Estimador de ingresos para personas naturales con actividad económica independiente en una entidad financiera

Eddison Andrés Jiménez Santana

Nestor Yamith Gómez Bermúdez

Trabajo de grado para optar al título de Especialista en estadística

Director:

Carlos Alfonso Mantilla Duarte

Magister en estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2020

Agradecimientos

En ocasiones se requiere más que perseverancia,
se requiere pasión para cumplir las metas
y permitir que otras personas se priven de nuestro tiempo,
para jugar y compartir.
Gracias Robert Yamith por acompañarme en la travesía, te amo hijo.

Yamigo

A mi esposa Ginereth Rojas Carreño y mi hija Liz Margarita Jiménez Rojas,
quienes con su amor y paciencia hicieron parte de este logro.

Eddison J.

A nuestra cooperativa,
gracias por el apoyo y oportunidad de adquirir nuevos conocimientos.

Resumen

Título: ESTIMADOR DE INGRESOS PARA PERSONAS NATURALES CON ACTIVIDAD ECONÓMICA INDEPENDIENTE EN UNA ENTIDAD FINANCIERA[1]

Autor: NESTOR YAMITH GÓMEZ BERMÚDEZ – EDDISON ANDRÉS JIMÉNEZ SANTANA [**]

Palabras Claves: Estimador, Ingresos, Independientes, Correspondencia Múltiple, Regresión lineal, Regresión Robusta

Descripción:

En el sector financiero resulta de gran importancia conocer los datos de una persona para realizar la medición del riesgo que trae el otorgar un crédito, un dato clave es el ingreso, pues a través de este se pueden crear escenarios para definir un monto que se ajuste a la persona.

El presente documento tiene como objetivo diseñar un estimador de ingresos para personas con actividad económica independiente, para lo cual se realiza en primera instancia un análisis de correspondencia múltiple para identificar grupos según variables sociodemográficas asociadas a los rangos de ingresos, en una segunda parte del documento se realiza regresión lineal múltiple y regresión robusta en forma paralela utilizando variables sociodemográficas como variables predictoras.

Como resultado final del proyecto se optó por elegir el modelo de regresión robusta dadas las bondades de ajuste que este aplica sobre los valores atípicos, logrando una predicción no muy alejada a la desviación estándar que tenían los ingresos en la muestra utilizada para el estudio.

---

[1] Trabajo de grado

[**] Facultad de Ciencias. Escuela de Matemáticas. Director: Carlos Alfonso Mantilla Duarte. Magister en Estadística

Abstract

Title: INCOME ESTIMATOR FOR NATURAL PERSONS WITH INDEPENDENT ECONOMIC ACTIVITY IN A FINANCIAL INSTITUTION[2]

Author: NESTOR YAMITH GÓMEZ BERMÚDEZ – EDDISON ANDRÉS JIMÉNEZ SANTANA [**]

Keywords: Estimator, Income, Independent, Multiple correspondence, Regression line, Regression Robust

Description:

In the financial sector it is of great importance to know the data of a person to measure the risk of granting a loan, a key data is income, because through this you can create scenarios to define an amount that adjusts to the person.

The objective of this document is to design an income estimator for people with independent economic activity, for which a multiple correspondence analysis is first performed to identify groups according to sociodemographic variables associated with income ranges, in a second part of the document Multiple linear regression and robust regression are performed in parallel using sociodemographic variables as predictor variables.

As a result of the project, we opted to choose the robust regression model given the goodness of fit that it applies to outliers, achieving a prediction not too far from the standard deviation of income in the sample used for the study.

[2] Project of grade

[**] Faculty Science. School Mathematical. Director: Carlos Alfonso Mantilla Duarte. Magister statistics

Tabla de Contenido

Lista de tablas

Lista de figuras

Apéndice

Introducción

La variable ingresos juega un papel importante en la solicitud de crédito, ayuda a establecer la capacidad de endeudamiento y amortización de los pagos que debe realizar un cliente en la solicitud de crédito.

En este sentido, validar los ingresos de los clientes se convierte en una tarea critica en el proceso análisis de crédito, especialmente para las personas que tienen actividad económica independiente o a cuenta propia; el presente estudio aborda tres etapas para explicar la relación de los ingresos con respecto a las variables capturadas en el momento de realizar la solicitud de crédito.

La primera etapa consiste en realizar un perfilamiento de las variables seleccionadas y capturadas en el formulario de solicitud del crédito en función de la variable respuesta, lo que permite tener inferir en la composición de la muestra y características sociodemográficas de los perfiles evaluados.

La segunda etapa utiliza técnicas de análisis de correspondencia múltiple para lograr un agrupamiento de variables que permitan maximizar el poder explicativo y de asociación de datos que se recolectan en las solicitudes de crédito, allí podemos observar el procedimiento metodológico aplicado, el aporte de las variables a las dimensiones del modelo y la fuerza de asociación de cada una de estas, hasta lograr el mejor modelo explicativo de correspondencia múltiple.

La tercera etapa confronta el análisis de correspondencia múltiple mediante un modelo de regresión lineal múltiple que se valida simultáneamente con un modelo de regresión robusta, en esta etapa se busca establecer cual técnica de regresión tiene el mejor poder explicativo y se puede aplicar como solución al estimador de ingresos.

En este sentido, la estructura del documento inicia describiendo los componentes técnicos que se abordan en cada uno de los apartados; luego se describen las variables mediante un análisis gráfico de barras, caja y bigotes en función de los ingresos; posteriormente se realiza validación del análisis de correspondencia múltiple para evaluar el poder explicativo de las variables hasta obtener el mejor agrupamiento; finalmente se estable un modelo de regresión lineal múltiple que brinda el mejor poder explicativo, el cual es comparado con el método de regresión robusta evaluando uno a uno de los atributos, finalmente se decide el modelo a utilizar para el estimador de ingresos en trabajadores independientes.

En los anexos se comparte el código utilizado en R-Studio para desarrollar cada uno de los apartados técnicos; de esta manera se logra establecer las variables y capacidad de explicación con el método de regresión robusta para estimar el nivel de ingresos en personas naturales con actividad económica independiente.

1. Justificación

Las entidades financieras tienen la necesidad de validar y/o establecer un nivel en ingresos de efectivo para los clientes que no tienen la capacidad de soportar ingresos fijos, pues mitigar el riesgo y asegurar la recuperabilidad del préstamo otorgado permite garantizar los niveles de solvencia establecidos por los entes reguladores.

En este sentido, el área encargada de evaluación, análisis y aprobación de los créditos de la entidad financiera evaluada evalúa tres perfiles: Empresariales, empleados e independientes. El perfil empresarial agrupa los asociados con una estructura formal y tienen capacidad de presentar soportes que demuestran la experiencia en una actividad económica, al igual que niveles de endeudamiento.

El perfil empleado agrupa las personas que dependen de ingresos fijos denominados “asalariados” los cuales pueden o no tener experiencia crediticia y bajo unas métricas se establece un rango de endeudamiento máximo tolerado por el solicitante del crédito, de esta manera el monto desembolsado para el perfil empleados dependerá de los ingresos que pueda soportar el asociado.

Finalmente, los asociados con perfil independiente no tienen una actividad económica estable y los ingresos pueden depender de una o más actividades laborales. Debido a la inestabilidad de estos trabajos no es fácil justificar el nivel de ingresos recibidos, sin embargo, dependiendo de la actividad principal que desarrolla el solicitante es posible estimar unos ingresos teniendo en cuenta

nivel de activos del negocio, experiencia en la actividad económica, entre otras; no obstante, el riesgo de crédito permanece cuando no se puede validar los ingresos.

El trabajo desarrollado pretende estimar el nivel de ingresos mediante el análisis de las variables capturadas en la solicitud de crédito, la entidad financiera minimiza el riesgo de crédito debido a que previamente ha validado el factor "ingresos del solicitante" y con ello la decisión de crédito estará sujeta a otro tipo de análisis partiendo del hecho que los ingresos están validados.

En este contexto mediante técnicas de análisis de correspondencia múltiple, regresión lineal múltiple y regresión robusta se valida un estimador de ingresos para asociados con perfil independiente de una entidad financiera.

Las técnicas previamente seleccionadas tienen la capacidad de agrupar variables explicativas, es así como en el análisis de correspondencia múltiple – ACM, logra establecer agrupaciones por niveles de asociación y fuerza de explicación para seleccionar el mejor modelo a explicar la variable respuesta.

La regresión lineal múltiple establece variables explicativas en función de la variable respuesta, teniendo en cuenta el principio de parsimonia que relaciona la optimización de las variables explicativas en función de obtener la mayor explicación; finalmente la regresión robusta reduce la influencia de datos atípicos para aumentar el poder explicativo, sin embargo, este método no refleja el coeficiente de correlación – R2, típicamente utilizado para selección del modelo óptimo.

2. Objetivos

2.1. Objetivo general

Diseñar un modelo de estimación de ingresos de personas naturales con actividad económica independiente en una entidad financiera.

2.2. Objetivos específicos

Preparar los datos necesarios para el diseño del modelo mediante una clasificación de la muestra.

Comparar técnicas de clasificación y regresión que sean útiles para estimar los ingresos de una persona natural.

Elegir el método estadístico que mejor explique los ingresos que reciben las personas naturales con actividad económica independiente.

3. Antecedentes

Los estudios de estimación de ingresos para personas que trabajan de manera independiente y que solicitan crédito al sector financiero son escasos, sin embargo existen estudios que brindan una perspectiva bajo esta premisa, es el caso de (Fernández, 2018) que intenta estimar los ingresos de personal empleado en función del perfil de los clientes vinculados a una entidad financiera con una muestra de 814.964 observaciones presentando técnicas de segmentación agrupadas por variables tales como: ingreso medio, sexo, edad, provincia y situación laboral, resaltando que para las personas con ingresos a cuenta propia "*son más difíciles de calcular ya que no se registran con conceptos de nómina o pensión*".

Así mismo se identificó un método scoring para clientes sin referencias crediticias (Espin & Rodríguez, 2013) en la cual indica que no es común encontrar metodologías de este tipo, el estudio es aplicado a una pequeña institución financiera mexicana y aborda una muestra de 4064 observaciones utilizando como técnica principal CHAID (Chi-squared automatic interaction detection) la cual se diferencia de otros algoritmos binarios debido a que puede formar segmentos con más de dos categorías al mismo tiempo, el estudio concluye que la metodología aplicada se puede extender al campo comercial como modelo de propensión al consumo, en este sentido el modelo resalta que "*la información sociodemográfica proveniente de las solicitudes de crédito, es altamente probable de encontrar falsedad en la información*", y esta afirmación puede tener mayor hincapié cuando no existe forma de validar la información brindada por un cliente que tiene actividades a cuenta propia.

Adicionalmente, (Guataquí, García, & Rodrigruez, 2009) realiza una estimación de las diferencias de ingresos entre asalariados y trabajadores a cuenta propia, utilizando como base la Gran Encuesta Entegrada de Hogares (GEIH) para el año 2007 en Colombia con una muestra de 92.025 observaciones, a través de operaciones mincerianas en la que demuestra que es un error realizar estimación de ingresos en un mismo grupo entre asalariados y trabajadores a cuenta propia debido al riesgo de selección, sesgo de estimación y efecto diferencial sobre el ingreso.

En este sentido estudiar cómo se comportan los ingresos en trabajadores independientes permitirá tener precisión al momento de generar resultados en los modelos de crédito evitando el sobreendeudamiento del cliente, disminuyendo el riesgo de la operación y generando mayor inclusión financiera a personas que realizan actividades de carácter independiente, así mismo se establecerán bases para el análisis de crédito cuando no sea posible soportar los ingresos.

4. Marco teórico

4.1. Trabajador independiente

El (Código Sustantivo del Trabajo, 2020) en el artículo 34 define los contratistas independientes como “verdaderos patronos y no representantes ni intermediarios…, que prestan servicios en beneficios de terceros, por un precio determinado, asumiendo todos los riesgos, para realizarlos con sus propios medios y con libertad y autonomía técnica y directiva”, al no tener una estructura comercial formal y promocionando los servicios a cuenta propia, el inventario que requiere para ejecutar la actividad un independiente estará sujeto bajo demanda.

En (El empleo, 2020) para Colombia resumen la resolución 5858 de 2016 del Ministerio de Salud y Protección Social, indicando que “el calificativo trabajador independiente es usado para el pago de los aportes al sistema de seguridad social integral y parafiscales de una persona natural que realiza una actividad económica o presta sus servicios de manera personal y por cuenta propia”, clasificando tres tipos de trabajadores independientes y según el tipo de afiliación:

- Trabajador independiente con contrato de prestación se servicio superior a un mes
- Independiente con contrato de prestación de servicios
- Independiente voluntario

La política interna de la entidad financiera define a los trabajadores independientes como “aquella persona que no está vinculada mediante contrato de trabajo, y que su remuneración

consiste básicamente en honorarios, comisiones y servicios", de acuerdo con estas definiciones se utilizará el concepto de la entidad que no tiene en cuenta los contratos de trabajo para clasificar a una persona natural como independiente.

4.2. Ingresos

El ingreso está definido por la (Real Academia Española, 2020) como "ganar cierta cantidad dinero regularmente por algún concepto", según (Alvarado & Pinos, 2017) "es una aproximación cuantificable del bienestar de las personas", de esta manera a mayor frecuencia de actividades realizadas mayor cantidad de dinero recibido.

En este sentido, "el ingreso está sometido en el corto plazo a las oscilaciones de la oferta y la demanda. Crecen con los auges y se comprimen con las crisis". Según López 1996. citado por (Guataquí, García, & Rodrigruez, 2009) en la cual estudian el determinante de los ingresos laborales usando los datos de la Gran Encuesta Integrada de Hogares del Departamento Administrativo Nacional de Estadística -DANE para el año 2017, en la cual seleccionaron las 7 principales ciudades del estudio, con 92025 observaciones aplicando como técnica principal ecuaciones mincerianas y resaltando los siguientes hallazgos: La experiencia no es significativa como determinante de los ingresos, a mayor nivel educativo se reduce la probabilidad de que los hombres sean trabajadores a cuenta propia, los jóvenes tienen menor aversión al riesgo por lo tanto tienden a generar mayor emprendimiento a cuenta propia, teniendo limitantes de capital.

Con la definición anterior es válido resaltar que los ingresos de los independientes no serán

constantes en el tiempo, ya que se debe evaluar el sector económico en la cual ejercen la actividad y por ende determinar ingresos máximos, los cuales obedecerán a una franja de tiempo en la cual la actividad se encuentra en auge, por ejemplo: los comerciantes que venden comida tradicional como lo son tamales en época navideña; así mismo los ingresos mínimos se dan por factores externos que afectan los ingresos, como sucede a los comerciantes que venden alimentos tipo fritos (empanadas – arepas …) de manera ambulante, un ejemplo de ello se aplica en las universidades, en la cual las ventas dependen de la cantidad de estudiantes que circulen en la zona de influencia.

En este contexto los ingresos de los independientes no siempre se pueden promediar, debido a que las estructuras se adaptan a las circunstancias económicas y del entorno, aun así, el mercado de los independientes es muy apetecido debido a las tasas de interés que se manejan.

Para este estudio, teniendo en cuenta las circunstancias anteriores se entenderán los ingresos como: *"El dinero que recibe una persona por realizar alguna actividad y proporcionan un beneficio personal, ya sea de carácter ocasional o fijo, sin tener en cuenta los costos y gastos ocasionados para cumplir la actividad generadora del rubro".*

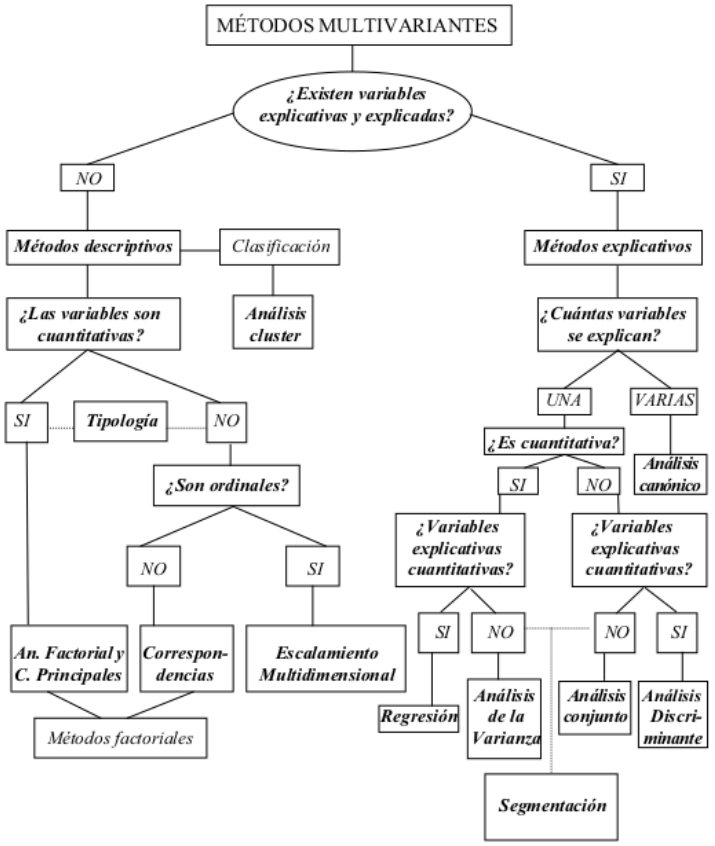
4.3. Análisis multivariado

El análisis multivariado consiste en emplear métodos matemáticos y estadísticos con el fin de analizar la relación entre variables de un conjunto de datos de manera agrupada, estos métodos son capaces de explicar un fenómeno con múltiples variables que los métodos estadísticos univariantes

y bivariantes son incapaces de explicar.

La técnica utilizada para realizar un análisis multivalente depende si se tiene una variable respuesta, el tipo de variable respuesta y/o variables explicativas, en la Figura 1, se puede observar su clasificación dependiendo de las características que contiene el conjunto de datos a analizar.

Figura 1: Métodos Multivariantes



Adaptado de: Pérez, 2004. Técnicas de Análisis Multivariante de Datos Aplicaciones con SPSS. Madrid, Pearson Prentice hill.

4.4. Análisis de correspondencia

Una de las técnicas estadísticas más utilizadas en el análisis de datos, es el análisis de tablas de

contingencia, esta es utilizada para evidenciar que tan relacionadas se encuentran dos variables, a través de la distribución de porcentajes de las categorías de una variable sobre las categorías de otra variable, para esto es necesario identificar las variables dependientes y e independiente con el objetivo de ubicar la primera en las filas y la segunda en las columnas. Figura 2.

Figura 2: Tabla de Contingencia

	A	No A	Total
B	a	b	a+b
No B	c	d	c+d
Total	a+c	b+d	

Adaptado de: Cañadas, Batanero, Contreras, & Arteaga. Estrategias en el estudio de la asociación en tablas de contingencia por estudiantes de psicología. Scielo 2011

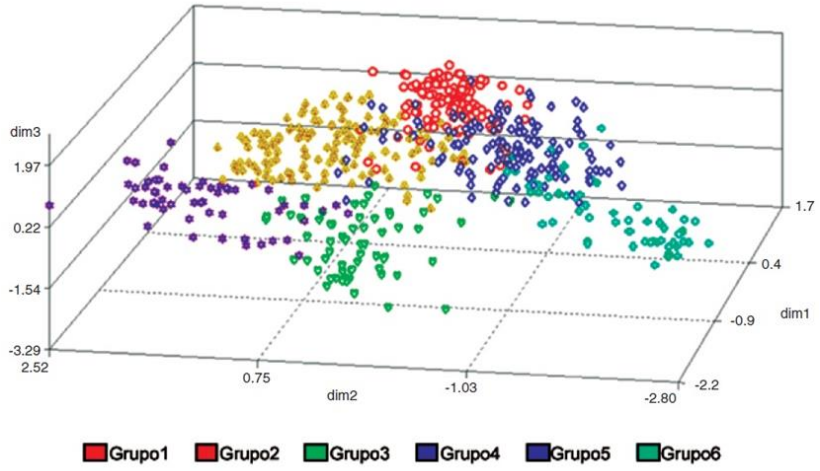
El análisis de correspondencia es una técnica utilizada para representar las tablas de contingencia, es muy similar a la de componentes principales, pero utilizando como fuente de información no los individuos sino las tablas de contingencia, usando como distancia chi-cuadrado.

Las variables utilizadas en esta técnica se deben conformar por categorías, por lo que deben ser variables cualitativas, en las tablas de contingencia estas variables se cruzan con las otras variables para indicar la frecuencia en cada categoría.

Una vez aplicada la técnica, las variables categóricas se evidencian en conglomerados a través de la varianza, formados por las categorías iniciales con varianza mínima entre ellos y máxima entre los grupos. El objetivo es poner gráficamente las relaciones de las categorías en las variables,

en el análisis de correspondencia se muestra un punto por cada fila y un punto por cada columna, siendo estos la proyección de las categorías en la tabla de contingencia, en un espacio de dos o tres dimensiones. Las relaciones se pueden observar con la formación de los conglomerados, como se muestra en la gráfica 4, basada en un estudio realizado a la variabilidad genética de la yuca cultivada por pequeños agricultores de la región Caribe de Colombia. Figura 3.

Figura 3: Análisis de correspondencia



Adaptado de: Alzate, Vallejo, Lascano, Pérez, & Fregene, 2010. Variabilidad genética de la yuca cultivada por pequeños agricultores de la región Caribe de Colombia. Revistas Unal – Acta Agronómica, 385.

4.5. Modelos de regresión

La regresión se utiliza para referirse a la explicación que tiene la predicción de una variable sobre otra u otras (variables x) cuando estas cambian, a esta última se conoce como variable respuesta (variable y) y puede ser cuantitativa o dicotómica, para variables cuantitativas se emplea la regresión lineal que predice el valor medio de la variable “y”, para variables dicotómicas que tiene como valores solamente la ausencia o presencia de una característica del sujeto se utiliza la regresión logística que predice la proporción de una de las dos categorías. Para nuestro caso se

utiliza el modelo de regresión lineal.

4.6. Regresión lineal

Cuando se tienen dos variables, la variable x en el eje horizontal y la variable y en el eje vertical del plano cartesiano y en estas se observan una nube de puntos distribuidos de forma lineal, se puede determinar si existe una relación entre las dos variables, para ello se calcula el coeficiente de correlación, el cual es propuesto de muchas formas y uno de los más usados es el de Pearson, representado en la siguiente formula (Hernández, y otros, 2018) :

$$r = \frac{\frac{1}{n} * \sum(x_i - x_m) * (y_i - y_m)}{\sqrt{\left(\frac{1}{n} * \sum(x_i - x_m)^2\right) * \left(\frac{1}{n} * \sum(y_i - y_m)^2\right)}}$$

En el numerador se calcula la covarianza multiplicando cada valor de x menos su media (x_m) y a su vez se multiplica cada valor de “y” menos su media, para dividir el resultado sobre el número de individuos de la muestra. El denominador se calcula multiplicando la varianza de “x” y de “y”, para calcular su raíz cuadrada.

El resultado de la correlación son valores $-1 < r < 1$, si este es mayor a 0 la correlación es positiva, y es más fuerte cuando más se acerque a 1. Si el valor es menor a 0 indica que la correlación es negativa y es más fuerte cuando más se acerque a -1. Si el valor es 0 entonces no existe correlación entre las variables.

Si existe una correlación significativa entre las variables, se puede determinar la recta que mejor se ajuste a la nube de puntos en el plano cartesiano, a través de la regresión lineal, esta recta está definida por la siguiente formula:

$$\hat{y} = a + bx$$

Donde "ŷ" es la variable dependiente explicada por la variable "x" que es independiente, los valores en a y b son parámetros definidos para ajustar la mejor recta que explique la variable y.

El parámetro "a" es el valor de "y" cuando "x" vale 0 y es el punto donde la recta cruza el eje vertical. Este parámetro se calcula con la siguiente formula:

$$a = \bar{y} + b\bar{x}$$

El parámetro b determina la inclinación de la recta y se calcula así (Universidad Nacional Autónoma de México, 2020):

$$b = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}$$

4.7. Valores tipificados

Cuando se tienen varias variables en un conjunto de datos, las variables cuantitativas pueden tomar valores en distintas medidas, por lo que los análisis que se hagan estos datos pueden ser sesgados por estos valores que difieren de una variable a otra. Los valores tipificados son el resultado de una técnica que consiste en transformar una variable en valores que se encuentren en una escala donde la media es 0 y el 99.73% de los datos se encuentren en valores entre -3 y 3, si se aplica la técnica a todas las variables cuantitativas del conjunto de datos, se obtienen valores en

la misma escala. La fórmula para aplicar esta técnica es la siguiente (Hernandez, 2012):

$$Z = \frac{X - \bar{x}}{s'}$$

Donde $\bar{x}$ es la media de la variable y s' la desviación estándar de la variable.

5. Descripción de la muestra

Los datos disponibles corresponde a una muestra de 31.940 registros de personas que tienen crédito vigente en la entidad financiera y obtuvieron el desembolso del crédito en un horizonte de tiempo de 4 años (2015-2019), los sujetos seleccionados pertenecen a líneas de crédito de trabajadores independientes y reportaron ingresos mensuales desde un 70% hasta 10 Salarios Mínimos Mensual Legal Vigente - SMMLV, manteniendo la relación del salario mínimo para cada año evaluado, con antigüedad en la actividad laboral de 1 hasta 50 años, ingresos de la actividad desde 70% de un SMMLV hasta 50 SMMLV y activos de 1 a 500 SMMLV.

En la descripción se han seleccionado 13 variables de interés que ayudaran a explicar cómo estimar los ingresos de los trabajadores independientes:

Variable 1 – Ingresos: Es la variable respuesta y registra la cantidad de dinero que recibe un trabajador independiente de manera mensual, el dato esta expresado en SMMLV del año desembolso. Tabla 1.

Tabla 1
Comportamiento de los ingresos

Ingresos	
Mínimo	0,70
1° Cuartil	1,45
Mediana	2,12
Media	2,69
3° Cuartil	3,37
Máximo	10,00

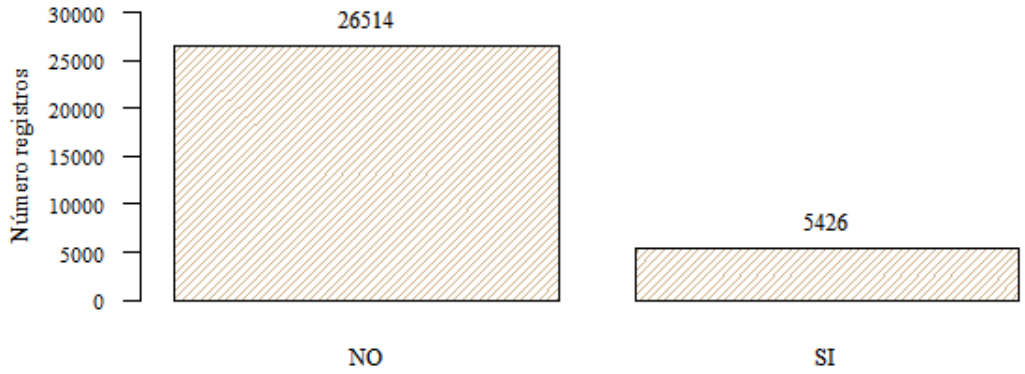
El 40.15% de la muestra corresponde a personas con ingresos de 1 a 2 SMMLV y el 82.15% de los registros tiene ingresos de hasta 4 salarios. Tabla 2.

Tabla 2
Frecuencia individual y acumulada de los ingresos

Rango Ingresos	Cantidad	% Part.	% Part. Acumulada
(< 1)	1.964	6,15%	6,15%
(1 - 2)	12.823	40,15%	46,30%
(>2 - 3)	7.553	23,65%	69,94%
(>3 - 4)	3.899	12,21%	82,15%
(>4 - 7)	4.474	14,01%	96,16%
(>7 - 10)	1.227	3,84%	100,00%
Total	31.940	100,00%	

Variable 2 – Otros ingresos: Corresponde a honorarios, salarios y otros ingresos, que no pueden ser soportados o no son certificados según la actividad económica. La variable es dicotómica e indica si un sujeto posee o no otros ingresos, en la figura 4 se observa que la muestra contiene 26.514 personas que no poseen ingresos diferentes a la actividad económica principal.

Figura 4: Distribución de otros ingresos



La participación de las personas que tienen otros ingresos es del 16,99%, la cual es baja respecto a las personas sin otros ingresos que tiene una participación de 83.01%. (Tabla 3).

Tabla 3
Frecuencia individual y acumulada de Otros ingresos

Otros Ingresos	Cantidad	% Part	% Part. Acumulada
NO	26.514	83,01%	83,01%
SI	5.426	16,99%	100,00%
Total	31.940	100,00%	

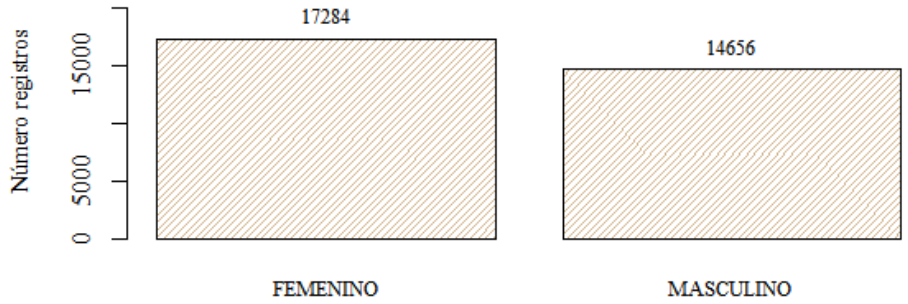
Variable 3 – Género: Establece el género del solicitante, la composición femenina es del 54.11% y masculina del 45.89%. (Tabla 4)

Tabla 4
Frecuencia individual y acumulada del Género

Género	Cantidad	% Part	% Part. Acumulada
Femenino	17.284	54,11%	54,11%
Masculino	14.656	45,89%	100,00%
Total	31.940	100,00%	

En la distribución gráfica se observa que los datos femeninos corresponden a 17.284 registros y masculino a 14.656 respectivamente. (Figura 5)

Figura 5: Distribución género



En este sentido en tabla 5, se encuentra que el 66.47% de los registros femeninos se distribuyen entre 1 y 3 SMMLV y la participación en SMMLV superiores (mayor a 7) es del 2.93%; para el género masculino la concentración entre 1 y 3 SMMLV es del 60.63% y SMMLV mayores a 7 es del 4.91%; ligeramente superior al de las mujeres.

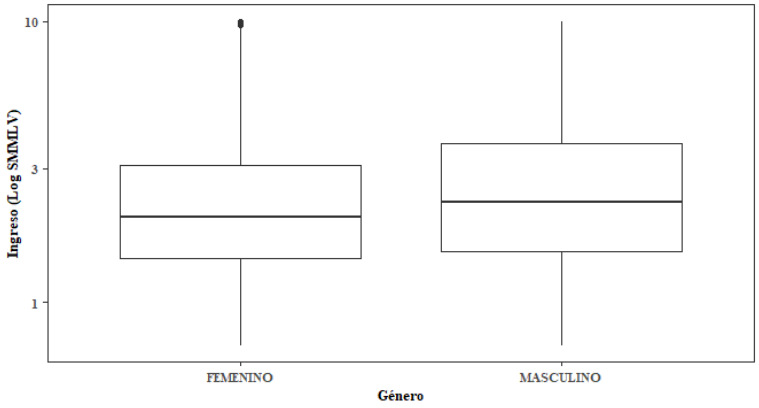
Tabla 5
Distribución de los ingresos respecto al género

Ingreso / Género	Femenino	Masculino
(< 1)	7,21%	4,89%
(1 - 2)	42,37%	37,52%
(>2 - 3)	24,10%	23,11%
(>3 - 4)	11,59%	12,94%
(>4 - 7)	11,79%	16,63%
(>7 - 10)	2,93%	4,91%
Total	100,00%	100,00%

En la figura 6, se observa un rango intercuartílico superior para el género masculino y el género

femenino con una mediana inferior a la masculina pero una cola más pesada.

Figura 6: Boxplot género



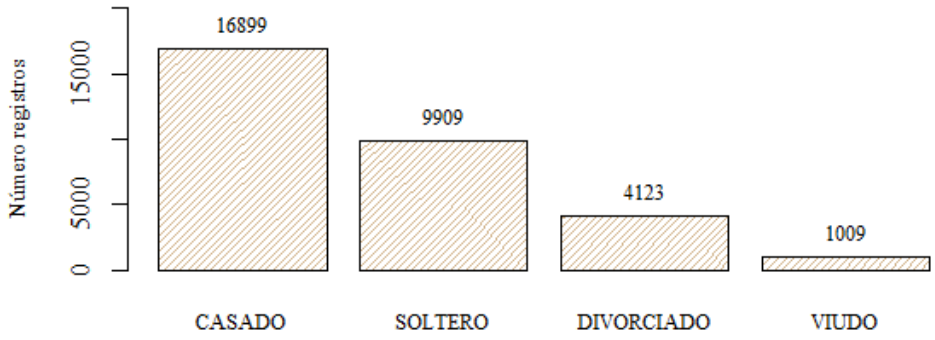
Variable 4 -Estado Civil: Indica el estado marital del trabajador al momento de la solicitud del crédito, la variable se compone de cuatro categorías (Tabla 6), Casado que concentra el 52.91%, Soltero (31.02%), Divorciado (12.91%) y Viudo (3.16%), el 83.93% de los registros se concentran entre Casados y Solteros.

Tabla 6
Frecuencia individual y acumulada del estado civil

Estado Civil	Cantidad	% Part	% Part. Acumulada
Casado	16.899	52,91%	52,91%
Soltero	9.909	31,02%	83,93%
Divorciado	4.123	12,91%	96,84%
Viudo	1.009	3,16%	100,00%
Total	31.940	100,00%	

En la figura 7, los Casados predominan con 16.899 registros de los 31.940 totales, seguido de los Solteros 9.909, Divorciado 4.123 y finalmente Viudos con 1.009 registros.

Figura 7: Distribución del estado civil



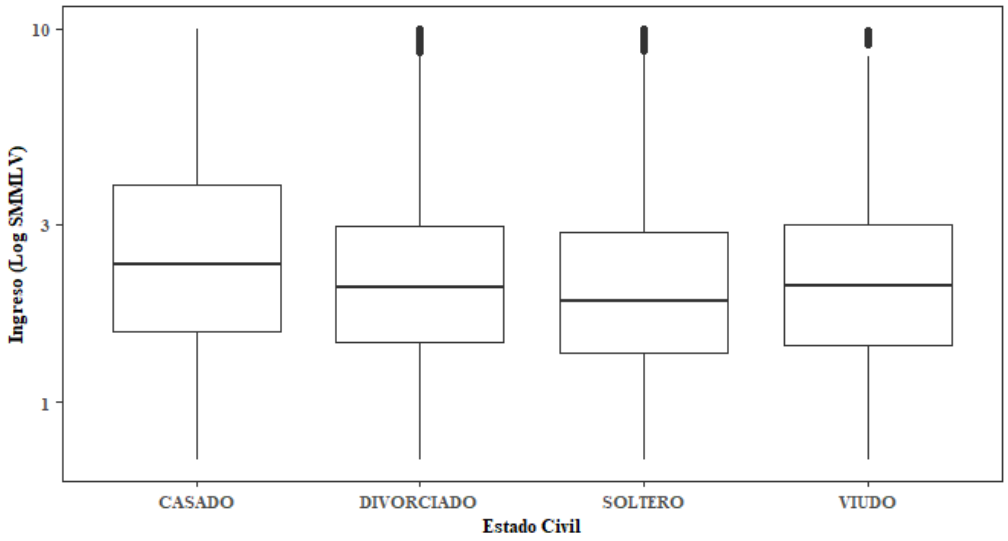
La tabla 7 evalúa la distribución de ingresos respecto al estado civil, el perfil fila con menor participación a un salario mínimo tienen (5.23%) y corresponde a los casados, en perfil columna de casados el 58.56% participa con ingresos entre 1 y 3 SMMLV, en las demás categorías para perfil columna se observan concentraciones del 70% de los ingresos inferiores a 3 SMMLV.

Tabla 7
Distribución de los ingresos respecto al estado civil

Ingreso / Estado Civil	**Casado**	**Divorciado**	**Soltero**	**Viudo**
(< 1)	5,23%	6,35%	7,39%	8,52%
(1 - 2)	35,24%	42,64%	47,51%	39,84%
(>2 - 3)	23,32%	26,66%	22,64%	26,76%
(>3 - 4)	13,51%	11,86%	10,22%	11,30%
(>4 - 7)	17,44%	10,16%	9,99%	11,69%
(>7 - 10)	5,26%	2,33%	2,25%	1,88%
Total	**100,00%**	**100,00%**	**100,00%**	**100,00%**

La figura 8, muestra que los casados mantienen una mediana superior respecto a otros estados civiles, con una variación intercuartílica igualmente superior, para las demás categorías hay solapamiento, con datos atípicos y colas pesadas.

Figura 8: Boxplot estado civil



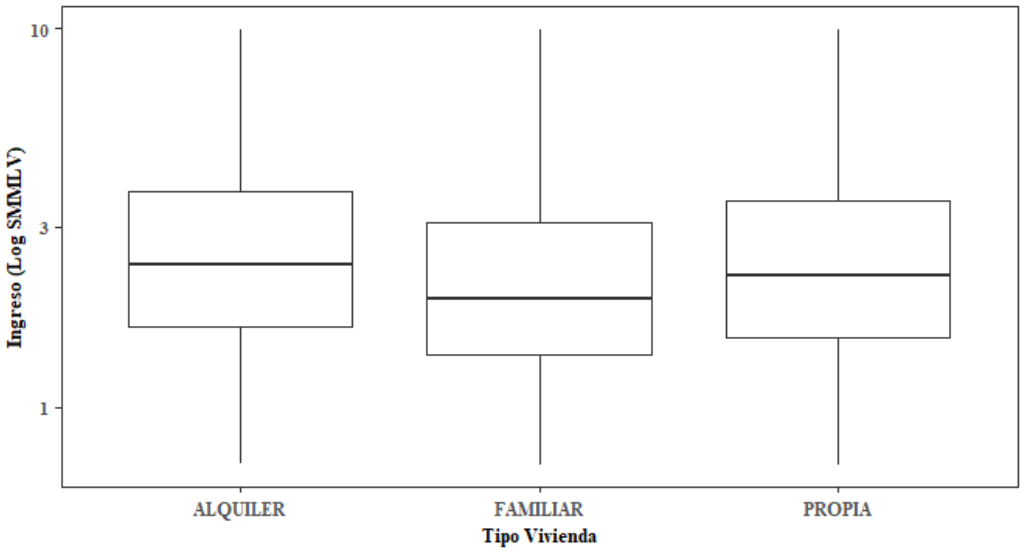
Variable 5 – Tipo de Vivienda: Variable cualitativa que indica el tipo de apoderamiento sobre la vivienda actual del deudor, la tabla 8 evidencia que el 49.40% de los registros poseen vivienda propia, el 41.91% vivienda familiar y el 8.69% de vivienda en alquiler.

Tabla 8
Frecuencia individual y acumulada del tipo de vivienda

Tipo de Vivienda	Cantidad	% Part	% Part. Acumulada
Propia	15778	49,40%	49,40%
Familiar	13386	41,91%	91,31%
Alquiler	2776	8,69%	100,00%
Total	31940	100,00%	

En la figura 9 se observa que la categoría con menos dispersión está en registros con vivienda familiar, además de que es la categoría con el primer cuartil más bajo, las otras categorías presentan solapamiento y una mediana similar.

Figura 9: Boxplot tipo de vivienda



Evaluando los ingresos respecto el tipo de vivienda (tabla 9), el perfil columna vivienda familiar tiene mayor participación en rango de ingresos de 1 a 2 SMMLV con el 44.72% y en rangos superiores a 7 SMMLV es del 3.19%; el perfil columna de la variable alquiler tiene menor participación en los rangos de ingresos hasta 2 SMMLV y en perfil columna de la variable vivienda propia tiene ingresos superiores mayores a 7 SMMLV.

Tabla 9
Distribución de los ingresos respecto al tipo de vivienda

Ing. / estado	Alquiler	Familiar	Propia
(< 1)	4,18%	7,43%	5,41%
(1 - 2)	34,58%	44,72%	37,25%
(>2 - 3)	25,50%	21,75%	24,93%
(>3 - 4)	14,19%	10,89%	12,97%
(>4 - 7)	17,58%	12,03%	15,06%
(>7 - 10)	3,96%	3,19%	4,37%
Total	**100,00%**	**100,00%**	**100,00%**

Variable 6 – Estrato: Establece el nivel socioeconómico de la vivienda actual del trabajador

independiente, la tabla 10 muestra concentración de registros en el estrato 2 (39.11%) y estrato 3 (33.88%), agrupando el 72.99% de los trabajadores independientes, los tres primeros estratos concentran el 85.22% de los registros totales, con poca participación de los estratos 5 (1.13%) y estrato 6 (0.23%).

Tabla 10
Frecuencia individual y acumulada del estrato socioeconómico

Estrato Socioeconómico	Cantidad	% Part	% Part. Acumulada
Est - 1	3.906	12,23%	12,23%
Est - 2	12.491	39,11%	51,34%
Est - 3	10.822	33,88%	85,22%
Est - 4	4.286	13,42%	98,64%
Est - 5	360	1,13%	99,77%
Est - 6	75	0,23%	100,00%
Total	31.940	100,00%	

Los ingresos respecto a los estratos se muestran en la tabla 11, el perfil columna estrato 5 concentra el 30.28% en ingresos entre 4 y 7 SMMLV, siendo el predominante en todas las filas; para el perfil columna de estratos 1 al 4 el nivel de ingresos entre 1 y 2 SMMLV se ubica un rango entre 38.89% y el 41.18% superior a los estratos 5 y 6, el estrato 6 no registra datos con ingresos inferiores a 1 SMMLV.

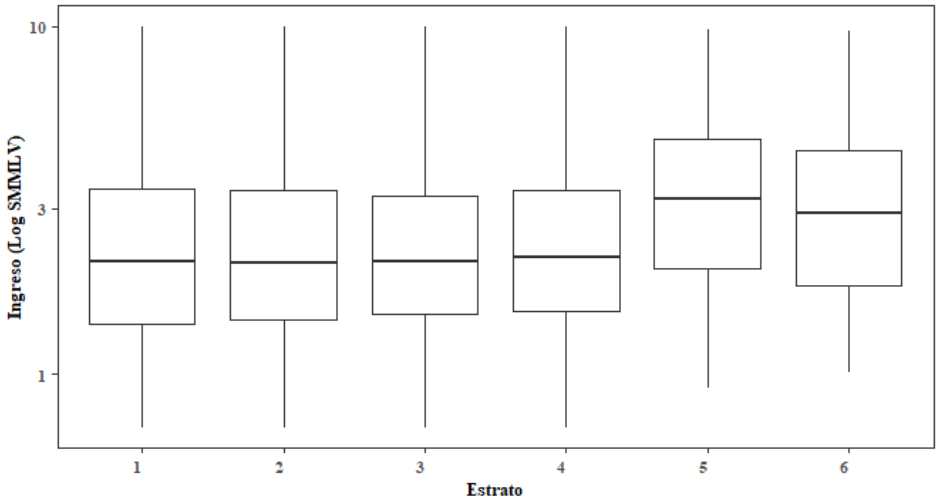
Tabla 11
Distribución de los ingresos respecto al estrato socioeconómico

Ingreso / estrato	Est - 1	Est - 2	Est - 3	Est - 4	Est - 5	Est - 6
(< 1)	8,58%	7,16%	4,96%	4,55%	0,83%	0,00%
(1 - 2)	38,89%	40,43%	41,18%	39,31%	24,17%	34,67%
(>2 - 3)	21,22%	22,46%	25,01%	26,08%	22,50%	17,33%
(>3 - 4)	13,13%	12,09%	11,94%	11,95%	16,39%	17,33%
(>4 - 7)	15,16%	14,20%	12,99%	13,51%	30,28%	18,67%

Ingreso / estrato	Est - 1	Est - 2	Est - 3	Est - 4	Est - 5	Est - 6
(>7 - 10)	3,02%	3,67%	3,92%	4,60%	5,83%	12,00%
Total	100,00%	100,00%	100,00%	100,00%	100,00%	100,00%

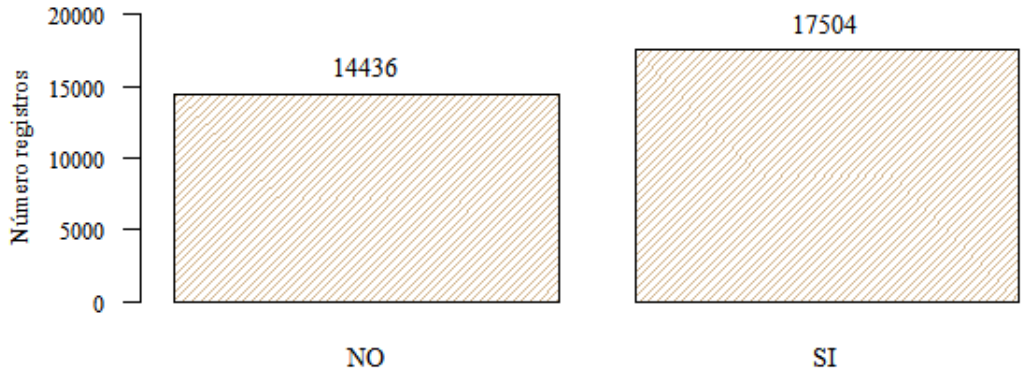
La figura 10 denominada distribución por estrato refleja que los estratos 1 al 4 presentan rangos intercuartílicos y mediana similares, a diferencia del estrato 5 y 6 que presentan una mediana superior comparada con los estratos inferiores.

Figura 10: Boxplot estrato socioeconómico



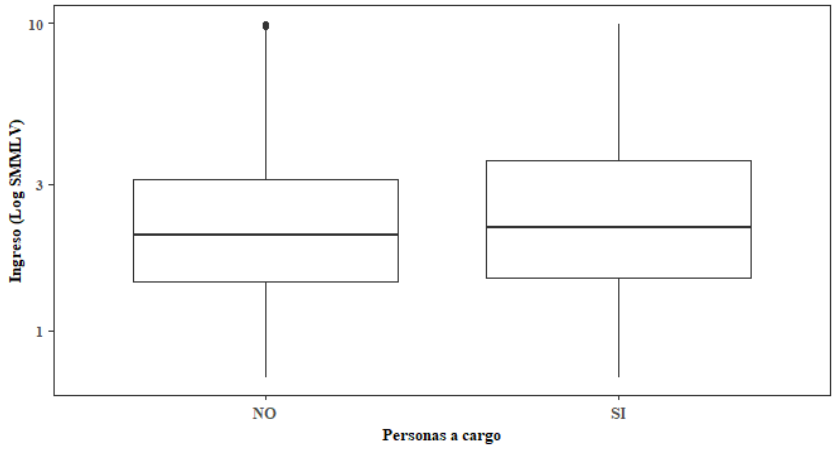
Variable 7 – Personas a cargo: La variable es dicotómica e indica si la persona tiene o no personas a cargo. Se realiza gráfico de barras (figura 11) donde se observa que predomina los registros con personas a cargo y 17.504 registros del total de la muestra.

Figura 11: Distribución de personas a cargo



En la figura 12 se observa que los registros con personas a cargo presentan más dispersión que los que no tienen personas a cargo, aunque la mediana se encuentra ubicada en un nivel similar.

Figura 12: Boxplot personas a cargo



La concentración de Personas a cargo en los perfiles columna se encuentran en ingresos de 1 a 2 SMMLV, seguido de las personas con ingresos entre 2 y 3 SMMLV. (tabla 12)

Tabla 12
Distribución de los ingresos respecto a tener personas a cargo

Rango Utilidades	Con personas a cargo	Sin personas a cargo
(< 1)	6,11%	6,20%
(1 - 2)	38,34%	42,34%
(>2 - 3)	22,62%	24,90%
(>3 - 4)	12,57%	11,77%
(>4 - 7)	15,76%	11,88%
(>7 - 10)	4,60%	2,92%
Total	100,00%	100,00%

Variable 8 -Escolaridad: Esta determinada por el nivel educativo de los solicitantes, según (tabla 13) la educación Básica concentra el 85,31% del total de registros, los registros restantes se distribuyen en Técnica (8.35%), Profesional (5.38%) y sin Ningún tipo de educación (0.96%).

Tabla 13
Frecuencia individual y acumulada en nivel de escolaridad

Escolaridad	Cantidad	% Part	% Part. Acumulada
Ninguna	306	0,96%	0,96%
Básica	27.248	85,31%	86,27%
Técnico	2.667	8,35%	94,62%
Profesional	1.719	5,38%	100,00%
Total	31.940	100,00%	

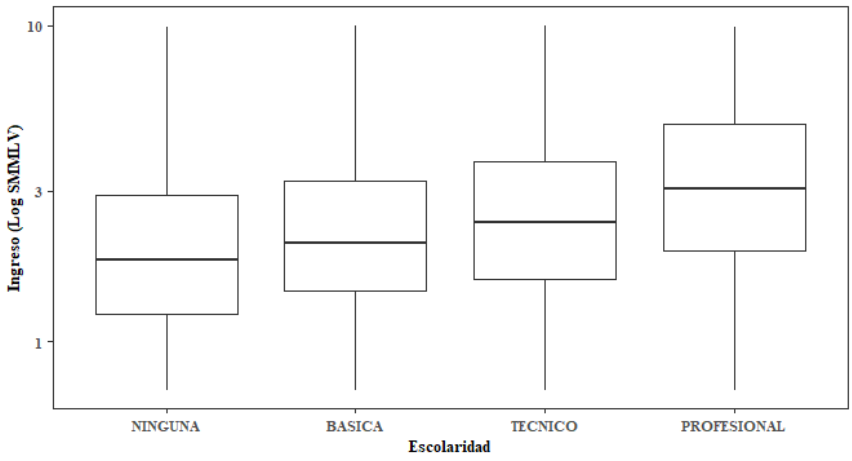
La tabla 14 permite observar que el perfil columna denominada Ninguna tiene un 13.07% de registros que ganan menos de 1 SMMLV, el perfil fila mayor a 4 y entre 7 salarios a medida que aumenta el nivel de escolaridad aumenta los ingresos y el perfil fila para SMMLV superior 7 en trabajadores sin ningún tipo de educación es del 4.25%, mayor que la educación básica.

Tabla 14
Distribución de los ingresos respecto al nivel de escolaridad

Ingreso / escolaridad	Ninguna	Básica	Técnico	Profesional
(< 1)	13,07%	6,48%	4,16%	2,73%
(1 - 2)	41,18%	41,57%	36,07%	23,73%
(>2 - 3)	21,90%	23,69%	24,22%	22,45%
(>3 - 4)	7,19%	11,87%	14,21%	15,42%
(>4 - 7)	12,42%	13,02%	17,06%	25,13%
(>7 - 10)	4,25%	3,37%	4,27%	10,53%
Total	100,00%	100,00%	100,00%	100,00%

El Boxplot figura 13 identifica que la categoría educación profesional tiene una mediana superior respecto a las demás categorías, el nivel técnico, básico y ninguna educación tiene colas pesadas y datos atípicos dominantes.

Figura 13: Boxplot distribución por escolaridad



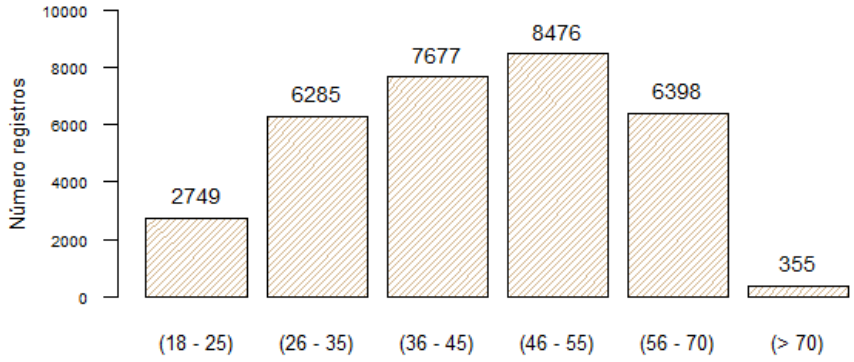
Variable 9 – Edad: Para acceder a productos crediticios es necesario ser mayor de edad – 18 años, la muestra contiene personas desde 18 hasta 81 años, la media de la muestra es de 44.16 años y se encuentra muy cercana a la mediana que es de 45 años (tabla 15).

Tabla 15
Distribución edad

Edad	
Mínimo	18,00
1° Cuartil	34,00
Mediana	45,00
Media	44,16
3° Cuartil	54,00
Máximo	81,00

Cuando se realizan rangos de edad a la muestra y se grafica en barras, se observa (figura 14) que las personas con menor participación son aquellas que tienen edad mayor a 70 años, seguido de las personas con edad entre 18 y 25 años, quienes tiene mayor participación son los registros entre 46 y 55 años.

Figura 14: Distribución de las edades



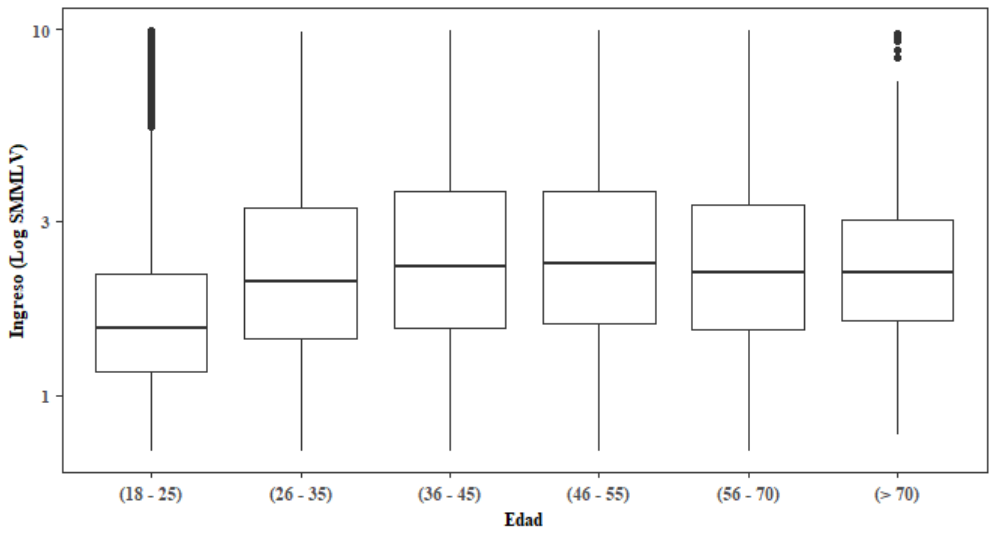
Para confirmar lo que se observa en el gráfico de barras, se construye tabla de frecuencia (tabla 16) donde se reitera que las personas con edades entre 46 y 55 años tienen una participación de 26.54% sobre el total de la muestra y las personas entre 18 y 25 años tienen el 8.61%. Se evidencia que edad mayor a 70 años tiene la menor participación con el 1.11% del total.

Tabla 16
Frecuencia individual y acumulada en rango de Edad

Edad	Cantidad	% Part.	% Part. Acumulado
(18 - 25)	2.749	8,61%	8,61%
(26 - 35)	6.285	19,68%	28,28%
(36 - 45)	7.677	24,04%	52,32%
(46 - 55)	8.476	26,54%	78,86%
(56 - 70)	6.398	20,03%	98,89%
(> 70)	355	1,11%	100,00%
Total	31.940	100,00%	

En la figura 15, se realizó diagrama de caja y bigotes con las variables edad e ingresos, la mediana tiene un comportamiento similar para los rangos mayores a 25 años, también se observa que los rangos de edades de 26 a 35 y de 56 a 70 tienen dispersión y mediana similar, al igual que los rangos de 36 a 45 y 46 a 55 años. Los sujetos con edad de 18 a 25 años presentan la menor dispersión de los datos.

Figura 15: Boxplot de los rangos de edad



Variable 10 – Antigüedad laboral: Está determinada en meses e indica el tiempo que un

trabajador ha ejercido la actividad de independiente, en la tabla 17 se observa que más del 67.21% de los datos se encuentran en rangos superiores a 60 meses, siendo el rango de 61 a 120 meses el de mayor participación con el 28.83%.

Tabla 17
Frecuencia individual y acumulada en rango de antigüedad laboral

Rango antigüedad	Cantidad	% Part.	% Part. Acumulado
(Hasta 24)	3.193	10,00%	10,00%
(25 - 60)	7.282	22,80%	32,80%
(61 - 120)	9.207	28,83%	61,62%
(121 - 240)	8.356	26,16%	87,78%
(>240)	3.902	12,22%	100,00%
Total	31.940	100,00%	

El nivel de ingresos versus la antigüedad se observa en la tabla 18, de allí se infiere que a medida que aumenta la experiencia incrementa el nivel de ingresos, en el perfil fila mayor a 4 SMMLV se observa la relación de crecimiento gradual.

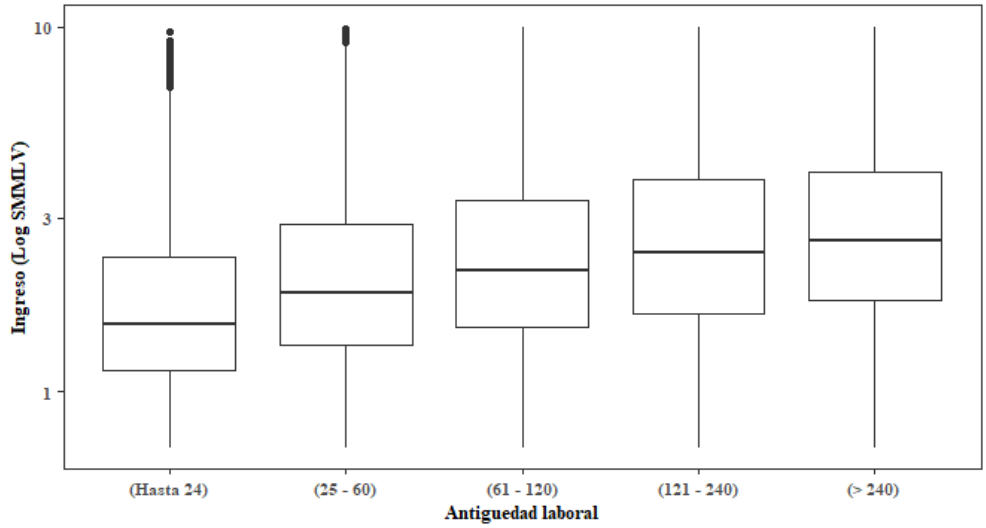
Tabla 18
Distribución de los ingresos respecto a antigüedad laboral

Rango Ingresos	(Hasta 24)	(25 - 60)	(61 - 120)	(121 - 240)	(>240)
(< 1)	14,59%	8,12%	5,09%	3,78%	3,13%
(1 - 2)	52,74%	46,72%	40,64%	34,36%	28,81%
(>2 - 3)	17,16%	21,88%	24,49%	25,06%	27,24%
(>3 - 4)	7,02%	10,19%	12,27%	14,16%	15,89%
(>4 - 7)	7,05%	10,63%	13,47%	17,86%	19,04%
(>7 - 10)	1,44%	2,47%	4,03%	4,79%	5,89%
Total	100,00%	100,00%	100,00%	100,00%	100,00%

En la figura 16, Boxplot por antigüedad laboral se observa que la mediana incrementa con la antigüedad laboral y dispersión de los datos, la antigüedad laboral hasta 24 meses tiene el 75 de

los datos con ingresos inferiores a 2.5 SMMLV.

Figura 16: Boxplot antigüedad laboral



Variable 11 – Activos: Indica el valor en salarios mínimos al momento de la solicitud del crédito de las propiedades del solicitante, los valores oscilan entre 1 y 500 SMMLV, la media se encuentra alejada de la mediana, aunque se encuentra dentro del 1° y 3° cuartil. (tabla 19).

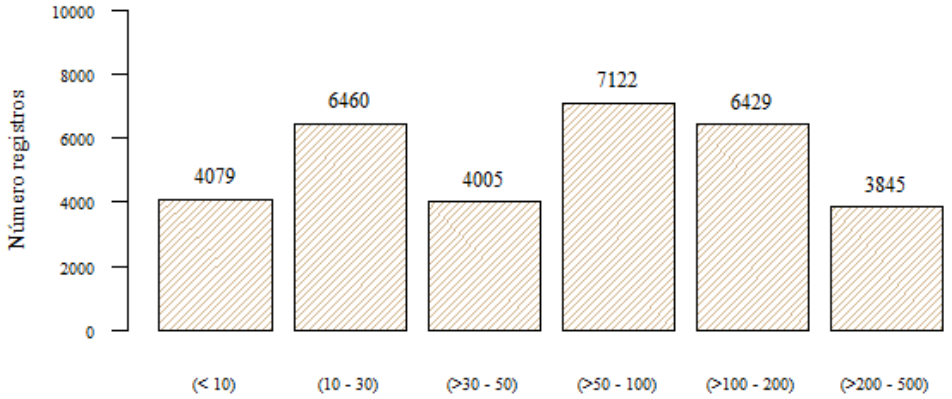
Tabla 19
Distribución activos

Activos	
Mínimo	1,00
1° Cuartil	20,49
Mediana	58,42
Media	90,51
3° Cuartil	120,76
Máximo	500,00

Se observa que no existe un rango que predomine (figura 17), sin embargo, se encuentra que

existe una gran participación en los activos entre 51 y 200 SMMLV.

Figura 17: Distribución por rango de activos



El rango de activos entre 50 y 100 salarios concentra el 22.30% de los registros, seguido de activos con rango entre 10 y 30, la menor participación es para activos menores a 10 con el 12.77%. (tabla 20).

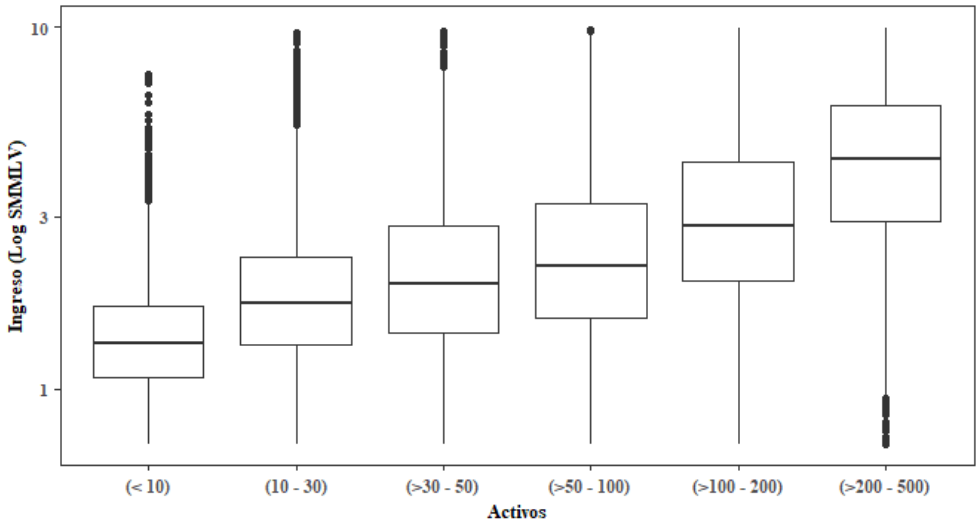
Tabla 20
Frecuencia individual y acumulada de los activos

Rango Activos	Cantidad	% Part.	% Part. Acumulado
(< 10)	4.079	12,77%	12,77%
(10 - 30)	6.460	20,23%	33,00%
(>30 - 50)	4.005	12,54%	45,54%
(>50 - 100)	7.122	22,30%	67,83%
(>100 - 200)	6.429	20,13%	87,96%
(>200 - 500)	3.845	12,04%	100,00%
Total	31.940	100,00%	

En el Boxplot (figura 18), se observa que los registros con activos superiores a 200 SMMLV tienen valores atípicos de ingresos bajos, aunque su variabilidad es notable, así mismo se observa un comportamiento que indica que a mayores ingresos mayor es la variabilidad de los ingresos y

la mediana se posiciona en valores más altos.

Figura 18: Boxplot rango de activos



Variable 12 – Actividad económica: Se encuentra registrada con códigos CIIU (Clasificación Industrial Internacional Uniforme) y agrupa las actividades en cinco categorías principales, la mayor participación es para Comercio con el 44.38% de registros, las demás categorías mantienen una distribución similar. Tabla 21.

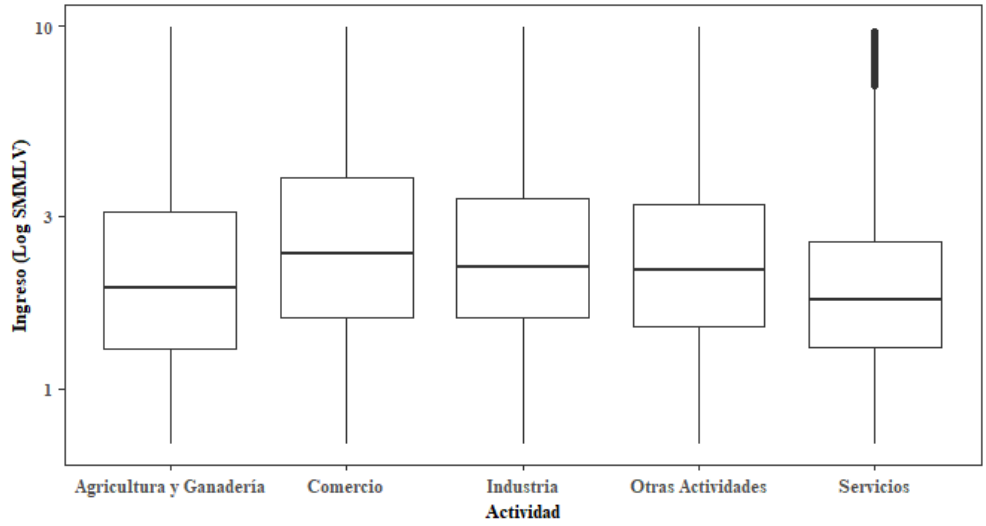
Tabla 21
Frecuencia individual y acumulada en agrupación CIIU

Nombre Actividad	Cantidad	% Part.	% Part. Acumulado
Agricultura y Ganadería	5.028	15,74%	15,74%
Comercio	14.175	44,38%	60,12%
Industria	3.634	11,38%	71,50%
Otras Actividades	4.290	13,43%	84,93%
Servicios	4.813	15,07%	100,00%
Total	31.940	100,00%	

En la figura 19 se observa que, la actividad con mayor variabilidad es comercio además de que

tiene la mediana más alta, por otro lado, la actividad con menor variabilidad es servicios, aunque son notables los valores atípicos que esta presenta.

Figura 19: Boxplot de la actividad económica



Todas las actividades económicas tienen concentrada la mayor participación en los ingresos entre 1 a 2 SMMLV, sin embargo, la actividad servicios contiene el 52.52% de los datos concentrados en este rango lo que explica que sea la categoría con menos variabilidad. (tabla 22).

Tabla 22
Distribución de los ingresos respecto actividad económica

Rango Ingresos	Agricultura y Ganadería	Comercio	Industria	Otras Actividades	Servicios
(< 1)	9,7%	5,28%	4,13%	4,20%	8,21%
(1 - 2)	43,4%	34,64%	39,05%	41,59%	52,52%
(>2 - 3)	20,7%	23,49%	27,13%	25,59%	22,79%
(>3 - 4)	11,3%	13,77%	11,45%	12,49%	8,85%
(>4 - 7)	12,3%	17,58%	14,12%	12,49%	6,54%

Rango Ingresos	Agricultura y Ganadería	Comercio	Industria	Otras Actividades	Servicios
(>7 - 10)	2,5%	5,24%	4,13%	3,64%	1,08%
Total	100,0%	100,00%	100,00%	100,00%	100,00%

Variable 13 – Zona: La variable Zona está distribuida en ocho (8) categorías y agrupa las agencias de la cooperativa según criterios propio de la entidad. Principalmente predomina la zona AMB o Área Metropolitana de Bucaramanga con una participación del 25.91% y es donde mayor presencia de agencias tiene la cooperativa, en segundo lugar la zona Cesar con una participación individual del 20.09% y una participación acumulada del 46.01%, es decir entre estas dos zonas se concentra alrededor del 50% de los datos; los valores restantes están distribuidos en las otras seis (6) zonas, Magdalena (15.94%), Centro (13.95%), Norte Santander (13.64%), Boyacá (8.57%), Cundinamarca (1.40%) y Atlántico (0.49%). Tabla 23.

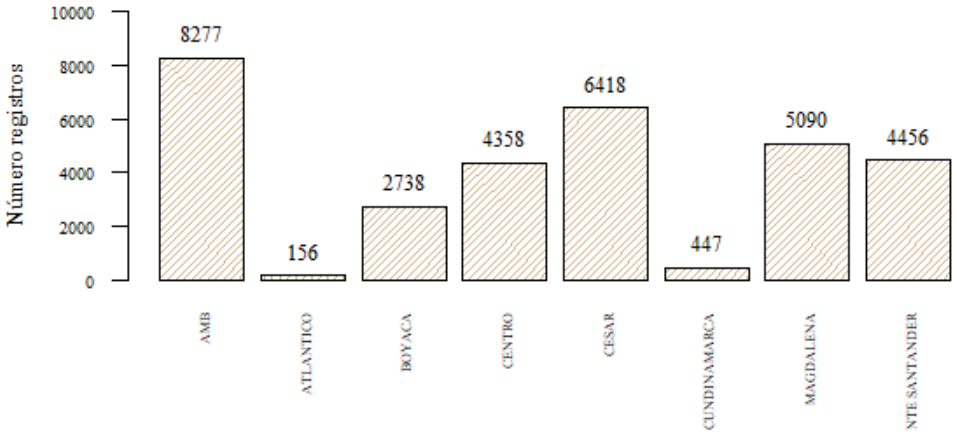
Tabla 23
Frecuencia individual y acumulada de las zonas

Zona	Cantidad	% Part	% Part. Acumulada
Amb	8.277	25,91%	25,91%
Cesar	6.418	20,09%	46,01%
Magdalena	5.090	15,94%	61,94%
Centro	4.456	13,95%	75,90%
Nte. Santander	4.358	13,64%	89,54%
Boyacá	2.738	8,57%	98,11%
Cundinamarca	447	1,40%	99,51%
Atlántico	156	0,49%	100,00%
Total	31.940	100,00%	

La figura 20, gráfica la distribución zonal, la zona AMB (8277) y Cesar (6418) son las principales zonas en número de registros, en contraste Atlántico (156) y Cundinamarca (447) son

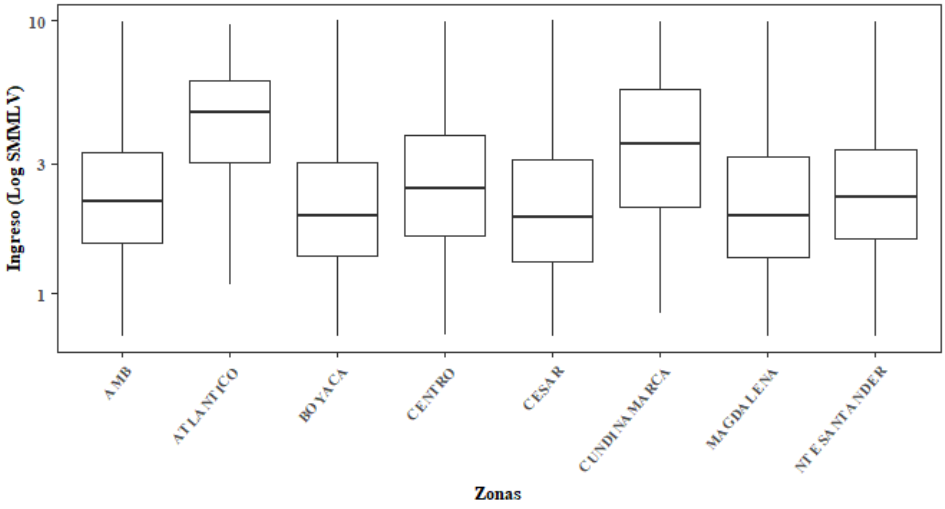
las que menor número de registro en trabajadores independientes.

Figura 20: Distibución de registros por zona



La figura 21 muestra el diagrama de caja y bigotes de zonas versus ingresos, la zona Atlántico tiene una mediana superior y sin cola superior pesada al igual que Cundinamarca, aunque esta última presenta una mayor variabilidad de los ingresos, por otra parte, se evidencia también que las zonas de Boyacá, Cesar y Magdalena presentan una distribución y mediana similar.

Figura 21: Boxplot de las zonas



6. Análisis de correspondencia múltiple

El proceso de transformación de variables aumenta la precisión del modelo, previamente se describieron 13 variables para realizar el análisis de correspondencia múltiple - ACM, sin embargo, en el proceso de construcción se identifica que en algunas variables se deben modificar las categorías de las variables iniciales (Tabla 24); a continuación, se detalla cada variable con sus respectivas categorías para aplicar en el modelo.

Tabla 24
Transformación de variables

Número de Variable	Nombre de Variable	Categorías iniciales	Categorías ajustadas
1	Rango Ingresos	(<1) (1 - 2) (> 2 - 3) (> 3 - 4) (> 4 - 7) (> 7 - 10)	(< 2 Ing) (2 - 3 Ing) (>3 - 4 Ing) (>4 Ing)
2	Otros Ingresos	Si No	Si No
3	Genero	Femenino Masculino	Femenino Masculino
4	Estado Civil	Casado Divorciado Soltero Viudo	Casado Divorciado Soltero Viudo
5	Tipo de Vivienda	Alquiler Familiar Propia	Alquiler Familiar Propia
6	Estrato Socioeconómico	Est - 1 Est - 2 Est - 3	Est 1-2 Est 3-4

Número de Variable	Nombre de Variable	Categorías iniciales	Categorías ajustadas
7	Personas a Cargo	Est - 4 Est - 5 Est - 6	Est 5-6
		Si	Si
		No	No
8	Escolaridad	Ninguna Básica	Básica
		Técnico	Técnico
		Profesional	Profesional
9	Edad	(18 - 25) (26 - 35)	(18 - 35 Años)
		(36 - 45) (46 - 55)	(36 - 55 Años)
		(56 - 70) (> 70)	(> 56 Años)
10	Antigüedad Laboral	(Hasta 24)	(Hasta 24 Ant)
		(25 - 60)	(25 - 60 Ant)
		(61 - 120)	(61 - 120 Ant)
		(121 - 240) (>240)	(>120 Ant)
11	Activos	(< 10) (10 - 30)	(< 30 SM Act)
		(>30 - 50) (>50 - 100)	(30 - 100 SM Act)
		(>100 - 200) (>200 - 500)	(>100 SM Act)
12	Actividad Económica	Agricultura y Ganadería	Agricultura y Ganadería
		Comercio	Comercio
		Industria	Industria
		Otras Actividades	Otras Actividades
		Servicios	Servicios
13	Zona	Amb	Amb
		Atlántico	Atlántico
		Boyacá	Boyacá
		Centro	Centro
		Cesar	Cesar
		Cundinamarca	Cundinamarca

Número de Variable	Nombre de Variable	Categorías iniciales	Categorías ajustadas
		Magdalena Nte Santander	Magdalena Nte Santander

En este sentido se realiza un proceso que va construyendo paso a paso la metodología hasta lograr agrupar y explicar de manera disiente un modelo que se ajuste a la problemática planteada, para ello se realiza el análisis en varios escenarios, observando resultados de la siguiente manera:

En el primer escenario se utilizan todas las variables (13 variables) para construir el modelo, esta etapa se ejecuta el ACM con las categorías ajustadas de las variables según tabla 24, con excepción de la variable ingresos y activos que son incluidas con las categorías iniciales; de ello se obtiene un poder explicativo total del 11,01% en las 2 primeras dimensiones, pero no logra representar grupos claramente definidos.

En el segundo escenario se excluye la variable Zona teniendo en cuenta la cantidad de categorías que esta contiene, allí se observa que el poder explicativo total del modelo en las dos primeras dimensiones es 13,14%, sin embargo, las variables ingresos y activos se agrupan en un mismo rango, razón por la cual se reagrupan según tabla 24 (Categoría ajustada).

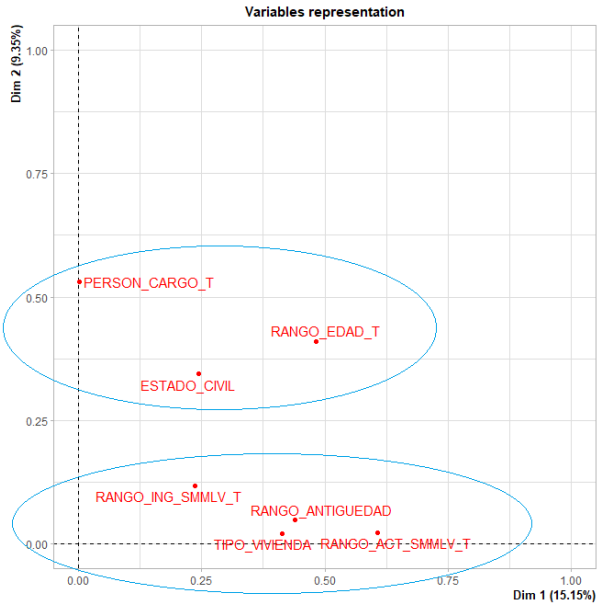
En el tercer escenario se observa el aporte de las variables a las tres (3) primeras dimensiones del modelo, donde se evidencia que la variable Otros Ingresos no brinda un aporte significativo en las dimensiones y la variable actividad tiene mayor aporte solo en la tercera dimensión, por lo que se decide ejecutar otro escenario excluyéndola la variable del análisis.

En un siguiente escenario se verifica el límite de contribución con el promedio de todas las contribuciones (Aldas & Uriel, 2017), asumiendo que todas las variables contribuyan igual; de esta forma la variable estrato no sobrepasa este límite en ninguna de sus categorías en las 3 primeras dimensiones, por lo cual se excluye.

Finalmente se observa un mejor entendimiento del comportamiento de los trabajadores independientes contrastando sus características sociodemográficas y económicas, sin embargo, hay variables que tienen bajo aporte: Género y Escolaridad, por lo que se decide ejecutar un escenario final excluyendo estas variables.

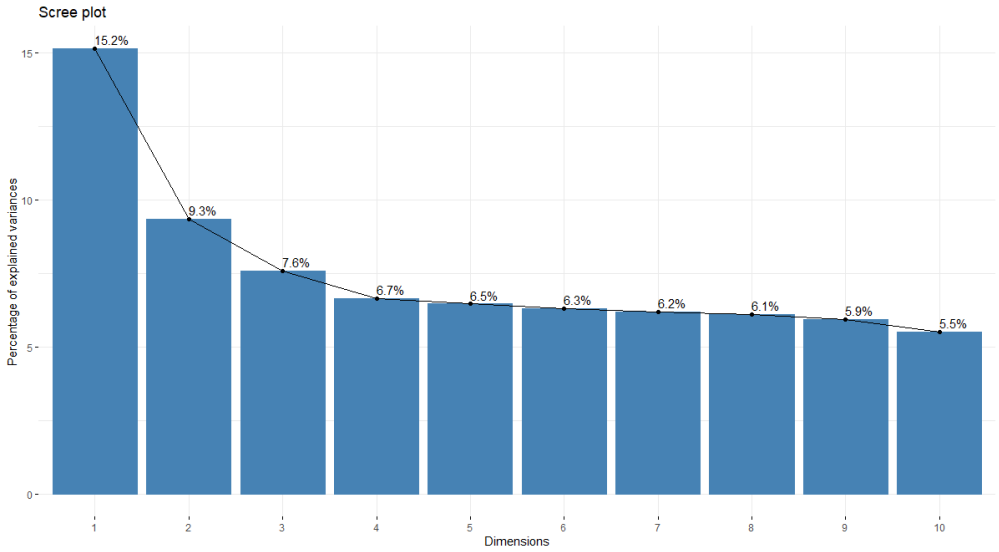
Se ejecuta el modelo con 7 variables de las 13 inicialmente planteadas (figura 22), con dos agrupaciones claramente definidas, la primera sociodemográfica y la segunda económica, con un poder explicativo del 24.50% en las primeras 2 dimensiones.

Figura 22: ACM en 2 dimensiones



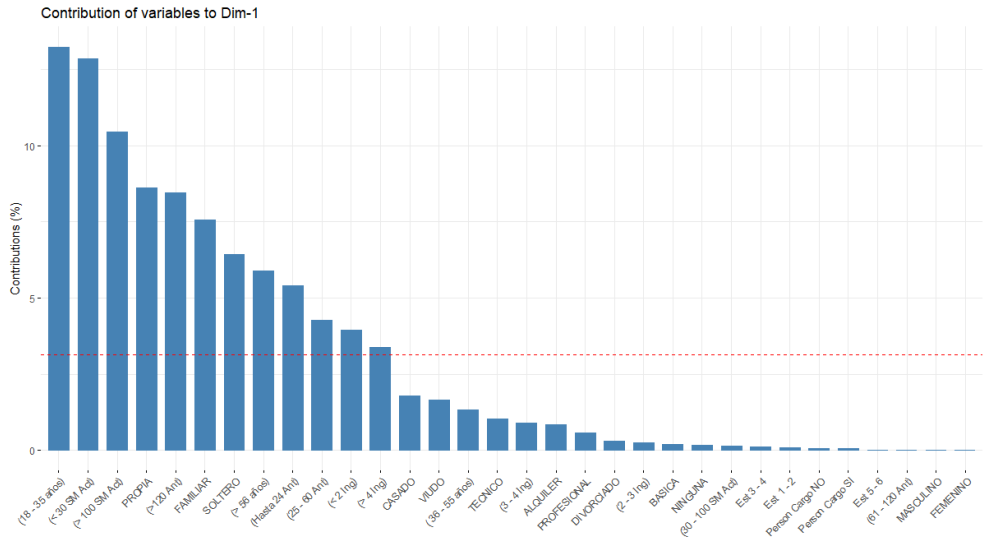
En la figura 23 se muestra la distribución de las dimensiones y el modelo presenta un mejor ajuste explicando en las tres primeras dimensiones el 32,2%.

Figura 23. Explicación del modelo en sus dimensiones



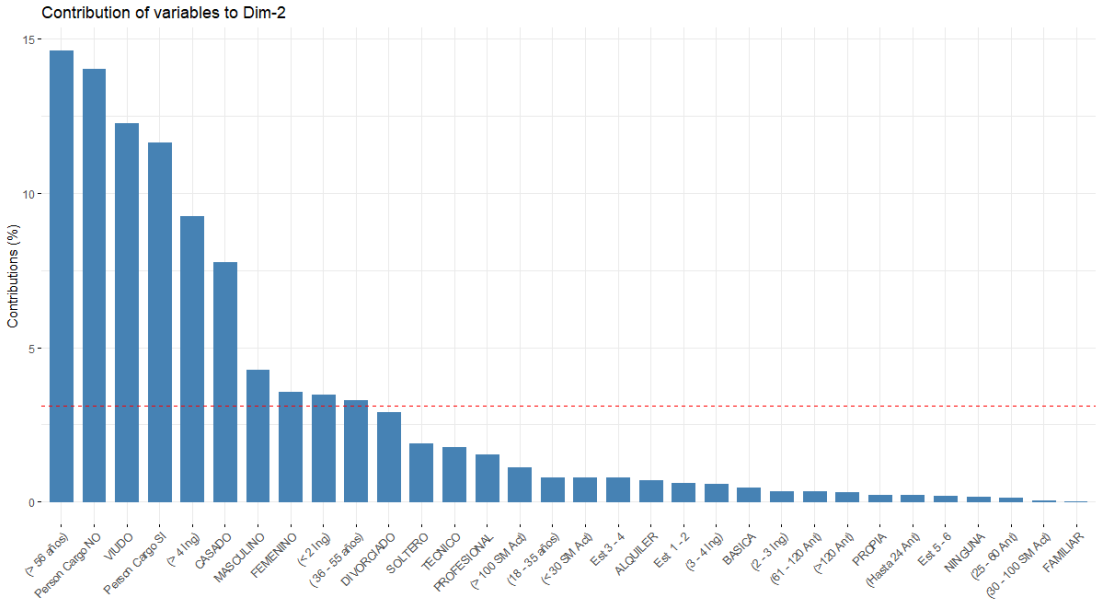
La figura 24 determina el aporte de las categorías en la primera dimensión, se genera un perfil que determina personas con edad de 18 a 35 años y activos menores a 30 o superiores a 10 salarios mínimos, con vivienda propia o familiar y antigüedad en la actividad superior a 10 años.

Figura 24. Categorías en la primera dimensión



En la figura 25 se muestra el aporte de las variables en sus categorías sobre la segunda dimensión, se puede determinar un perfil de personas con edad superior a 56 años, casados o viudos con ingresos superiores a cuatro (4) salarios mínimos.

Figura 25. Categorías en la segunda dimensión

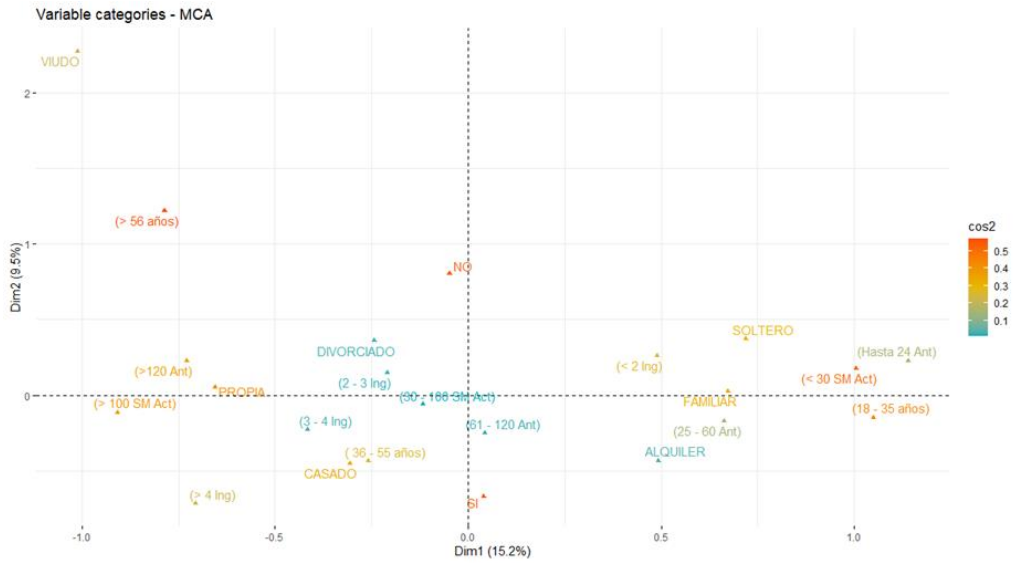


En la figura 26, se evidencia la distribución de las variables de las categorías en las dos primeras dimensiones, donde se observa que se forman grupos sobre la variable ingresos así:

- Personas con ingresos menores a dos (2) salarios mínimos en el cuadrante superior derecho.
- Personas con ingresos de dos (2) a tres (3) salarios mínimos en el cuadrante superior izquierdo.
- Personas con ingresos mayores a tres (3) salarios mínimos en el cuadrante inferior izquierdo.

Sin embargo, cuando se observa la fuerza de las categorías reflejada en el coseno cuadrado (Figura 26), estas no presentan una fuerza suficiente en todos los rangos de la variable.

Figura 26: Coseno cuadrado



En el análisis de correspondencia múltiple, las agrupaciones no reflejan fuerza de asociación al tener un bajo poder explicativo. Las variables diferentes a la variable que se busca explicar

(ingresos) tienen mayor influencia en las 2 primeras dimensiones, por lo que no se considera retirar más variables. Por lo anterior se decide realizar un modelo de regresión lineal que permita evaluar el nivel de predicción de las variables.

7. Modelos de regresión lineal

La regresión lineal múltiple es una extensión de la regresión lineal simple, manteniendo la condición de generar un modelo lineal con más de una variable predictora.

En este sentido, se lleva a cabo dos procedimientos de regresión lineal paralelos, un modelo de regresión lineal múltiple y un modelo de regresión robusta, teniendo en cuenta los valores atípicos observados en la descripción de la muestra.

7.1. Regresión robusta

La regresión robusta según (Amat R. J., 2020) “consigue reducir la influencia de los valores atípicos en el ajuste del modelo”, los datos atípicos pueden influir en el modelo, para ello se puede emplear dos métodos:

M- estimation: Se emplea para atenuar el peso de las observaciones extremas, disminuyendo la influencia de valores atípicos siempre y cuando sean pocos.

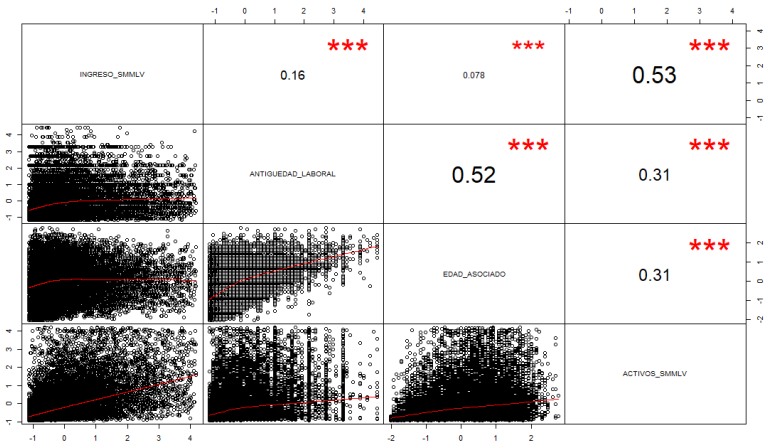
Least Trimmed Squares: Ajusta el modelo mediante mínimos cuadrados, pero empleando los *q* residuos de menor tamaño ignorando el resto de las observaciones.

7.2. Análisis del modelo de regresión

El proceso de regresión inicia con el escalamiento de variables cuantitativas utilizando la muestra de entrenamiento que previamente se estableció en el modelo de análisis de correspondencia.

Seguidamente se realiza la verificación de la correlación sobre las variables cuantitativas, con el objetivo de evaluar las variables que resultan prácticas para comprobar en el modelo aplicado, en este sentido la correlación realizada (Figura 27) refleja que la variable antigüedad laboral y edad asociado tienen una correlación fuerte (0.52); por lo cual se decide utilizar la variable Edad asociado ya que es un dato fácil de validar.

Figura 27. Correlación para regresión



Se realiza planteamiento de escenarios para determinar el mejor coeficiente de determinación – R2, identificando el siguiente modelo lineal múltiple:

$$Ingreso\ Smmlv = Activos\ Smmlv + Otros\ Ingresos + Tipo\ Vivienda + Estado\ Civil + Escolaridad + Actividad\ + Personas\ a\ Cargo + Zona$$

El modelo lineal múltiple obtiene resultados con un error estándar de 1.388 y un R2 ajustado de 0.3639.(Tabla 25)

Tabla 25
Coeficientes modelo regresión lineal múltiple

Estadístico	Valor
Error residual estándar	1.388
R-cuadrado	0.3648
R-cuadrado Ajustado	0.3639
Estadístico F	416.2
P-Valor	2.20E-16

De forma paralela se ejecuta el modelo robusto utilizando las mismas variables y genera un error residual estándar de 0.9723, este modelo no calcula el coeficiente de determinación (Tabla 26).

Tabla 26
Coeficiente regresión robusta

Estadístico	Valor
Error residual estándar	0.9723

7.1. Validación de supuestos

7.1.2. Independencia: Se aplica el estadístico Durbin-Watson (DW) para determinar el grado de independencia que existe entre los residuos, donde la oscilación del estadístico se encuentra entre 0 y 4, toma el valor 2 cuando los residuos son independientes. Los valores menores a 2 indican autocorrelación positiva y los mayores a 2 autocorrelación negativa. Se puede asumir independencia con valores entre 1,5 y 2,5.

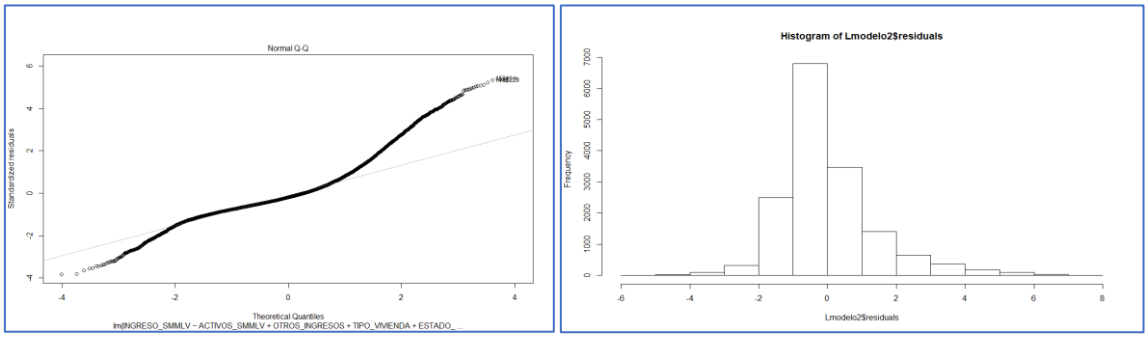
En este sentido, el modelo de regresión lineal múltiple (Lmodelo2) y la regresión robusta (Rmodelo2) – Tabla 27, indican autocorrelación negativa o no están correlacionados y el p-valor es superior al nivel de significancia indicando que hay independencia de los residuos.

Tabla 27
Test Durbin-Watson

Modelo	Autocorrelación	Estadístico	P-Valor
Regresión Lineal Múltiple	-0.006153483	2.012197	0.414
Regresión Robusta	0.009627815	1.98059	0.23

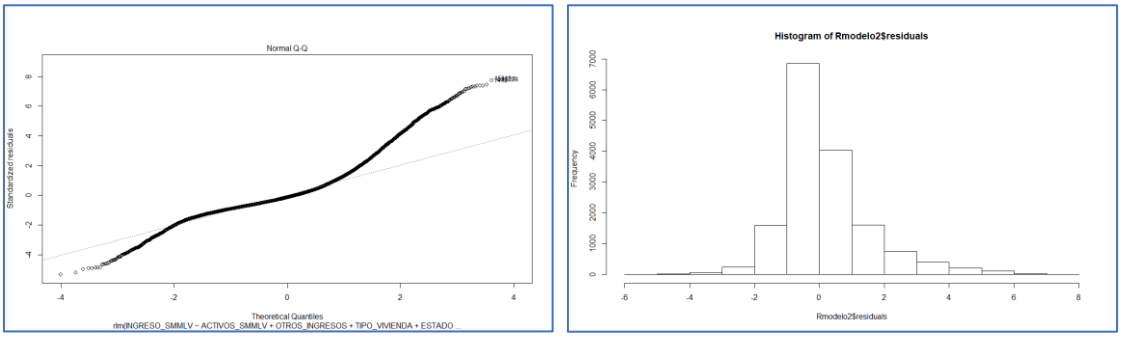
7.1.3. Normalidad: La figura 28, grafica la normalidad de los residuos para el modelo lineal múltiple (Lmodelo2) en el cual se aprecia un sesgo notable a la derecha.

Figura 28. Gráficos normalidad regresión múltiple



La figura 29, grafica la prueba de normalidad para el modelo Robusto (Rmodelo2), el cual mantiene las proporciones similares al modelo lineal múltiple.

Figura 29. Gráficos normalidad regresión robusta



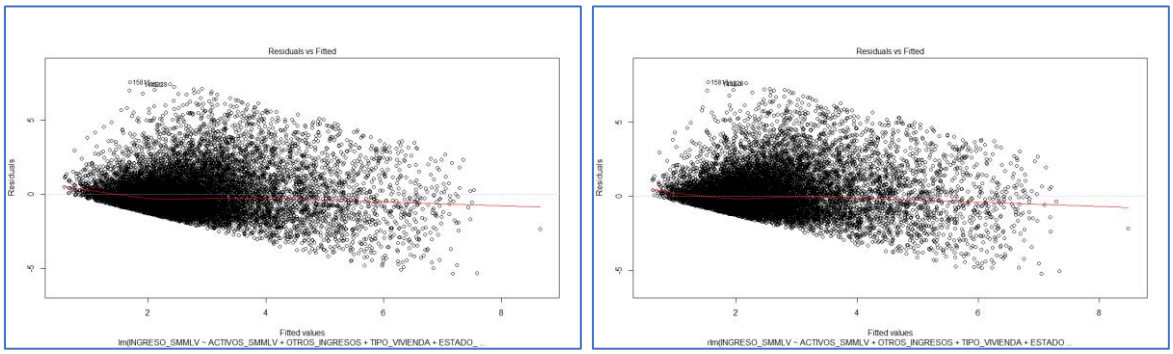
Para validar la normalidad se aplica la prueba de Anderson-Darling (Tabla 28) a los dos modelos, donde se encuentra evidencia estadística para rechazar la hipótesis nula que indica que los residuos provienen de una distribución normal.

Tabla 28
Test Anderson-Darling regresión multiple

Modelo	Estadístico	P-Valor
Regresión Lineal Múltiple	430.63	2.20E-16
Regresión Robusta	476.08	2.20E-16

7.1.4. Homocedasticidad: La figura 30: (izquierda regresión lineal múltiple, derecha regresión robusta), gráficamente son similares y muestran un patrón en forma de cono indicando una distribución de los residuos, razón por la cual se realiza la prueba de Breush-Pagan que establece como hipótesis nula la homogeneidad de la varianza.

Figura 30. Prueba homocedasticidad



Se aplica la prueba (Tabla 29) a los modelos y el resultado es rechazo de hipótesis nula, debido a que los residuos presentan heterocedasticidad.

Tabla 29
Test Breusch-Pagan

Modelo	Estadístico	Grados Libertad	P-Valor
Regresión Lineal Múltiple	1355,1	22	2.20E-16
Regresión Robusta	1355,1	22	2.20E-16

7.1.5. Multicolinealidad: Se verifica la multicolinealidad entre las variables predictoras, pero se observa que no tienen problema que afecte el modelo (Tabla 30) , teniendo en cuenta que la inflación de la varianza no presenta valores altos.

Tabla 30
Inflación de la varianza de variables predictoras

Variable	Regresión Lineal Múltiple	Regresión Robusta
ACTIVOS_SMMLV	1,284477	1,284477
OTROS_INGRESOS	1,087113	1,087113
TIPO_VIVIENDA	1,278552	1,278552
ESTADO_CIVIL	1,189169	1,189169
ESCOLARIDAD	1,086758	1,086758
ACTIVIDAD	1,191268	1,191268
PERSON_CARGO	1,096486	1,096486
ZONA	1,205197	1,205197

Finalmente, se realiza validación para elegir el mejor modelo entre regresión lineal múltiple y regresión robusta; teniendo en cuenta el error de los modelos (Tabla 31) y el efecto de los valores atípicos se elige el modelo robusto para continuar el análisis.

Tabla 31
Comparación de modelos

Estadístico	Regresión Lineal Múltiple	Regresión Robusta
Observaciones	15970	15970
R-cuadrado	0.365	
R-cuadrado Ajustado	0.364	
Error Residual Estándar	1.388	0.972
Estadístico F	416.216	

Los coeficientes del modelo de regresión robusta son los siguientes:

Tabla 32
Coeficientes del modelo de regresión robusta

Variable	Valor	Error Estándar	Valor t
Intercepto	2.9315	0.0428	68.4468
ACTIVOS_SMMLV	1.0071	0.0097	104.185

Variable	Valor	Error Estándar	Valor t
OTROS_INGRESOS	-0.5868	0.0235	-24.9256
TIPO_VIVIENDAFAMILIAR	-0.3179	0.0314	-10.1136
TIPO_VIVIENDAPROPIA	-0.6965	0.032	-21.7771
ESTADO_CIVILDIVORCIADO	-0.1598	0.0265	-6.0207
ESTADO_CIVILSOLTERO	-0.1526	0.0205	-7.4481
ESTADO_CIVILVIUDO	-0.3335	0.0478	-6.9742
ESCOLARIDADNINGUNA	-0.141	0.0905	-1.5579
ESCOLARIDADPROFESIONAL	0.5288	0.039	13.5743
ESCOLARIDADTECNICO	0.1672	0.0315	5.3013
ACTIVIDADComercio	0.2556	0.0261	9.7919
ACTIVIDADIndustria	0.1345	0.0344	3.9099
ACTIVIDADOtras Actividades	-0.0455	0.0324	-1.4038
ACTIVIDADServicios	-0.1178	0.0319	-3.6917
PERSON_CARGO	0.1303	0.0178	7.3244
ZONAATLANTICO	1.3107	0.1234	10.6227
ZONABOYACA	0.055	0.0343	1.6016
ZONACENTRO	0.0987	0.0287	3.439
ZONACESAR	-0.0851	0.0259	-3.2876
ZONACUNDINAMARCA	0.7856	0.0727	10.809
ZONAMAGDALENA	-0.0544	0.0275	-1.9773
ZONANTE SANTANDER	0.0977	0.0284	3.4355

Se evalúa la varianza del modelo en las variables (Tabla 33) y difieren entre sí, por consiguiente, se confirma que las variables seleccionadas en el modelo se ajustan al estudio, rechazando la hipótesis nula debido a que hay diferencia entre las medias de las variables.

Tabla 33
Coeficientes del modelo de regresión robusta

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ACTIVOS_SMMLV	1	13587	13587	7047.91	< 2e-16
OTROS_INGRESOS	1	1234	1234	640.12	< 2e-16
TIPO_VIVIENDA	2	1175	587	304.74	< 2e-16
ESTADO_CIVIL	3	271	90	46.81	< 2e-16
ESCOLARIDAD	3	362	121	62.53	< 2e-16

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ACTIVIDAD	4	604	151	78.33	< 2e-16
PERSON_CARGO	1	116	116	60.3	8.62E-15
ZONA	7	304	43	22.51	< 2e-16
Residuos	15947	30743	2		

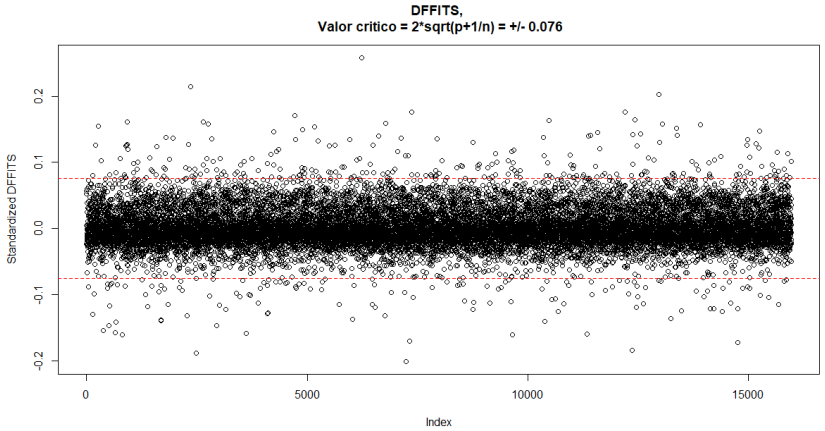
Con el fin de encontrar la homogeneidad de los residuos se realiza análisis de influencia sobre el modelo de regresión robusta, detectando que existen varios valores influyentes en diferencia de ajustes (dffits), razón por la cual se aplica la ecuación para determinar el límite (Stanford, 2020):

$$DFFITS = 2\sqrt{\frac{p+1}{n}}$$

Donde *p* equivale a los parámetros del modelo, para la evaluación *p* equivale a 22 parámetros y *n* equivale al tamaño de la muestra con 15970 observaciones.

El valor límite de DFFITS es de 0.076, donde 249 observaciones sobrepasan este valor (Figura 31), por lo que se ejecuta el modelo excluyendo estas observaciones.

Figura 31. Análisis de influencia



7.2. Validación del modelo de regresión robusta sin observaciones influyentes

7.2.2. Independencia: Se realiza la prueba de Durbin Watson (Tabla 34) para evaluar la independencia de las variables, las cuales muestran independencia aceptando la hipótesis nula.

Tabla 34
Prueba DW sin datos influyentes

Modelo	Autocorrelación	Estadístico	P-Valor
Regresión Robusta	0.007459354	1.984893	0.344

7.2.3. Normalidad: A pesar de retirar los valores influyentes el modelo continúa rechazando la hipótesis nula que indica normalidad en los residuos (Tabla 35).

Tabla 35
Test normalidad sin datos influyentes

Modelo	Estadístico	P-Valor
Regresión Robusta	421.74	2.20E-16

7.2.4. Homocedasticidad: Se realiza la prueba de Breusch Pagan y nuevamente se rechaza la hipótesis nula que indica homogeneidad de los residuos (Tabla 36), presentando heterocedasticidad.

Tabla 36
Test homocedasticidad sin datos influyentes

Modelo	Estadístico	Grados Libertad	P-Valor
Regresión Robusta	1365.7	22	2.20E-16

7.2.5. Multicolinealidad: No se observa multicolinealidad en las variables predictoras (Tabla 37) teniendo en cuenta que la inflación de la varianza no presenta valores altos.

Tabla 37
Inflación de la varianza de variables predictoras

Variable	VIF
ACTIVOS_SMMLV	1.28723
OTROS_INGRESOS	1.0873
TIPO_VIVIENDA	1.27787
ESTADO_CIVIL	1.18976
ESCOLARIDAD	1.08151
ACTIVIDAD	1.18831
PERSON_CARGO	1.09674
ZONA	1.20397

7.3. Predicción del modelo de regresión robusta

Se Selecciona la muestra de evaluación y se realiza una predicción del modelo que genera el error cuadrático estándar, la predicción versus la variable de valor ingreso en la muestra es la siguiente (Tabla 38):

Tabla 38
Error MSE del modelo

Modelo según muestra	MSE
Resultado Modelo con muestra de entrenamiento	1,980540
Resultado Modelo sin valores influyentes en la muestra de entrenamiento	2,012352

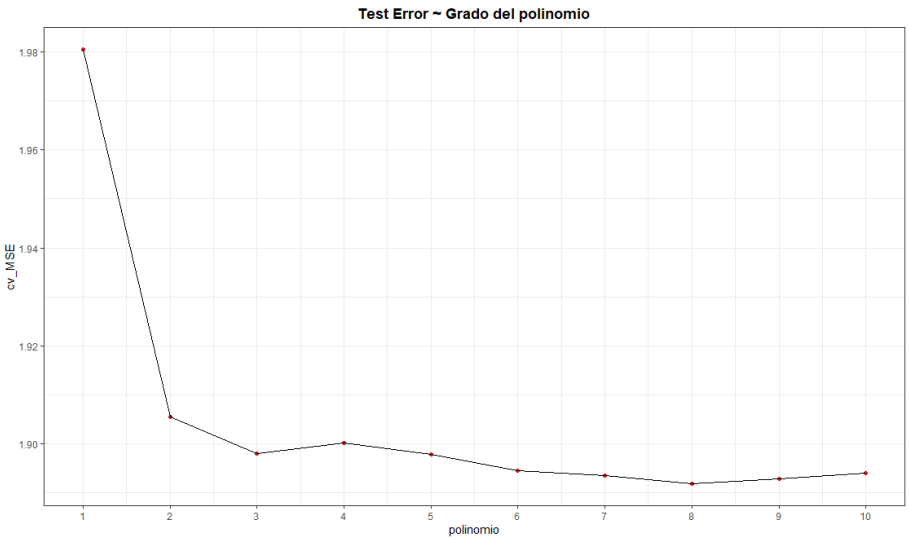
Teniendo en cuenta el error cuadrático en la predicción se continua la evaluación del modelo con la muestra de entrenamiento y muestra un error cuadrático de 1.98 que implica tener una diferencia de casi 2 salarios mínimos en la predicción de los ingresos.

Se realiza la generación de 10 modelos ajustando la variable activos siendo esta la única variable

predictora continua que contiene el modelo mediante polinomios de grado 1 hasta 10, utilizando validación cruzada simple. (Amat R. J., 2020). (Figura 32).

El resultado muestra que aplicando un polinomio de grado 3 se mejora la relación de las variables y por ende la predicción.

Figura 32.Gráfico polinomios



Al aplicar el ajuste al modelo, el error medio cuadrático es de 1.8980, logrando una diferencia de 0.09032 respecto al modelo sin el ajuste a la variable activos, siendo está diferencia no muy significativa, por lo que se decide no aplicar el ajuste al modelo.

Al comparar la desviación estándar en la muestra de entrenamiento y la muestra evaluada se observa un valor de 1.74, valor cercano al error estándar del modelo de regresión robusta con 1.98, lo que indica que la predicción es aceptable.

Tabla 39
Estadísticos de las muestras

Muestra	Media	Desviación Estándar
Muestra de Entrenamiento	2.6873	1.740867
Muestra de Evaluación	2.6888	1.745896

8. Conclusiones

La calidad de los datos recolectados juega un rol esencial en el proceso de análisis y transformación de variables, en este sentido, detallar cada una de las variables estudiadas mediante un análisis descriptivo permitió identificar que el 54.11% de la muestra es género femenino, el 52.91% del total tiene estado civil casado, un 91.31% viven en vivienda propia o familiar, el 72.99% de los registros se concentra en estratos socioeconómicos 2 y 3, el 86.27% de los registros tienen nivel de educación básica y el 91.40% de los trabajadores independientes son mayores de 25 años, vale resaltar que el 67.21% de los datos analizados tienen mas de 60 meses de antigüedad laboral en la actividad independiente, que el 44.38% de la actividad que se ejerce es Comercio y que el 83.01% de los trabajadores independientes no cuentan con otros ingresos adiciones, tales como arriendo, honorarios entre otros.

El análisis de correspondencia múltiple no expresa claramente fuerza de asociación de las variables del estudio, aunque se realizó varios escenarios para explicar la asociación de las 13 variables inicialmente planteadas, el modelo construido logró explicar el 32.2% con siete variables agrupadas en dos dimensiones, la primera es la dimensión demográfica que agrupa las variables Edad, Personas a cargo y Estado Civil; la segunda dimensión es económica que agrupa Rango de ingresos en salarios mínimos, Rango de antigüedad laboral, Tipo de vivienda y Rango de activos, aun así la asociación de la variable respuesta con las variables explicativas permite una agrupación poco efectivo del análisis de correspondencia.

El modelo de regresión lineal múltiple estableció el mejor coeficiente de determinación con un R2 de 0.3639 para explicar la variable ingresos en función de ocho variables explicativas, este modelo fue comparado en simultaneo con una regresión robusta, sin embargo, una de las características de la regresión robusta es que no calcula el coeficiente de determinación, no obstante, al excluir las observaciones influyentes y validar los supuestos se estableció el modelo de regresión robusta como motor estimador de ingresos para trabajadores independientes logrando predecir con un error estándar de 1.98 la variable respuesta.

Finalmente, el modelo de regresión robusta aunque no cumplió los supuestos de normalidad y homocedasticidad de los residuos y teniendo en cuenta la diversidad de factores que pueden incidir en el perfilamiento de una persona en el estudio, brinda ventajas técnicas al reducir los residuos, siendo preciso y sencillo de comparar; al validar los datos utilizados para el estudio se evidencia una desviación estándar de 1.74 y el modelo seleccionado tiene un error estándar de 1.98 es decir una diferencia de 0.24 por encima de la desviación estándar, indicando que el modelo seleccionado se considera aceptable.

Bibliografía

Aldas, M. J., & Uriel, J. E. (2017). *Análisis multivariante aplicado con R.* Madrid: Madrid Paraninfo 2017.

Alvarado, J. L., & Pinos, O. A. (2017). Estimación de ingresos de la población ecuatoriana. Una propuesta desde la regresión cuantílica. *Cuestiones Económicas*, Vol. 27, No. 2:2, .

Amat, R. J. (27 de 04 de 2020). *https://rpubs.com/*. Obtenido de https://rpubs.com/Joaquin_AR/238251

Amat, R. J. (27 de 04 de 2020). *www.rpubs.com*. Obtenido de https://rpubs.com/Joaquin_AR/226291

Código Sustantivo del Trabajo. (07 de 03 de 2020). *Ministerio del Trabajo.* Obtenido de http://www.mintrabajo.gov.co/: http://www.mintrabajo.gov.co/normatividad/leyes-y-decretos-ley/codigo-sustantivo-del-trabajo

El empleo. (8 de 03 de 2020). *El Empleo*. Obtenido de https://www.elempleo.com/co: https://www.elempleo.com/co/noticias/investigacion-laboral/tipos-de-trabajadores-independientes-5942

Espin, G. O., & Rodríguez, C. C. (2013). Metodología para un scoring de clientes sin referencias crediticias. *Cuadernos de Economía*, 137-162.

Fernández, A. F. (18 de Enero de 2018). Trabajo fin de Master: Modelo de estimación de ingresos para clientes poco o nada vinculados con Abanca. Coruña, España.

Guataquí, J. C., García, A. F., & Rodrigruez, M. (2009). Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para aslariados y cuenta

propia. *Universidad del Rosario - Facultad de Economía*, 1-23.

Gutierrez, P. H., & De la Vara, S. R. (2013). *Control estadístico de la calidad y seis sigma.* Mexico: Mc Graw Hill.

Hernández, L. J., Espinosa, C. J., Peñaloza, T. M., Rodriguez, J. E., Chacon, R. J., Toloza, S. C., . . . Bermudez, P. V. (2018). Sobre el uso adecuado del coeficiente de correlación de Pearson: definición, propiedades y suposiciones. *Revista AVFT*, 587-595.

Hernandez, M. Z. (2012). *Método de Análisis de datos: Apuntes.* Logroño: Universidad la Rioja.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R.* New York: Springer.

Monterey Institute. (s.f.). Recuperado el 19 de Noviembre de 2019, de https://www.montereyinstitute.org/courses/DevelopmentalMath/TEXTGROUP-1-8_RESOURCE/U01_L2_T3_text_final_es.html

Real Academia Española. (14 de 02 de 2020). Obtenido de https://www.rae.es/: https://dle.rae.es/ingresar?m=form

Real Academia Española. (14 de 02 de 2020). Recuperado el 19 de Noviembre de 2019, de https://www.rae.es/: https://dle.rae.es/?w=estimar

Stanford, U. (03 de 05 de 2020). *https://web.stanford.edu/*. Obtenido de https://web.stanford.edu/class/stats191/notebooks/Diagnostics_for_multiple_regression.html

Universidad Nacional Autónoma de México. (15 de Febrero de 2020). Obtenido de http://www.cuautitlan.unam.mx/: http://asesorias.cuautitlan2.unam.mx/Laboratoriovirtualdeestadistica/CARPETA%203%20INFERENCIA_ESTADISTICA/DOC_%20INFERENCIA/TEMA%204/09%20REGRE

SION%20Y%20CORRELACION%20LINEAL%20SIMPLE.pdf

Apéndice A

Código R-Studio con las debidas librerías ejecutadas para realizar el análisis de correspondencia múltiple y la regresión lineal múltiple y robusta.

```
library(gplots)
library(fdth)
library(readxl)
library(dplyr)
library(sqldf)
library(ggplot2)
library(extrafont)
library(corrplot)
library(PerformanceAnalytics)
library(factoextra)
library(FactoMineR)
library(car)
library(stargazer)
library(MASS)
library(lmtest)
library(nortest)

#Carga de la muestra
base_ini=read_excel("D:/Drive/Estadistica/Proyecto
Estadística/Base_Proyecto_final.xlsx")

#se toman las variables necesarias
base=sqldf("select
ZONA,RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO,COD_ESTRATO
_T,TIPO_VIVIENDA,EDAD_ASOCIADO,RANGO_EDAD_T,ACTIVIDAD,TOT_ACTIVOS,RANGO_ACT_S
MMLV,RANGO_ACT_SMMLV_T,INGRESO_SMMLV,RANGO_ING_SMMLV,RANGO_ING_SMMLV_T,OTROS_
INGRESOS,PERSON_CARGO_T from base_ini")


#####Extracción de las muestras de entrenamiento y evaluación#####
#####################################################################

#Muestra N
N=nrow(base)
```

```
#Semilla
set.seed(1)

#División aleatoria
aleatorio=sample(N,N/2,replace=FALSE)

#Extracción de la muestra de entrenamiento incluyendo los registros de la
variable aleatorio
muestra1=data.frame(base[aleatorio,])

#Extracción de la muestra de evaluación excluyendo los registros de la
variable aleatorio
muestra2=data.frame(base[-aleatorio,])

#-----------------------------------------------------------------------#
#---------------Análisis de correspondencia múltiple----------------#
#-----------------------------------------------------------------------#

#Selección de las variables transformadas para el modelo
muestra1_2=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO_T,TIPO_VIVIENDA,
RANGO_EDAD_T,ACTIVIDAD,RANGO_ACT_SMMLV,RANGO_ING_SMMLV,OTROS_INGRESOS,PERSON_
CARGO_T,ZONA from muestra1")

#Aplicación del modelo
resmca=MCA(muestra1_2)
var <- get_mca_var(resmca)

#Se excluye la variable zona
muestra1_3=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO_T,TIPO_VIVIENDA,
RANGO_EDAD_T,ACTIVIDAD,RANGO_ACT_SMMLV,RANGO_ING_SMMLV,OTROS_INGRESOS,PERSON_
CARGO_T from muestra1")

#Aplicación del modelo
resmca2=MCA(muestra1_3)

#Categorías
fviz_mca_var(resmca2,repel = TRUE,ggtheme = theme_minimal())

#Se cambia la variable Ingresos y activos con las transformadas
muestra1_4=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO_T,TIPO_VIVIENDA,
```

```
RANGO_EDAD_T,ACTIVIDAD,RANGO_ACT_SMMLV_T,RANGO_ING_SMMLV_T,OTROS_INGRESOS,PER
SON_CARGO_T from muestra1")

#Aplicación del modelo
resmca3=MCA(muestra1_4)

#Categorías
fviz_mca_var(resmca3,repel = TRUE,ggtheme = theme_minimal())

summary(resmca3)

#Se excluye la variable otros ingresos
muestra1_5=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO_T,TIPO_VIVIENDA,
RANGO_EDAD_T,ACTIVIDAD,RANGO_ACT_SMMLV_T,RANGO_ING_SMMLV_T,PERSON_CARGO_T
from muestra1")

#Aplicación del modelo
resmca4=MCA(muestra1_5)

#Dimensiones
get_eigenvalue(resmca4)
fviz_screeplot(resmca4, addlabels = TRUE)

#Se excluye la variable actividad económica
muestra1_6=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,COD_ESTRATO_T,TIPO_VIVIENDA,
RANGO_EDAD_T,RANGO_ACT_SMMLV_T,RANGO_ING_SMMLV_T,PERSON_CARGO_T from
muestra1")

#Aplicación del modelo
resmca5=MCA(muestra1_6)

#Categorías en las dimensiones
fviz_contrib(resmca5, choice ="var", axes = 1)
fviz_contrib(resmca5, choice ="var", axes = 2)
fviz_contrib(resmca5, choice ="var", axes = 3)

#Se excluye la variable estrato
muestra1_7=sqldf("select
RANGO_ANTIGUEDAD,GENERO,ESTADO_CIVIL,ESCOLARIDAD,TIPO_VIVIENDA,RANGO_EDAD_T,R
ANGO_ACT_SMMLV_T,RANGO_ING_SMMLV_T,PERSON_CARGO_T from muestra1")
```

```r
#Aplicación del modelo
resmca6=MCA(muestra1_7)

#Categorías
fviz_mca_var(resmca6, repel = TRUE,ggtheme = theme_minimal())

#Se excluyen las variables género y escolaridad
muestra1_8=sqldf("select
RANGO_ANTIGUEDAD,ESTADO_CIVIL,TIPO_VIVIENDA,RANGO_EDAD_T,RANGO_ACT_SMMLV_T,RA
NGO_ING_SMMLV_T,PERSON_CARGO_T from muestra1")

#Aplicación del modelo
resmca7=MCA(muestra1_8)

#Dimensiones
fviz_screeplot(resmca7, addlabels = TRUE)

#Categorías en las dimensiones
fviz_contrib(resmca5, choice ="var", axes = 1)
fviz_contrib(resmca5, choice ="var", axes = 2)
fviz_contrib(resmca5, choice ="var", axes = 3)

#Coseno cuadrado en las categorías
fviz_mca_var(resmca7, axes = c(1, 2), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # Avoid text overlapping
             ggtheme = theme_minimal())

#-------------------------------------------------------------------#
#-----------------Regresión Lineal múltiple y robusta---------------#
#-------------------------------------------------------------------#

#transformación de variables
base_ini=base_ini %>%
  mutate(OTROS_INGRESOS = ifelse(OTROS_INGRESOS == "SI", 1, 0))

base_ini=base_ini %>%
  mutate(PERSON_CARGO = ifelse(PERSON_CARGO == "SI", 1, 0))

base_ini$COD_ESTRATO=factor(base_ini$COD_ESTRATO)
```

```
#se escalan las variables continuas
base_escalada=scale(base_ini[,c(5,14,21)],center=T,scale=T)

#unión de variables escaladas y variables categóricas
base_escalada=cbind(base_escalada,base_ini[,c(4,7:9,11,13,18,28,31,25)])
str(base_escalada)

#-------Selección de las muestras con variables escaladas
muestra1=data.frame(base_escalada[aleatorio,])
muestra2=data.frame(base_escalada[-aleatorio,])

attach(muestra1)

#Modelo con todas las variables
modelo=lm(INGRESO_SMMLV ~ .,data = muestra1)
summary(modelo)

#Regresión lineal con variables utilizadas en el análisis de correspondencia
LmodeloMC=lm(INGRESO_SMMLV~EDAD_ASOCIADO+ACTIVOS_SMMLV+TIPO_VIVIENDA+ESTADO_C
IVIL+OTROS_INGRESOS+ESTADO_CIVIL+ESCOLARIDAD+TIPO_VIVIENDA+PERSON_CARGO,data=
muestra1)
summary(LmodeloMC)

#Regresión robusta con variables utilizadas en el análisis de correspondencia
RmodeloMC=rlm(INGRESO_SMMLV~EDAD_ASOCIADO+ACTIVOS_SMMLV+TIPO_VIVIENDA+ESTADO_
CIVIL+OTROS_INGRESOS+ESTADO_CIVIL+ESCOLARIDAD+TIPO_VIVIENDA+PERSON_CARGO,data
=muestra1)
summary(RmodeloMC)

######################Modelo 1#################################
#regresión lineal múltiple excluyendo la variable edad
Lmodelo1=lm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_C
IVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA+COD_ESTRATO)
summary(Lmodelo1)

#regresión robusta excluyendo la variable edad
Rmodelo1=rlm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_
CIVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA+COD_ESTRATO)
summary(Rmodelo1)

######################Modelo 2#################################
#regresión lineal múltiple excluyendo estrato y edad
Lmodelo2=lm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_C
```

```
IVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA)
summary(Lmodelo2)

#regresión robusta múltiple excluyendo estrato y edad
Rmodelo2=rlm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_
CIVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA)
summary(Rmodelo2)

par(mfrow=c(1,2))
attach(muestra1)

#Histograma de residuos del modelo de regresión lineal múltiple
hist(Lmodelo2$residuals)
#Histograma de residuos del modelo de regresión robusta
hist(Rmodelo2$residuals)
par(mfrow=c(1,1))

#Graficas de residuos del modelo de regresión lineal múltiple
plot(Lmodelo2)
#Graficas de residuos del modelo de regresión robusta
plot(Rmodelo2)

#-----Test de autocorrelación Durbin Watson
#modelo de regresión lineal múltiple
durbinWatsonTest(Lmodelo2)
#modelo de regresión robusta
durbinWatsonTest(Rmodelo2)

#-----Test de normalidad de residuos Anderson-Darling normality test
#modelo de regresión lineal múltiple
ad.test(Lmodelo2$residuals)
#modelo de regresión robusta
ad.test(Rmodelo2$residuals)

#-----Test de homocedasticidad de Breusch-Pagan
#modelo de regresión lineal múltiple
bptest(Lmodelo2)
#modelo de regresión robusta
bptest(Rmodelo2)

#-----Multicolinealidad Inflación de la varianza
#modelo de regresión lineal múltiple
vif(Lmodelo2)
```

```r
#modelo de regresión robusta
vif(Rmodelo2)

#-----Comparación de los modelos de regresión lineal múltiple y Robusta
stargazer(Lmodelo2, Rmodelo2, type = "text", model.numbers = FALSE,
title="Comparación de modelo OLS y Robusto")
par ( mfrow  =  c ( 2 , 2 ),  oma  =  c ( 0 ,  0 ,  1.1 ,  0 ))
plot (Lmodelo1,  las  =  1 )
plot (Rmodelo1,  las  =  1 )

par(mfrow=c(1,1))

#-----Varianza del modelo de regresión robusto
summary(aov(Rmodelo2))

#-----Análisis de influencia sobre modelo de regresión robusta

summary(influence.measures(Rmodelo2))

#Campo de DFFITS en la muestra
muestra1$dffits=dffits(Rmodelo2)

#-----Tamaño de la muestra
n <- nrow(muestra1)
#-----Número de Parámetros
k <- length(Rmodelo2$coefficients)-1
#-----Limite de DFFITS
cv <- 2*sqrt((k+1)/n)

#-----Gráfica DFFITS valores estandarizados
plot(dffits(Rmodelo2),
     ylab = "Standardized DFFITS", xlab = "Index",
     main = paste("DFFITS, \n Valor critico = 2*sqrt(p+1/n) = +/-",
round(cv,3)))
abline(h=0.076, lty=2, col="red"); abline(h=-0.076, lty=2, col="red")

#Selección de la muestra sin las observaciones influyentes
muestra1_2=sqldf(paste("select * from muestra1 where dffits<=",cv))

#Modelo sin las observaciones influyentes
Rmodelo2=rlm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_
CIVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA, data=muestra1_2)
```

```r
#-----Test de autocorrelación Durbin Watson
durbinWatsonTest(Rmodelo2)

#-----Test de normalidad de residuos Anderson-Darling normality test
ad.test(Rmodelo2$residuals)

#-----Test de homocedasticidad de Breusch-Pagan
bptest(Rmodelo2)

#-----Multicolinealidad Inflación de la varianza
vif(Rmodelo2)

predic = predict(object = Rmodelo2, newdata = muestra2)

#Error medio de los errores
mean((muestra2$INGRESO_SMMLV - predic)^2)

attach(muestra1)

########################Validación del modelo 2############################

#Reasignación del modelo con toda la muestra de entrenamiento
Rmodelo2=rlm(INGRESO_SMMLV~ACTIVOS_SMMLV+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_
CIVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA)
summary(Rmodelo2)
#Predicción del modelo sobre la muestra de evaluación
predic = predict(object = Rmodelo2, newdata = muestra2)

#Error medio de los errores
mean((muestra2$INGRESO_SMMLV - predic)^2)

#Variables
cvMSE <- rep(NA,10)
#Semilla
set.seed(1)

#Matrix con error medio aplicando polinomio a la variable Activos
for (i in 1:10) {
  mod = rlm(INGRESO_SMMLV ~
poly(ACTIVOS_SMMLV,i)+OTROS_INGRESOS+TIPO_VIVIENDA+ESTADO_CIVIL+ESCOLARIDAD+A
CTIVIDAD+PERSON_CARGO+ZONA, data = muestra1)
  predic = predict(object = mod, newdata = muestra2)
  cvMSE[i] = mean((muestra2$INGRESO_SMMLV - predic)^2)
```

```
}

#Grafico de errores
ggplot(data = data.frame(polinomio = 1:10, cvMSE = cvMSE),
       aes(x = polinomio, y = cvMSE)) +
  geom_point(colour = c("firebrick3")) +
  geom_path() +
  scale_x_continuous(breaks = c(0:10)) +
  theme_bw() +
  labs(title  =  'Test Error - Grado del polinomio (RModelo2)') +
  theme(plot.title = element_text(hjust = 0.5, face = 'bold'))

#Modelo aplicando polinomio de grado 3 a la variable activos
Rmodelo2=rlm(INGRESO_SMMLV~poly(ACTIVOS_SMMLV,3)+OTROS_INGRESOS+TIPO_VIVIENDA
+ESTADO_CIVIL+ESCOLARIDAD+ACTIVIDAD+PERSON_CARGO+ZONA)

#Predicción del modelo
predicciones = predict(object = Rmodelo2, newdata = muestra2)

#Media de los errores
mean((muestra2$INGRESO_SMMLV - predicciones)^2)

#Descripción de la muestra 2
summary(muestra2$INGRESO_SMMLV)

#Desviación estándar de la muestra 2
sd(muestra2$INGRESO_SMMLV)

knitr::opts_chunk$set(echo = TRUE)
```