Tarea 1 - Modelamiento Predictivo

Víctor Loyola, Alejandra Neely y Marco Vergara

26-07-2020

# Introducción

En el contexto de la asignatura de Modelamiento Predictivo del Magíster en Ciencia de Datos UAI, se presenta el desarrollo de un modelo predictivo con un set de datos asociados a la banca. Las variables utilizadas definen características bancarias de sus clientes, además de su edad. El objetivo del modelo es predecir cómo estos hábitos contribuyen a los ingresos del banco. Los datos analizados corresponden a 7.420 clientes, con 16 variables. El desarrollo de este proyecto ha sido estructurado en 6 pasos, que se decribirán a continuación:

#### Análisis exploratorio de las variables

Tiene como objetivo conocer los datos, es decir, observar la distribución de la variable dependiente y de las variables independientes numéricas. Evaluar la distribución de variables categóricas, es decir la frecuencia con la que se encuentra presente cada categoría de respuesta, detectando posibles desbalances. El desarrollo de un modelamiento predictivo se ve favorecido cuando las variables con las cuales se trabaja se distribuyen de forma normal, en el caso de las variables numéricas, y de forma balanceada, en el caso de las variables categóricas. En el caso de variables numéricas, se pueden aplicar transformaciones no lineales que mejoren sus propiedades para el trabajo posterior. También, se pueden transformar a categóricas variables que tienen una distribución muy asimétrica, de tal forma de agrupar los valores que podrían estar representando diferentes grupos de la población. Se someterá además cada variable independiente a un análisis de regresión simple para evaluar el porcentaje de varianza de la variable independiente que es explicado por cada una.

#### Análisis de correlaciones

Se evalúa el nivel de correlación entre las variables independientes y la variable dependiente, así como también entre las variables independientes. Esto nos permitirá tener una idea de aquellas variables que podrían estar midiendo constructos muy similares. Es necesario utilizar diferentes estadísticos dependiendo de la naturaleza de las variables. Así, para evaluar correlaciones entre variables numéricas se utilizará la correlación de Pearson, mientras que para evaluar correlaciones entre variables categóricas se utilizará la correlación de Cramer. Sin embargo, también es necesario evaluar la correlación entre las variables categóricas y las variables numéricas, dentro de las cuales se encuentra la variable dependiente. Así, se evaluó primero la correlación Biserial entre las variables categóricas y la variable dependiente que corresponde a los ingresos que percibe el banco respecto de cada cliente. Al utilizar la correlación de tipo Biserial, notamos que el valor de la correlación era igual al valor entregado por la correlación de Pearson cuando se analizan los mismos pares de variables. La razón de esto parece ser que la correlación Biserial es un caso especial de la correlación de Pearson. Se demuestra en el desarrollo que los valores de ambos estadísticos para cada par de variables es igual. Posteriormente, y producto de la comprobación previa, se analizan las correlaciones entre todas las variables utilizando la correlación de Pearson.

#### Especificación del modelo

Se separa el dataset en un subset de entrenamiento (que contiene un 70% de los datos) y un subset de testeo (que contiene el 30% restante de los datos), y se especifica el modelo de regresión múltiple incluyendo todas las variables que no presentaban correlación absoluta entre ellas analizando también el nivel de colinealidad entre las variables.

#### Selección de variables

En este paso se busca seleccionar aquellas variables que sean más significativas para el análisis y que explican y se ajustan mejor a los datos para predecir la variable objetivo. Así, el modelo especificado previamente se someterá a análisis utilizando el Coficiente de Determinación Ajustado, el CP de Mallows, BIC, Paso a paso forward, Paso a paso backward, y Paso a paso Stepwise. Estos coeficientes utilizan diferentes variaciones de cálculo para encontrar el conjunto de variables que presentan un mejor ajuste a los datos. Las técnicas anteriores pueden entregar más de un set de variables, lo que permitirá contar con uno o más modelos posibles a evaluar en el paso siguiente.

#### Evaluación de la calidad predictiva

Una vez obtenidas las sugerencias de modelos por parte de los coeficientes analizados en el punto previo, estos se someterán a otro grupo de estadísticos que permiten evaluar la calidad predictiva. La primera función a aplicar es la validación cruzada PRESS, la que evalúa el comportamiento de la selección dejando en forma iterativa un dato afuera armando el modelo con el resto de los datos, aplicados a la muestra de entrenamiento, siendo el que entregue un menor valor el que demostrará un mejor rendimiento. Adicionalmente, para descartar sobre ajuste, se comparan los estadísticos MAE, MAPE y RMSE, aplicados a la data de entrenamiento y data de testeo, esperándose valores similares. Además, se evaluarán los supuestos de homocedasticidad y normalidad de residuos de los modelos, los cuales permiten validar que los test de hipótesis aplicados al modelo son válidos.

#### Interpretación de coeficientes y conclusiones

Se interpretan los coeficientes aplicados a la realidad del negocio.

# 1. Análisis Exploratorio de las Variables

El análisis exploratorio de variables tiene como objetivo conocer la información del dataset, identificando las características individuales de cada variable, su relación con la variable target a través de regresiones lineales simples, y la presencia de datos faltantes. Estos aspectos serán de gran importancia al momento de tomar decisiones con cada una de ellas a nivel global. Así, se comenzará por identificar la variable objetivo y analizar su distribución. Luego separaremos las variables categóricas de las variables numéricas, analizando cada variable desde el balance de sus categorías y sus distribuciones, respectivamente.

## 1.1 Variable objetivo y variables explicativas

De acuerdo al análisis preliminar de los datos, cada variable ha adoptado la siguiente tipificación:

1. Rev\_Total (Variable Objetivo): Numérica.
2. Bal\_Total: Numérica.
3. Offer: Categórica.
4. AGE: Numérica.
5. CHQ: Categórica.
6. CARD: Categórica.
7. SAV1: Categórica.
8. LOAN: Categórica.
9. MORT: Categórica.
10. INSUR: Categórica.
11. PENS: Categórica.
12. Check: Categórica.
13. CD: Categórica.
14. MM: Categórica.
15. Savings: Categórica.
16. AccountAge: Numérica.

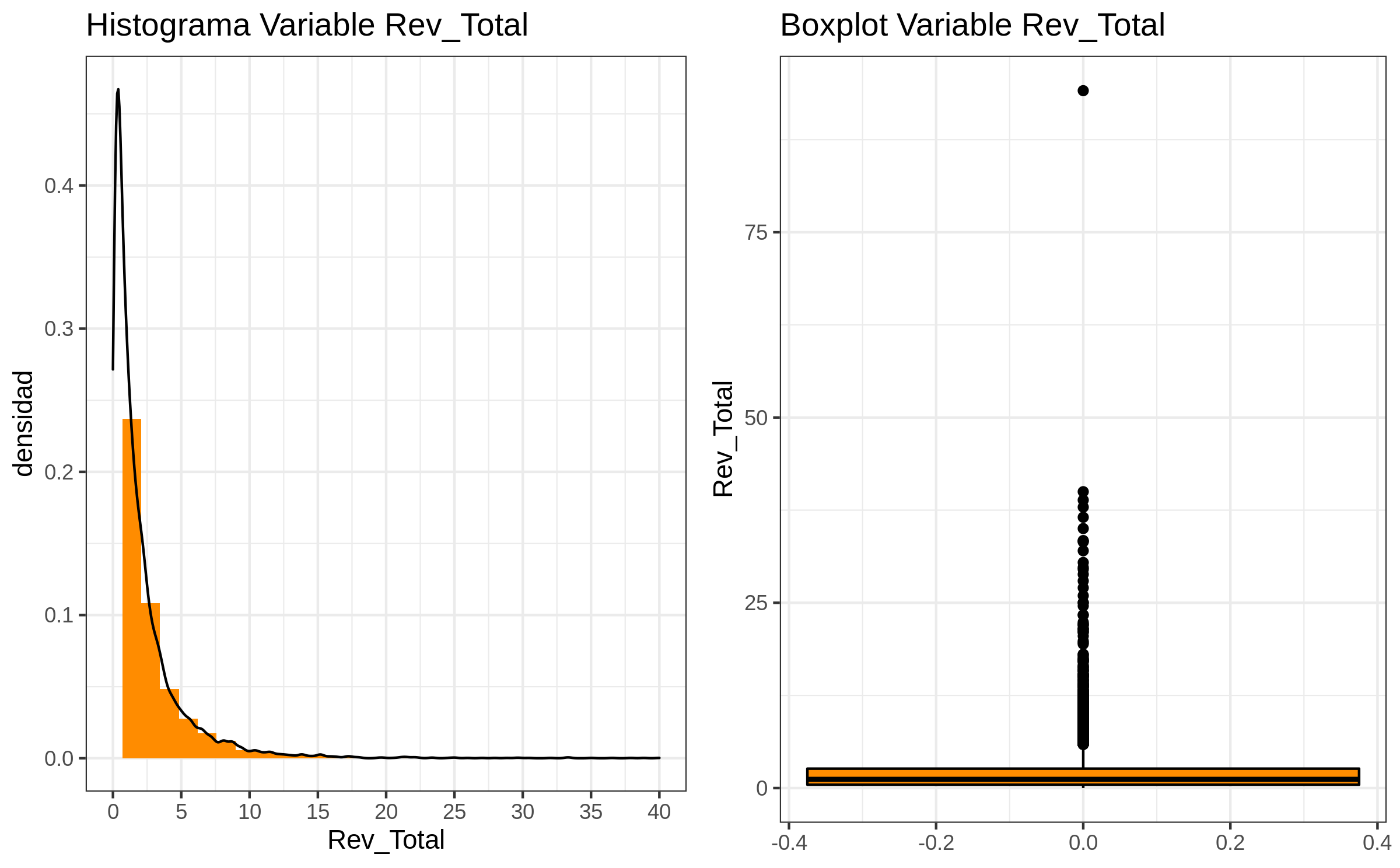
El total de registros alcanza los 7420.

## 1.2 Datos Faltantes

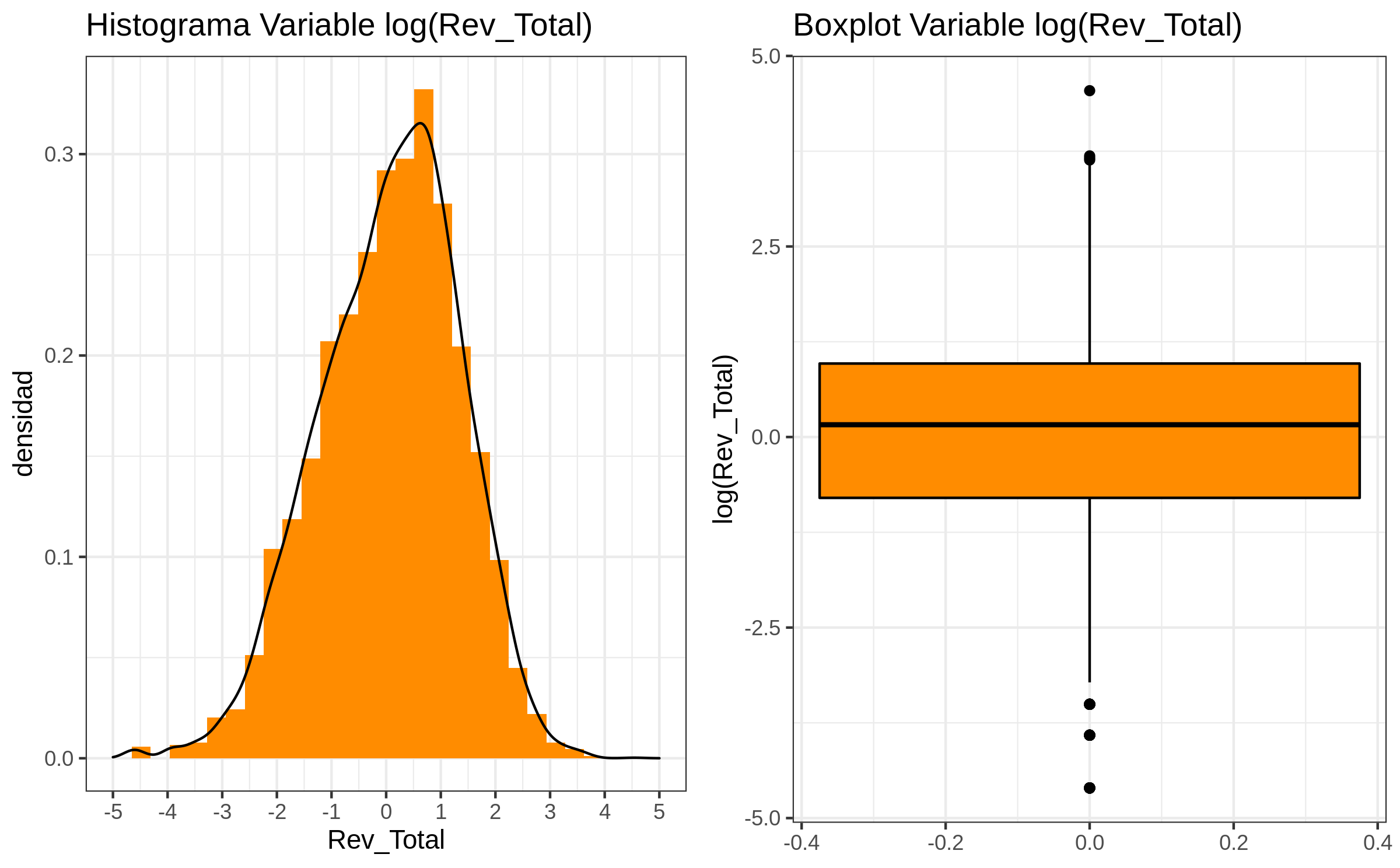
Luego de aplicar funciones de detección de datos faltantes no se encuentran, por lo que se descarta aplicar eliminación de registros incompletos en esta etapa.

## 1.3 Distribución de la Variable Objetivo.

La variable objetivo (Rev\_Total) que corresponde a la rentabilidad percibida por el banco respecto de cada cliente. Lo primero que observaremos es como se comporta ésta en un histograma y en un diagrama de cajas, como se aprecia en la figura siguiente. Al mirar el ésta se observa en el histograma una distribución exponencial, que se confirma en el diagrama de cajas por la existencia de varios outliers. Lo anterior complicará la modelación, dada principalmente por la asimetría. Al obtener valores de kurtosis de 117.9 y de skewness de 6.8, éstos representan valores muy distintos en relación a una distribución normal (como referencia se tiene una kurtosis de 3 y skewness de 0).



Por lo anterior y con el objetivo de modelar una variable objetivo que presente simetria en su distribución, se aplica la transformación no lineal de logaritmo, apreciándose en el siguiente gráfico que la distribución queda muy cercana a una distribución normal, con valores que facilitan la realización de test paramétricos posteriores. La kurtosis esta vez es de 3 y la skewness de -0.3. Así también, la cantidad de outliers se reduce en forma importante. La única desventaja de esta transformación es que la posterior interpretación de los coefientes será menos directa.

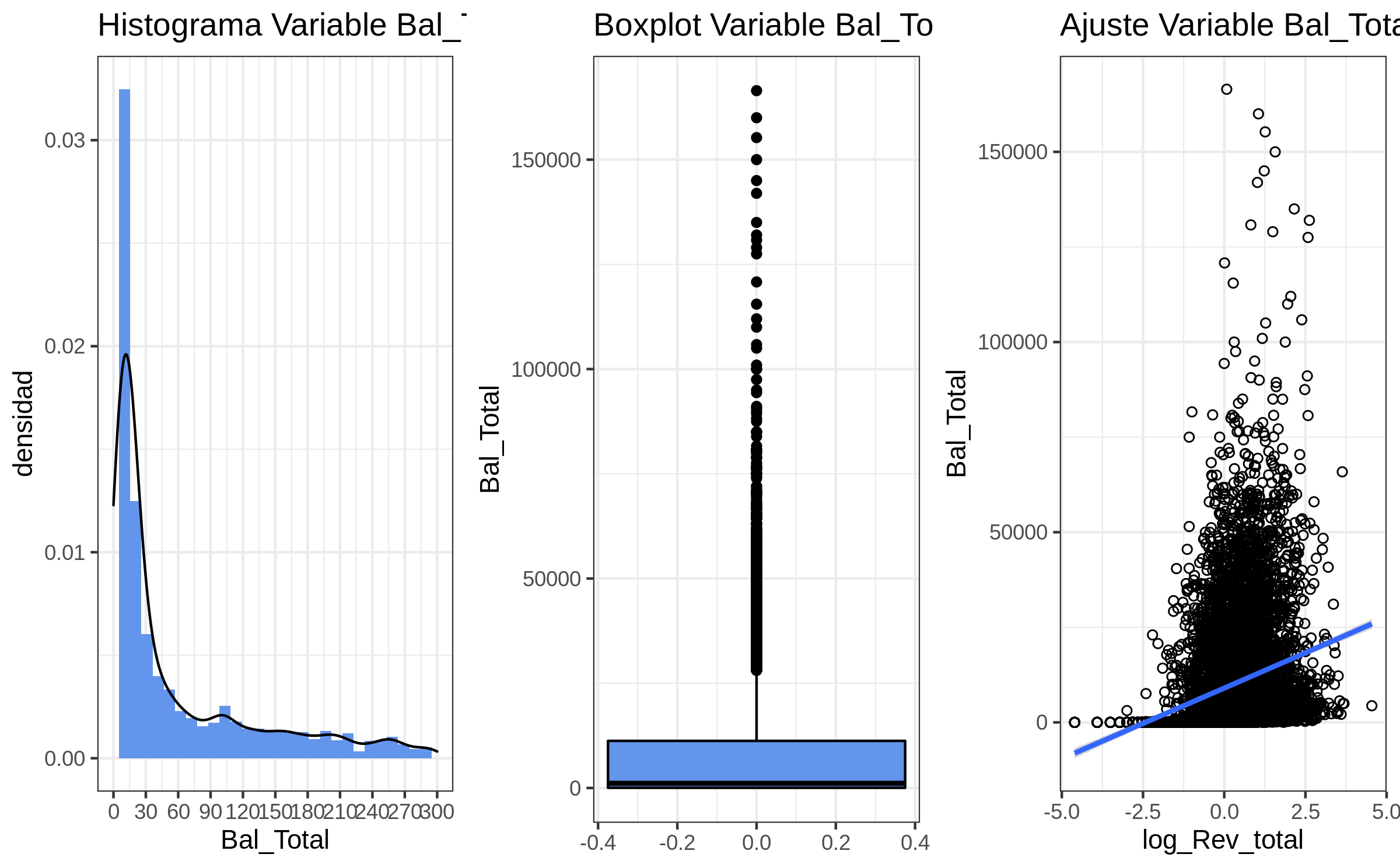


## 1.4 Distribución de Variables Numéricas y regresiones lineales simples en relación a la variable de interés.

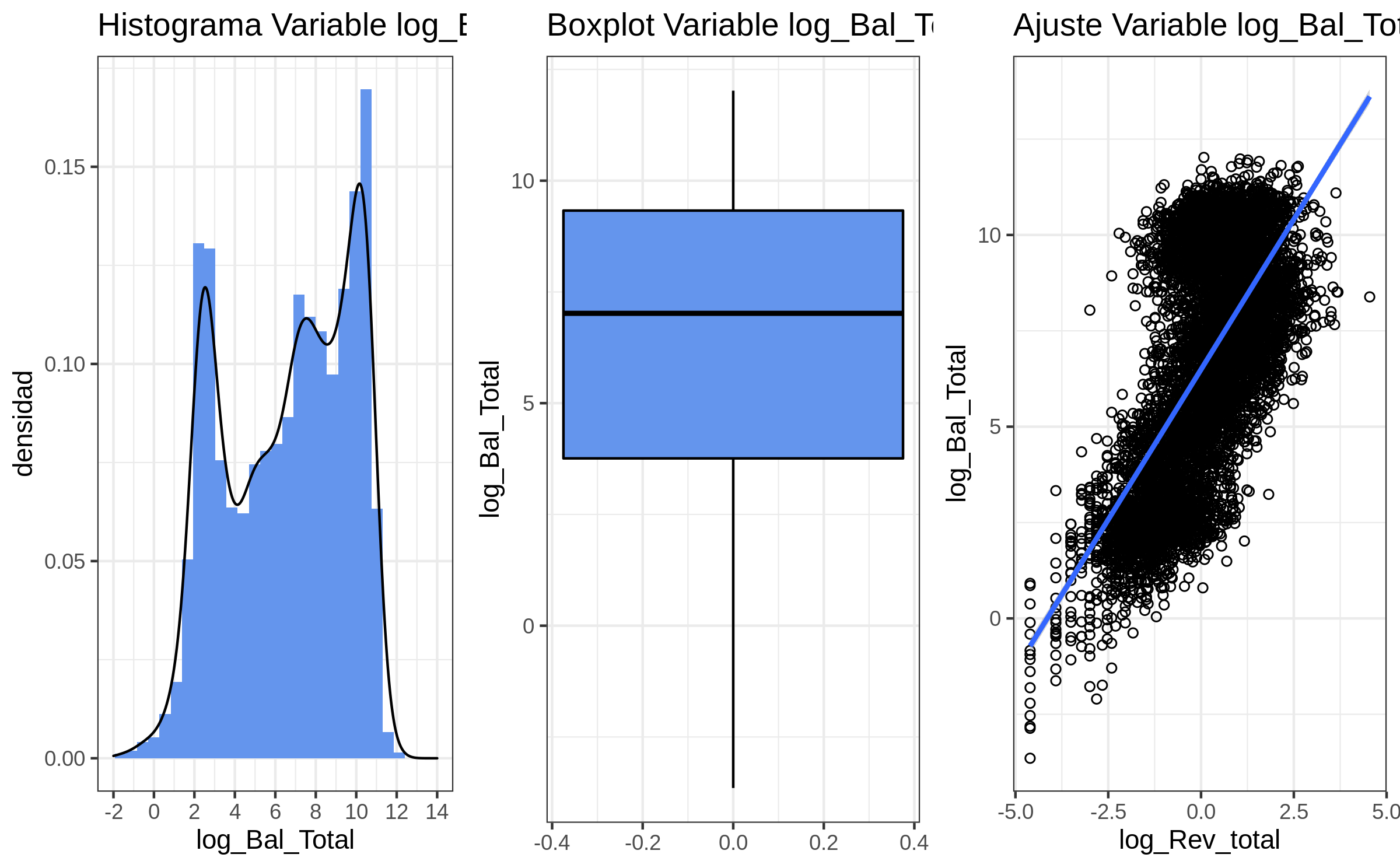
Nuestro foco para estas variables, además de su distribución, es su relación con la variable objetivo, para lo cual se realizarán regresiones simples. Si bien en este caso particular la cantidad de variables es pequeña, no siendo necesario una selección en esta etapa, se espera obtener señales preliminares de cuales variables serán las más importantes para el modelo.

### 1.4.1 Total de todas las cuentas mantenidas por el cliente (variable Bal\_Total)

Se observa que la varialbe Bal\_Total cuenta con una situación similar al de la variable objetivo si se observan los gráficos siguientes. Entonces nuevamente obtenemos los valores de kurtosis de 14.2 y skewness de 2.8, valores lejanos a una distribución normal simétrica, como se sospechaba. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.08 con un valor-p < 0.05.



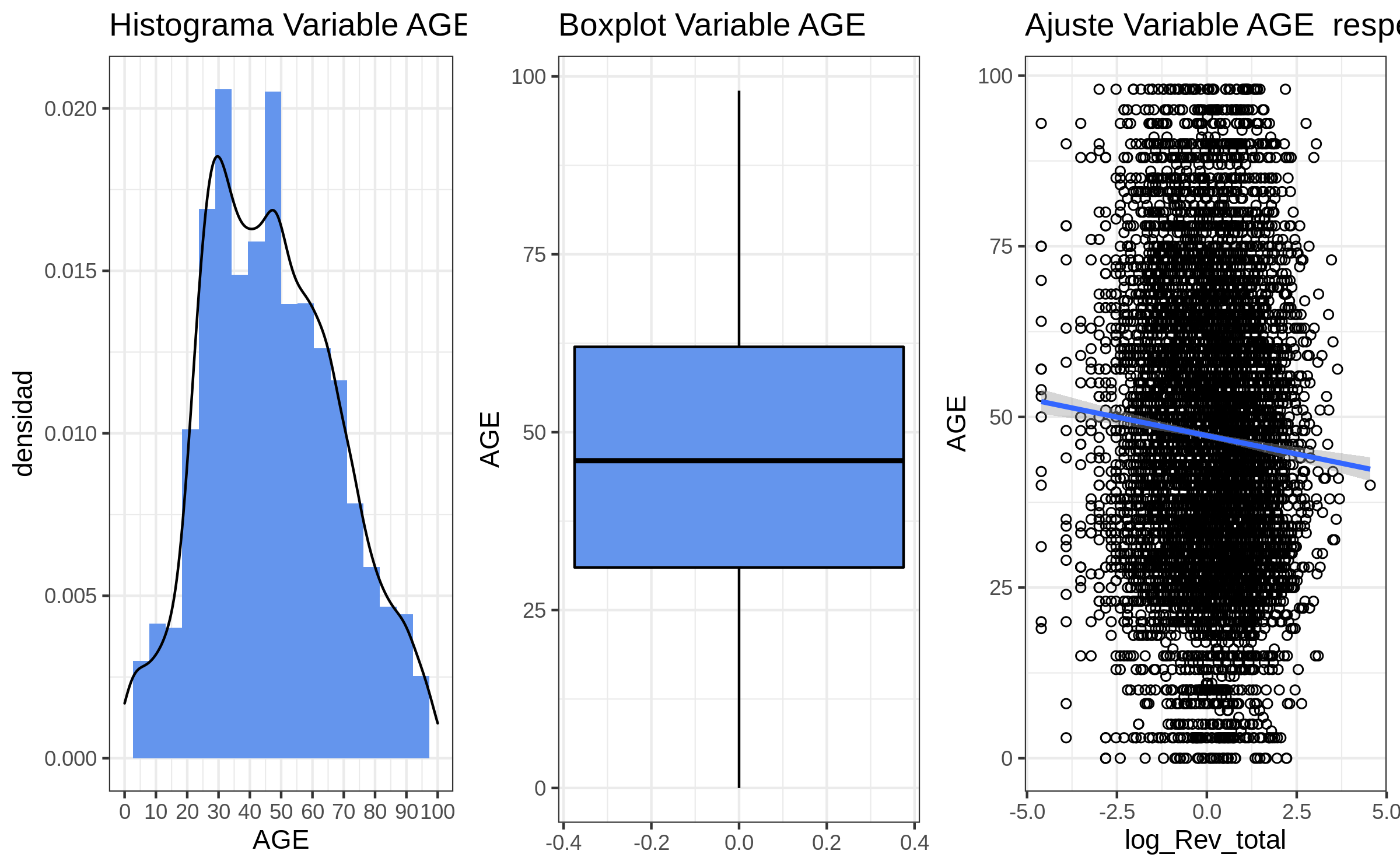
Dado lo anterior, aplicamos también la misma transformación logarítmica (viendo el resultado en los gráficos siguientes), lo que mejora los valores con una kurtosis de 1.9 y un skewness de -0.3, valores más cercanos a una distribución normal simétrica; aún cuando el resultado ha sido menos exitoso que el realizado con la variable objetivo.



Al realizar la regresión lineal simple, se observa que el coeficiente asociado es significativo, con un valor p < 0.05 y con un valor de R cuadrado de 0.4337, el cual es muy superior al que tuvo la variable posterior a la transformación no lineal. Es por lo anterior que se mantendrá la variable transformada y se espera que ésta sea parte del modelo final.

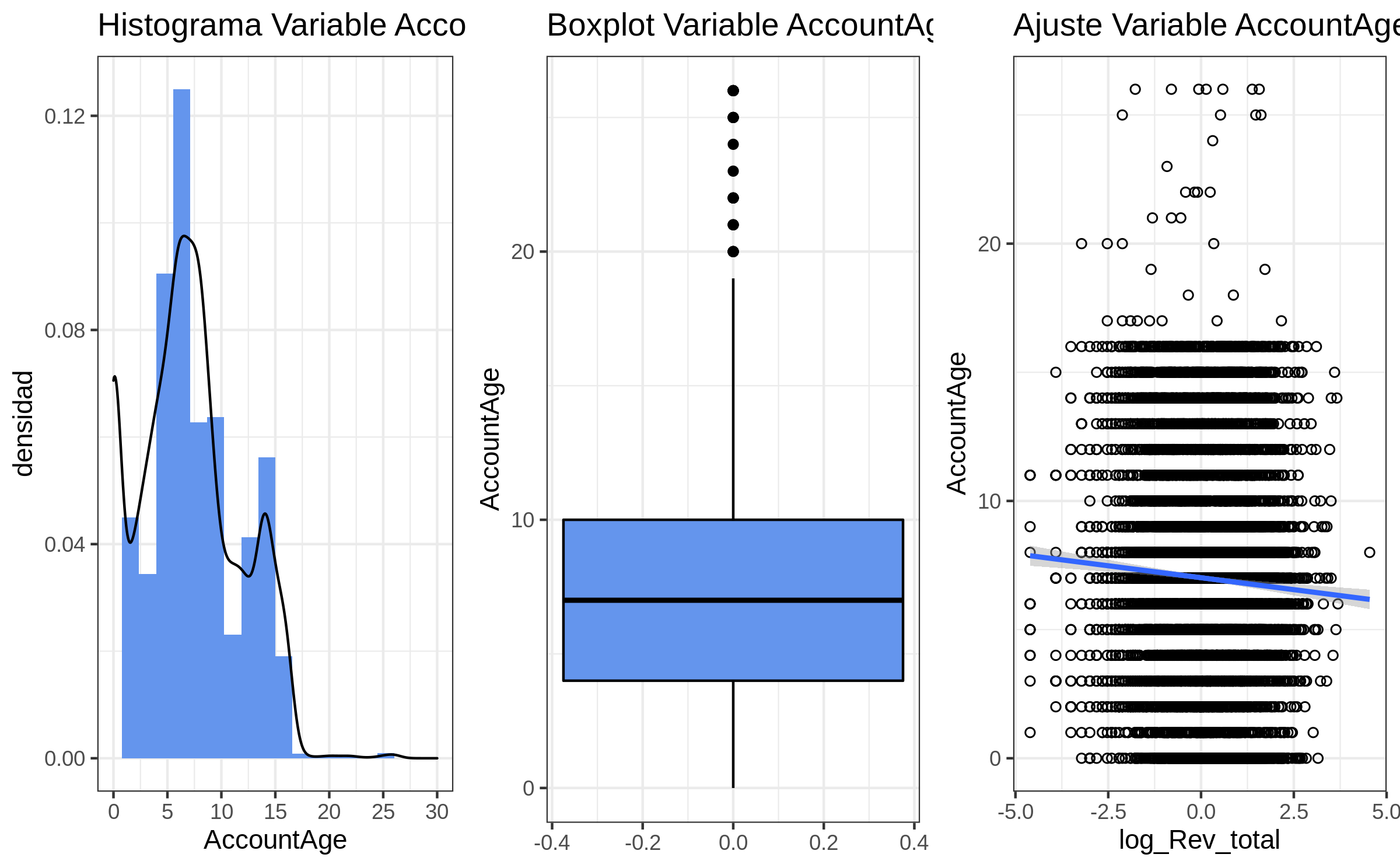
### 1.4.2 Edad del Cliente (AGE)

Variable Numérica. Se observa una distribución cercana a la normal con una kurtosis de 2.5 y un skewness de 0.2, valores que son muy superiores a los de una normal (3 y 0) respectivamente. No se realizarán transformaciones no lineales sobre esta variable, debido a lo anterior. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.004 con un valor-p < 0.05. El valor presenta un bajo nivel de ajuste, por lo que no existe seguridad que esta variable sea parte del modelo.



### 1.4.3 Antiguedad de la cuenta (AccountAge)

Variable Numérica. Se observa una distribución cercana a la normal con coeficientes de kurtosis de 2.7 y un skewness de 0.4, los que son levemente distintos a los valores de una normal (3 y 0) respectivamente, por lo que no se realizarán transformaciones no lineales sobre esta variable. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.003 con un valor-p < 0.05. Dado que el valor bajo no existe seguridad que esta variable sea parte del modelo final.

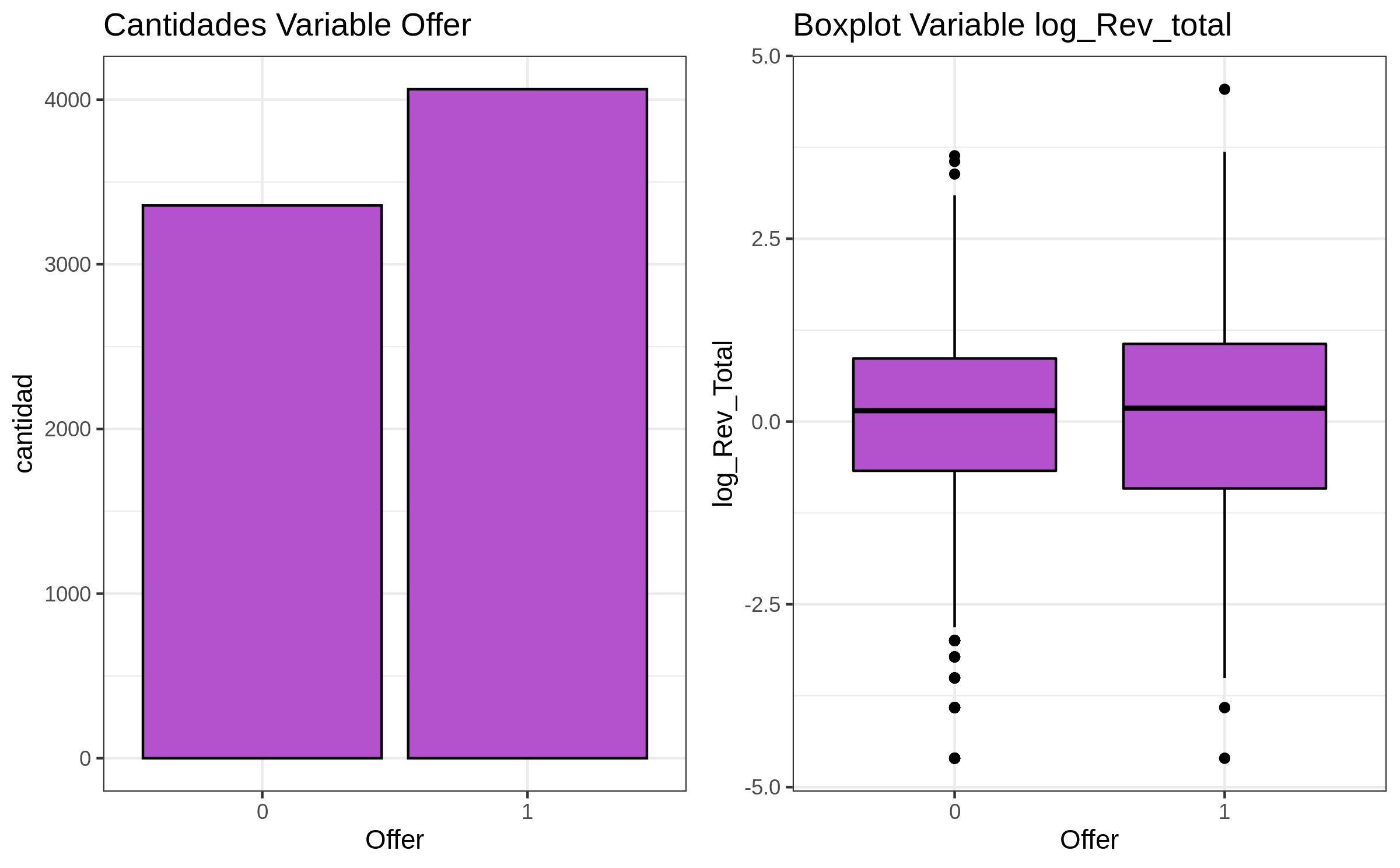


## 1.5 Distribución de Variables Categóricas y regresiones lineales simples en relación a la variable de interés

En esta sección se revisarán las variables categóricas, en términos de su balance y asociación visual, y su relación con la variable a explicar para adelantar posibles asociaciones, mediante modelos de regresión lineal simple con la variable objetivo.

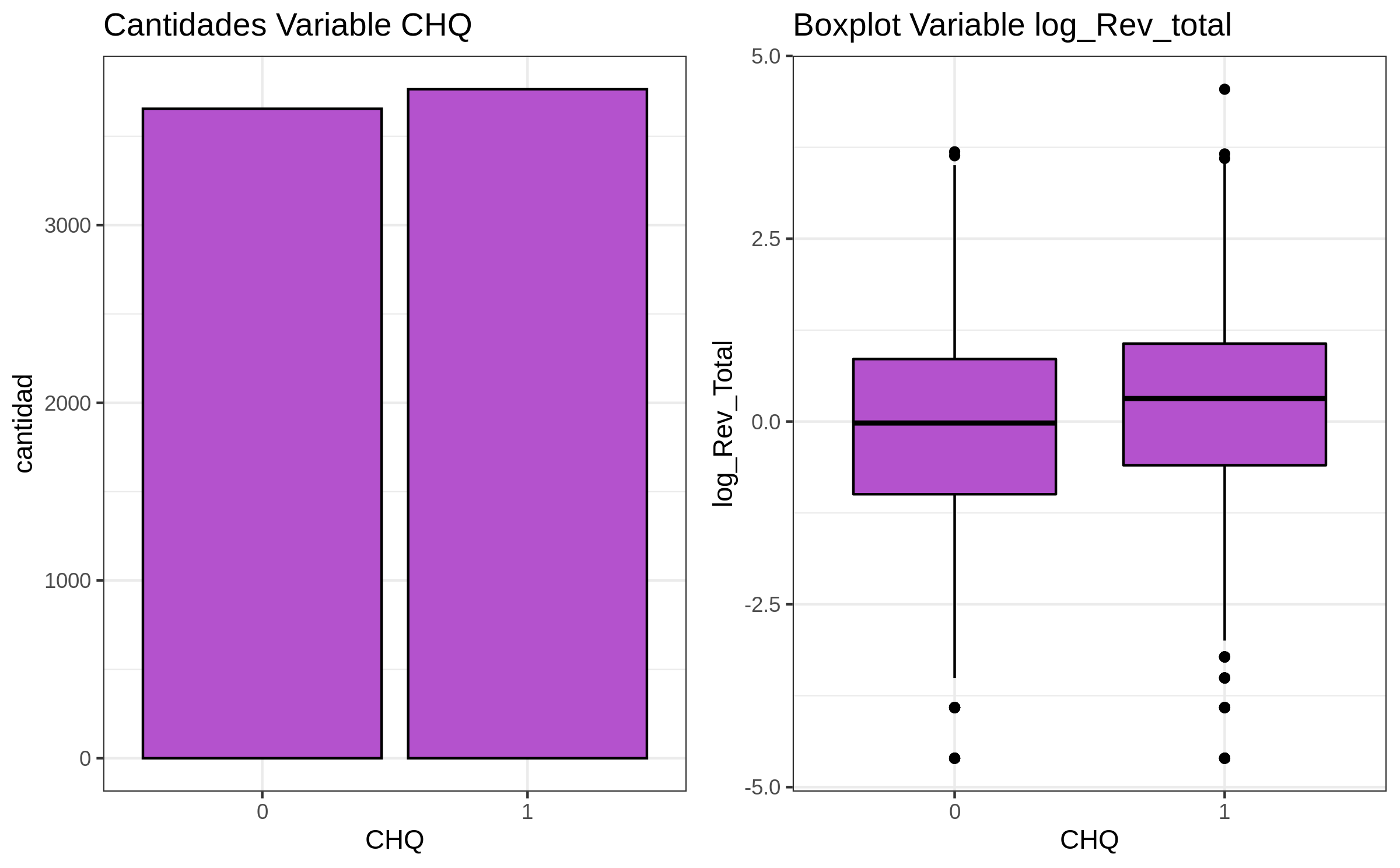
### 1.5.1 Indicador de recepción o no recepción de oferta (Offer)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0001 con un valor-p > 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final.



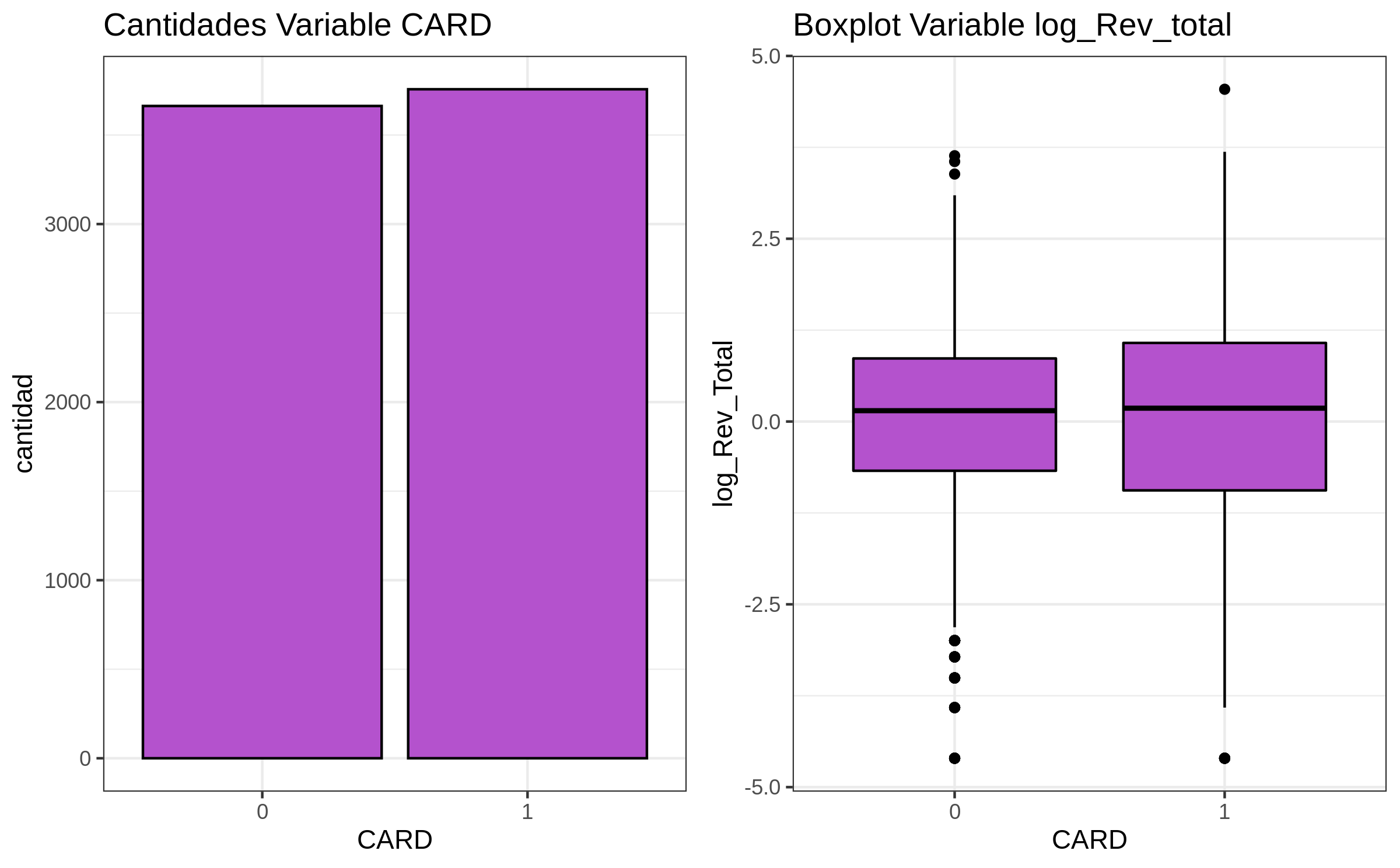
### 1.5.2 Indicador de alta o baja actividad en la tarjeta de débito (CHQ)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0132 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable sea parte del modelo final.



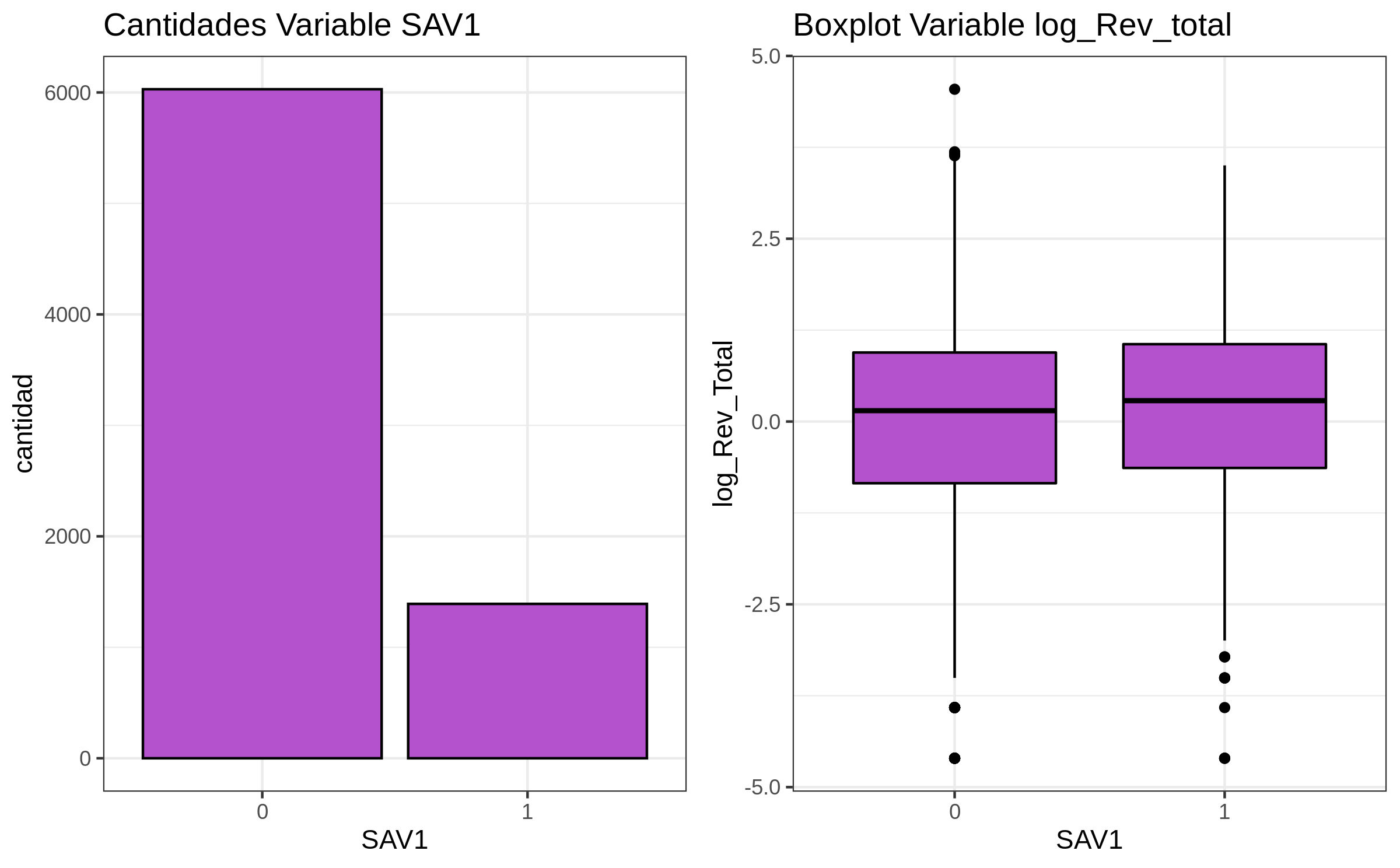
### 1.5.3 Indicador de alta o baja actividad en la tarjeta de crédito (CARD)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de -0.0001 con un valor-p > 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final.



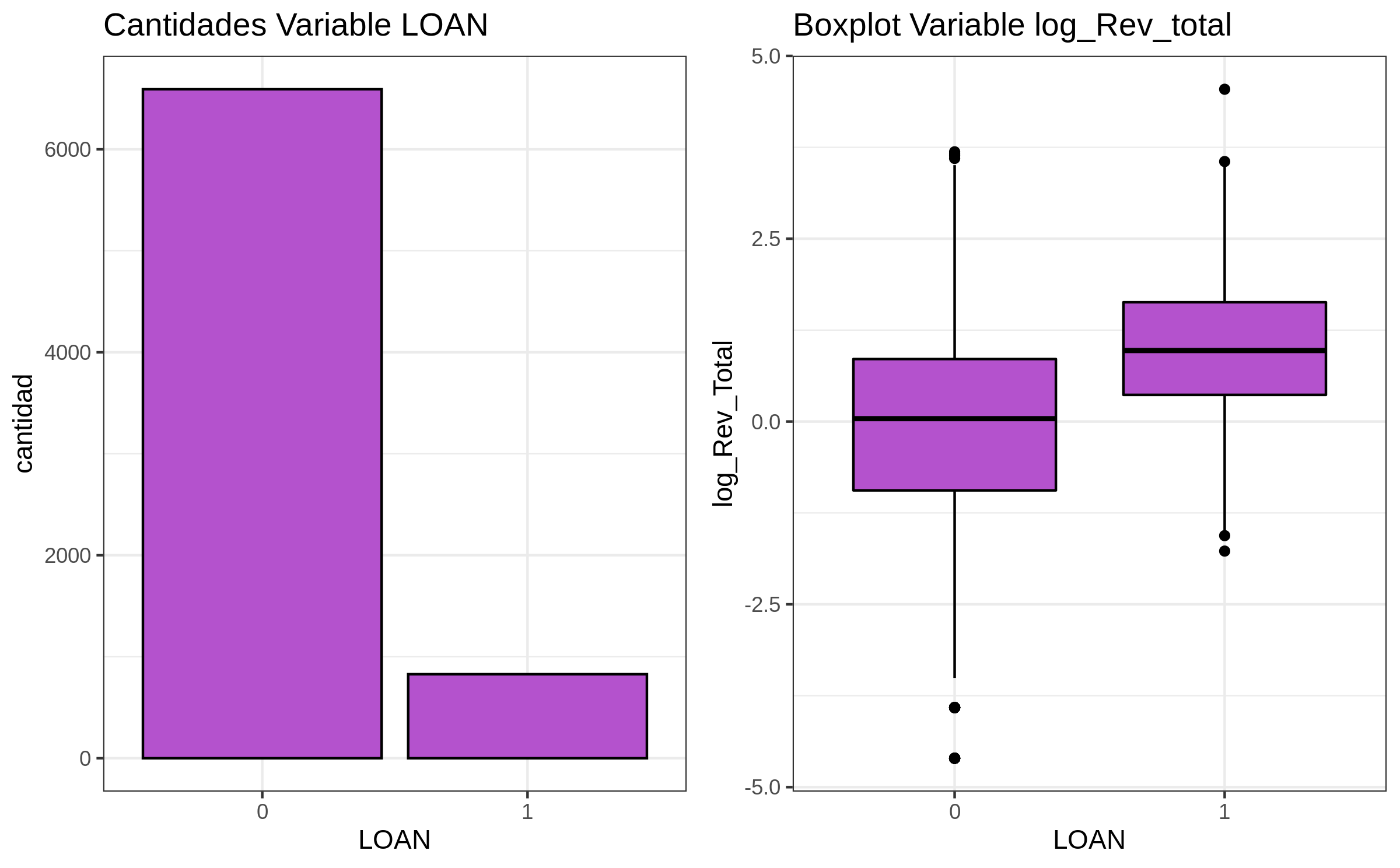
### 1.5.4 Indicador de alta o baja actividad en la cuenta de ahorro primaria (SAV1)

Se observa, en el siguiente gráfico (izquierda), que las categorías presentan un fuerte desbalance entre ellas, lo que puede presentar un problema en el modelo. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0014 con un valor-p < 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final.



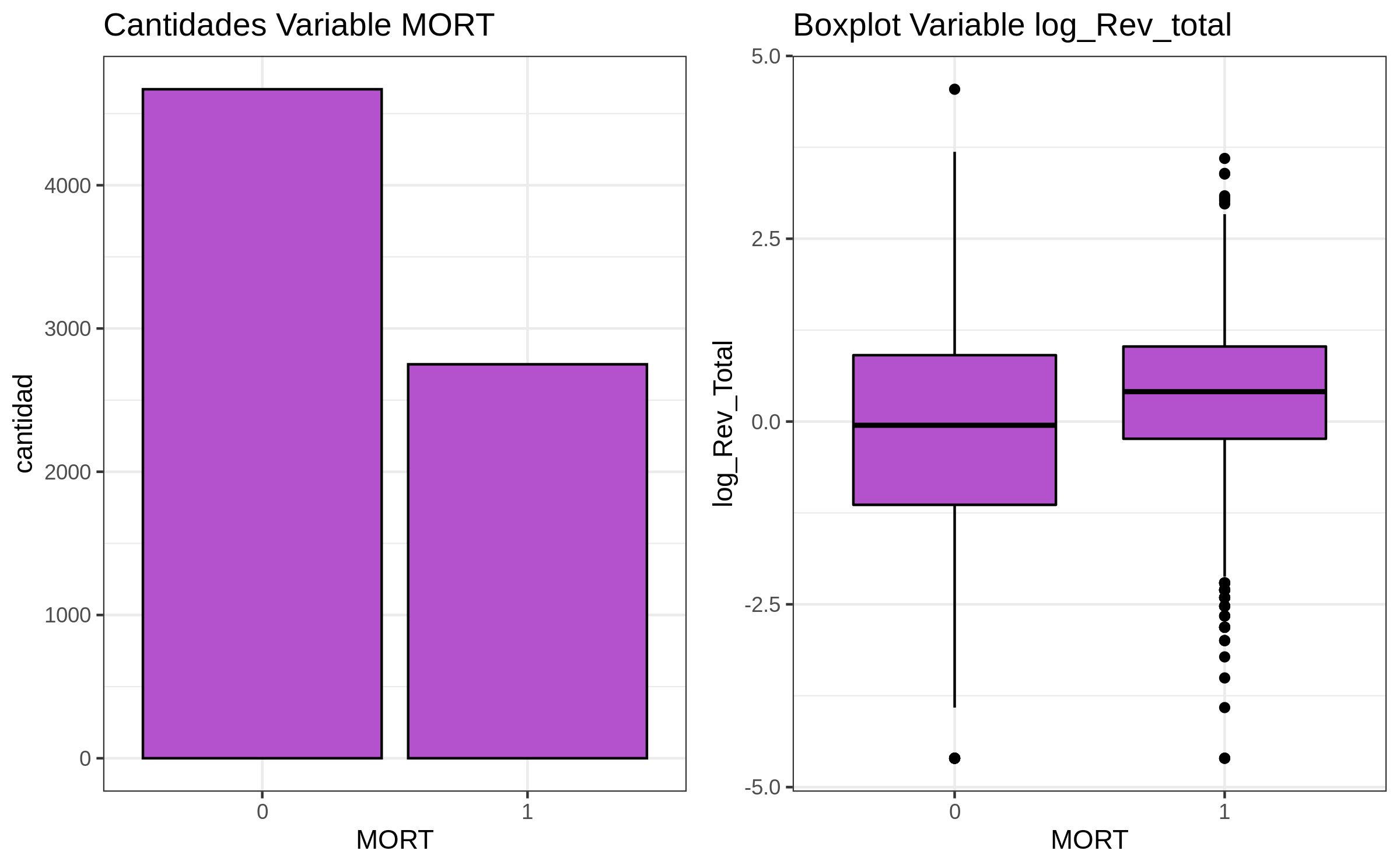
### 1.5.5 Indicador de alta o baja actividad en la cuenta de préstamo personal (LOAN)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, observa que existe una media distinta en función de la variable a determinar, compartiendo solo una parte del rango cuartil, por lo que se espera un ajuste relativamente importante. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0637 con un valor-p < 0.05. Dado el valor de ajuste, se espera que esta variable sea parte del modelo final.



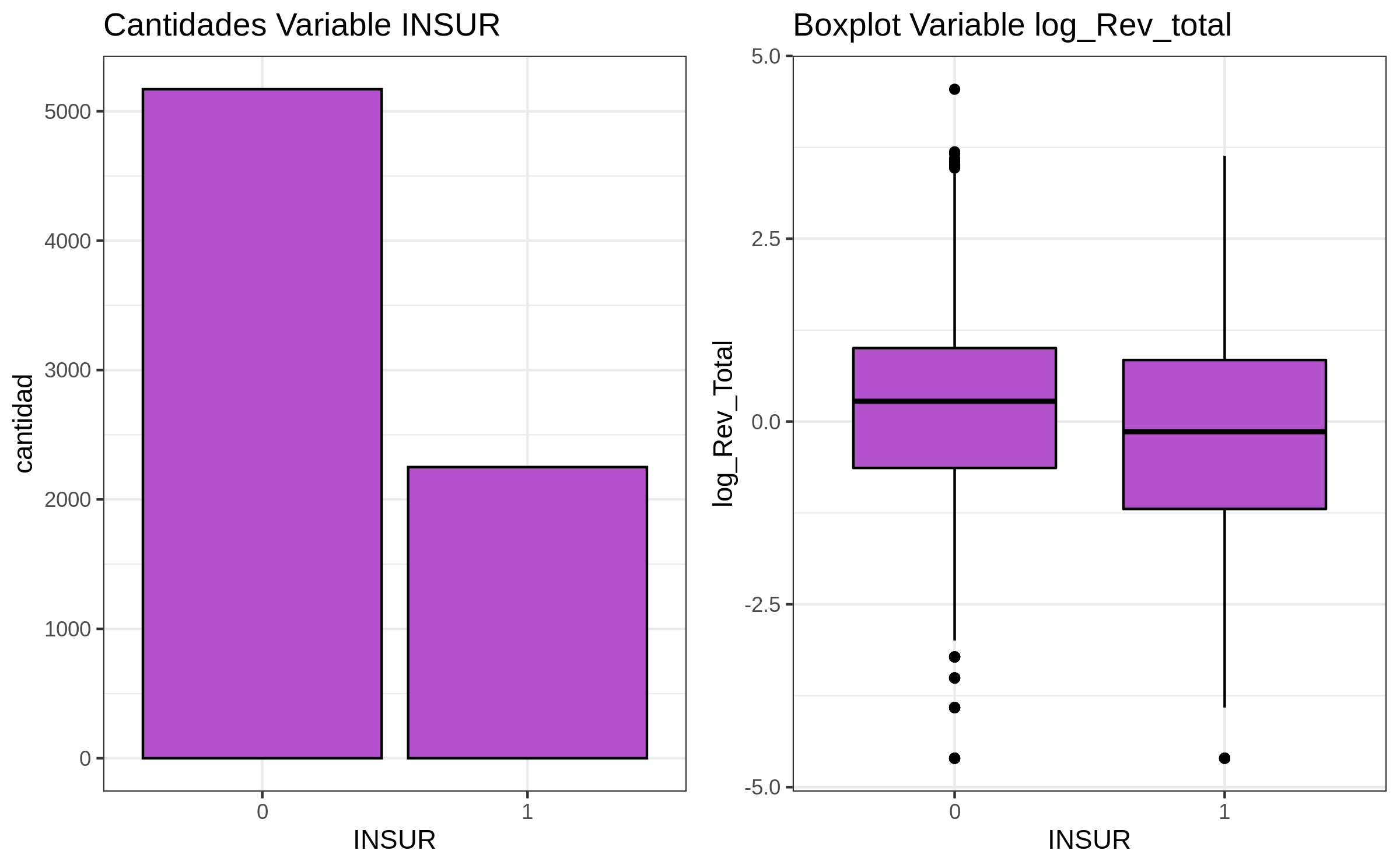
### 1.5.6 Indicador del Nivel de importancia de la cuenta hipotecaria (MORT)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0326 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable sea parte del modelo final.



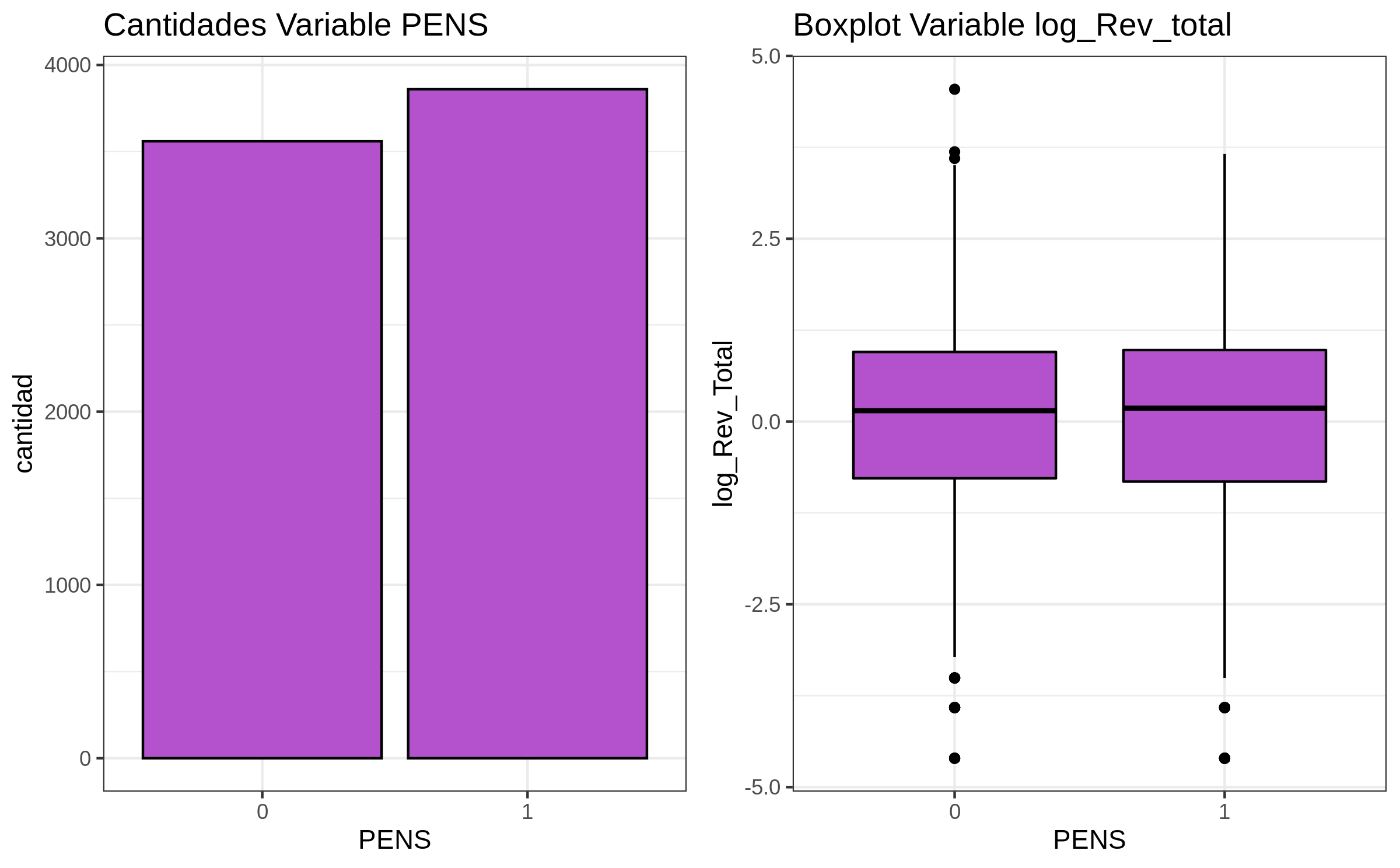
### 1.5.7 Indicador de alta o baja actividad en la cuenta de seguros (INSUR)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.015 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable sea parte del modelo final.



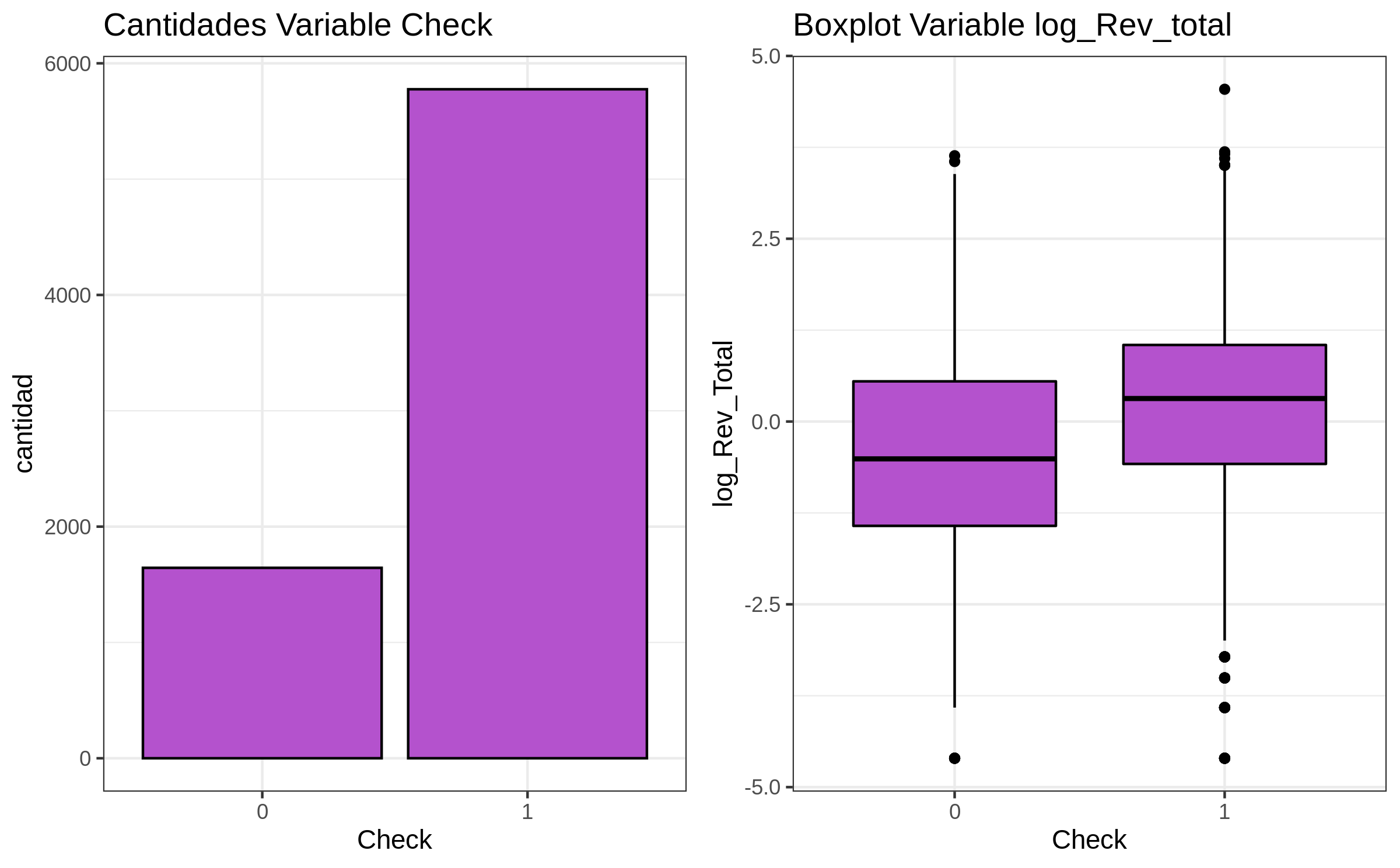
### 1.5.8 Indicador del Nivel de ahorro para la jubilación (PENS)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de -0.0001 con un valor-p > 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final.



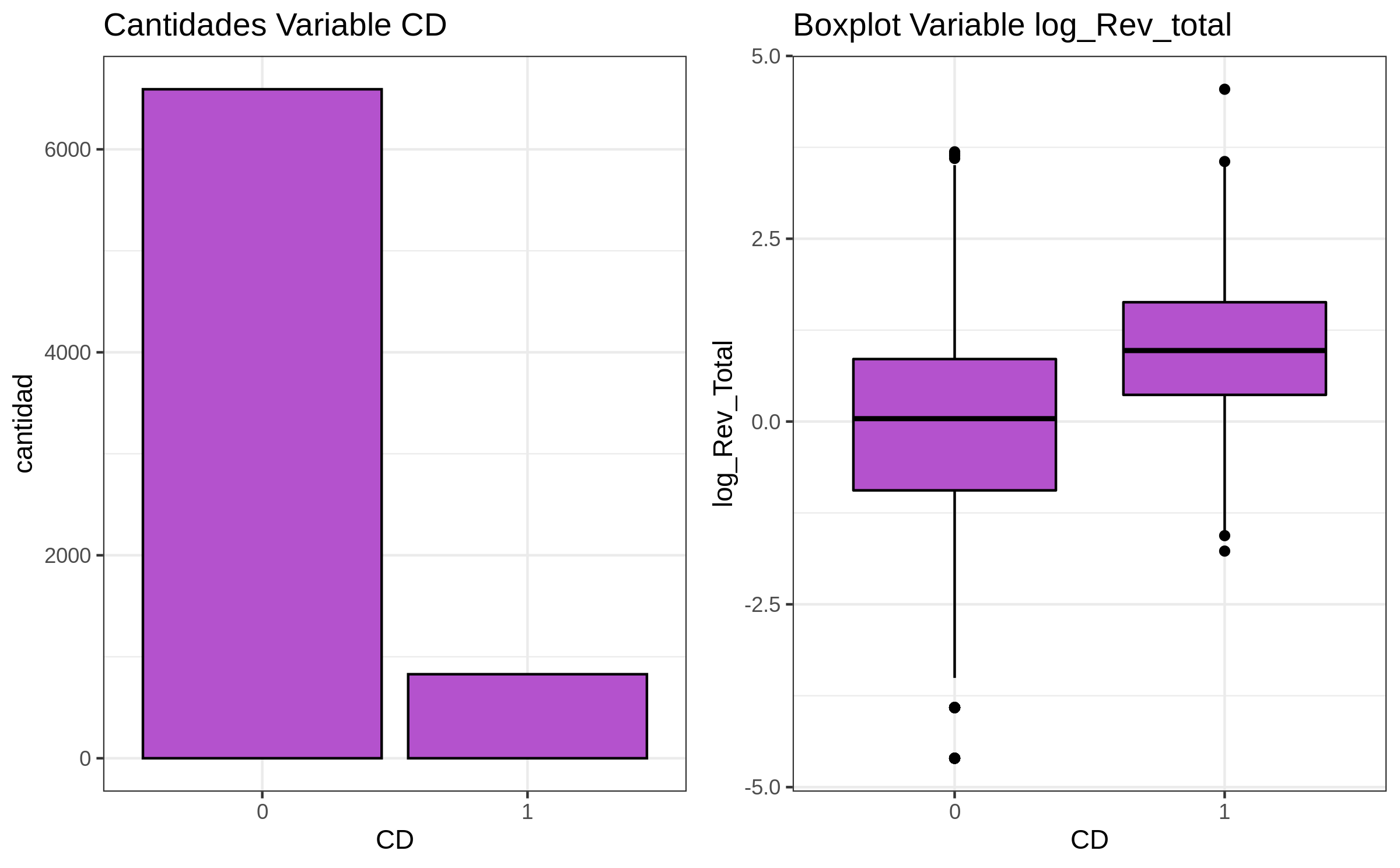
### 1.5.9 Indicador de alta o baja actividad en la cuenta corriente (Check)

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.043 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable sea parte del modelo final.



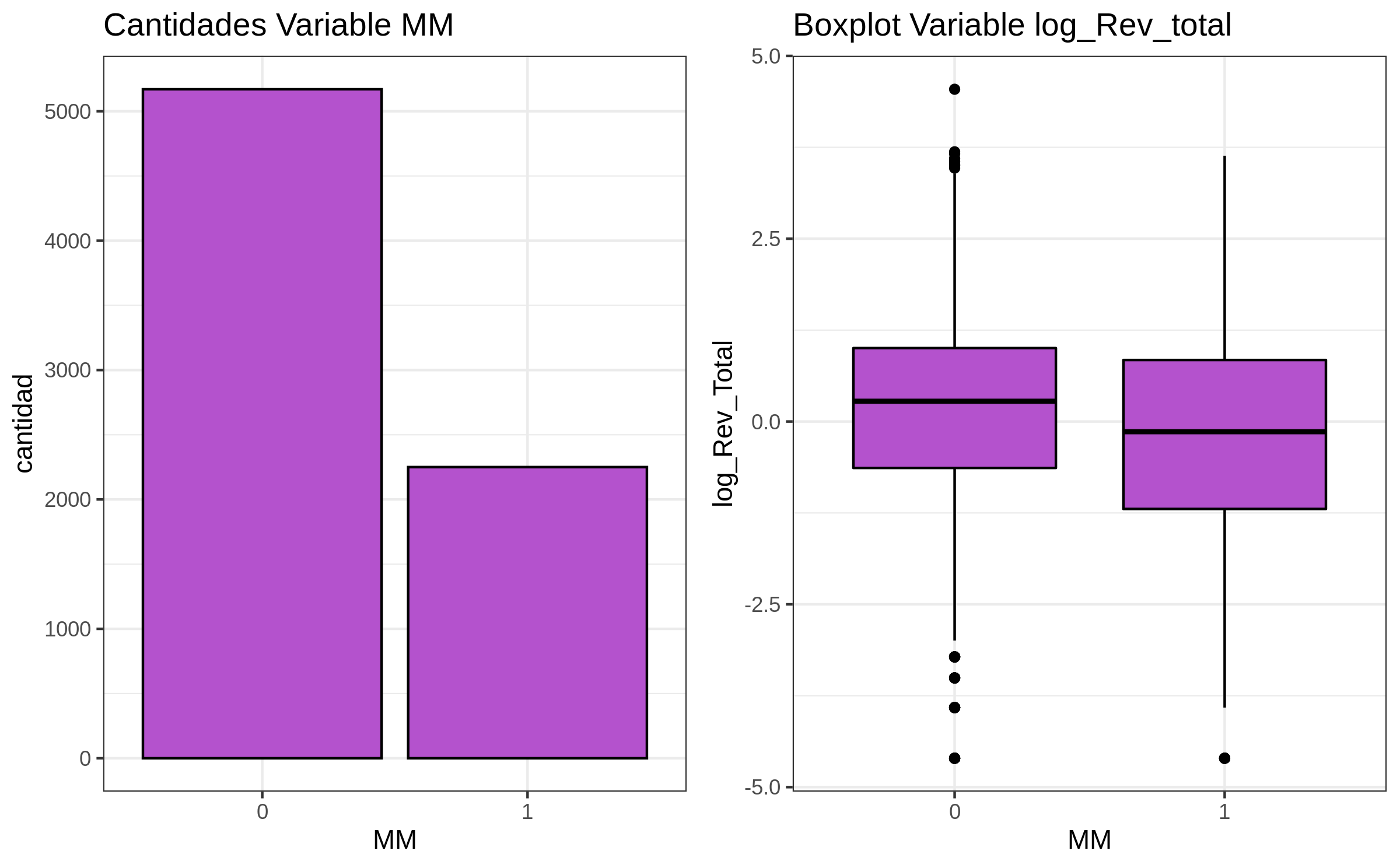
### 1.5.10 Indicador de importancia del nivel de certificado de depósito (CD)

Se observa, en el siguiente gráfico (izquierda), que las categorías presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, observa que existe una media distinta en función de la variable a determinar, compartiendo solo una parte del rango cuartil, por lo que se espera un ajuste relativamente importante. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0637 con un valor-p < 0.05. Dado el valor de ajuste, se espera que esta variable sea parte del modelo final.



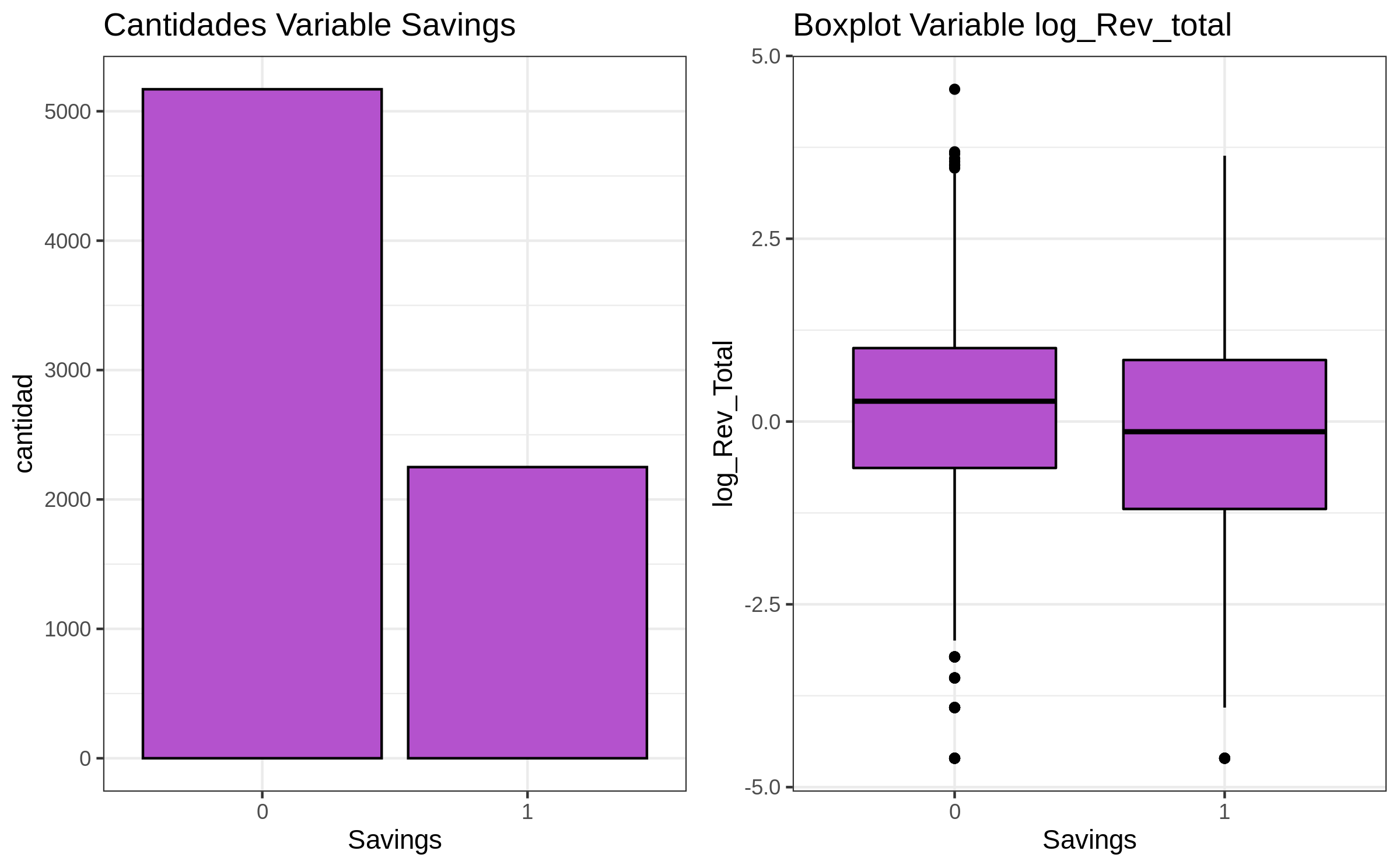
### 1.5.11 Indicador del nivel de actividad en la cuenta del mercado monetario (MM)

Se observa, en el siguiente gráfico (izquierda), que las categorías presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.015 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable haga parte del modelo final.



### 1.5.12 Indicador del nivel de actividad en la cuenta de ahorro que no sea primaria (Savings)

Se observa, en el siguiente gráfico (izquierda), que las categorías presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una una media distinta, pero no absoluta, en función de la variable a determinar, debido a que comparten valores en el rengo cuartil 25-50, por lo que se espera un valor bajo de ajuste. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.015 con un valor-p < 0.05. Si bien el ajuste es mayor que el caso anterior sigue siendo bajo, por lo que no se puede asegurar que esta variable haga parte del modelo final.

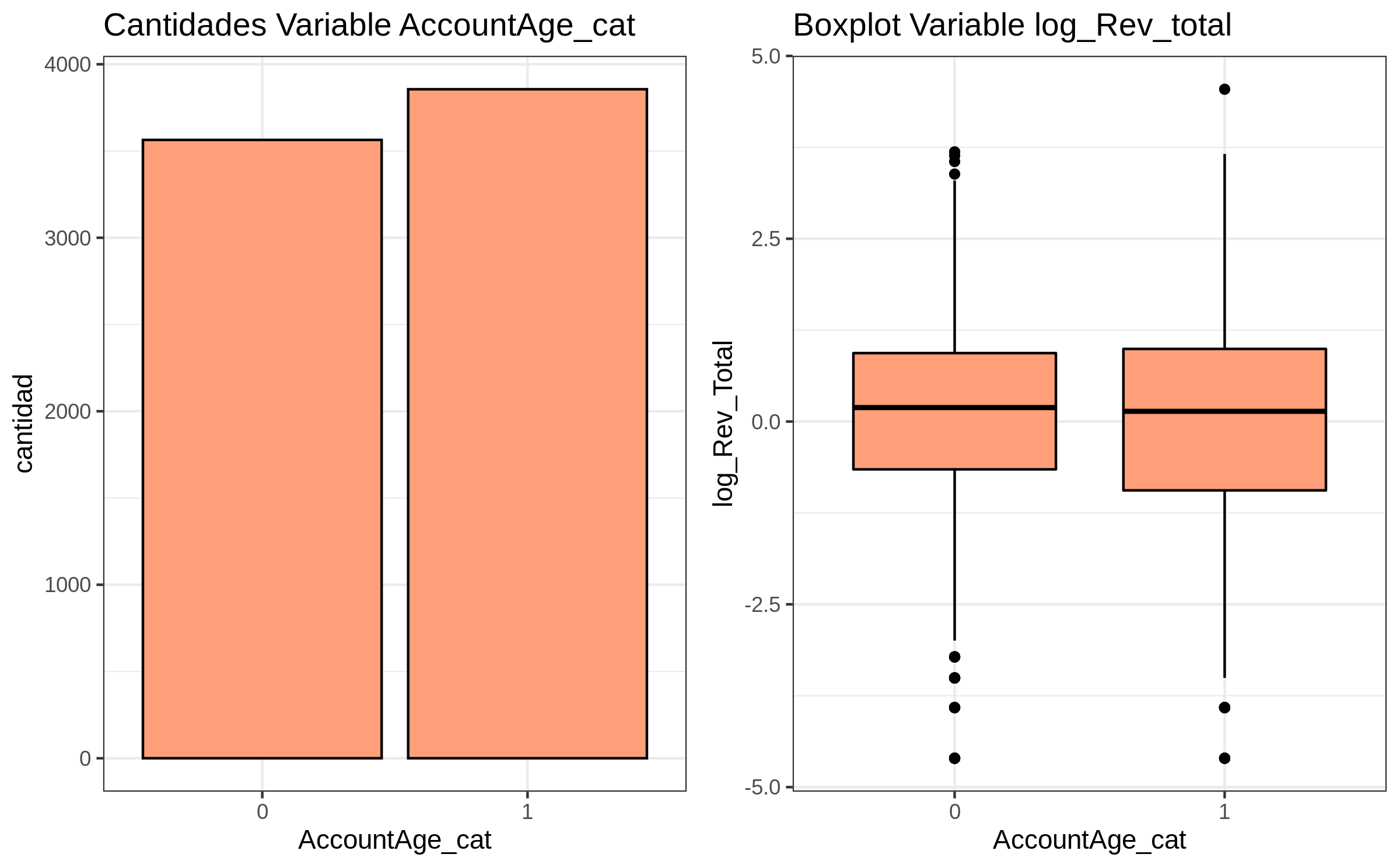


## 1.6 Distribución de variables numéŕicas convertidas a categóricas y regresiones simples en relación a la variable de interés.

A continuación se convertirán las variables numéricas en variables categóricas con el objetivo de modelar su distribución no normal y analizar si esto mejora la capacidad predictiva del modelo en etapas posteriores, medida en esta etapa en base a la comparación entre regresiones lineales simples.

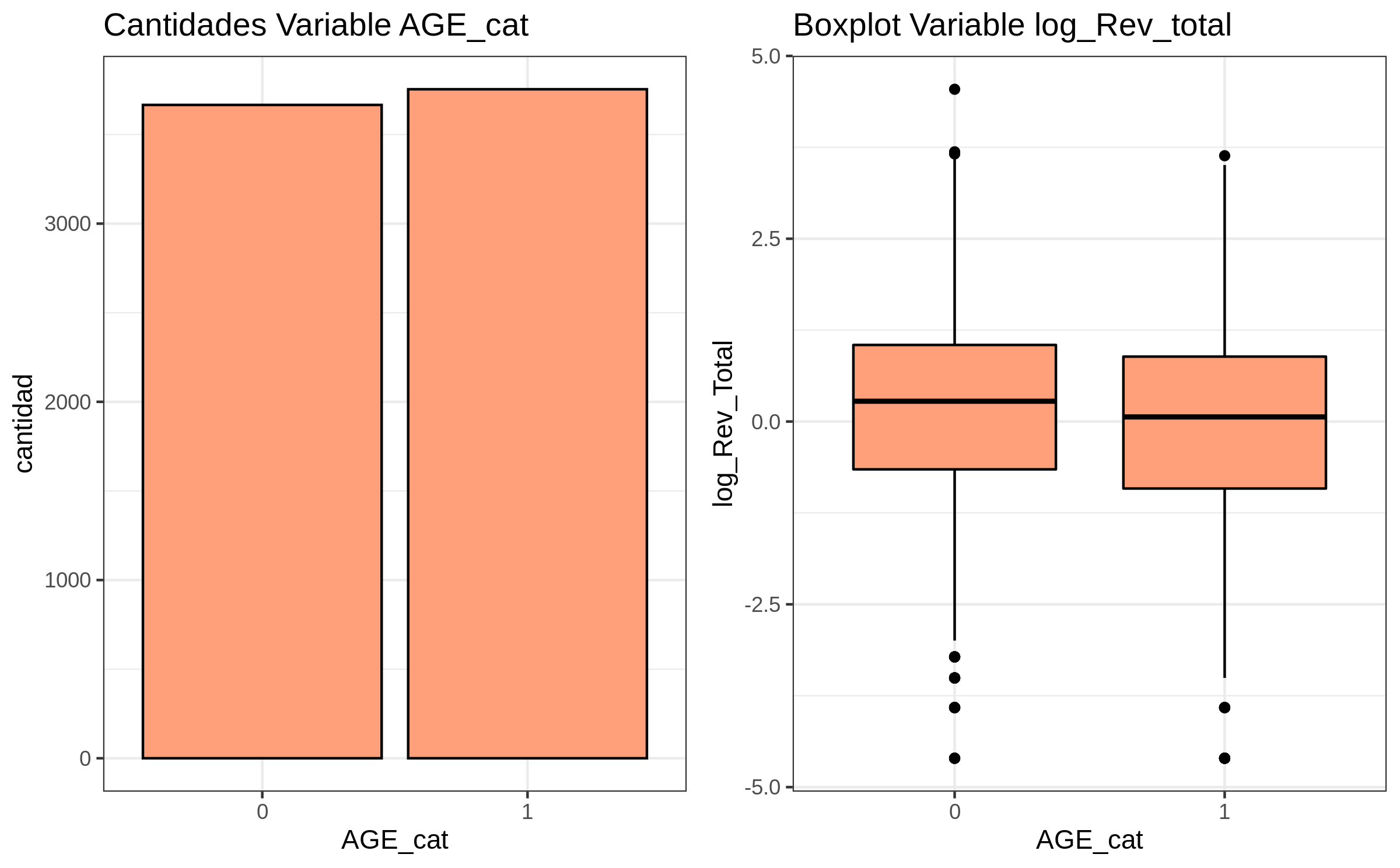
### 1.6.1 Variable Antiguedad de la cuenta “Account Age”

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0006 con un valor-p < 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final. Dado que no se provoca mejora, se mantendrá la variable original.



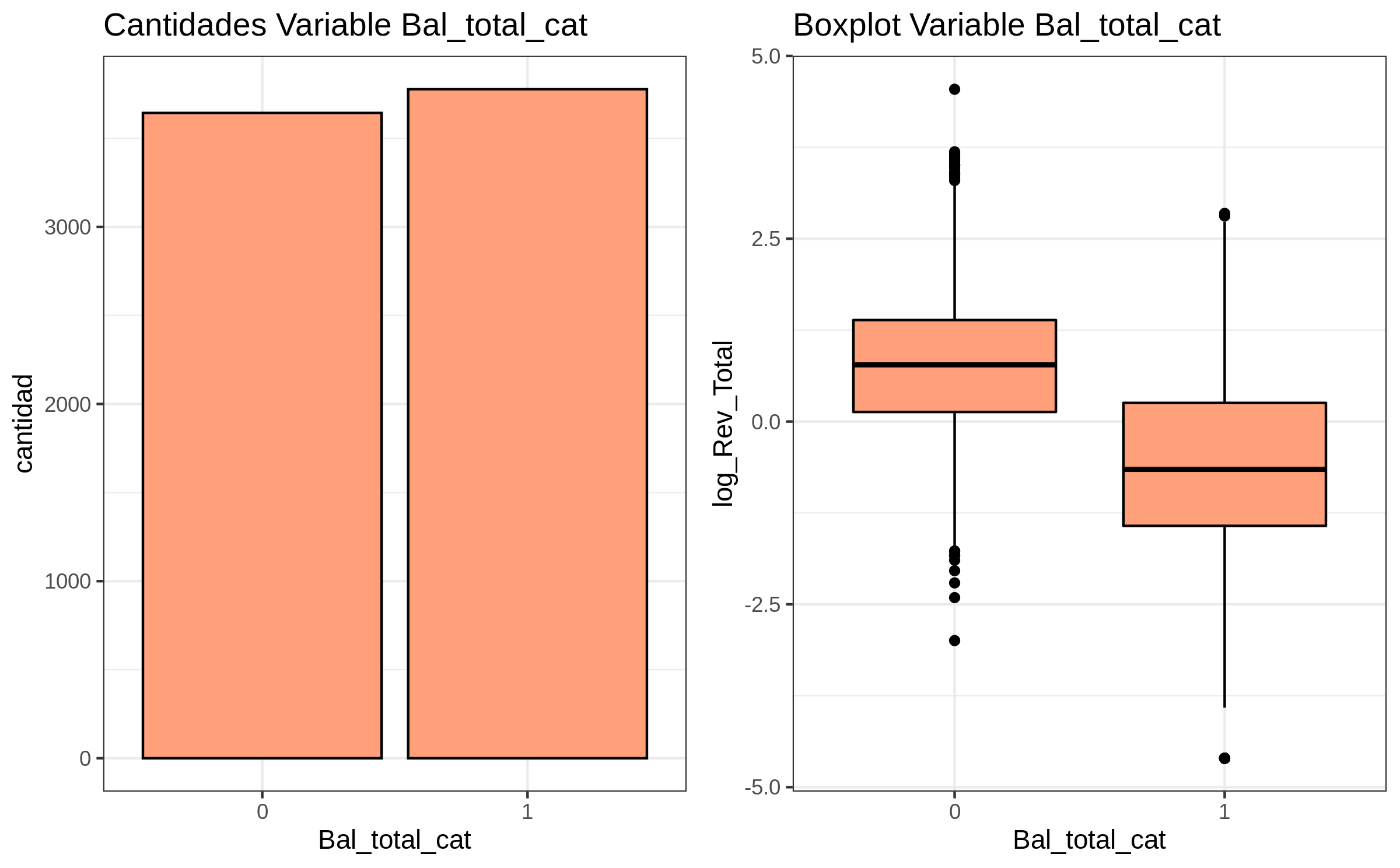
### 1.6.2 Variable Edad del Cliente “Age”

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, no se observa que existe una media distinta en función de la variable a determinar, por lo que no se espera una relación de ajuste alta. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.0039 con un valor-p < 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final. Dado que no se provoca mejora, se mantendrá la variable original.



### 1.6.3 Variable Total de todas las cuentas mantenidas por el cliente “Bal\_total”

Se observa, en el siguiente gráfico (izquierda), que las categorías no presentan un fuerte desbalance entre ellas. Al observar el gráfico de la derecha, se observa una diferencia con muy poco rango intercuartil en común con la separación realizada. Al evaluar esta variable en una regresión simple junto a la variable dependiente entrega un ajustado de 0.298 con un valor-p < 0.05. Dado el bajo valor de ajuste no se espera que esta variable sea parte del modelo final. Si bien el valor es alto, éste no supera al valor de la variable transformada mediante logaritmo, por lo que se mantiene la transformación original.



## 1.7 Conclusiones Analisis Exploratorio.

Luego del análisis exploratorio, se ha establecido para la variable objetivo una transformación no lineal (logaritmo), con la finalidad de acercar la distribución a una normal centrada. Lo anterior permite mejorar la predicción con el costo de una interpretación directa de los ingresos (variable Rev\_total).

Por otra parte, para las variables explicativas se han separado, a partir del que entrega cada modelo de regresión simple para la variable objetivo, en términos de su probabilidad de participación del modelo final:

* Altamente Probables ( > 0.05): Bal\_Total, log\_Bal\_Total,Bal\_total\_cat, LOAN, CD
* Probables (0.05 > > 0.01): CHQ, MORT, INSUR, Check, MM, Savings
* Poco Probables ( < 0.01): Age, AccountAge, Offer, CARD, SAV1, PENS, AccountAge\_cat, AGE\_cat

Se observa que las transformaciones de las variables (ya sea mediante aplicación de logaritmo o categorización), no mejora su capacidad de ajuste respecto de la variable de interés. La única transformación que se mantiene es la no lineal de logaritmo, ya que existe un indicador de ajuste mayor.

A continuación se observan los valores beta de las regresiones simples entre cada variable independiente y la variable dependiente. Estos valores se analizarán con los coeficientes obtenidos en la regresión múltiple del modelo final, dentro de la interpretación de coeficientes.

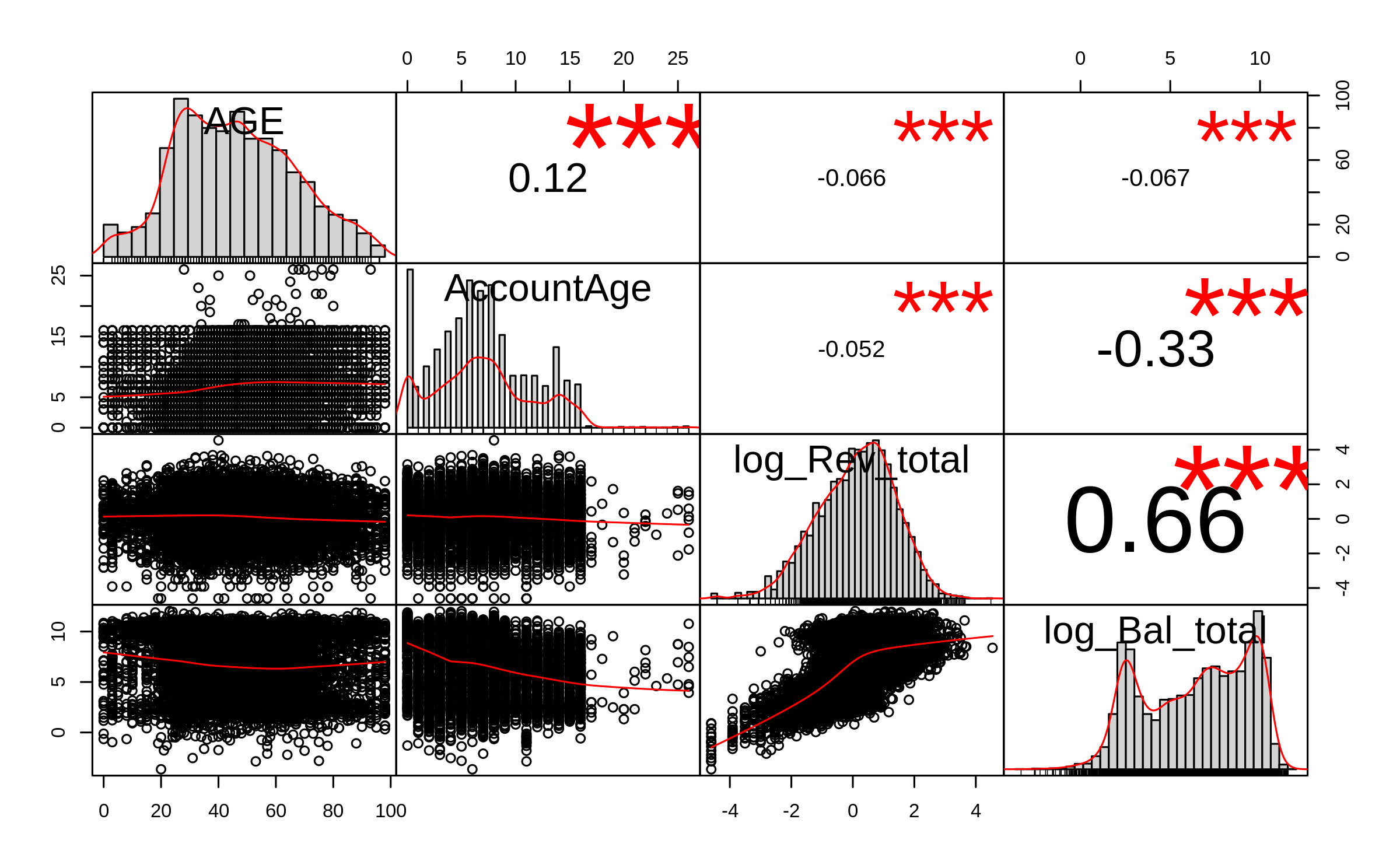
library(dplyr)  
betas %>% dplyr::select(2) %>% pander()

|  |  |
| --- | --- |
|  | beta\_ind |
| **Bal\_Total** | 0.00002287 |
| **log\_Bal\_total** | 0.2769 |
| **AGE** | -0.003987 |
| **AccountAge** | -0.01452 |
| **Offer1** | 0.04156 |
| **CHQ1** | 0.2953 |
| **CARD1** | 0.01461 |
| **SAV11** | 0.1268 |
| **LOAN1** | 1.026 |
| **MORT1** | 0.4788 |
| **INSUR1** | -0.3419 |
| **PENS1** | 0.004865 |
| **Check1** | 0.6396 |
| **CD1** | 1.026 |
| **MM1** | -0.3419 |
| **Savings1** | -0.3419 |
| **AccountAge\_cat1** | -0.0711 |
| **AGE\_cat1** | -0.1619 |
| **Bal\_total\_cat1** | -1.397 |

# 2. Análisis de correlaciones

Para el análisis de correlaciones primero nos centraremos en las variables numéricas, con el objetivo de identificar si existen, preliminarmente, relaciones fuertes o débiles entre las variables, puestas en la siguiente figura.

Considerando que la variable a predecir es el logaritmo de los ingresos totales percibidos a partir de cada cliente (log\_Rev\_total), observamos que si bien todas las correlaciones entre variables numéricas son significativas, hay una correlación fuerte y positiva (valor 0.66) con el logaritmo de las cuentas mantenidas por el cliente (log\_Bal\_total), con un gráfico cercano a una relación lineal. La siguiente relación, de un 0.33, es negativa entre el logaritmo de las cuentas mantenidas por el cliente (log\_Bal\_total) y la antiguedad de la cuenta (AccountAge), con un comportamiento no lineal. Dado que éstas son variables explicativas, podrían producir un efecto de inflación por colinealidad. Luego se observa una relación más débil (0.12) entre las variables edad del cliente (AGE) y antiguedad de la cuenta (AccountAge).



Respecto de las variables categóricas, se analiza la correlación existente entre ellas con el estadístico de Cramer. Se observa una correlación perfecta entre algunas de ellas (“LOAN” y “CD”, “INSUR” y “MM”, “INSUR” y “Savings” y entre “MM” y “Savings”), lo cual indica que no deberían ser ingresadas al modelo aquellas que están mostrando dicha redundancia, es decir, CD, MM y Savings. Offer y CARD también presentan una alta correlación, de .87, sin embargo esperaremos a analizar colinealidad antes de descatar alguna de dichas variables del modelo.

Table continues below

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Offer | | | CHQ | CARD | | SAV1 | | | LOAN | | MORT |
| **Offer** | 1 | | | 0.09553 | 0.8697 | | 0.003938 | | | 0.05021 | | 0.3559 |
| **CHQ** | 0.09553 | | | 1 | 0.105 | | 0.2543 | | | 0.2773 | | 0.09188 |
| **CARD** | 0.8697 | | | 0.105 | 1 | | 0.001167 | | | 0.03445 | | 0.3982 |
| **SAV1** | 0.003938 | | | 0.2543 | 0.001167 | | 1 | | | 0.07324 | | 0.0511 |
| **LOAN** | 0.05021 | | | 0.2773 | 0.03445 | | 0.07324 | | | 1 | | 0.09206 |
| **MORT** | 0.3559 | | | 0.09188 | 0.3982 | | 0.0511 | | | 0.09206 | | 1 |
| **INSUR** | 0.2457 | | | 0.0735 | 0.264 | | 0.1527 | | | 0.04649 | | 0.2586 |
| **PENS** | 0.01335 | | | 0.09302 | 0.006178 | | 0.06247 | | | 0.09447 | | 0.08467 |
| **Check** | 0.4676 | | | 0.06114 | 0.5403 | | 0.009816 | | | 0.1974 | | 0.3745 |
| **CD** | 0.05021 | | | 0.2773 | 0.03445 | | 0.07324 | | | 1 | | 0.09206 |
| **MM** | 0.2457 | | | 0.0735 | 0.264 | | 0.1527 | | | 0.04649 | | 0.2586 |
| **Savings** | 0.2457 | | | 0.0735 | 0.264 | | 0.1527 | | | 0.04649 | | 0.2586 |
|  | | INSUR | PENS | | | Check | | CD | MM | | Savings | |
| **Offer** | | 0.2457 | 0.01335 | | | 0.4676 | | 0.05021 | 0.2457 | | 0.2457 | |
| **CHQ** | | 0.0735 | 0.09302 | | | 0.06114 | | 0.2773 | 0.0735 | | 0.0735 | |
| **CARD** | | 0.264 | 0.006178 | | | 0.5403 | | 0.03445 | 0.264 | | 0.264 | |
| **SAV1** | | 0.1527 | 0.06247 | | | 0.009816 | | 0.07324 | 0.1527 | | 0.1527 | |
| **LOAN** | | 0.04649 | 0.09447 | | | 0.1974 | | 1 | 0.04649 | | 0.04649 | |
| **MORT** | | 0.2586 | 0.08467 | | | 0.3745 | | 0.09206 | 0.2586 | | 0.2586 | |
| **INSUR** | | 1 | 0.1142 | | | 0.623 | | 0.04649 | 1 | | 1 | |
| **PENS** | | 0.1142 | 1 | | | 0.1012 | | 0.09447 | 0.1142 | | 0.1142 | |
| **Check** | | 0.623 | 0.1012 | | | 1 | | 0.1974 | 0.623 | | 0.623 | |
| **CD** | | 0.04649 | 0.09447 | | | 0.1974 | | 1 | 0.04649 | | 0.04649 | |
| **MM** | | 1 | 0.1142 | | | 0.623 | | 0.04649 | 1 | | 1 | |
| **Savings** | | 1 | 0.1142 | | | 0.623 | | 0.04649 | 1 | | 1 | |

Además, se analizó la correlación biserial entre cada variable categórica dicotómica y la variable objetivo. Todas las variables categóricas presentaron una baja correlación con la variable objetivo. Luego de cada correlación biserial entre cada par de variables (introducida por “[1]”) se añade la correlación de Pearson, ya que la literatura sugiere que ambos dan los mismos valores. Esto se comprueba, por lo que posteriormente se correlacionan todas las variables entre sí, utilizando la correlación de Pearson.

## [1] 0.01617846

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$Offer)  
## t = 1.3936, df = 7418, p-value = 0.1635  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.006578031 0.038918193  
## sample estimates:  
## cor   
## 0.01617846

## [1] 0.1154562

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$CHQ)  
## t = 10.011, df = 7418, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.09294634 0.13784815  
## sample estimates:  
## cor   
## 0.1154562

## [1] 0.005712216

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$CARD)  
## t = 0.49199, df = 7418, p-value = 0.6227  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.01704406 0.02846258  
## sample estimates:  
## cor   
## 0.005712216

## [1] 0.03871994

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$SAV1)  
## t = 3.3374, df = 7418, p-value = 0.0008499  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.01597995 0.06141989  
## sample estimates:  
## cor   
## 0.03871994

## [1] 0.2527005

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$LOAN)  
## t = 22.495, df = 7418, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2312763 0.2738798  
## sample estimates:  
## cor   
## 0.2527005

## [1] 0.1808468

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$MORT)  
## t = 15.837, df = 7418, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1587460 0.2027665  
## sample estimates:  
## cor   
## 0.1808468

## [1] -0.122909

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$INSUR)  
## t = -10.667, df = 7418, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.1452569 -0.1004359  
## sample estimates:  
## cor   
## -0.122909

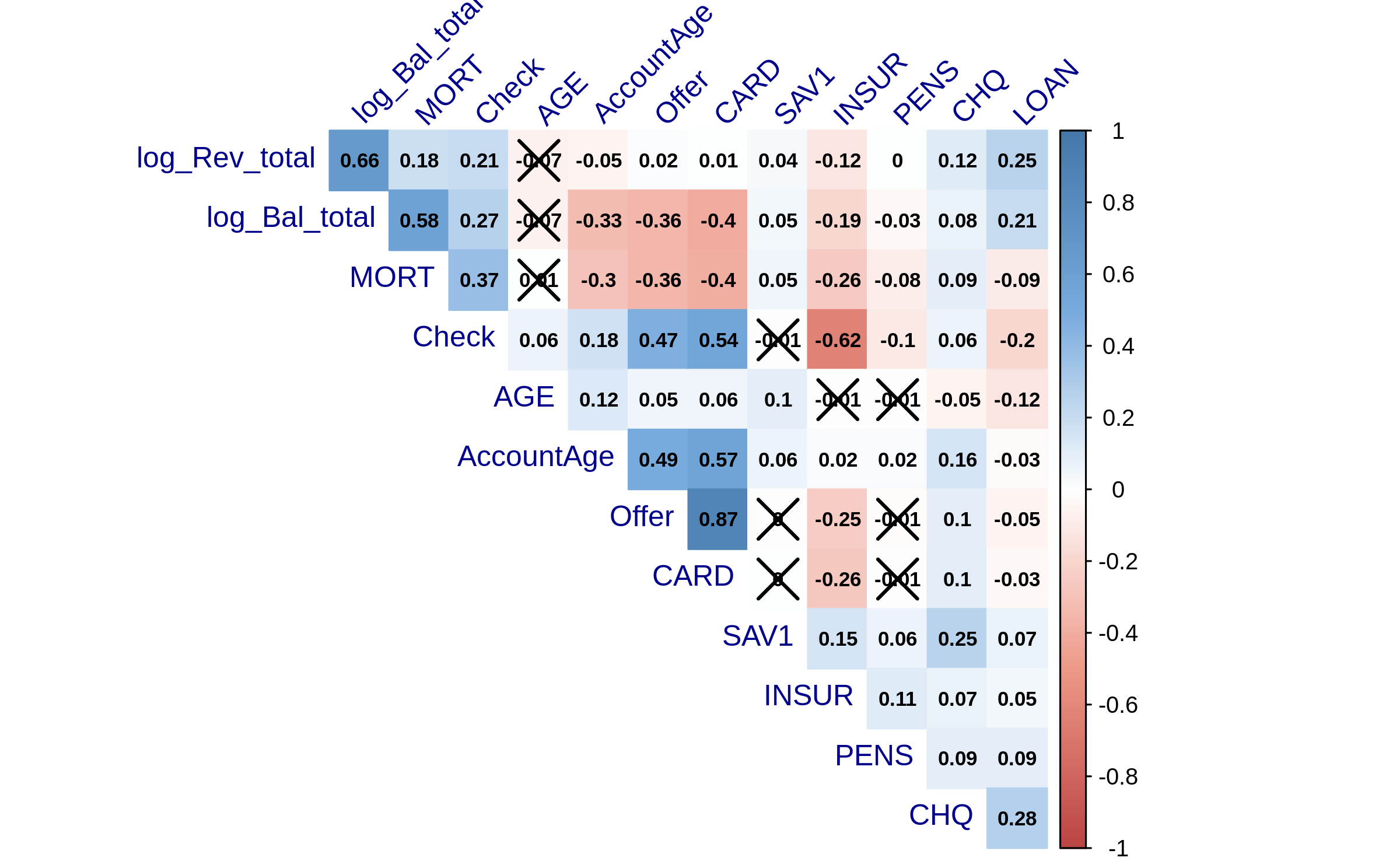
## [1] 0.001901021

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$PENS)  
## t = 0.16373, df = 7418, p-value = 0.8699  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.02085395 0.02465402  
## sample estimates:  
## cor   
## 0.001901021

## [1] 0.2077446

##   
## Pearson's product-moment correlation  
##   
## data: df\_1$log\_Rev\_total and as.numeric(df\_1$Check)  
## t = 18.292, df = 7418, p-value < 0.00000000000000022  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.1858692 0.2294143  
## sample estimates:  
## cor   
## 0.2077446

Ahora, que hemos analizado correlaciones entre variables numéricas, entre variables categóricas, y entre las variables categóricas y la variable objetivo que era de carácter numérico, faltaría determinar la correlación entre las variables categóricas y el resto de las variables numéricas. Ya que la correlación de Pearson demostró dar los mismos resultados que la correlación Biserial, utilizaremos una matriz de Pearson para visualizar una última vez las correlaciones entre todas las variables, para desde allí analizar las correlaciones entre variables numéricas y variables categóricas.



La matriz de correlaciones muestra que hay algunas correlaciones altas entre AccountAge y Offer, entre AccountAge y CARD, y entre log\_Bal\_Total y MORT. Sin embargo, ya que no son correlaciones absolutas, las dejaremos por ahora, esperando analizarlas en mayor profundidad a través del análisis de colinealidad. A modo de conclusión del análisis de correlaciones, la variable de interés tiene una correlación fuerte y positiva con las variables log\_Bal\_Total (0.66), Check (0.27), y LOAN (0.21), lo que es coherente con los valores expresados en las conclusiones del análisis exploratorio (punto 1.7).

# 3. Especificación del modelo

El modelo que se establecerá, es un modelo lineal de la forma siguiente:

Y = beta \* X

con

Y = log\_Rev\_total

y

X = [log\_Bal\_total,AGE,AccountAge,Offer,CHQ,CARD,SAV1,LOAN,MORT,INSUR,PENS,Check]

Se separa la data de entrenamiento en 70%, para luego aplicar la regresión lineal sobre estos datos, obteniendo los estadísticos del modelo. Se observa que el estadístico F rechaza la hipótesis nula que no existe una relación entre las variables que explican y la variable objetivo. Se observa también que los valores VIF se encuentran bajo 10, que es uno de los criterios usualmente utilizados para determinar ausencia de multicolinealidad. Sin embargo, otro punto de corte típico para este estadístico es 5, sobre el cual se encuentran las variables CARD (9.2), y Check (6.5). De esta forma y, de acuerdo a este último punto de corte, CARD y Check sí presentan problemas de colinealidad. Para abordar el problema anterior, se elimina primero la variable de mayor VIF (CARD) y se vuelve a probar el modelo.

Obteniéndose que ahora ninguna variable supera el valor 5 de VIF (siendo la mayor Check con un valor de 3.6). Lo anterior valida que las variables que se analizarán para el siguiente proceso son las siguientes:

1. log\_Bal\_Total: Numérica.
2. Offer: Categórica.
3. AGE: Numérica.
4. CHQ: Categórica.
5. SAV1: Categórica.
6. LOAN: Categórica.
7. MORT: Categórica.
8. INSUR: Categórica.
9. PENS: Categórica.
10. Check: Categórica.
11. AccountAge: Numérica.

# 4. Selección de variables

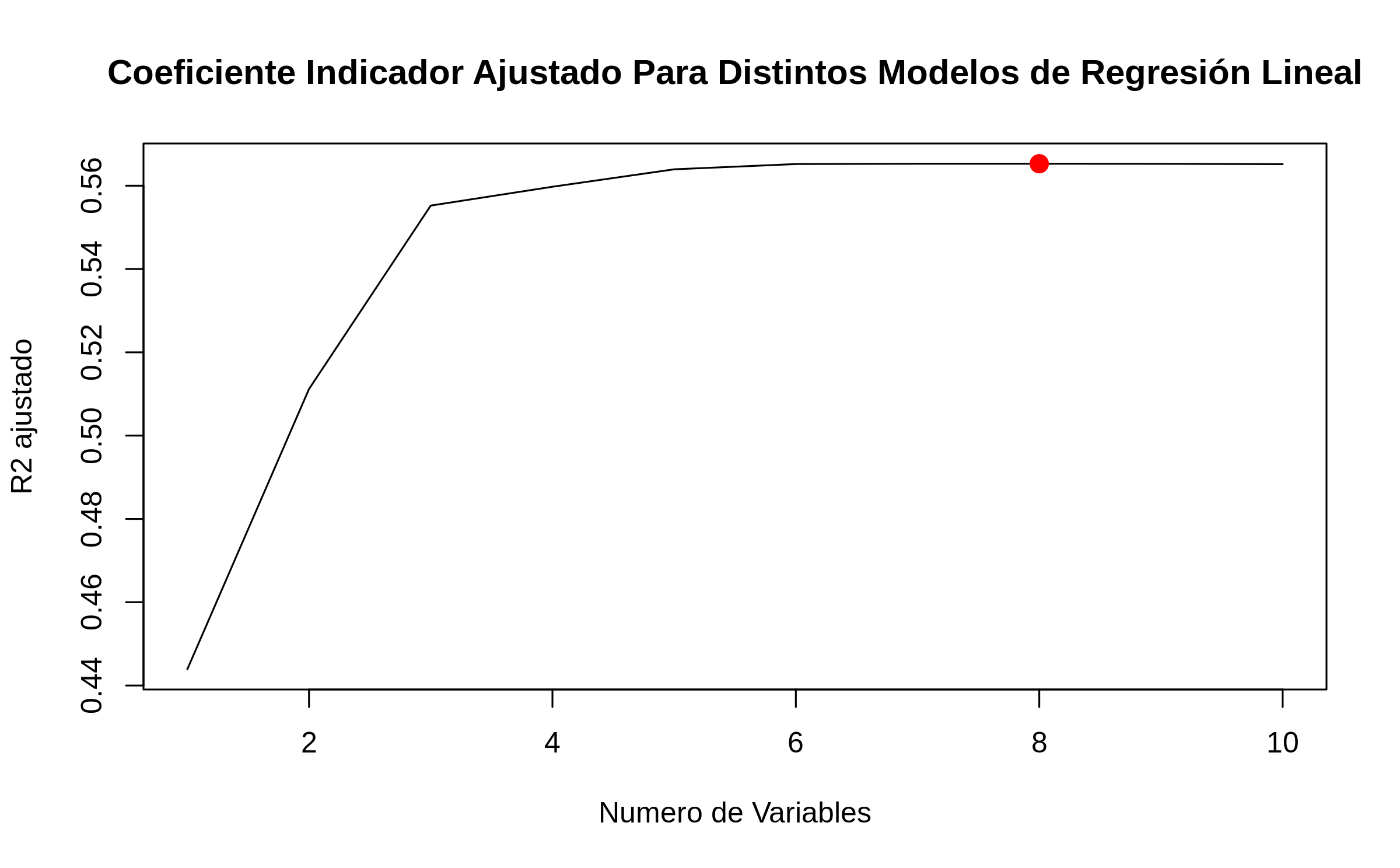
La selección de variables se puede realizar a través de diferentes técnicas, siendo las utilizadas en el desarrollo de este análisis las siguientes:

1. Coeficiente de determinación ajustado
2. Cp de Mallows
3. BIC (Schwartz’ Bayesian Criterion)
4. Paso a paso Forward
5. Paso a paso Backward
6. Paso a paso Stepwise

Cada una de estas técnicas sugiere un grupo específico de las variables ingresadas que mejor se ajustan a los datos. Cada modelo sugerido por cada técnica será también evaluado a través del estadístico PRESS. Esto, considerando que este estadístico indica la capacidad predictiva y no solo la bondad de ajuste. Aquel modelo sugerido que tenga un mayor valor PRESS será el modelo elegido para continuar el análisis. También cada modelo sugerido será sometido a los estadísticos MAPE, MAE Y RMSE con las muestras de entrenamiento (train) y validacíon (test), para descartar un sobreajuste.

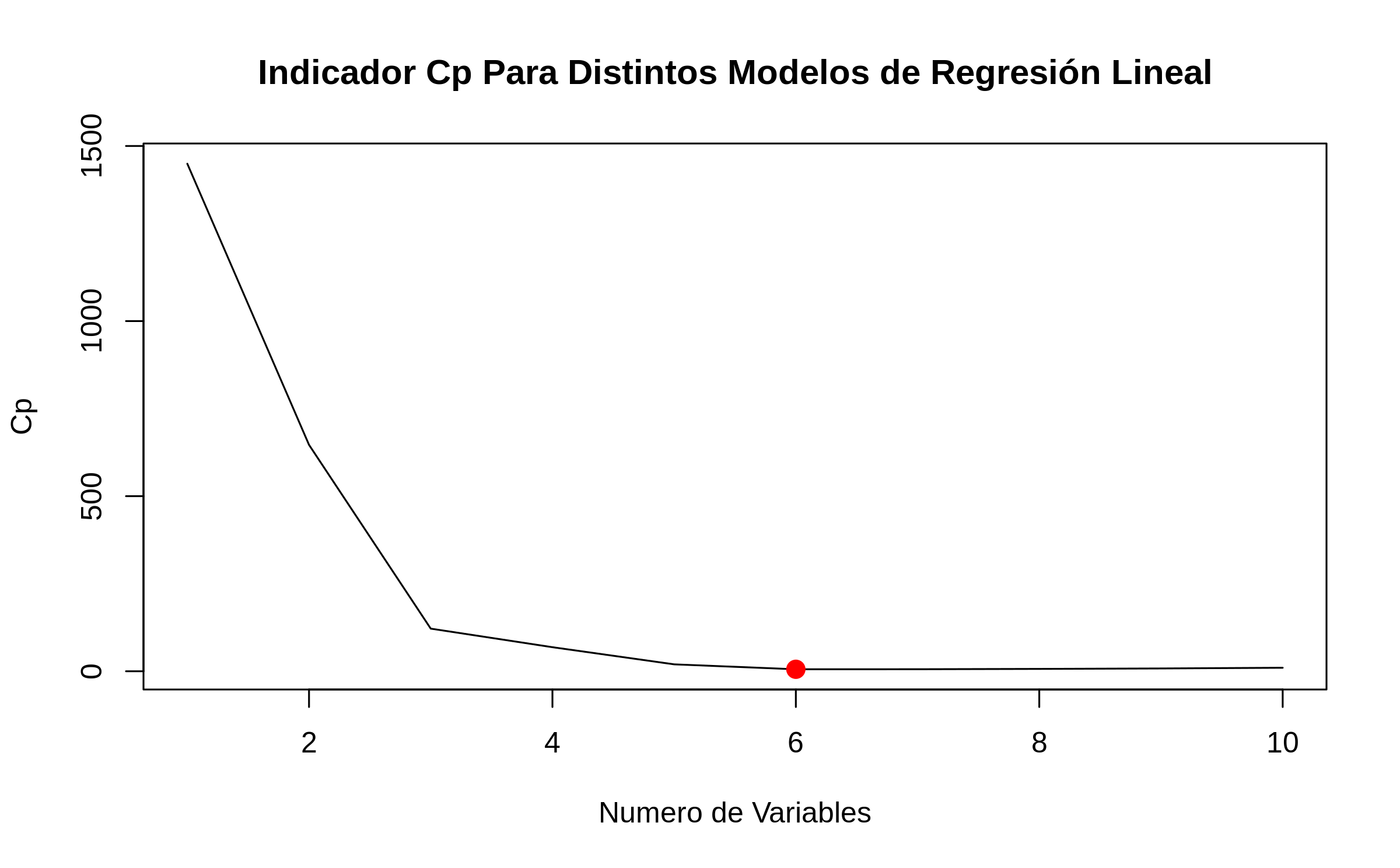
## 4.1 Coeficiente de Determinación Ajustado

El coeficiente de determinación ajustado nos entrega el siguiente gráfico, observándose que el modelo que maximiza el indicador cuenta con 8 variables, las que son: Offer1, AGE, CHQ1, LOAN1, MORT1, Check1, AccountAge, log\_Bal\_total. El modelo sugerido mostró un indicador PRESS de 3753.505.



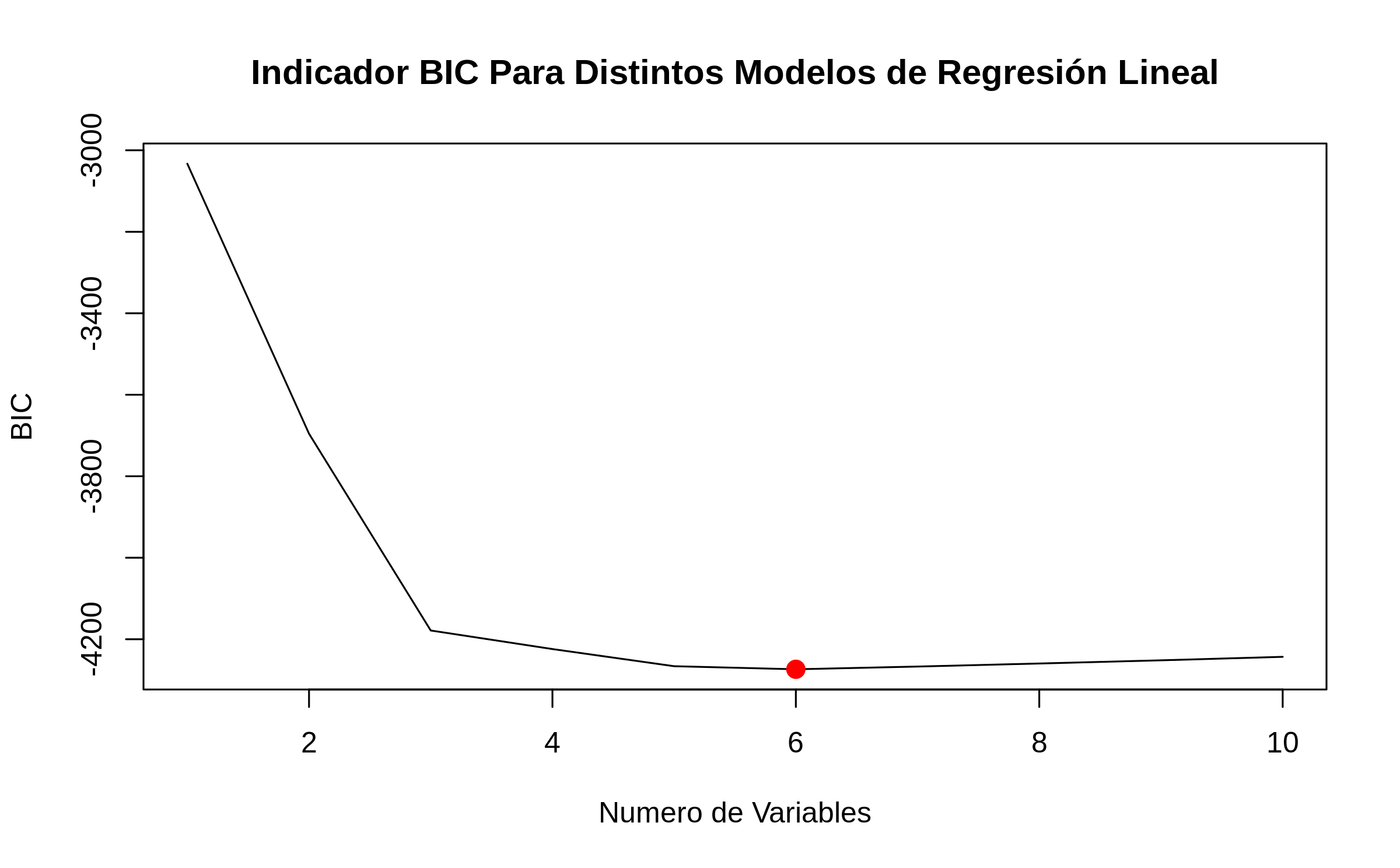
## 4.2 CP de Mallows

Para el indicador CP de Mallows el resultado del modelo sugerido contiene 6 variables, siendo este el que minimiza el indicador (ver figura siguiente). Las variables seleccionadas son: Offer, LOAN, MORT, Check, AccountAge, log\_Bal\_total. Presenta además un valor PRESS de 3752.775



## 4.3 BIC

El modelo resultante tiene 6 variables, lo que es idéntico al indicador CP de Mallows, es decir sugiere las mismas variables con el mismo valor PRESS. En el gráfico siguiente muestra que el conjunto de variables minimiza el indicador.



## 4.4 Paso a paso Forward

Para el método forward el resultado del modelo queda igual que los modelos sugeridos por el CP de Mallows y BIC.

## 4.5 Paso a paso Backward

Para el método forward el resultado del modelo queda igual que los modelos sugeridos por el método Forward, CP de Mallows y BIC.

## 4.6 Paso a paso Forward y Backward

Para el método forward el resultado del modelo queda igual que los modelos sugeridos por el método Forward, Backward, CP de Mallows y BIC.

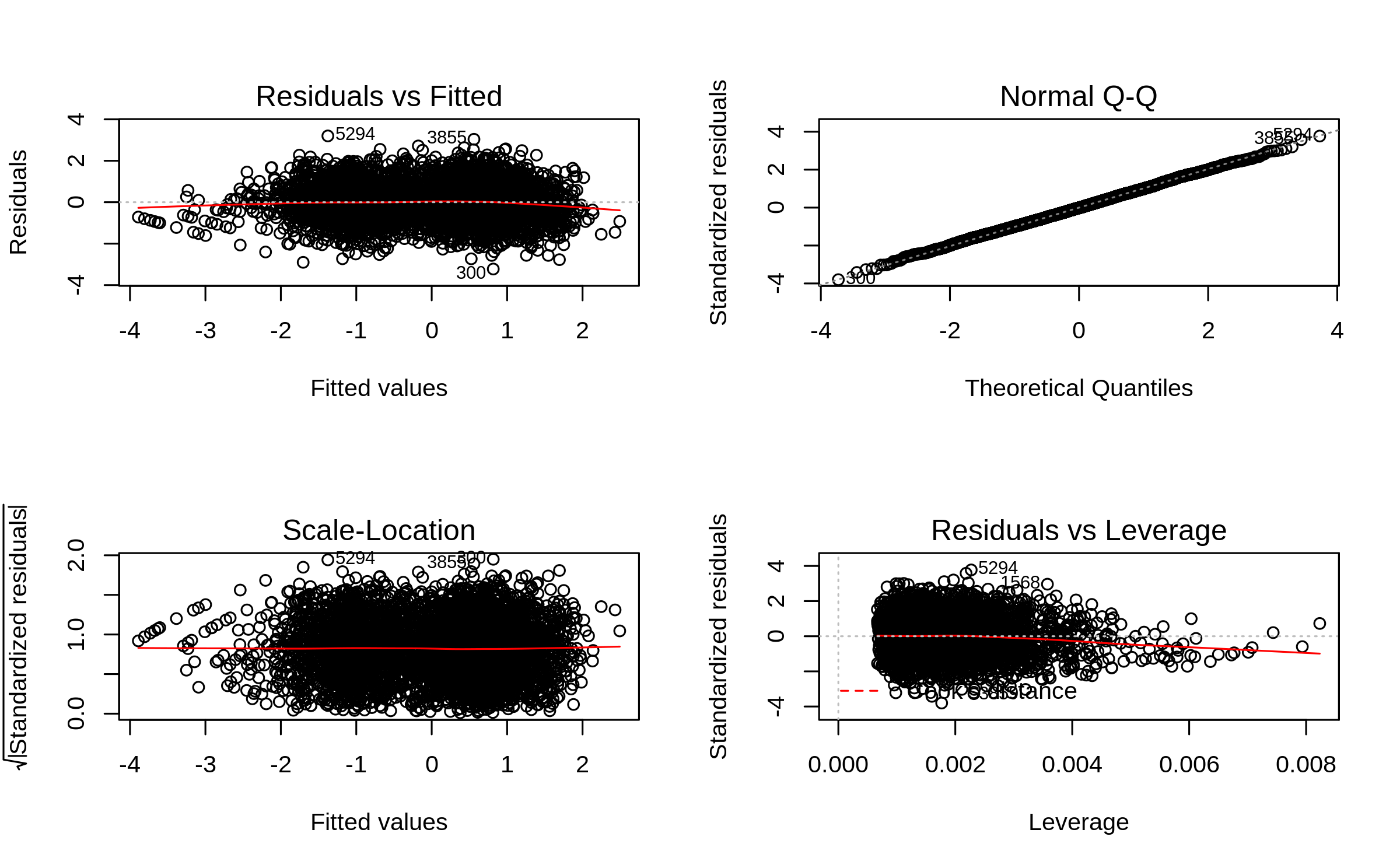
# 5. Análisis de calidad predictiva de los modelos

En base al punto anterior se tienen dos modelos y sus respectivos indicador PRESS:

* Modelo de 8 variables (1 criterio lo apoya): 3754
* Modelo de 6 variables (5 criterios lo apoyan): 3753

De acuerdo al valor PRESS es el menor, se sugiere seleccionar el modelo de 6 variables (Offer, LOAN, MORT, Check, AccountAge, log\_Bal\_total) en lugar del modelo de 8 variables (Offer, AGE, CHQ, LOAN, MORT, Check, AccountAge, log\_Bal\_total). De todas formas, es necesario evaluar ambos modelos a partir de los estadísticos MAE, MAPE y RMSE, para descartar sobreajuste. Adicionalmente, se chequeará si los residuos presentan homocedasticidad.

## 5.1 Modelo de 8 variables



Se tiene como resumen del modelo lo siguiente:

summary(modeloFinal1)

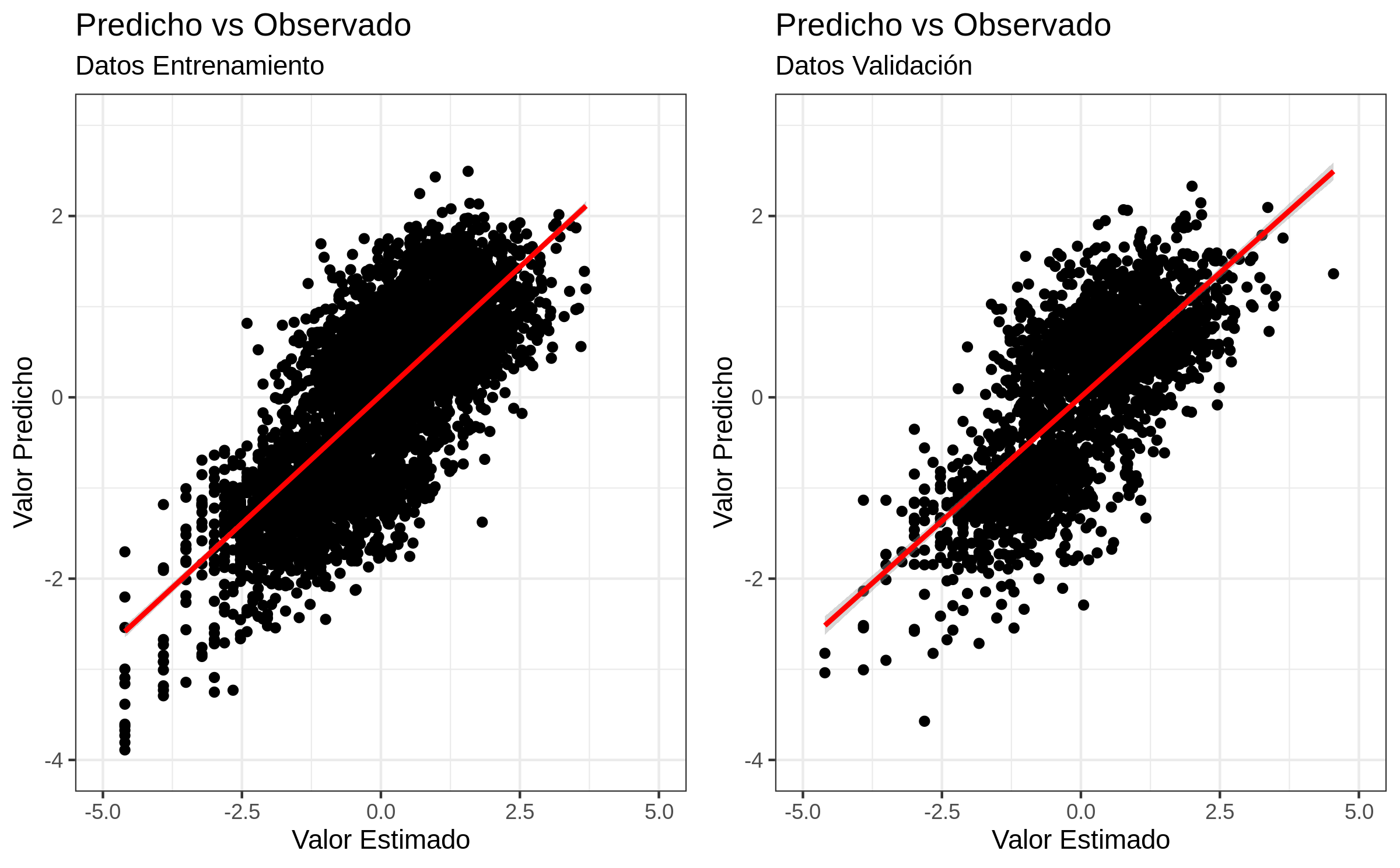
##   
## Call:  
## lm(formula = log\_Rev\_total ~ Offer + AGE + LOAN + MORT + Check +   
## CHQ + AccountAge + log\_Bal\_total, data = df\_2\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2243 -0.5825 -0.0094 0.5814 3.1996   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.5912320 0.0489331 -52.955 < 0.0000000000000002 \*\*\*  
## Offer1 0.7030620 0.0393318 17.875 < 0.0000000000000002 \*\*\*  
## AGE -0.0007434 0.0005668 -1.312 0.189746   
## LOAN1 0.1480759 0.0430720 3.438 0.000591 \*\*\*  
## MORT1 -0.5253501 0.0369054 -14.235 < 0.0000000000000002 \*\*\*  
## Check1 -0.3000408 0.0465561 -6.445 0.0000000001263 \*\*\*  
## CHQ1 0.0262071 0.0255288 1.027 0.304671   
## AccountAge 0.0209452 0.0030999 6.757 0.0000000000157 \*\*\*  
## log\_Bal\_total 0.3862693 0.0054678 70.645 < 0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8491 on 5187 degrees of freedom  
## Multiple R-squared: 0.566, Adjusted R-squared: 0.5653   
## F-statistic: 845.4 on 8 and 5187 DF, p-value: < 0.00000000000000022

Se observa que el estadistico F rechaza la hipótesis nula y que no todas las variables son significativas, no siendo rechazada la hipótesis nula de los beta de AGE y CHQ. Lo anterior sustenta adicionalmente la decisión guiada por el estadístico PRESS de preferir el modelo con menos variables. Si se observan los gráficos de control del modelo, se nota una leve heterocedasticidad, así como algunos puntos que en los extremos de la normal Q-Q no se ajustan a lo esperado.

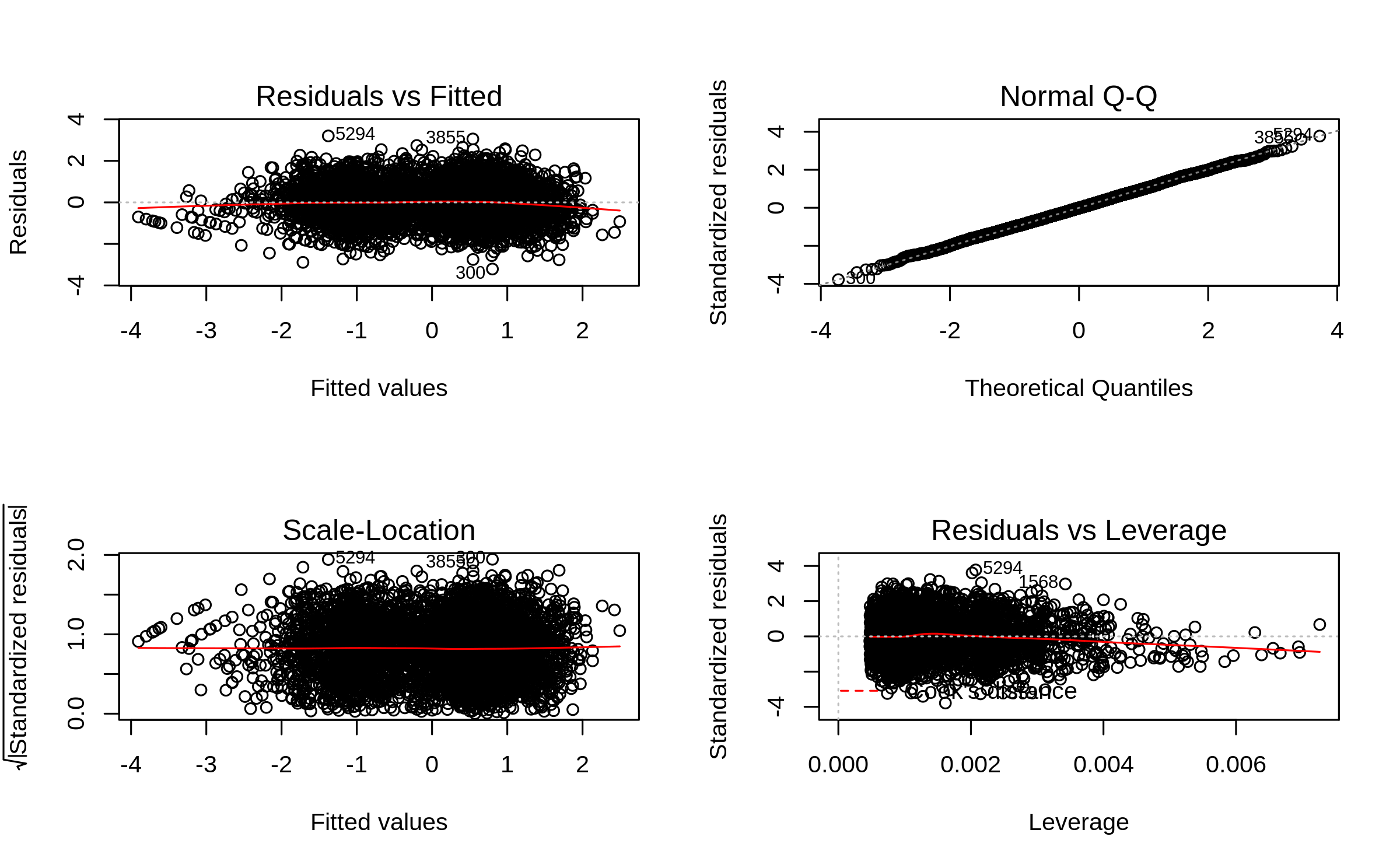
Finalmente para comparar el desempeño del modelo se comparan los estadísticos MAE, MAPE y RMSE en su desempeño en el entrenamiento y en el test obteniéndose valores similares, con lo que se ratifica que el modelo no cuenta con un sobreajuste observado. Dado que el indicador mape se indefine no se considerará dentro de la conclusión anterior. Las razones de esta indefinición están fuera del alcance de este trabajo. El MAE indica que el valor típico de los residuos es de aproximadamente 0.67 puntos de la variable dependiente que regiere a los ingresos totales. Respecto del RMSE varía también muy poco entre el subset de entrenamiento y de testeo, e indica que la variable dependiente puede tiene una varianza no explicada que fluctúe en 0.85 unidades de ingreso aproximadamente

|  |  |  |
| --- | --- | --- |
| indicador | train | test |
| MAE | 0.6786 | 0.6768 |
| MAPE | Inf | Inf |
| RMSE | 0.8484 | 0.8561 |

Ahora observaremos la relación entre los datos modelados y los predichos tanto en el data set de entrenamiento como en el de testeo. En ambos casos se observa una relación lineal.



## 5.2 Modelo de 6 variables



Se tiene como resumen del modelo lo siguiente:

summary(modeloFinal2)

##   
## Call:  
## lm(formula = log\_Rev\_total ~ log\_Bal\_total + Offer + MORT + AccountAge +   
## Check + LOAN, data = df\_2\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.2104 -0.5806 -0.0067 0.5787 3.2018   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.619574 0.041400 -63.274 < 0.0000000000000002 \*\*\*  
## log\_Bal\_total 0.386571 0.005465 70.739 < 0.0000000000000002 \*\*\*  
## Offer1 0.708385 0.039155 18.092 < 0.0000000000000002 \*\*\*  
## MORT1 -0.519821 0.036331 -14.308 < 0.0000000000000002 \*\*\*  
## AccountAge 0.021016 0.003028 6.941 0.00000000000438 \*\*\*  
## Check1 -0.303698 0.046499 -6.531 0.00000000007142 \*\*\*  
## LOAN1 0.165382 0.041236 4.011 0.00006140455090 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8492 on 5189 degrees of freedom  
## Multiple R-squared: 0.5657, Adjusted R-squared: 0.5652   
## F-statistic: 1127 on 6 and 5189 DF, p-value: < 0.00000000000000022

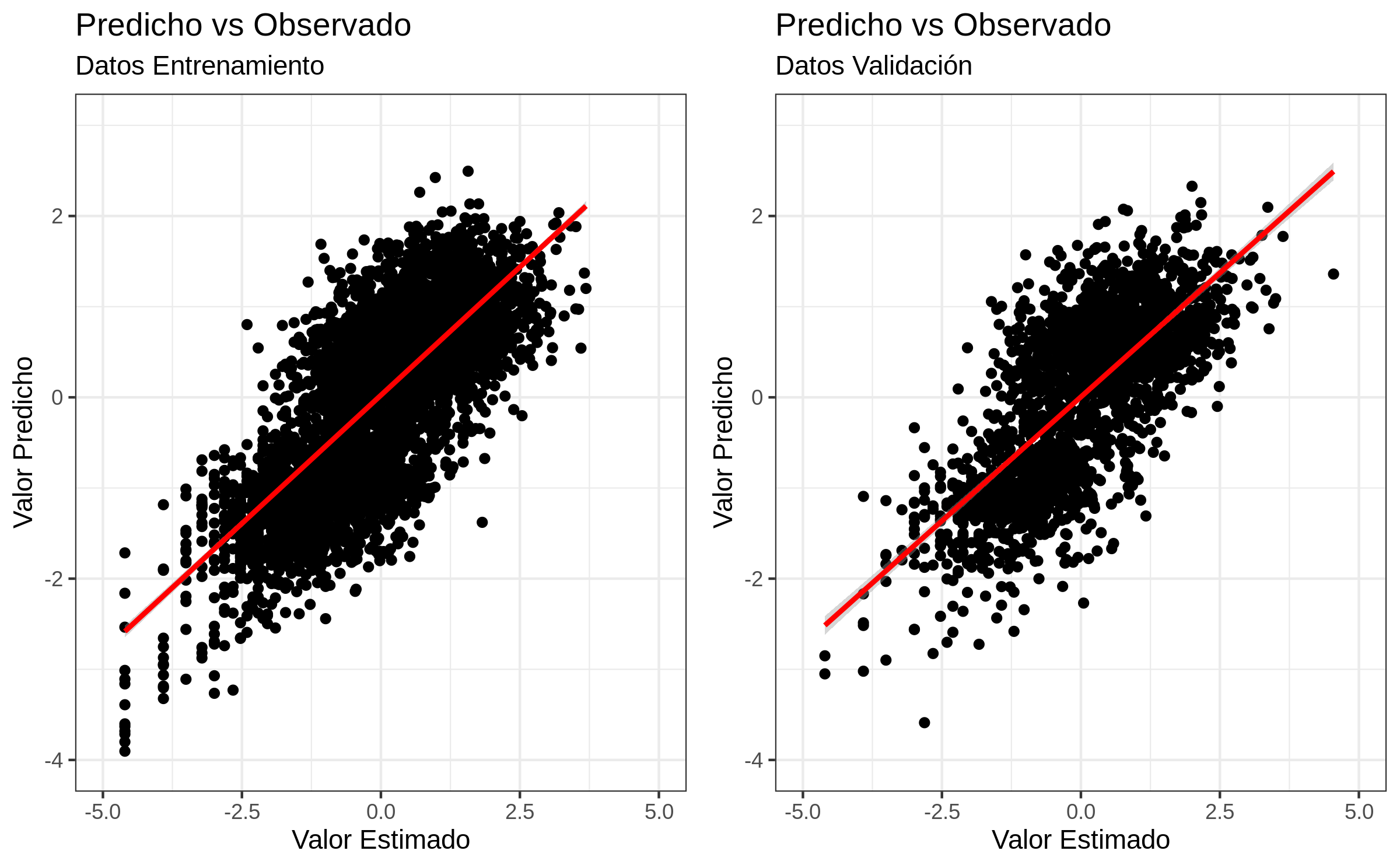
Se observa que el estadistico F rechaza la hipótesis nula y que todos las variables son significativas, a diferencia del modelo anterior.

Si se observan los gráficos de control del modelo, se nota una leve heterocedasticidad, así como algunos puntos que en los extremos de la normal Q-Q no se ajustan a lo esperado, al igual que para el modelo de 8 variables.

Finalmente para comparar el desempeño del modelo se comparan los estadísticos mae, mape y rmse en su desempeño en el entrenamiento y en el test obteniéndose valores similares, con lo que se ratifica que el modelo no cuenta con un sobreajuste observado. Se observa el mismo fenómeno asociado a la indefinición del mape. El MAE indica que el valor típico de los residuos es de aproximadamente 0.67 puntos de la variable dependiente que regiere a los ingresos totales. Respecto del RMSE varía también muy poco entre el subset de entrenamiento y de testeo, e indica que la variable dependiente predicha puede tener una varianza no explicada que fluctúa en 0.85 unidades aproximadamente.

|  |  |  |
| --- | --- | --- |
| indicador | train | test |
| MAE | 0.6786 | 0.6774 |
| MAPE | Inf | Inf |
| RMSE | 0.8486 | 0.8569 |

Ahora observaremos la relación entre los datos modelados y los predichos tanto en el data set de entrenamiento como en el de testeo. En ambos casos se observa una relación lineal.



# 6. Interpretación de coeficientes y conclusiones

Dado que en ambos modelos evaluados hay muy poca variación en sus diferentes coeficientes de evaluación, y dado que ambos cumplen con los supuestos requeridos, se opta por elegir el modelo que tiene la menor cantidad de variables, ya que implica un costo de mantención de medición menor. Las variables elegidas (Offer, LOAN, MORT, Check, AccountAge, log\_Bal\_total) son consistentes con los valores anticipados en el punto 1.7, siendo solo las excepciones Offer y Age, que individualmente presentaron muy bajos valores de ajuste.

En resumen, el modelo elegido cuenta con un buen ajuste de acuerdo a cuatro coeficientes de selección de variables. De acuerdo al CP de Mallows, la cantidad de variables sugeridas en el modelo es similar al número de variables ingresadas más la constante, lo cual indica que no se encuentra sesgado en la estimación de los verdaderos coeficientes de regresión y la inclusión de respuestas futuras. Por su parte el BIC también sugirió este modelo. El BIC agrega un componente de penalización a su fórmula de cálculo para lograr un balance entre un modelo que obtiene buen ajuste y el número de parámetros. El modelo en cuestión también fue sugeridos por el Coeficiente de Determinación de ajuste, el cual entrega el modelo que presenta un mayor ajuste a los datos, observado en un mayor valor de ajustado. Sin embargo, el valor alcanzado fue de tan solo 0.56, indicando que solo un 56% de la varianza de los ingresos totales se explican por las variables incluidas en el modelo. Finalmente el valor PRESS, que corresponde a la suma cuadrada de los residuos del modelo, indicó que ambos modelos evaluados en la etapa final del proyecto presentaban una calidad predictiva similar.

Para eliminar la existencia de sobreajuste, se usaron los estadísticos MAE y RSE, los cuales confirmaron la similitud entre ambos modelos finales, ya que ninguno presentó desviaciones importantes entre la muestra de entrenamiento y validación. Adicionalmente ambos cumplían los supuestos de homocedasticidad y normalidad de residuos. En el caso del modelo elegido finalmente, este obtuvo un Error Absoluto Medio (MAE) aproximado de 0.67 tanto al ser evaluado en la fase de entrenamiento como en la fase de testeo, lo cual significa que típicamente el valor predicho de los ingresos totales podría errar en 0.67 unidades. Por su parte, el RMSE indica que la varianza de los ingresos predicha por el modelo tiene una desviación estándar de aproximadamente 0.85 unidades, de nuevo valor que casi no varía entre la fase de entrenamiento y la de testeo.

A modo de conclusión, los dos modelos evaluados son muy similares en cuanto a su calidad predictiva, evaluados por el estadístico PRESS, y al cumplimiento de supuestos. Sin embargo, la mayoría de los coeficientes de bondad de ajuste sugieren el modelo de 6 variables, y además por el hecho de contener un menor número de variables independientes tendría un menor costo de implementación. Por estas razones se elige este último modelo. Es importante recalcar que si bien el modelo cumple con los supuestos y no se encuentra sobreajustado, no presenta un alto nivel de ajuste, tal como señala el indicador de ajustado, y presenta alta variabilidad según indican los estadísticos MAE y RMSE, por lo tanto se sugiere a los stakeholders evaluar si el porcentaje de varianza explicado del modelo elegido cumple con las expectativas y requerimientos del negocio.

Visto entonces el resultado del mejor modelo predictivo según el trabajo realizado, algunas interpretaciones prácticas de estos valores recordando que la variable objetivo corresponde al logaritmo de la rentabilidad (por lo tanto los coeficientes se deben multiplicar por 100) serian:

log\_Bal\_total 0.386571 \* 100 -> 39%

Por cada peso adicional que tiene el cliente en su saldo total, la rentabilidad aumenta en un 39%. Se observa que el valor de beta obtenido de la regresión múltiple (0.39) mantiene el mismo signo y magnitud similar al de la regresión simple (0.28), debido a que esta variable es la que mayor correlación lineal tiene con la variable a explicar.

Offer1 0.708385 -> 0.708385 \* 100 -> 71%

Los clientes que han recibido una oferta promocional especial el mes anterior, aumentan su rentabilidad en aproximadamente un 71%, comparados con aquellos clientes que no han recibido oferta promocional. Se observa que el valor de beta obtenido de la regresión múltiple (0.71) mantiene el mismo signo, pero una magnitud muy superior al de la regresión simple (0.04). Lo anterior se debe, probablemente, a que ésta relación se potencia con la presencia las otras variables.

MORT1 -0.519821 -> -0.519821 \* 100 -> -52%

Las cuentas de Hipotecarias con una gran importancia para el banco, disminuyen en aproximadamente un 52% la rentabilidad del cliente, comparada con aquellas cuentas Hipotecarias de baja importancia para el banco. Se observa que el valor de beta obtenido de la regresión múltiple (-0.52) tiene un cambio de signo bajo una magnitud superior al de la regresión simple (0.48). Es muy relevante este cambio, pues las conclusiones entre ambos acercamientos difieren enormemente.

AccountAge 0.021016 -> 0.021016 \* 100 -> 2.1%

Por cada año de antigüedad de un cliente la rentabilidad aumenta en un 2.1%. Se observa que el valor de beta obtenido de la regresión múltiple (0.02) tiene un cambio de signo bajo una magnitud superior al de la regresión simple (-0.01), pero con un valor bajo respecto a las otras variables seleccionadas, siendo consistente con el valor de ajuste individual.

Check1 -0.303698 -> -0.303698 \* 100 -> -30%

Si el cliente tiene una alta actividad en la cuenta corriente, su rentabilidad disminuye en un 30%, comparada con aquellas cuentas corrientes con una baja actividad. Se observa que el valor de beta obtenido de la regresión múltiple (-0.30) cambia el signo y la magnitud superior al de la regresión simple (0.63), bajo un fuerte efecto confundente.

LOAN1 0.165382 -> 0.165382\*100 -> 17%

Si el cliente tiene una alta actividad en la cuenta de préstamo personal, su rentabilidad aumenta en un 17%, comparada con aquellas cuentas de préstamo personal con una baja actividad. Se observa que el valor de beta obtenido de la regresión múltiple (0.17) mantiene el mismo signo y considerablemente menor que para la regresión lineal simple (1.026). Era esperable que esta variable no cambiase de signo debido a la relación con la variable a explicar mostrada en el análisis exploratorio.

De manera general la interpretación nos dice que la posibilidad de enviar una oferta a los clientes, su mayor antigüedad como cliente del banco, una alto endeudamiento en la cuenta de préstamo personal y un alto nivel de dinero en la cuenta total; contribuyen a un aumento en la rentabilidad del cliente. De la misma forma, un alto nivel de deuda en las cuentas hipotecarias y una alto nivel de actividad en la cuenta corriente contribuyen a una disminución de la misma rentabilidad. Además, que el efecto confundente provocó cambios de signos y magnitud, siendo la variable menos afectada (log\_Bal\_total) la que presentó la mayor correlación lineal con la variable objetivo, como era esperable.

# 7. Conclusiones personales del equipo

El presente trabajo nos permitió explorar una metodología completa, explicada en la introducción, para la aplicación de regresión lineal múltiple a un set de datos. Dentro de cada paso fuimos capaces de ir realizando preselecciones y transformaciones de variables para obtener el mejor modelo posible, el cual deberá ser evaluado dentro del contexto y expectativas del negocio de la banca, si cumple o no con los objetivos planteados.

Si bien el informe es secuencial, la realidad del trabajo fue completamente distinta. Dado que esta disciplina tiene una parte de arte, fue necesario realizar más de una vez el camino completo, buscando mejoras al modelo planteado. Sin duda, que deben existir mejores modelos que el planteado al final de este informe, pues las opciones de transformación y construcción son infinitas. Sin embargo, estamos satisfechos con contar con un modelo con 6 variables que fuimos capaces de interpretar a cabalidad.

Otro aspecto relevante es entender, que dado el tamaño de la base de datos y la cantidad de variables, pudimos realizar múltiples iteraciones desde inicio a fin, sin tener que esperar más allá de minutos. Frente a una situación real, con una base de datos de millones de registros y una cantidad de variables que superan las 1000, es necesario que los primeros procesos de selección, sobre todo en la construcción de modelos de regresión lineal simple, cobren más protagonismo, para ir acotando la cantidad de variables a revisar en las etapas de análisis de correlación y selección de variables.

Finalmente, la interpretación de variables nos ayudó a acercarnos a bajar desde la modelación al negocio, que cobra mucha relevancia en términos que será lo que finalmente experimente el usuario final. Quisiéramos cerrar estas conclusiones indicando el valor pedagógico del trabajo, por sobre otras instancias, ya que permitió a este grupo profundizar en los conocimientos de los conceptos y técnicas de modelamiento.