante) incluidos en el modelo son ceros.

Es estadístico Chi – Cuadrado para este contraste es la diferencia entre el valor de $(-2lnL)$.

$$D = RV = \chi^2 = (-2lnL_{Modelo0}) - (-2lnL_{Modelo3}) = 403{,}170 - 84{,}149 = 319{,}021$$

Entonces la Devianza o Razón de Verosimilitud sirve para evaluar si las variables tomadas en conjunto, contribuyen efectivamente a explicar las modificaciones que se producen en la probabilidad $P(Y = 1)$, y en nuestro caso, estas variables son significativas.

Tabla 48:
Resumen del Modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	83,106[a]	,621	,880
2	83,573[a]	,620	,880
3	84,149[a]	,620	,879

a. La estimación ha terminado en el número de iteración 8 porque las estimaciones de parámetro han cambiado en menos de ,001.

En la Tabla 48. observamos 3 medidas resumen de los modelos, para evaluar de forma global su validez. Los coeficientes Mc Fadden, Cox y Snell y Nagelkerke tienen los valores de 79,1%; 62,0% y 87,9% respectivamente, e indican una excelente calidad de ajuste del modelo, por estar sobre los valores del intervalo $0 \leq R^2 \leq 1$ y por encima del intervalo $0{,}2 \leq R^2 \leq 0{,}4$.

Tabla 49:
Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	10,897	8	,208
2	10,010	8	,264
3	9,608	8	,294

Hosmer y Lemeshow muestra que existe un buen ajuste global, ya que se obtiene significación estadística cuando el $p - valor = 0{,}294 > 0{,}05$.

Tabla 50:
Tabla de Contingencia para la Prueba de Hosmer y Lemeshow

		Morosidad del Cliente = No		Morosidad del Cliente = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	33	32,999	0	,001	33
	2	33	32,993	0	,007	33
	3	33	32,973	0	,027	33
	4	32	32,910	1	,090	33
	5	33	32,698	0	,302	33
	6	32	31,878	1	1,122	33
	7	28	26,075	5	6,925	33
	8	7	8,069	26	24,931	33
	9	0	,392	33	32,608	33
	10	0	,013	33	32,987	33
Paso 2	1	33	32,999	0	,001	33
	2	33	32,993	0	,007	33
	3	33	32,972	0	,028	33
	4	32	32,902	1	,098	33
	5	33	32,679	0	,321	33
	6	32	31,891	1	1,109	33
	7	28	26,019	5	6,981	33
	8	7	8,131	26	24,869	33
	9	0	,400	33	32,600	33
	10	0	,014	33	32,986	33
Paso 3	1	33	32,999	0	,001	33
	2	33	32,993	0	,007	33
	3	33	32,971	0	,029	33
	4	32	32,898	1	,102	33
	5	33	32,683	0	,317	33

6	32	31,845	1	1,155	33
7	28	26,141	5	6,859	33
8	7	7,965	26	25,035	33
9	0	,487	33	32,513	33
10	0	,018	33	32,982	33

La Bondad de Ajuste resultó excelente, basta notar la similitud entre valores esperados y observados en el procedimiento de Hosmer y Lemeshow.

Tabla 51:
Tabla de Clasificacióna

			Pronosticado		
			Morosidad del Cliente		
	Observado		No	Si	Corrección de porcentaje
Paso 1	Morosidad del Cliente	No	225	6	97,4
		Si	7	92	92,9
	Porcentaje global				96,1
Paso 2	Morosidad del Cliente	No	226	5	97,8
		Si	7	92	92,9
	Porcentaje global				96,4
Paso 3	Morosidad del Cliente	No	224	7	97,0
		Si	8	91	91,9
	Porcentaje global				95,5

a. El valor de corte es ,500

La morosidad observada $8 + 91$ y no morosidad $224 + 7$ (paso 3); pero estas cantidades fueron clasificados por los valores predictivos; es decir, el modelo matemático clasificó a los grupos sin morosidad y con morosidad pero de todos los clientes con morosidad solamente a 91 los clasificó como clientes morosos y a los 8 los clasificó como clientes no morosos por eso la sensibilidad es de 91,9% es decir $\frac{91}{91+8}$ y de los 231 no morosos a 224 los clasificó como no morosos, es decir, al 97,0% y esta es la especificidad del modelo matemático para predecir el diagnóstico de un cliente moroso.

Tabla 52*:*
Matriz de Correlaciones

		Constante	Edad	Sexo(1)	Actividad(1)	Actividad(2)	Actividad(3)	Actividad(4)	Antiguedad	Destinoprestamo(1)	Destinoprestamo(2)	Destinoprestamo(3)	Antecedente(1)	Pasivofinan(1)	Capital	Numcuota
Paso 1	Constante	1,000	-,736	-,281	-,113	-,070	,002	,033	,491	-,148	-,039	,080	-,302	-,118	,123	-,327
	Edad	-,736	1,000	,016	-,145	-,104	-,083	-,079	-,863	,025	-,093	-,157	-,046	-,064	-,014	,025
	Sexo(1)	-,281	,016	1,000	-,134	,036	,016	-,146	,144	,080	,077	-,154	,161	,367	-,423	,339
	Actividad(1)	-,113	-,145	-,134	1,000	,543	,528	,488	,108	-,250	,042	-,006	,198	-,064	,117	-,197
	Actividad(2)	-,070	-,104	,036	,543	1,000	,558	,283	,142	-,354	-,125	-,217	,203	-,063	-,222	,076
	Actividad(3)	,002	-,083	,016	,528	,558	1,000	,318	,102	-,277	-,259	-,233	-,001	-,109	-,225	,029
	Actividad(4)	,033	-,079	-,146	,488	,283	,318	1,000	-,109	,009	,180	,263	,038	-,169	,258	-,345
	Antigüedad	,491	-,863	,144	,108	,142	,102	-,109	1,000	-,078	-,014	,005	,069	,251	-,205	,101
	DestPrest.(1)	-,148	,025	,080	-,250	-,354	-,277	,009	-,078	1,000	,480	,444	-,055	,125	,001	,064
	DestPrest. (2)	-,039	-,093	,077	,042	-,125	-,259	,180	-,014	,480	1,000	,528	-,031	,063	,030	-,103
	DestPrest. (3)	,080	-,157	-,154	-,006	-,217	-,233	,263	,005	,444	,528	1,000	-,186	-,132	,367	-,315
	Antecedent(1)	-,302	-,046	,161	,198	,203	-,001	,038	,069	-,055	-,031	-,186	1,000	,215	-,134	,223
	Pasivfinan(1)	-,118	-,064	,367	-,064	-,063	-,109	-,169	,251	,125	,063	-,132	,215	1,000	-,492	,158
	Capital	,123	-,014	-,423	,117	-,222	-,225	,258	-,205	,001	,030	,367	-,134	-,492	1,000	-,703
	Numcuota	-,327	,025	,339	-,197	,076	,029	-,345	,101	,064	-,103	-,315	,223	,158	-,703	1,000
Paso 2	Constante	1,000		-,391	-,329	-,219	-,088	-,050	-,416	-,195	-,146	-,057	-,488	-,241	,161	-,452
	Sexo(1)	-,391		1,000	-,135	,029	,018	-,141	,304	,081	,065	-,160	,152	,364	-,409	,331
	Actividad(1)	-,329		-,135	1,000	,542	,528	,492	-,044	-,236	,051	,003	,183	-,082	,128	-,211
	Actividad(2)	-,219		,029	,542	1,000	,556	,284	,100	-,340	-,126	-,220	,195	-,080	-,210	,060
	Actividad(3)	-,088		,018	,528	,556	1,000	,324	,048	-,262	-,246	-,225	-,003	-,111	-,216	,007
	Actividad(4)	-,050		-,141	,492	,284	,324	1,000	-,355	,018	,184	,270	,031	-,177	,264	-,341
	Antigüedad	-,416		,304	-,044	,100	,048	-,355	1,000	-,109	-,191	-,274	,060	,384	-,436	,244
	DestPrest.(1)	-,195		,081	-,236	-,340	-,262	,018	-,109	1,000	,486	,454	-,059	,135	-,006	,062
	DestPrest. (2)	-,146		,065	,051	-,126	-,246	,184	-,191	,486	1,000	,530	-,061	,052	,040	-,127
	DestPrest. (3)	-,057		-,160	,003	-,220	-,225	,270	-,274	,454	,530	1,000	-,194	-,143	,373	-,324
	Antecedent(1)	-,488		,152	,183	,195	-,003	,031	,060	-,059	-,061	-,194	1,000	,208	-,129	,222
	Pasivfinan(1)	-,241		,364	-,082	-,080	-,111	-,177	,384	,135	,052	-,143	,208	1,000	-,493	,163
	Capital	,161		-,409	,128	-,210	-,216	,264	-,436	-,006	,040	,373	-,129	-,493	1,000	-,699
	Numcuota	-,452		,331	-,211	,060	,007	-,341	,244	,062	-,127	-,324	,222	,163	-,699	1,000
Paso 3	Constante	1,000		-,347	-,285	-,151	-,097	-,058	-,413	-,276	-,219	-,203		-,152	,077	-,386
	Sexo(1)	-,347		1,000	-,174	-,008	,022	-,145	,271	,101	,076	-,127		,329	-,385	,306
	Actividad(1)	-,285		-,174	1,000	,532	,538	,501	-,072	-,223	,063	,045		-,125	,169	-,270
	Actividad(2)	-,151		-,008	,532	1,000	,566	,297	,087	-,332	-,125	-,172		-,125	-,163	-,005
	Actividad(3)	-,097		,022	,538	,566	1,000	,333	,033	-,259	-,238	-,229		-,091	-,213	-,003
	Actividad(4)	-,058		-,145	,501	,297	,333	1,000	-,366	,024	,183	,286		-,179	,270	-,355
	Antigüedad	-,413		,271	-,072	,087	,033	-,366	1,000	-,111	-,198	-,277		,372	-,426	,233
	DestPrest.(1)	-,276		,101	-,223	-,332	-,259	,024	-,111	1,000	,496	,467		,152	-,004	,073
	DestPrest. (2)	-,219		,076	,063	-,125	-,238	,183	-,198	,496	1,000	,544		,065	,033	-,110
	DestPrest. (3)	-,203		-,127	,045	-,172	-,229	,286	-,277	,467	,544	1,000		-,095	,353	-,287
	Pasivfinan(1)	-,152		,329	-,125	-,125	-,091	-,179	,372	,152	,065	-,095		1,000	-,485	,121
	Capital	,077		-,385	,169	-,163	-,213	,270	-,426	-,004	,033	,353		-,485	1,000	-,684
	Numcuota	-,386		,306	-,270	-,005	-,003	-,355	,233	,073	-,110	-,287		,121	-,684	1,000

La Tabla 52, Matriz de Correlaciones, nos indica que las correlaciones entre las variables que están en el modelo no son significativas, por lo tanto no existe colinealidad o interacción entre las variables.

Tabla 53*:*
Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Edad	,042	,061	,474	1	,491	1,043	,925	1,175
	Sexo(1)	-2,289	,682	11,279	1	,001	,101	,027	,386
	Actividad			7,818	4	,098			
	Actividad(1)	1,128	,898	1,577	1	,209	3,088	,531	17,944
	Actividad(2)	-,042	1,268	,001	1	,974	,959	,080	11,515
	Actividad(3)	,251	1,083	,054	1	,817	1,285	,154	10,729
	Actividad(4)	3,580	1,368	6,848	1	,009	35,889	2,457	524,319
	Antiguedad	-,260	,085	9,327	1	,002	,771	,653	,911
	Destinoprestamo			18,746	3	,000			
	Destinoprestamo(1)	-,996	,936	1,131	1	,288	,369	,059	2,314
	Destinoprestamo(2)	,000	,880	,000	1	1,000	1,000	,178	5,616
	Destinoprestamo(3)	3,021	,959	9,916	1	,002	20,511	3,129	134,461
	Antecedente(1)	-,511	,656	,608	1	,436	,600	,166	2,168
	Pasivofinan(1)	-3,749	,833	20,240	1	,000	,024	,005	,121
	Capital	,000	,000	24,857	1	,000	1,000	1,000	1,000
	Numcuota	-,234	,059	15,775	1	,000	,791	,705	,888
	Constante	2,313	2,025	1,305	1	,253	10,109		
Paso 2[a]	Sexo(1)	-2,316	,676	11,724	1	,001	,099	,026	,371
	Actividad			8,116	4	,087			
	Actividad(1)	1,229	,891	1,904	1	,168	3,419	,596	19,611
	Actividad(2)	,052	1,247	,002	1	,967	1,053	,091	12,121
	Actividad(3)	,305	1,079	,080	1	,778	1,356	,164	11,237
	Actividad(4)	3,669	1,374	7,130	1	,008	39,198	2,654	579,029
	Antiguedad	-,210	,043	24,335	1	,000	,811	,746	,881
	Destinoprestamo			20,848	3	,000			
	Destinoprestamo(1)	-1,015	,919	1,219	1	,270	,362	,060	2,196
	Destinoprestamo(2)	,054	,883	,004	1	,951	1,056	,187	5,958
	Destinoprestamo(3)	3,145	,947	11,019	1	,001	23,220	3,626	148,705
	Antecedente(1)	-,488	,649	,566	1	,452	,614	,172	2,189
	Pasivofinan(1)	-3,735	,822	20,652	1	,000	,024	,005	,120
	Capital	,000	,000	26,000	1	,000	1,000	1,000	1,000
	Numcuota	-,237	,059	15,930	1	,000	,789	,703	,887
	Constante	3,350	1,361	6,057	1	,014	28,509		
Paso 3[a]	Sexo(1)	-2,259	,665	11,542	1	,001	,104	,028	,385
	Actividad			8,499	4	,075			
	Actividad(1)	1,367	,882	2,404	1	,121	3,923	,697	22,082
	Actividad(2)	,234	1,192	,039	1	,844	1,264	,122	13,058
	Actividad(3)	,309	1,067	,084	1	,772	1,363	,168	11,026
	Actividad(4)	3,737	1,370	7,437	1	,006	41,973	2,861	615,713
	Antiguedad	-,210	,042	24,839	1	,000	,811	,746	,880
	Destinoprestamo			21,226	3	,000			
	Destinoprestamo(1)	-1,062	,925	1,320	1	,251	,346	,056	2,117
	Destinoprestamo(2)	,013	,884	,000	1	,988	1,013	,179	5,735
	Destinoprestamo(3)	3,036	,929	10,674	1	,001	20,831	3,370	128,767
	Pasivofinan(1)	-3,641	,794	21,040	1	,000	,026	,006	,124
	Capital	,000	,000	26,392	1	,000	1,000	1,000	1,000
	Numcuota	-,229	,057	16,284	1	,000	,796	,712	,889
	Constante	2,869	1,182	5,896	1	,015	17,620		

a. Variables especificadas en el paso 1: Edad, Sexo, Actividad, Antiguedad, Destinoprestamo, Antecedente, Pasivofinan, Capital, Numcuota.

Para la predicción del modelo tomaremos en cuenta los resultados del paso 3 indicados en la Tabla 53. Variables en la Ecuación.

5.1.4. Predicción del Modelo

Para mostrar la predicción del modelo inicialmente hallamos la estimación como se indica en el ítem 5.1.3. y ésta es:

$$\hat{\boldsymbol{L}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{AntNeg} + \hat{\beta}_2 \boldsymbol{PasFin} + \hat{\beta}_3 \boldsymbol{CapPres} + \hat{\beta}_4 \boldsymbol{DestCred_3} + \hat{\beta}_5 \boldsymbol{NumCuo} + \hat{\beta}_6 \boldsymbol{Sexo} + \hat{\beta}_7 \boldsymbol{ActClie_4}$$

De probabilidad p,

$$p = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{AntNeg} + \hat{\beta}_2 \boldsymbol{PasFin} + \hat{\beta}_3 \boldsymbol{CapPres} + \hat{\beta}_4 \boldsymbol{DestCred_3} + \hat{\beta}_5 \boldsymbol{NumCuo} + \hat{\beta}_6 \boldsymbol{Sexo} + \hat{\beta}_7 \boldsymbol{ActClie_4})}}$$

$$p = \frac{1}{1 + e^{-(2{,}869 - 0{,}210 \boldsymbol{AntNeg} - 3{,}641 \boldsymbol{PasFin} + 0{,}000 \boldsymbol{CapPres} + 3{,}036 \boldsymbol{DestCred_3} - 0{,}229 \boldsymbol{NumCuo} - 2{,}259 \boldsymbol{Sexo} + 3{,}737 \boldsymbol{ActClie_4})}}$$

donde:

Variables	$X_1 = \boldsymbol{AntNeg}$ Antigüedad del Negocio		$X_2 = \boldsymbol{PasFin}$ Pasivo Financiero		$X_3 = \boldsymbol{CapPres}$ Capital del Préstamo		$X_4 = \boldsymbol{DestCred_3}$ Destino del Crédito (Compra de bien mueble)	
Valores	2	30	$Si = 1$	$No = 0$	5000	40000	$Si = 1$	$No = 0$
Variables	$X_5 = \boldsymbol{NumCuo}$ Número de Cuotas		$X_6 = \boldsymbol{Sexo}$ Sexo del Cliente		$X_7 = \boldsymbol{ActClie_4}$ Actividad del Cliente (Empleado Técnico)			
valores	12	48	$Mas = 1$	$Fem = 0$	$Si = 1$	$No = 0$		

$\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\boldsymbol{\beta}}_4, \hat{\boldsymbol{\beta}}_5, \hat{\boldsymbol{\beta}}_6, \hat{\boldsymbol{\beta}}_7$, son los parámetros estimados cuyos valores son:

Valores				
	$\hat{\beta}_0 = 2{,}869$	$\hat{\beta}_1 = -0{,}210$	$\hat{\beta}_2 = -3{,}641$	$\hat{\beta}_3 = 0{,}000$
	$\hat{\beta}_4 = 3{,}036$	$\hat{\beta}_5 = -0{,}229$	$\hat{\beta}_6 = -2{,}259$	$\hat{\beta}_7 = 3{,}737$

Como se puede observar con el método “Backward” se obtuvo el mismo modelo que con el método “Forward”, con la única diferencia que con el método Forward se tuvieron 7 pasos y con el método Backward únicamente 3 pasos.

5.1.5. Medidas de Asociación

Ahora analizaremos la asociación entre la variable respuesta, Riesgo Crediticio (Morosidad) y las variables independientes cuyos datos, son categóricos, tanto dicotómicas como politómicas. Las medidas de asociación utilizadas son Odds Ratio, Q de Yule y Chi-Cuadrado, y sus respectivos cálculos se detallan en Medidas de Asociación para variables categóricas del **Anexo C.**

5.1.5.1. Variables Dicotómicas:

i. Asociación mediante OR:

Tabla 54:
Asociación mediante "OR" entre las Variables, Morosidad e Independientes (Dicotómicas)

Variables (respuesta & independiente)	OR	Intervalo de OR		Decisión	Tipo de Asociación
		LI	LS		
Morosidad & Sexo del Cliente	4,029	2,408	6,742	Existe asociación	Significativa Factor de Riesgo
Morosidad & Antecedentes en Clearing	0,741	0,460	1,196	No existe asociación	No significativa
Morosidad & Pasivo Financiero	11,782	6,638	20,911	Existe asociación	Significativa Factor de Riesgo

La Tabla muestra un resumen de valores de asociación mediante Odds Ratio, de la variable Morosidad con Sexo del Cliente, Morosidad con Antecedentes en el Clearing y Morosidad con Pasivo Financiero, que son las variables independientes categóricas, de las cuales podemos decir que:

- El $OR = 11{,}782 > 1$ indica que, si el cliente presenta pasivo financiero tiene 11 veces más riesgo de presentar morosidad, respecto de aquellos clientes que no tengan pasivo financiero, además el valor de 11,782, muestra un grado de asociación alta respecto a las demás variables, y se considera factor de riesgo.
- El $OR = 4.029 > 1$ indica que, si el cliente es de sexo masculino tienen 4 veces más riesgo de presentar morosidad, respecto de aquellos clientes que sean de sexo femenino,

además el valor de 4.029, muestra un grado de asociación, y se considera factor de riesgo.

- El $OR = 0{,}741 < 1$, indica el hecho de presentar antecedentes en el clearing no necesariamente incidirá en morosidad, además el valor 0,741 no muestra grado de asociación, y se considera no significativo.

Los cálculos se detallan en las tablas **C.1, C.2 y C.3 del Anexo C**.

ii. Asociación mediante Coeficiente "Q de Yule"

Tabla 55:
Asociación mediante "Q de Yule" entre la Variable Morosidad con Sexo de Cliente, Antecedentes en Clearing y Pasivo Financiero

Variable (Respuesta vs Independientes)	Q de Yule	Tipo de Asociación
Morosidad & Sexo del Cliente	0.6	Asociación Positiva
Morosidad & Antecedentes en Clearing	-0.149	Asociación Negativa
Morosidad & Pasivo Financiero	0.844	Asociación Positiva

La Tabla 55, muestra que la medida de asociación mediante el coeficiente Q de Yule, existe una asociación positiva entre la variable Morosidad y las variables Sexo del Cliente y Pasivo Financiero, mientras que entre la variable Morosidad y la variable Antecedentes en el Clearing no existe asociación.

iii. Asociación mediante Prueba Chi-Cuadrado:

Tabla 56*:*
Asociación mediante "Chi-Cuadrado" entre la Variable Morosidad y las Variables Independientes Dicotómicas

Variables (Respuesta vs Independientes)	Chi – Cuadrado χ_c^2	$\chi^2_{(\alpha,g.l.)}$	Decisión
Morosidad & Sexo del Cliente	30,006	3,84	Existe Relación
Morosidad & Antecedentes en Clearing	1,511	3,84	No Existe Relación
Morosidad & Pasivo Financiero	85,368	3,84	Existe Relación

En conclusión, al analizar la asociación entre la variable Morosidad y las Variables Independientes dicotómicas, de forma individual y utilizando las Medidas de Asociación

de OR, Q de Yule y Chi-Cuadrado; encontramos que Morosidad está relacionado con las dos variables que son Sexo del Cliente y Pasivo Financiero.

5.1.5.2. Variables Politómicas

Tabla 57:
Medidas de Asociación entre la Variable Morosidad y las Variables Independientes Politómicas

Variables (Respuesta vs Independientes)	Chi – Cuadrado χ_c^2	$\chi^2_{(\alpha,g.l.)}$	Decisión
Morosidad vs Actividad del Cliente	20,028	9,49	Existe Relación
Morosidad vs Destino del Préstamo	126,760	7,82	Existe Relación

En la Tabla 57 Apreciamos que existe relación entre las variable Morosidad y las variables independientes politómicas Actividad del Cliente y Destino del Préstamo.

Ver Anexo C, Medidas de Asociación de variables politómicas, Tablas 78. y 79.

VI. CONCLUSIONES

- El Modelo de Elección Binaria que mejor se ajusta al Riesgo Crediticio (Morosidad) en la Caja Municipal de Ahorro y Crédito Cusco, es el modelo de Regresión Logística, cuya ecuación esta dada por:

$$p = \frac{1}{1 + e^{-(2{,}869 - 0{,}210\boldsymbol{AntNeg} - 3{,}641\boldsymbol{PasFin} + 0{,}000\boldsymbol{CapPres} + 3{,}036\boldsymbol{DestPres_3} - 0{,}229\boldsymbol{NumCuo} - 2{,}259\boldsymbol{Sexo} + 3{,}737\boldsymbol{ActClie_4})}}$$

- Los factores asociados al Riesgo Crediticio (Morosidad) de la Caja Municipal de Ahorro y Crédito Cusco son: Antigüedad del Negocio, Pasivo Financiero, Capital del Préstamo, Destino del Préstamo, Número de Cuotas, Sexo del Cliente y Actividad del Cliente.
- Los factores que no están asociados al Riesgo Crediticio (Morosidad) de la Caja Municipal de Ahorro y Crédito Cusco son: Antecedentes en el Clearing y Edad del Cliente.
- Los métodos Forward y Backward establecieron que los factores Antigüedad del Negocio, Pasivo Financiero, Capital del Préstamo, Destino del Préstamo, Número de Cuotas, Sexo del Cliente y Actividad del Cliente están asociados y los factores Antecedentes en el Clearing y Edad del Cliente no están asociados al Riesgo Crediticio (Morosidad) de la Caja Municipal de Ahorro y Crédito Cusco.
- Según las medidas de asociación Odd Ratio, Q de Yule y Chi - Cuadrado los factores dicotómicos Sexo y Pasivo Financiero están asociados al riesgo crediticio en la Caja Municipal de Ahorro y Crédito Cusco además de ser considerados factores de riesgo, mientras que Antecedentes en Clearing no está asociado al riesgo crediticio en la Caja Municipal de Ahorro y Crédito Cusco.

.

VII. RECOMENDACIONES

- El análisis de Regresión Logística, es útil en estudios de variables dependientes binarias, sin embargo se recomienda un análisis más detallado de las variables independientes, ya sean cuantitativas o cualitativas, esto con el propósito de encontrar los factores que realmente influyen en la variable de estudio.
- Recomendamos a la oficina del Área de Estadística o la que haga sus veces de la Caja Municipal de Ahorro y Crédito Cusco, que maneja la información completa de los clientes, tener cuidado con los historiales crediticos de los clientes para evitar duplicidad de historiales de un mismo cliente y la falta de datos en algunos expedientes de crédito que dificulta la recolección de información de interés del investigador. Que pueden resolver mediante capacitaciones al personal encargado de llenar los expedientes de los historiales crediticios.
- Sugerimos tomar en cuenta este trabajo de investigación para estudios posteriores en los que participe un grupo multidisciplinario, con conocimiento especializado en temas de créditos, y consideren las variables Garantía presentada por el Cliente, Ingreso Mensual del Cliente, Carga Familiar, entre otros considerados también como posibles factores asociados. Todo esto acompañado de una encuesta que contenga la variable de interés.
- En vista de un alto porcentaje de morosidad, recomendamos a la Caja Municipal de Ahorro y Crédito Cusco y las demás entidades financieras propicien campañas permanentes para capacitación de su personal, en temas de evaluación de expedientes crediticios.
- Por último recomendamos el uso de la ecuación del modelo de Regresión Logística estimado:

$$p = \frac{1}{1 + e^{-(2{,}869 - 0{,}210\boldsymbol{AntNeg} - 3{,}641\boldsymbol{PasFin} + 0{,}000\boldsymbol{CapPres} + 3{,}036\boldsymbol{DestCred_3} - 0{,}229\boldsymbol{NumCuo} - 2{,}259\boldsymbol{Sexo} + 3{,}737\boldsymbol{ActClie_4})}}$$

Para identificar a un cliente con alto, medio o bajo riesgo crediticio, en la Caja Municipal de Ahorro y Crédito Cusco.

VIII. REFERENCIAS

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Gainesville, Florida: Wiley Interscience.

Ballón Beltran, D. D., & Bernabé Ponte, E. (2015). *Análisis Clasificatorio de las Gestantes según Vía de Culminación del Parto aplicando Regresión Logistica Binaria.* Lima.

Bedregal, J., & Barba, H. (2017). *Informe de Clasificación de Riesgo.* Lima.

Caja Municipal Cusco. (2017). *Memoria Anual 2016.* Cusco.

Camarero Rioja, L., Almazán Llorente, A., & Mañaz Ramirez, B. (2013). *Regresión Logística: Fundamentos y Aplicación en Investigación Sociológica.* Madrid: UNED.

Cea D'Ancona, M. A. (2004). *Análisis Multivariable. Teoría y Práctica en la Investigación Social.* Madrid: Síntesis S.A.

Contreras Vilca, N. (2012). *Análisis de Votos Electorales usando Modelos de Regresión para datos de conteo.* Lima.

Cruz Zuluaga, M. N. (2014). *Fundamentos de Estadística para las Ciencias Económico - Administrativas.* Medellín: Esumer.

Chitarroni, H. (2002). *La Regresión Logística.* Buenos Aires.

De la Fuente Fernández, S. (2011). Regresión Logística. 1 - 27.

Figueroa Arboccó, G. T. (2005). *La Fecundidad y su relacion con Variables Socioeconómicas, Demográficas y Educativas aplicando en Modelo de Regresion Poisson.* Lima.

Flores Manrique, L. (2002). *Análisis Estadístico de los Factores de Riesgo que influyen en la enfermedad de Angina de Pecho.* Lima.

Gonzales King-Keé, K. C. (2001). *Metodo de Mínimos Cuadrados Ponderados para la estimacion de los Modelos Lineales Generalizados.* Lima - Perú.

Hernandez Sampieri, R., Fernandez Collado , C., & Baptista Lucio, P. (2014). *Metodología de la Investigación.* México: Mc GRAW- HILL/ INTERAMERICANA EDITORES, S.A. DE C.V.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression.* Massachusetts: Columbus Ohio.

Iglesias Cabo, T. (2013). *Métodos de Bondad de Ajuste en Regresión Logística.* Granada.

Lopez Gonzales, E., & Ruiz Soler, M. (2011). Análisis de datos con el Modelo Lineal Ganeralizado. Una aplicacion con R. *Revista Española de Pedagogía*, 59 - 80.

Martín Martín, Q., Cabero Morán, T., & De Paz Santana, Y. (2008). *Tratamiento Estadístico de Datos con SPSS Prácticas Resueltas y Comentadas.* Madrid: Thomson Editores Spain.

Martínez Gómez, M., & Marí Benlloch, M. (2002). La Distribución Binomial. *Estadística, Investigación Operativa Aplicadas y Calidad*, 1 - 8.

Medina Moral, E. (2003). *Modelos de Elección Discreta.* Madrid.

Meza Saldaña, E., Reyes Cervantes, H., Pérez Salvador, B. R., & Tajonar Sanabria, F. S. (2006). Evaluación del Riesgo Crediticio, a través de Credit Scoring mediante Regresión logística: Un caso de estudio.

Moral Pelaez, I. (2006). Modelos de Regresipón: Lineal Simple y Regresión Logística. *Seden*, 195 - 214.

Moya C., R. (2004). *Probabilidad e Inferencia Estadística.* Jesus Maria: San Marcos.

Moya C., R., & Saravia A. , G. (2002). *Probabilidad e Inferencia Estadística.* Jesus Maria: San Marcos.

Moya C., R., & Saravia A., G. (2002). *Probabilidad e Inferencia Estadística.* Jesus Maria: San Marcos.

Peña, D. (2002). *Análisis de Datos Multivariantes.* Desconocida.

Rodas Guizado, E. (2011). *Factores Asociados de Riesgo para que una Persona Muera o sea Diagnosticada con el Virus H1N1 mediante el Modelo de Regresión Logística, en el Departamento del Cusco 2009.* Cusco.

Salcedo Poma, C. M. (2002). *Estimación de la ocurrencia en la incidencias en declaraciones de pólizas de importación* . Lima.

Salmerón Gomez, R. (s.f.). *El Modelo Lineal Uniecuacional Múltiple.*

Tomaiconza Ataulluco, B. F., & Pari Sallo, A. S. (2014). *Prediccion con Regresión Logística Aplicado a Partos Prematuros en el Hospital Regional de Cusco, 2010 - 2011.* Cusco.

Vásconez E., G. (2010). El Riesgo de Crédito en las Microfinanzas. *DGRV,* Asunción.

Vela Zavala, S., & Caro Anchay, A. (2010). *Herramientas Financieras en la Evaluación del Riesgo Crediticio.* Lima: Fondo Editorial de la Universidad Inca Garcilazo de la Vega.

Yampasi, M. G. (2017). *Factores que determinan el otorgamiento de crédito de la financiera Credinka en la ciudad de Ayaviri, 2015.* Puno.

IX. ANEXOS

Los anexos adjuntos al presente trabajo, están divididos en 2 partes, la primera parte consta de los números aleatorios usados para tomar datos de la muestra de la población en estudio y la matriz de consistencia y la segunda parte considera la corrida de datos en el programa SPSS (v. 22), así como sus respectivas interpretaciones.

ANEXO A

A.1. Números aleatorios

Tabla 58:
Números aleatorios comprendidos entre 1 y 950 para la selección de clientes que entran en la Muestra.

44	131	247	717	99	761	184	890	260	136	693	691	793	495	277
784	459	592	19	859	171	419	287	259	90	781	472	836	101	300
826	710	317	539	693	317	825	618	814	724	838	224	825	38	579
201	833	502	545	24	905	317	762	651	330	147	550	144	764	241
626	565	673	575	487	127	386	871	389	131	873	37	414	821	343
334	424	552	22	473	372	374	44	412	408	346	165	930	498	431
275	856	56	51	600	495	577	728	696	741	252	814	383	591	164
498	313	752	19	943	770	53	649	264	397	67	416	595	219	153
283	845	644	194	614	771	620	17	180	630	829	659	675	138	296
747	3	827	409	211	356	534	241	842	105	346	472	277	284	678
484	287	397	41	831	117	921	551	383	616	711	418	910	415	917
684	735	488	810	598	428	703	916	870	475	248	485	239	91	659
630	178	266	407	341	112	75	395	805	442	663	931	174	6	509
54	848	22	493	730	931	69	763	636	659	670	187	697	938	276
671	904	676	395	585	591	95	537	935	251	900	735	631	499	70
441	360	712	329	832	101	108	140	508	845	91	651	380	684	782
415	92	470	244	33	409	619	694	763	758	931	102	147	33	721
881	695	544	387	389	281	146	494	808	715	538	206	877	866	200
454	632	932	898	385	798	872	692	50	468	148	454	922	155	449
576	572	690	293	323	857	845	437	514	247	357	534	482	428	939
643	516	498	249	91	2	487	296	190	562	360	666	73	228	816
908	602	537	703	388	944	829	259	246	589	332	483	895	245	660

Tabla 59:
A.2. *Matriz de Consistencia*

PROBLEMAS	OBJETIVOS	HIPÓTESIS	METODOLOGÍA
Problema General ¿Qué modelos de elección binaria es el que mejor se ajusta al Riesgo Crediticio de la Caja Municipal de Ahorro y Crédito Cusco, año 2014? **Problema Específico** Qué factores están asociados al Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco, año 2014?	**Objetivo General** Determinar el modelo de elección binaria que mejor se ajusta al Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco, año 2014. **Objetivos Específicos** - Elaborar el marco teórico apropiado para la elección del modelo que mejor se ajusta al Riesgo Crediticio en la Caja Municipal de Ahorro y Crédito Cusco, año 2014. - Determinar los factores asociados al Riesgo Crediticio de la Caja Municipal de Ahorro y Crédito Cusco, año 2014, utilizando el modelo de elección binaria que mejor se ajusta.	**Hipótesis general** Existe un modelo de regresión binaria que se ajusta al riesgo crediticio de la Caja Municipal de Ahorro y Crédito Cusco. **Hipótesis Específica** Existen factores asociados al riesgo crediticio en la Caja Municipal de Ahorro y Crédito Cusco.	- Cuantitativa porque nuestro propósito es responder preguntas de investigación haciendo uso de técnicas estadísticas con el fin de cumplir los objetivos de estudio. - Descriptiva porque en el presente trabajo de investigación se han seleccionado 9 variables que nos permitirá predecir la existencia del riesgo crediticio. - Correlacional porque se analiza el grado de asociación entre la variable dependiente y las variables independientes. - Explicativa porque a partir de las medidas de asociación (Odds Ratio, Q de Yule y Chi-Cuadrado) explicamos la existencia o no de asociación entre variable dependiente y las variables independientes. - Transversal porque en el presente trabajo de investigación se analiza datos de variables recopiladas en un periodo de tiempo (año 2014).

ANEXO B

B.1. Selección de Variables, según Método "HACIA FORWARD"

Tabla 60:
Resumen de Procesamiento de Casos

Casos sin ponderar[a]		N	Porcentaje
Casos seleccionados	Incluido en el análisis	330	100,0
	Casos perdidos	0	0,0
	Total	**330**	**100,0**
Casos no seleccionados		0	0,0
	Total	**330**	**100,0**

a. Si la ponderación está en vigor, consulte la tabla de clasificación para el número total de casos.

La Tabla 60. muestra el resumen del número de datos introducidos, los que fueron seleccionados para el análisis y los excluídos (casos perdidos, que hayan tenido algún faltante); en nuestro caso 330 datos son considerados con un porcentaje del 100% y los casos perdidos son cero con una porcentaje del 0% y el total de datos a analizar son 330 que representa el 100%, los datos no seleccionados son cero.

Tabla 61:
Codificación de la Variable Dependiente

Valor original	Valor interno
No (0)	0
Si (1)	1

La Tabla 61. indica la codificación de la variable dependiente o respuesta, el valor interno codificado por el programa coincide con el valor externo que le asignamos; esto es, uno (1) para el resultado de nuestra evaluación, Riesgo Crediticio (Morosidad); y cero (0) para No Morosidad. Es importante mencionar la codificación, ya que el modelo clasificará o tratará de predecir, al cliente que presente morosidad.

Tabla 62:
Codificaciones de Variables Categóricas

			Codificación de parámetroS			
		Frecuencia	**(1)**	**(2)**	**(3)**	**(4)**
Actividad del Cliente	Indep.sin estudios Sup.	100	1,000	,000	,000	,000
	Indep.con estudiosSup.	71	,000	1,000	,000	,000
	Empleado profesional	66	,000	,000	1,000	,000
	Empleado técnico	27	,000	,000	,000	1,000
	Empleado con oficio	66	,000	,000	,000	,000
	Implement. de Negocio	106	1,000	,000	,000	
Destino del Préstamo	Adquis. bien inmueble	83	,000	1,000	,000	
	Adquis. bien mueble	68	,000	,000	1,000	
	Construcc. de vivienda	73	,000	,000	,000	
Pasivo Financiero	No	193	1,000			
	Si	137	,000			
Antecedentes en Clearing	No	130	1,000			
	Si	200	,000			
Sexo del Cliente	Femenino	166	1,000			
	Masculino	164	,000			

La Tabla 62. muestra la codificación que usa el SPSS para las variables independientes categóricas (cinco en nuestro estudio), de las cuales, 3 son dicotómicas y su esquema de codificación interna coincide con el extremo, y para el caso de variables con más de dos categorías, ejemplo para la variable, Actividad del Cliente verifica que tiene 5 categorías y en cada categoría el SPSS recodifica las variables dando valores 0 y 1 que funcionan de la siguiente manera, 1 indica la presencia del cliente en una categoría y 0 la no presencia, de este modo se han creado 4 nuevas variables. También podemos apreciar como se distribuye la frecuencia en cada categoría.

En el bloque inicial o bloque 0, el programa calcula la verosimilitud del modelo que inicialmente tiene solo el término constante β_0. Puesto que la verosimilitud 1, es un número muy pequeño (comprendido entre 0 y 1), se suele ofrecer el logaritmo natural de la

verosimilitud (LL), que es un número negativo, $(-2LL)$, o menos dos veces el logaritmo natural de la verosimilitud, que es un número positivo.

Bloque 0, Bloque inicial

Tabla 63*:*
Historial de Iteracionesa,b,c

Iteración		Logaritmo de la verosimilitud -2	Coeficientes Constante
Paso 0	1	403,326	-,800
	2	403,170	-,847
	3	403,170	-,847

a. La constante se incluye en el modelo.
b. Logaritmo de la verosimilitud -2 inicial: 403,170
c. La estimación ha terminado en el número de iteración 3 porque las estimaciones de parámetro han cambiado en menos de 0,001.

El estadístico $(-2lnL)$ mide hasta que punto un modelo se ajusta bien a los datos. El resultado de esta mediación recibe también el nombre de "desviación". Cuanto más pequeño sea el valor, mejor será el ajuste. Como en este primer paso solo se ha introducido el término constante en el modelo β_0. El proceso ha terminado en 3 iteraciones o cliclos para estimar correctamente el término constante, porque la variación de $(-2lnL)$ que se da entre el segundo y tercer ciclo, ha cambiado en menos del criterio fijado por el programa $(0{,}001)$. Lo importante es que también nos muestra el valor del parámetro calculado, que es de $\beta_0 = -0{,}847$.

Tabla 64:
Tabla de Clasificacióna,b

			Pronosticado		
			Morosidad del Cliente		
	Observado		No	Si	Corrección de porcentaje
Paso 0	Morosidad del Cliente	No	231	0	100,0
		Si	99	0	0,0
	Porcentaje global				70,0

a. La constante se incluye en el modelo.
b. El valor de corte es 0,500

La Tabla 64. de clasificación, permite evaluar el ajuste del modelo de regresión que hasta este momento tiene un solo parámetro en la ecuación, esto lo hace comparando los valores pronosticados con los valores observados. Por defecto se emplea un punto de corte, de 0,5 de la probabilidad de *Y* para clasificar a los clientes; es decir, que los clientes con una probabilidad $< 0{,}5$ se clasifican como cero (no presentan morosidad), mientras que si la probabilidad resulta $\geq 0{,}5$ se clasifican como 1 (si presenta morosidad).

En esta tabla se muestran los casos bien clasificados en la diagonal principal, y los casos mal clasificados en la segunda diagonal, por esto es que se dice que, de los $231 + 0$ clientes que NO presentan Morosidad, 231 fueron pronosticados como que no presentan Morosidad; es decir, hay un porcentaje de aciertos del $\frac{231}{231+0} = 100\%$.

De los $99 + 0$ clientes que SI presentan Morosidad, 0 fueron pronosticados como que presentan Morosidad; es decir, hay un porcentaje de aciertos del $\frac{0}{99+0} = 0\%$.

El porcentaje global de aciertos es del $\frac{231+0}{231+0+99+0} = \frac{231}{330} = 70{,}0\%$.

Por lo que la variable en la ecuación será solo la constante ingresada.

Tabla 65*:*
Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 0	Constante	-,847	,120	49,751	1	,000	,429

Tabla 66:
Las variables no están en la ecuacióna

			Punt. Eficient. Rao	gl	Sig.
Paso 0	Variables	Edad	97,237	1	,000
		Sexo(1)	30,006	1	,000
		Actividad	20,028	4	,000
		Actividad(1)	2,460	1	,117
		Actividad(2)	,931	1	,335
		Actividad(3)	2,078	1	,149
		Actividad(4)	15,215	1	,000

Antigüedad	145,607	1	,000
Destinoprestamo	126,760	3	,000
Destinoprestamo(1)	34,404	1	,000
Destinoprestamo(2)	12,756	1	,000
Destinoprestamo(3)	118,154	1	,000
Antecedente(1)	1,511	1	,219
Pasivofinan(1)	85,368	1	,000
Capital	2,299	1	,129
Numcuota	58,980	1	,000

a. Los chi-cuadrados residuales no se calculan debido a redundancias.

La Tabla 63. presenta el parámetro estimado (B), su error estándar y su significación estadística con la prueba de Wald, que es un estadístico que sigue la ley Chi - Cuadrado con 1 grado de libertad. Y la estimación de la OR $(Exp(B))$. Podemos observar que en este primer bloque (Paso 0), en la ecuación de regresión solo aparece la constante, quedando fuera las demás variables que se analizan, que son todas las que muestra la Tabla 65 y que no están en la ecuación, para seleccionar las variables se procede de la siguiente forma:

1. Se busca la variable candidata a entrar.
2. Se comprueba si dicha variable realmente debe estar en el modelo.

Para elegir a la variable candidata nos basamos en el p-valor asociado a la Puntuación Eficiente de Rao, si es menor que 0.05 (p-valor fijado para este caso) esa variable entra en el modelo, y si se tienen varias candidatas, entra la variable que tenga una mayor puntuación, que en nuestro caso será la variable Antigüedad del Negocio, la incorporada a la ecuación.

Bloque 1, Metodo – Por pasos Forward (Wald)

Ahora se analiza con una secuencia automática (por pasos); un segundo paso (bloque 1), con el método hacia adelante y teniendo en cuenta el criterio de la razón de verosimilitud (RV) para contrastar las nuevas variables a introducir o sacar del modelo.

Tabla 67:
Historial de Iteraciones[a,b,c,d,e,f,g]

Iteración		Logaritmo de la verosimilitud -2	Coeficientes												
			Constante	Antigüedad	Pasivofinan(1)	Capital	Destinoprestamo(1)	Destinoprestamo(2)	Destinoprestamo(3)	Numcuota	Sexo(1)	Actividad(1)	Actividad(2)	Actividad(3)	Actividad(4)
Paso 1	1	262.548	1.817	-.146											
	2	246.180	2.473	-.203											
	3	245.212	2.669	-.220											
	4	245.206	2.685	-.222											
	5	245.206	2.685	-.222											
Paso 2	1	229.758	2.121	-.123	-1.230										
	2	203.704	3.079	-.177	-1.901										
	3	200.511	3.545	-.203	-2.270										
	4	200.415	3.643	-.208	-2.353										
	5	200.415	3.647	-.208	-2.356										
	6	200.415	3.647	-.208	-2.356										
Paso 3	1	209.277	1.191	-.119	-1.568	.000									
	2	171.902	1.618	-.182	-2.604	.000									
	3	164.001	1.907	-.225	-3.381	.000									
	4	163.281	2.047	-.243	-3.713	.000									
	5	163.272	2.068	-.246	-3.758	.000									
	6	163.272	2.069	-.246	-3.759	.000									
Paso 4	1	186.914	.415	-.090	-1.345	.000	-.204	-.118	1.272						
	2	143.531	.559	-.141	-2.341	.000	-.363	-.144	1.944						
	3	131.854	.617	-.181	-3.227	.000	-.524	-.122	2.499						
	4	130.028	.668	-.204	-3.751	.000	-.650	-.109	2.801						
	5	129.956	.688	-.209	-3.881	.000	-.690	-.108	2.867						
	6	129.956	.689	-.210	-3.888	.000	-.693	-.108	2.870						
	7	129.956	.689	-.210	-3.888	.000	-.693	-.108	2.870						
Paso 5	1	177.742	.649	-.077	-1.155	.000	-.250	-.185	1.173	-.035					
	2	127.724	1.119	-.118	-1.919	.000	-.421	-.203	1.724	-.075					
	3	111.326	1.598	-.151	-2.581	.000	-.599	-.138	2.173	-.123					
	4	107.700	1.974	-.173	-3.047	.000	-.763	-.075	2.487	-.158					
	5	107.412	2.121	-.181	-3.227	.000	-.840	-.048	2.607	-.171					
	6	107.409	2.137	-.182	-3.247	.000	-.850	-.045	2.620	-.172					
	7	107.409	2.137	-.182	-3.247	.000	-.850	-.045	2.620	-.172					
Paso 6	1	173.141	.780	-.074	-1.190	.000	-.222	-.187	1.118	-.031	-.431				
	2	119.914	1.424	-.113	-1.998	.000	-.436	-.261	1.623	-.068	-.849				

	3	100.452	2.189	-.145	-2.737	.000	-.729	-.313	1.993	-.115	-1.321				
	4	94.669	2.963	-.170	-3.338	.000	-1.054	-.405	2.244	-.159	-1.767				
	5	93.824	3.423	-.184	-3.670	.000	-1.258	-.483	2.380	-.184	-2.026				
	6	93.799	3.523	-.187	-3.742	.000	-1.304	-.504	2.410	-.189	-2.082				
	7	93.798	3.527	-.187	-3.745	.000	-1.305	-.505	2.412	-.189	-2.084				
	8	93.798	3.527	-.187	-3.745	.000	-1.305	-.505	2.412	-.189	-2.084				
Paso 7	1	170.012	.545	-.072	-1.139	.000	-.241	-.154	1.154	-.032	-.448	.268	.144	.115	.702
	2	115.010	1.052	-.112	-1.892	.000	-.473	-.211	1.685	-.070	-.873	.532	.231	.271	1.348
	3	93.397	1.683	-.147	-2.573	.000	-.726	-.186	2.128	-.122	-1.352	.831	.297	.390	2.152
	4	85.799	2.329	-.180	-3.146	.000	-.932	-.105	2.559	-.178	-1.827	1.108	.303	.391	2.978
	5	84.242	2.744	-.202	-3.516	.000	-1.036	-.022	2.903	-.216	-2.152	1.297	.260	.334	3.544
	6	84.149	2.861	-.209	-3.633	.000	-1.061	.011	3.027	-.228	-2.252	1.362	.236	.311	3.723
	7	84.149	2.869	-.210	-3.641	.000	-1.062	.013	3.036	-.229	-2.259	1.367	.234	.309	3.737
	8	84.149	2.869	-.210	-3.641	.000	-1.062	.013	3.036	-.229	-2.259	1.367	.234	.309	3.737

a. Método: Avanzar por paso (Wald)

b. La constante se incluye en el modelo.

c. Logaritmo de la verosimilitud -2 inicial: 403,170

d. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de ,001.

e. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de ,001.

f. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de ,001.

g. La estimación ha terminado en el número de iteración 8 porque las estimaciones de parámetro han cambiado en menos de ,001.

En la Tabla 67. se muestra el proceso de iteración, que ahora en cada paso se realiza a partir de 2 coeficientes.

En el paso 1, se considera la constante, incluída en el paso anterior y la variable Antigüedad del Negocio; aquí apreciamos que, al ingresar esta variable el proceso concluye en 5 iteraciones, además vemos como disminuye el $(-2lnL)$ respecto al paso anterior, en el bloque 0, cuyo valor era de $403{,}170$ y en el paso 1 es de 245.206. Los coeficientes calculados para la constante $\beta_0 = 2.685$ y para la variable Antigüedad del Negocio es, $\beta_1 = -0.222$.

Tabla 68:
Pruebas ómnibus de coeficientes de modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	157,964	1	,000
	Bloque	157,964	1	,000
	Modelo	157,964	1	,000
Paso 2	Escalón	44,791	1	,000
	Bloque	202,755	2	,000
	Modelo	202,755	2	,000
Paso 3	Escalón	37,143	1	,000
	Bloque	239,899	3	,000
	Modelo	239,899	3	,000
Paso 4	Escalón	33,316	3	,000
	Bloque	273,215	6	,000
	Modelo	273,215	6	,000
Paso 5	Escalón	22,547	1	,000
	Bloque	295,761	7	,000
	Modelo	295,761	7	,000
Paso 6	Escalón	13,611	1	,000
	Bloque	309,372	8	,000
	Modelo	309,372	8	,000
Paso 7	Escalón	9,650	4	,047
	Bloque	319,022	12	,000
	Modelo	319,022	12	,000

La Tabla 68. de la Prueba Omnibus sobre los coeficientes del modelo, muestra 3 entradas que son Escalón, Bloque y Modelo, como se describe a continuación.

La entrada "Bloque", corresponde al cambio de verosimilitud $(-2lnL)$ entre pasos sucesivos en la construcción del modelo, donde se constrasta la Hipótesis Nula de que los coeficientes de las variables añadidas en el último paso son cero. Así tenemos:

Hipótesis

$$H_0: \beta_1 = 0,$$

$$H_1: \beta_1 \neq 0$$

Construcción del contraste

$$Chi - Cuadrado = (-2lnL_{modelo0}) - (-2lnL_{modelo1})$$

$$Chi - Cuadrado = (403{,}170) - (245{,}206)$$

$$Chi - Cuadrado = 157{,}964$$

Luego calculando el estadístico de prueba:

$$\chi_c^2 = 157{,}964 > \chi_{0.05,1}^2 = 3{,}841$$

Por regla de decisión aceptamos H_0, ya que tenemos que:

$$Chi - Cuadrado = 157{,}964 > \chi_{0.05,1}^2 = 3{,}841$$

Concluimos que el modelo tiene un buen ajuste si se introduce la variable Antigüedad del Negocio.

La entrada "Bloque", mide el cambio de verosimilitud $(-2lnL)$ entre bloques de entradas sucesivas durante la construcción del modelo. Si las variables se introducen en un solo bloque, el Chi – Cuadrado del Bloque es el mismo que el Chi – Cuadrado del Modelo.

Para el Paso 2 se tiene:

$$Chi - Cuadrado = (-2lnL_{modelo0}) - (-2lnL_{modelo2})$$

$$Chi - Cuadrado = (403{,}170) - (200{,}415)$$

$$Chi - Cuadrado = 202{,}755$$

Luego calculando el estadístico de prueba:

$$\chi_c^2 = 202{,}755 > \chi_{0.05,2}^2 = 5{,}991$$

Por regla de decisión aceptamos H_0, ya que tenemos que:

$$Chi - Cuadrado = 202{,}755 > \chi_{0.05,2}^2 = 5{,}991$$

Concluimos que el modelo tiene un buen ajuste si se introducen las variables Antigüedad del Negocio y Pasivo Financiero.

La entrada "Bloque", mide el cambio de verosimilitud $(-2lnL)$ entre bloques de entradas sucesivas durante la construcción del modelo. Si las variables se introducen en un solo bloque, el Chi – Cuadrado del Bloque es el mismo que el Chi – Cuadrado del Modelo.

Continuamos para los pasos 3, 4, 5, 6 de la misma manera y concluimos con el paso 7:

Para el Paso 7 se tiene:

$$Chi - Cuadrado = (-2lnL_{modelo0}) - (-2lnL_{modelo7})$$

$$Chi - Cuadrado = (403{,}170) - (84{,}149)$$

$$Chi - Cuadrado = 319{,}021$$

Luego calculando el estadístico de prueba:

$$\chi_c^2 = 319{,}021 > \chi_{0.05,7}^2 = 14{,}067$$

Por regla de decisión aceptamos H_0, ya que tenemos que:

$$Chi - Cuadrado = 319{,}021 > \chi_{0.05,7}^2 = 14{,}067$$

Concluimos que el modelo tiene un buen ajuste si se introducen las variables Antigüedad del Negocio, Pasivo Financiero, Capital del Préstamo, Destino del Préstamo, Número de Cuotas, Sexo del Cliente, y Actividad del Cliente.

B.2. Medidas de Bondad de Ajuste

La siguiente tabla, muestra 3 medidas resumen de los modelos, para cada paso, que complementan a la anterior, para verificar su validez; el primero es el valor del estadístico $(-2lnL)$ y los otros dos son Coeficientes de Determinación (R^2). Si el modelo es perfecto el valor de $(-2lnL)$ debe de ser muy pequeño, cercano a cero; y un coeficiente (R^2) cercano a uno, igual a uno en el mejor de los casos.

Tabla 69:
Resumen del Modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	245,206[a]	0,380	0,539
2	200,415[b]	0,459	0,651
3	163,272[b]	0,517	0,733
4	129,956[c]	0,563	0,798
5	107,409[c]	0,592	0,839
6	93,798[d]	0,608	0,863
7	84,149[d]	0,620	0,879

a. La estimación ha terminado en el número de iteración 5 porque las estimaciones de parámetro han cambiado en menos de .001.
b. La estimación ha terminado en el número de iteración 6 porque las estimaciones de parámetro han cambiado en menos de .001.
c. La estimación ha terminado en el número de iteración 7 porque las estimaciones de parámetro han cambiado en menos de .001.
d. La estimación ha terminado en el número de iteración 8 porque las estimaciones de parámetro han cambiado en menos de .001.

En la Tabla 69, el $(-2lnL)$, llamada también "Devianza" mide hasta que punto un modelo se ajusta bien a los datos, cuanto más pequeño sea el valor, mejor será el ajuste, como podemos apreciar, para cada paso, el estadístico disminuye de 245,206 a 84,149.

Los Coeficientes de Determinación Pseudos (R^2), son altos y aumentan desde el segundo al séptimo modelo. El coeficiente de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1, que para nuestro caso en el último modelo es $0{,}879$, podemos decir que es un coeficiente alto, que un importante porcentaje de la varianza es explicada por las variables predictoras introducidas en el modelo.

Tabla 70:
Prueba de Hosmer y Lemeshow

Escalón	Chi-cuadrado	gl	Sig.
1	77,817	8	0,000
2	15,890	8	0,044
3	38,113	8	0,000
4	66,355	8	0,000
5	7,790	8	0,454
6	16,770	8	0,033
7	9,608	8	0,294

La Tabla 70, proporciona los resultados de la prueba de Hosmer y Lemeshow (otra prueba para evaluar la bondad del ajuste de un modelo de Regresión Logística). Cuyo fin es que si se tiene un valor alto de la probabilidad predicha (p-valor) se asocia con el resultado $Y = 1$ de la variable binaria dependiente, mientras que un valor bajo de p (cercano a cero) se asociará con el resultado $Y = 0$.

La hipótesis nula es que no hay diferencias entre los valores observados y los valores esperados; la hipótesis alterna es que si hay, y por lo tanto, el rechazo de esta prueba indica que el modelo no está bien ajustado. Lo que se desea en esta prueba es que no haya diferencia, lo contrario a lo que suele ser habitual.

Para nuestro caso el valor de $p = 0{,}294$, es no significativo, por lo que aceptamos la hipótesis nula y decimos que existe ajuste en el modelo.

Tabla 71:
Tabla de contingencia para la prueba de Hosmer y Lemeshow

		Morosidad del Cliente = No		Morosidad del Cliente = Si		
		Observado	Esperado	Observado	Esperado	Total
Paso 1	1	21	25,474	5	,526	26
	2	15	14,338	0	,662	15
	3	44	50,114	9	2,886	53
	4	33	30,696	0	2,304	33
	5	13	12,349	1	1,651	14
	6	40	38,318	5	6,682	45
	7	30	24,357	1	6,643	31
	8	29	21,014	4	11,986	33
	9	1	4,666	14	10,334	15
	10	5	9,673	60	55,327	65
Paso 2	1	24	24,714	1	,286	25
	2	33	34,319	2	,681	35
	3	38	36,844	0	1,156	38
	4	29	29,348	2	1,652	31
	5	31	32,318	4	2,682	35
	6	25	21,502	0	3,498	25
	7	20	23,089	9	5,911	29
	8	19	18,074	14	14,926	33
	9	11	7,752	18	21,248	29
	10	1	3,040	49	46,960	50
Paso 3	1	31	32,890	2	,110	33
	2	34	34,659	1	,341	35
	3	34	34,326	1	,674	35
	4	33	31,809	0	1,191	33
	5	33	31,446	1	2,554	34
	6	28	28,338	5	4,662	33
	7	26	23,250	7	9,750	33
	8	11	12,202	22	20,798	33
	9	1	1,752	34	33,248	35
	10	0	,329	26	25,671	26
Paso 4	1	31	32,940	2	,060	33
	2	33	32,839	0	,161	33
	3	32	31,655	0	,345	32
	4	33	32,354	0	,646	33
	5	33	31,673	0	1,327	33
	6	30	29,888	3	3,112	33
	7	27	26,100	6	6,900	33
	8	10	12,041	24	21,959	34
	9	2	1,398	32	32,602	34
	10	0	,114	32	31,886	32
Paso 5	1	33	32,990	0	,010	33
	2	33	32,941	0	,059	33
	3	32	32,830	1	,170	33
	4	32	32,636	1	,364	33
	5	34	33,175	0	,825	34
	6	31	31,320	2	1,680	33

	7	26	25,037	7	7,963	33
	8	10	8,846	23	24,154	33
	9	0	1,133	33	31,867	33
	10	0	,094	32	31,906	32
Paso 6	1	33	32,997	0	,003	33
	2	33	32,979	0	,021	33
	3	33	32,932	0	,068	33
	4	33	32,827	0	,173	33
	5	30	32,532	3	,468	33
	6	32	31,342	1	1,658	33
	7	29	26,027	4	6,973	33
	8	8	8,715	25	24,285	33
	9	0	,600	33	32,400	33
	10	0	,047	33	32,953	33
Paso 7	1	33	32,999	0	,001	33
	2	33	32,993	0	,007	33
	3	33	32,971	0	,029	33
	4	32	32,898	1	,102	33
	5	33	32,683	0	,317	33
	6	32	31,845	1	1,155	33
	7	28	26,141	5	6,859	33
	8	7	7,965	26	25,035	33
	9	0	,487	33	32,513	33
	10	0	,018	33	32,982	33

Tabla 72:
Tabla de Clasificación[a]

			Pronosticado		
			Morosidad del Cliente		Corrección de
	Observado		No	Si	porcentaje
Paso 1	Morosidad del Cliente	No	225	6	97,4
		Si	23	76	76,8
	Porcentaje global				91,2
Paso 2	Morosidad del Cliente	No	218	13	94,4
		Si	23	76	76,8
	Porcentaje global				89,1
Paso 3	Morosidad del Cliente	No	225	6	97,4
		Si	20	79	79,8
	Porcentaje global				92,1
Paso 4	Morosidad del Cliente	No	224	7	97,0
		Si	16	83	83,8
	Porcentaje global				93,0
Paso 5	Morosidad del Cliente	No	222	9	96,1
		Si	12	87	87,9
	Porcentaje global				93,6
Paso 6	Morosidad del Cliente	No	225	6	97,4
		Si	8	91	91,9
	Porcentaje global				95,8
Paso 7	Morosidad del Cliente	No	224	7	97,0
		Si	8	91	91,9
	Porcentaje global				95,5

a. El valor de corte es .500

La Tabla 72, muestra en su séptimo paso, una especificidad (probabilidad de que la prueba identifique como cliente que no tendrá morosidad al cliente que efectivamente no haya presentado morosidad) de 97,0% y una sensibilidad (probabilidad de que la prueba identifique como cliente que tendrá morosidad al cliente que efectivamente haya presentado morosidad) de 91,9%.

Tabla 73:
Variables en la Ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)	95% C.I. para EXP(B) Inferior	Superior
Paso 1[a]	Antigüedad	-,222	,023	95,775	1	,000	,801	,766	,838
	Constante	2,685	,379	50,274	1	,000	14,661		
Paso 2[b]	Antigüedad	-,208	,024	73,742	1	,000	,812	,774	,851
	Pasivo finan(1)	-2,356	,381	38,324	1	,000	,095	,045	,200
	Constante	3,647	,482	57,122	1	,000	38,343		
Paso 3[c]	Antigüedad	-,246	,031	63,849	1	,000	,782	,736	,831
	Pasivo finan(1)	-3,759	,550	46,651	1	,000	,023	,008	,069
	Capital	,000	,000	29,646	1	,000	1,000	1,000	1,000
	Constante	2,069	,580	12,741	1	,000	7,914		
aso 4[d]	Antigüedad	-,210	,034	38,857	1	,000	,811	,759	,866
	Destino préstamo			26,454	3	,000			
	Destino préstamo(1)	-,693	,681	1,035	1	,309	,500	,132	1,900
	Destino préstamo(2)	-,108	,626	,030	1	,863	,897	,263	3,063
	Destino préstamo(3)	2,870	,712	16,234	1	,000	17,642	4,367	71,274
	Pasivo finan(1)	-3,888	,636	37,350	1	,000	,020	,006	,071
	Capital	,000	,000	27,115	1	,000	1,000	1,000	1,000
	Constante	,689	,739	,869	1	,351	1,991		
Paso 5[e]	Antigüedad	-,182	,035	27,080	1	,000	,833	,778	,893
	Destino préstamo			23,237	3	,000			
	Destino préstamo(1)	-,850	,789	1,161	1	,281	,428	,091	2,005
	Destino préstamo(2)	-,045	,733	,004	1	,951	,956	,227	4,022
	Destino préstamo(3)	2,620	,739	12,552	1	,000	13,733	3,224	58,503
	Pasivo finan(1)	-3,247	,666	23,777	1	,000	,039	,011	,143
	Capital	,000	,000	34,754	1	,000	1,000	1,000	1,000
	Número cuota	-,172	,043	16,265	1	,000	,842	,774	,915
	Constante	2,137	,856	6,232	1	,013	8,471		
	Sexo(1)	-2,084	,621	11,263	1	,001	,124	,037	,420
	Antigüedad	-,187	,038	24,752	1	,000	,829	,770	,893
	Destino préstamo			20,306	3	,000			
	Destino préstamo(1)	-1,305	,834	2,447	1	,118	,271	,053	1,391
	Destino préstamo(2)	-,505	,781	,418	1	,518	,604	,131	2,788
Paso 6[f]	Destino préstamo(3)	2,412	,790	9,309	1	,002	11,152	2,369	52,500
	Pasivo finan(1)	-3,745	,770	23,652	1	,000	,024	,005	,107
	Capital	,000	,000	31,405	1	,000	1,000	1,000	1,000
	Número cuota	-,189	,049	14,818	1	,000	,828	,752	,912
	Constante	3,527	1,071	10,844	1	,001	34,024		

Paso 7[g]	Sexo(1)	-2,259	,665	11,542	1	,001	,104	,028	,385
	Actividad			8,499	4	,075			
	Actividad(1)	1,367	,882	2,404	1	,121	3,923	,697	22,082
	Actividad(2)	,234	1,192	,039	1	,844	1,264	,122	13,058
	Actividad(3)	,309	1,067	,084	1	,772	1,363	,168	11,026
	Actividad(4)	3,737	1,370	7,437	1	,006	41,973	2,861	615,713
	Antigüedad	-,210	,042	24,839	1	,000	,811	,746	,880
	Destino préstamo			21,226	3	,000			
	Destino préstamo(1)	-1,062	,925	1,320	1	,251	,346	,056	2,117
	Destino préstamo(2)	,013	,884	,000	1	,988	1,013	,179	5,735
	Destino préstamo(3)	3,036	,929	10,674	1	,001	20,831	3,370	128,767
	Pasivo finan(1)	-3,641	,794	21,040	1	,000	,026	,006	,124
	Capital	,000	,000	26,392	1	,000	1,000	1,000	1,000
	Número cuotas	-,229	,057	16,284	1	,000	,796	,712	,889
	Constante	2,869	1,182	5,896	1	,015	17,620		

a. Variables especificadas en el paso 1: Antiguedad.
b. Variables especificadas en el paso 2: Pasivofinan.
c. Variables especificadas en el paso 3: Capital.
d. Variables especificadas en el paso 4: Destinoprestamo.
e. Variables especificadas en el paso 5: Numcuota.
f. Variables especificadas en el paso 6: Sexo.
g. Variables especificadas en el paso 7: Actividad.

La Tabla 73, muestra las variables que entran en la ecuación, sus coeficientes de regresión con sus correspondientes errores estándar, el valor del estadístico de Wald para evaluar la Hipótesis Nula $\beta_i = 0$, la significación estadística asociada, y el valor de la $OR(exp(B))$ con sus intervalos de confianza.

Tabla 74:
Las Variables no están en la Ecuacióna

			Puntuación Eficiente de Rao	Gl	Sig.
Paso 1	Variables	Edad	7,034	1	,008
		Sexo(1)	13,066	1	,000
		Actividad	9,025	4	,060
		Actividad(1)	,790	1	,374
		Actividad(2)	1,087	1	,297
		Actividad(3)	,341	1	,559
		Actividad(4)	7,880	1	,005
		Destino préstamo	48,223	3	,000
		Destino préstamo (1)	12,554	1	,000
		Destino préstamo (2)	1,286	1	,257
		Destino préstamo (3)	46,847	1	,000
		Antecedente(1)	,442	1	,506
		Pasivo financiero(1)	49,985	1	,000
		Capital	9,156	1	,002

Paso 2		Numcuota	17,152	1	,000
	Variables	Edad	10,873	1	,001
		Sexo(1)	19,585	1	,000
		Actividad	10,528	4	,032
		Actividad(1)	2,706	1	,100
		Actividad(2)	,048	1	,826
		Actividad(3)	4,092	1	,043
		Actividad(4)	4,496	1	,034
		Destino préstamo	37,287	3	,000
		Destino préstamo (1)	12,427	1	,000
		Destino préstamo (2)	,240	1	,624
		Destino préstamo (3)	33,971	1	,000
		Antecedente(1)	,012	1	,913
		Capital	38,415	1	,000
		Numcuota	2,765	1	,096
Paso 3	Variables	Edad	5,629	1	,018
		Sexo(1)	20,085	1	,000
		Actividad	4,900	4	,298
		Actividad(1)	,537	1	,464
		Actividad(2)	,431	1	,511
		Actividad(3)	,402	1	,526
		Actividad(4)	4,043	1	,044
		Destino préstamo	34,590	3	,000
		Destino préstamo (1)	7,254	1	,007
		Destino préstamo (2)	1,436	1	,231
		Destino préstamo (3)	33,291	1	,000
		Antecedente(1)	,271	1	,603
		Numcuota	22,725	1	,000
	Estadísticos globales		63,661	11	,000
Paso 4	Variables	Edad	2,947	1	,086
		Sexo(1)	16,610	1	,000
		Actividad	4,753	4	,314
		Actividad(1)	,008	1	,929
		Actividad(2)	,439	1	,508
		Actividad(3)	,099	1	,753
		Actividad(4)	4,402	1	,036
		Antecedente(1)	,013	1	,911
		Número de cuotas	18,693	1	,000
	Estadísticos globales		35,007	8	,000
Paso 5	Variables	Edad	1,774	1	,183
		Sexo(1)	13,217	1	,000
		Actividad	9,029	4	,060
		Actividad(1)	,577	1	,447
		Actividad(2)	,500	1	,479
		Actividad(3)	,146	1	,702
		Actividad(4)	7,178	1	,007
		Antecedente(1)	,290	1	,590
	Estadísticos globales		23,385	7	,001

Paso 6	Variables	Edad	,951	1	,329
		Actividad	9,533	4	,049
		Actividad(1)	1,286	1	,257
		Actividad(2)	,403	1	,525
		Actividad(3)	,496	1	,481
		Actividad(4)	6,670	1	,010
		Antecedente(1)	1,202	1	,273
	Estadísticos globales		10,314	6	,112
Paso 7	Variables	Edad	,434	1	,510
		Antecedente(1)	,571	1	,450
	Estadísticos globales		1,057	2	,590

a. Los chi-cuadrados residuales no se calculan debido a redundancias.

Tabla 75:
Matriz de Correlaciones

		Constante	Antigüedad	Antigüedad	Pasivo finan(1)	Capital	Antigüedad Destino préstamo(1)	Destino préstamo(2)	Destino préstamo(3)	Pasivo finan(1)	Capital	Numcuota	Sexo(1)	Actividad(1)	Actividad(2) Actividad(3)		Actividad(4)
Paso	Constante	1,000	-,902														
	Antigüedad	-,902	1,000														
Paso 2	Constante	1,000		-,868	-,512												
	Antigüedad	-,868		1,000	,229												
	Pasivo finan(1)	-,512		,229	1,000												
Paso 3	Constante	1,000		-,616	-,235	-,244											
	Antigüedad	-,616		1,000	,462	-,481											
	Pasivo finan(1)	-,235		,462	1,000	-,653											
	Capital	-,244		-,481	-,653	1,000											
Paso 4	Constante	1,000					-,425	-,384	-,253	-,448	-,339	-,075					
	Antigüedad	-,425					1,000	-,093	-,155	-,155	-,440	,383					
	Destinoprést.(1)	-,384					-,093	1,000	,496	,420	-,007	,104					
	Destinoprést.(2)	-,253					-,155	,496	1,000	,444	-,095	-,001					
	Destinoprést.(3)	-,448					-,155	,420	,444	1,000	,273	-,206					
	Pasivo finan(1)	-,075					,383	,104	-,001	-,206	-,662	1,000					
	Capital	-,339					-,440	-,007	-,095	,273	1,000	-,662					
Paso 5	Constante	1,000					-,387	-,327	-,202	-,334	,028	-,064	-,447				
	Antigüedad	-,387					1,000	-,135	-,220	-,188	-,305	,358	,018				
	Destinoprést.(1)	-,327					-,135	1,000	,514	,458	-,097	,073	,067				
	Destinoprést.(2)	-,202					-,220	,514	1,000	,520	-,092	-,025	-,019				
	Destinoprést.(3)	-,334					-,188	,458	,520	1,000	,180	-,178	-,094				
	Pasivo finan(1)	-,064					,358	,073	-,025	-,178	-,468	1,000	-,023				
	Capital	,028					-,305	-,097	-,092	,180	1,000	-,468	-,610				
	Numcuota	-,447					,018	,067	-,019	-,094	-,610	-,023	1,000				
Paso 6	Constante	1,000					-,490	-,385	-,252	-,239	,228	-,224	-,536	-,465			
	Sexo(1)	-,465					,234	,190	,184	-,042	-,355	,357	,247	1,000			
	Antigüedad	-,490					1,000	-,010	-,100	-,129	-,351	,407	,087	,234			
	Destinoprést.(1)	-,385					-,010	1,000	,526	,415	-,219	,175	,145	,190			
	Destinoprést.(2)	-,252					-,100	,526	1,000	,466	-,241	,092	,069	,184			
	Destinoprést.(3)	-,239					-,129	,415	,466	1,000	,134	-,106	-,109	-,042			
	Pasivo finan(1)	-,224					,407	,175	,092	-,106	-,531	1,000	,041	,357			
	Capital	,228					-,351	-,219	-,241	,134	1,000	-,531	-,632	-,355			
	Número cuota	-,536					,087	,145	,069	-,109	-,632	,041	1,000	,247			
Paso 7	Constante	1,000					-,413	-,276	-,219	-,203	,077	-,152	-,386	-,347	-,285	-,151	-,097	-,058
	Sexo(1)	-,347					,271	,101	,076	-,127	-,385	,329	,306	1,000	-,174	-,008	,022	-,145
	Actividad(1)	-,285					-,072	-,223	,063	,045	,169	-,125	-,270	-,174	1,000	,532	,538	,501
	Actividad(2)	-,151					,087	-,332	-,125	-,172	-,163	-,125	-,005	-,008	,532	1,000	,566	,297
	Actividad(3)	-,097					,033	-,259	-,238	-,229	-,213	-,091	-,003	,022	,538	,566	1,000	,333
	Actividad(4)	-,058					-,366	,024	,183	,286	,270	-,179	-,355	-,145	,501	,297	,333	1,000
	Antigüedad	-,413					1,000	-,111	-,198	-,277	-,426	,372	,233	,271	-,072	,087	,033	-,366
	Destinoprést.(1)	-,276					-,111	1,000	,496	,467	-,004	,152	,073	,101	-,223	-,332	-,259	,024
	Destinoprést.(2)	-,219					-,198	,496	1,000	,544	,033	,065	-,110	,076	,063	-,125	-,238	,183
	Destinoprést.(3)	-,203					-,277	,467	,544	1,000	,353	-,095	-,287	-,127	,045	-,172	-,229	,286
	Pasivo finan(1)	-,152					,372	,152	,065	-,095	-,485	1,000	,121	,329	-,125	-,125	-,091	-,179
	Capital	,077					-,426	-,004	,033	,353	1,000	-,485	-,684	-,385	,169	-,163	-,213	,270
	Número cuota	-,386					,233	,073	-,110	-,287	-,684	,121	1,000	,306	-,270	-,005	-,003	-,355

ANEXO C

C.1. Medidas de Asociación para Variables Categóricas

En este anexo realizaremos los cálculos necesarios para hallar la asociación de todas las variables independientes categóricas con la variable respuesta.

C.1.1. Variables Dicotómicas

Morosidad & Sexo del Cliente

Tabla 76:
*Morosidad del Cliente*Sexo del Cliente*

		Sexo del Cliente		
		Femenino	**Masculino**	**Total**
Morosidad del cliente	**No**	139	92	231
		60,2%	39,8%	100,0%
	Si	27	72	99
		27,3%	72,7%	100,0%
	Total	**166**	**164**	**330**

En la tabla 76. se observa que los clientes de sexo masculino que tuvieron morosidad representan el 72,7%, que es una cantidad mayor a comparación de los clientes masculinos sin morosidad que representan un 39,8% y las clientes femeninas con morosidad con una representación del 27,3%.

Para verificar si existe asociación entre las variables, calculamos los valores de OR, Q de Yule y Chi-Cuadrado.

- Cálculo de OR e Intervalo de Confianza de OR

$$OR = \frac{139 \times 72}{27 \times 92} = 4.029$$

$$IC = OR^{\left(1 \mp \frac{Z}{Xhm}\right)} = 4.029^{\left(1 \mp \frac{1.96}{5,954}\right)}$$

$$LI = 2,408 \qquad LS = 6,742$$

Donde; $Xhm = \sqrt{\frac{(n-1)(a \times d - b \times c)^2}{(a+b)(c+d)(a+c)(b+d)}} = \sqrt{\frac{(330-1)(138 \times 72 - 27 \times 92)^2}{(231)(99)(166)(164)}} = 5,954$

Existe asociación entre las variables Morosidad y Sexo del Cliente. $OR > 1$ y los límites del intervalo $LI > 1$ y $LS > 1$, a la variable Sexo del Cliente se le considera Factor de Riesgo

- Q de Yule

$$Q = \frac{(138 \times 72) - (27 \times 92)}{(138 \times 72) + (27 \times 92)} = \frac{7452}{12420} = 0{,}6$$

Como $Q > 0$ significa que hay una asociación entre las variables y ésta es positiva.

- Cálculo de Chi-Cuadrado

Hipótesis:

H_0 : Las variables Morosidad y Sexo del Cliente son independientes

H_1 : Las variables Morosidad y Sexo del Cliente no son independientes

Cálculo del Estadístico:

$$e_{11} = \frac{231 \times 166}{330} = 116.2 \qquad e_{12} = \frac{231 \times 164}{330} = 114.8$$

$$e_{21} = \frac{99 \times 166}{330} = 49.2 \qquad e_{22} = \frac{99 \times 164}{330} = 49.2$$

$$\chi_c^2 = \frac{(139 - 116{,}2)^2}{116{,}2} + \frac{(92 - 114{,}8)^2}{114{,}8} + \frac{(27 - 49{,}8)^2}{49{,}8} + \frac{(72 - 49{,}2)^2}{49{,}2}$$

$$\chi_c^2 = 30{,}006$$

- Decisión:

$\chi_c^2 = 30{,}006 > \chi_{0.05,1}^2 = 3.84$ se acepta H_1; es decir, el Riesgo Crediticio (morosidad) depende del sexo del cliente.

Morosidad & Antecedentes en Clearing

Tabla 77:
Morosidad & Antecedentes en Clearing

			Antecedentes en Clearing		Total
			No	Si	
Morosidad del Cliente		No	86	145	231
			37,2%	62,8%	100,0%
		Si	44	55	99
			44,4%	55,6%	100,0%
		Total	**130**	**200**	**330**

En la Tabla 77. se aprecia que del grupo de clientes que no tienen Antecedentes en Clearing son el 37,2% que tampoco presentan morosidad, frente a los 55,6% de los clientes que si tienen Antecedentes en Clearing y también presentan morosidad.

Calculamos los valores de OR, Q de Yule y Chi Cuadrado.

- Cálculo de OR e Intervalo de Confianza de OR

$$OR = \frac{86 \times 55}{44 \times 145} = 0{,}741$$

$$IC = OR^{\left(1 \mp \frac{Z}{Xhm}\right)} = 0{,}741^{\left(1 \mp \frac{1.96}{1{,}024}\right)}$$

$$LI = 0{,}460 \qquad LS = 1{,}196$$

Donde; $Xhm = \sqrt{\frac{(n-1)(a\times d - b\times c)^2}{(a+b)(c+d)(a+c)(b+d)}} = \sqrt{\frac{(330-1)(86\times 55 - 44\times 145)^2}{(231)(99)(130)(200)}} = 1{,}024$

No existe asociación entre las variables Morosidad y Antecedentes en Clearing. $OR < 1$ y los límites del intervalo $LI < 1$ y $LS > 1$, a la variable Antecedentes en Clearing no se le considera Factor de Riesgo.

- Q de Yule

$$Q = \frac{(86 \times 55) - (44 \times 145)}{(86 \times 55) + (44 \times 145)} = \frac{-1650}{1110} = -0{,}149$$

Como $Q < 0$ significa que hay una asociación entre las variables y ésta es negativa.

- Cálculo de Chi-Cuadrado

Hipótesis:

H_0 : Las variables Morosidad y Antecedentes en Clearing son independientes

H_1 : Las variables Morosidad y Antecedentes en Clearing no son independientes

Cálculo del Estadístico:

$$e_{11} = \frac{231 \times 130}{330} = 91 \qquad e_{12} = \frac{231 \times 200}{330} = 140$$

$$e_{21} = \frac{99 \times 130}{330} = 39 \qquad e_{22} = \frac{99 \times 200}{330} = 60$$

$$\chi_c^2 = \frac{(86-91)^2}{91} + \frac{(145-140)^2}{140} + \frac{(44-39)^2}{39} + \frac{(55-60)^2}{60}$$

$$\chi_c^2 = 1{,}511$$

- Decisión:

$\chi_c^2 = 1{,}511 < \chi_{0.05,1}^2 = 3.84$ se rechaza H_1; es decir, la morosidad no depende de la variable Antecedentes en Clearing.

No existe asociación entre las variables Riesgo Crediticio (morosidad) y Antecedentes en Clearing, $OR < 1$, además los límites del intervalo $LI < 1$ y $LS > 1$, la variable Antecedentes en el Clearing no puede ser considerada Factor de Riesgo.

Morosidad & Pasivo Financiero

Tabla 78:
Morosidad & Pasivo Financiero

		Pasivo Financiero		Total
		No	**Si**	
Morosidad del Cliente	**No**	173	58	231
		74,9%	25,1%	100,0%
	Si	20	79	99
		20.2%	79,8%	100,0%
	Total	**193**	**137**	**330**

En la Tabla 78. se aprecia que los clientes que tienen Pasivo Financiero y no presentan Riesgo Crediticio (morosidad) son el 25,1%, frente a un 79,8% de los clientes que tienen Pasivo Financiero y también presentan morosidad.

Calculamos los valores de OR, Q de Yule y Chi Cuadrado.

- Cálculo de OR e Intervalo de Confianza de OR

$$OR = \frac{173 \times 79}{20 \times 58} = 11{,}782$$

$$IC = OR^{\left(1 \mp \frac{Z}{Xhm}\right)} = 11{,}782^{\left(1 \mp \frac{1.96}{7{,}697}\right)}$$

$$LI = 6{,}638 \qquad LS = 20{,}911$$

Donde; $Xhm = \sqrt{\frac{(n-1)(a\times d - b\times c)^2}{(a+b)(c+d)(a+c)(b+d)}} = \sqrt{\frac{(330-1)(173\times 79 - 20\times 58)^2}{(231)(99)(193)(137)}} = 7{,}697$

Existe asociación entre las variables Morosidad y Pasivo Financiero. $OR > 1$ y los límites del intervalo $LI > 1$ y $LS > 1$, a la variable Pasivo Financiero se le considera Factor de Riesgo.

- Q de Yule

$$Q = \frac{(173 \times 79) - (20 \times 58)}{(173 \times 79) + (20 \times 58)} = \frac{12507}{14827} = 0{,}844$$

Como $Q > 0$ significa que hay una asociación entre las variables y ésta es positiva.

- Cálculo de Chi-Cuadrado

Hipótesis:

H_0 : Las variables Morosidad y Pasivo Financiero son independientes

H_1 : Las variables Morosidad y Pasivo Financiero no son independientes

Cálculo del Estadístico:

$$e_{11} = \frac{231 \times 193}{330} = 135{,}1 \qquad e_{12} = \frac{231 \times 137}{330} = 95{,}9$$

$$e_{21} = \frac{99 \times 193}{330} = 57{,}9 \qquad e_{22} = \frac{99 \times 137}{330} = 41{,}1$$

$$\chi_c^2 = \frac{(173 - 135{,}1)^2}{135{,}1} + \frac{(58 - 95{,}9)^2}{95{,}9} + \frac{(20 - 57{,}9)^2}{57{,}9} + \frac{(79 - 41{,}1)^2}{41{,}1}$$

$$\chi_c^2 = 85{,}368$$

- Decisión:

$\chi_c^2 = 85{,}368 > \chi_{0.05,1}^2 = 3.84$ se acepta H_1; es decir, la morosidad depende del Pasivo Financiero.

Existe asociación entre las variables Morosidad y Pasivo Financiero, $OR > 1$, además los límites del intervalo $LI > 1$ y $LS > 1$, a la variable Pasivo Financiero puede ser considerado Factor de Riesgo.

Morosidad & Actividad del Cliente

Tabla 79:
Morosidad & Actividad del Cliente

		Actividad del Cliente					Total
		Independiente sin Estudios Superiores	Independiente con Estudios Superiores	Empleado Profesional	Empleado Técnico	Empleado con Oficio	
Morosidad del Cliente	No	76	53	51	10	41	231
		32,9%	22,9%	22,1%	4,3%	17,8%	100,0%
	Si	24	18	15	17	25	99
		24,2%	18,2%	15,2%	17,2%	25,2%	100,0%
	Total	**100**	**71**	**66**	**27**	**66**	**330**

- Hipótesis:

  H_0 : Las variables Morosidad y Actividad del Cliente son independientes

  H_1 : Las variables Morosidad y Actividad del Cliente no son independientes

- Cálculo del Estadístico:

$$e_{11} = \frac{231 \times 100}{330} = 70 \qquad e_{21} = \frac{99 \times 100}{330} = 30$$

$$e_{12} = \frac{231 \times 71}{330} = 49{,}7 \qquad e_{22} = \frac{99 \times 71}{330} = 21{,}3$$

$$e_{13} = \frac{231 \times 66}{330} = 46{,}2 \qquad e_{23} = \frac{99 \times 66}{330} = 19{,}8$$

$$e_{14} = \frac{231 \times 27}{330} = 18{,}9 \qquad e_{24} = \frac{99 \times 27}{330} = 8{,}1$$

$$e_{15} = \frac{231 \times 66}{330} = 46{,}2 \qquad e_{25} = \frac{99 \times 66}{330} = 19{,}8$$

$$\chi_c^2 = \frac{(76-70)^2}{70} + \frac{(53-49{,}7)^2}{49{,}7} + \frac{(51-46{,}2)^2}{46{,}2} + \frac{(10-18{,}9)^2}{18{,}9} + \frac{(41-46{,}2)^2}{46{,}2} + \frac{(24-30)^2}{30} + \frac{(18-21{,}3)^2}{21{,}3} + \frac{(15-19{,}8)^2}{19{,}8} + \frac{(17-8{,}1)^2}{8{,}1} + \frac{(25-19{,}8)^2}{19{,}8}$$

$$\chi_c^2 = 20{,}028$$

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(2-1)(5-1)} = \chi^2_{0.05,4} = 9{,}49$$

- Decisión:

$\chi_c^2 = 20{,}028 > \chi^2_{0.05,4} = 9{,}49$ se acepta H_1; es decir, el Riesgo Crediticio

(morosidad) depende de la Actividad del Cliente.

Morosidad & Destino del Préstamo

Tabla 80:
Morosidad & Destino del Préstamo

		Destino del Préstamo				
		Implementación de Negocio	**Adquisición de Bien Inmueble**	**Adquisición de Bien Mueble**	**Construcción de Vivienda**	**Total**
Morosidad del Cliente	**No**	97	71	11	52	231
		42,0%	30,7%	4,8%	22,5%	100,0%
	Si	9	12	57	21	99
		9,1%	12,1%	57,6%	21,2%	100%
	Total	**106**	**83**	**68**	**73**	**330**

- Hipótesis:

  H_0 : Las variables Morosidad y Destino del Préstamo son independientes

  H_1 : Las variables Morosidad y Destino del Préstamo no son independientes

- Cálculo del Estadístico:

$$e_{11} = \frac{231 \times 106}{330} = 74{,}2 \qquad e_{21} = \frac{99 \times 106}{330} = 31{,}8$$

$$e_{12} = \frac{231 \times 83}{330} = 58{,}1 \qquad e_{22} = \frac{99 \times 83}{330} = 24{,}9$$

$$e_{13} = \frac{231 \times 68}{330} = 47{,}6 \qquad e_{23} = \frac{99 \times 68}{330} = 20{,}4$$

$$e_{14} = \frac{231 \times 73}{330} = 51{,}1 \qquad e_{24} = \frac{99 \times 73}{330} = 21{,}9$$

$$\chi_c^2 = \frac{(97-74{,}2)^2}{74{,}2} + \frac{(71-58{,}1)^2}{58{,}1} + \frac{(11-47{,}6)^2}{47{,}6} + \frac{(52-51{,}1)^2}{51{,}1} + \frac{(9-31{,}8)^2}{31{,}8} + \frac{(12-24{,}9)^2}{24{,}9} + \frac{(57-20{,}4)^2}{20{,}4} + \frac{(21-21{,}9)^2}{21{,}9}$$

$$\chi_c^2 = 126{,}760$$

$$\chi^2_{\alpha,(r-1)(c-1)} = \chi^2_{0.05,(2-1)(4-1)} = \chi^2_{0.05,3} = 7{,}82$$

- Decisión:

$\chi_c^2 = 126{,}760 > \chi^2_{0.05,3} = 7{,}82$ se acepta H_1; es decir, el Riesgo Crediticio (morosidad) depende del Destino del Préstamo.