# ¿Cuales son los factores principales y en que proporcion intervienen en el gasto de vivienda de una familia mexicana?

*Act. Andrés Antonio Medina Landeros*

*2018-07-24*

# Contents



```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

# Chapter 1

# Abstracto

Esta investigacion tiene como objetivo conocer y cuantificar el impacto que tienen en el gasto en vivienda los factores mas relevantes del entorno socio economico y demografico de las familias mexicanas. Como resultado de este estudio encontre patrones linealmente positivos pero marginalmente decrecientes del gasto mensual y la edad del jefe de familia con respecto del gasto en vivienda. Encontre que el sexo del jefe de familia tambien afecta los patrones de consumo en vivienda, el estudio demuestra que los hombres gastan menos dinero en este concepto que las mujeres. Por ultimo, descubri que existe una relacion lineal positiva entre el estrato socio economico al que pertenece la unidad de analisis y su patron de consumo de vivienda, este efecto positivo puede ser descrito como que, entre mas alto el estrato al que pertenece, mas dinero destina esa familia a la vivienda. La metodoligia que implemente en este estudio es la construccion de un modelo de regresion lineal multiple generalizado.

# Chapter 2

# Introduccion

La teoria economica nos ha presentado un marco teorico en el cual se incluye la forma en la que el mercado de vivienda contribuye a propulsar el dinamismo economico. algunas de las formas mas importantes en las que la vivienda logra impactar a los mexicanos son, por ejemplo la derrama economica que genera en otros sectores de la economia, y en la medida en que permite el desarrollo de otros aspectos del ser humano que le permiten mejorar sus condiciones socio economicas, sin embargo la industria inmobiliaria, por diversos factores ha sido descrita como inflexible, inamovible, firme, e inapelable por lo que en muchos mercados, el sector no ha podido generar soluciones de valor para su consumidor final. Ha llegado el momento de implantarle un nuevo dinamismo a la industria mediante la implementacion de la inteligencia de negocios.

Muchos de los procesos que URBI utiliza para prospectar y cuantificar sus mercados meta y el valor de estos son obsoletos, apoyados en hipotesis que carecen de sustento frente a una realidad mexicana constanemente cambiante, estas hipotesis deben ser puestas a prueba y contrastadas con datos para encontrar nuevas ideas y soluciones que entreguen valor a nuestrso clientes.

El enfoque que presento en este trabajo esta centrado en la familia mexicana como unidad de analisis, quiero conocer y cuantificar los patrones de consumo en este bien, esto, con el proposito de que URBI, como empresa tenga un panorama general, apoyado en datos sobre quienes son las personas que destinan sus recursos en nuestras soluciones integrales de vivienda, de tal forma que URBI pueda adaptar su gama de productos y servicios a la realidad de nuestros clientes.

Es claro que la aproximacion que utilice en la presente investigacion no es del tipo exhaustivo, sino mas bien quiero presentarles mis hallazgos para que en conjunto tomemos conciencia del importante poder descriptivo y predictivo que nos provee esta clase de modelacion estadistica con miras a que en un futuro, mejoremos nuestros procesos de interacci?n con los clientes, procesos en los cuales recabemos, almacenemos y procesemos de forma valiosa la informacion.

# Chapter 3

# Datos

los datos utilizados en este proyecto son los proporcionados por la Encuesta Nacional Ingresos y Gastos de los hogares en su edicion 2016 (ENIGH 2016). El objetivo de esta encuesta es el de proporcionar un panorama estadistico del comportamiento de los ingresos y gastos de los hogares en cuanto a su monto, procedencia y distribucion, adicionalmente la encuesta ofrece caracteristicas ocupacionales y socio demograficas de la infraestructura de la vivienda y el equipamiento del hogar. El periodo de levantamiento de la encuesta fue del 21 de Agosto del 2016 al 28 de noviembre del 2016. La cobertura geografica de la encuesta es a nivel nacional.

Para conformar la base de datos de la investigacion utilice el archivo concentradohogar.csv que se encuentra disponible en la pestana de microdatos en el sitio web de la ENIGH

## 3.1 Descripcion de las variables

A continuacion, presento una tabla de las variables que conforman la base de datos utilizada en este analisis, el numero significa la columna en la que se encuentran en el archivo, su nombre, su etiqueta, su categoria (si es numerica o categorica) y en caso de ser categorica el numero de niveles que la conforman. El archivo se compone de 70,311 unidades observacionales, de las cuales seleccione 1000 mediante un muestreo aleatorio y omiti de la base aquellos cuyo gasto en vivienda reportado es del orden de centavos, ya que estos datos no tienen ningun sentido economico ni financiero y claramente son errores de captura.Gracias al procedimiento anteriormente mencionado me quedo con 969 unidades para realizar el analisis.

Table 3.1: Tabla de variables a utilizar.

| Numero | variable | etiqueta |
|---|---|---|
| 6 | estsocio | estrato socio economico |
| 11 | sexojefe | sexo del jefe de familia |
| 12 | educajefe | educacion formal del jefe de familia |
| 13 | totinteg | numero de integrantes del hogar |
| 14 | ingcor | ingreso corriente trimestral de la unidad de observacion |
| 58 | gastomon | gasto corriente trimestral de la unidad de observacion |
| 80 | vivienda | gasto en vivienda trimestral, se incluye renta o pago de hipoteca, gasto en servicios de reconstrucci |

Table 3.2: niveles de la variable categorica sexo del jefe de familia.

| Valor | Etiqueta |
|---|---|
| 0 | Mujer |
| 1 | Hombre |

Table 3.3: niveles de la variable categorica estrato socio-economico.

| Valor | Etiqueta |
|---|---|
| 1 | Bajo |
| 2 | Medio Bajo |
| 3 | Medio Alto |
| 4 | Alto |

Table 3.4: niveles de la variable categorica educacion formal del jefe de familia.

| Valor | Etiqueta |
|---|---|
| 1 | sin instruccion |
| 2 | preescolar |
| 3 | primaria incompleta |
| 4 | primaria completa |
| 5 | secundaria incompleta |
| 6 | secundaria completa |
| 7 | preparatoria incompleta |
| 8 | preparatoria completa |
| 9 | profesional incompleta |
| 10 | profesional completa |
| 11 | posgrado |

# Chapter 4

# Analisis exploratorio de los datos y estadistica descrpitiva

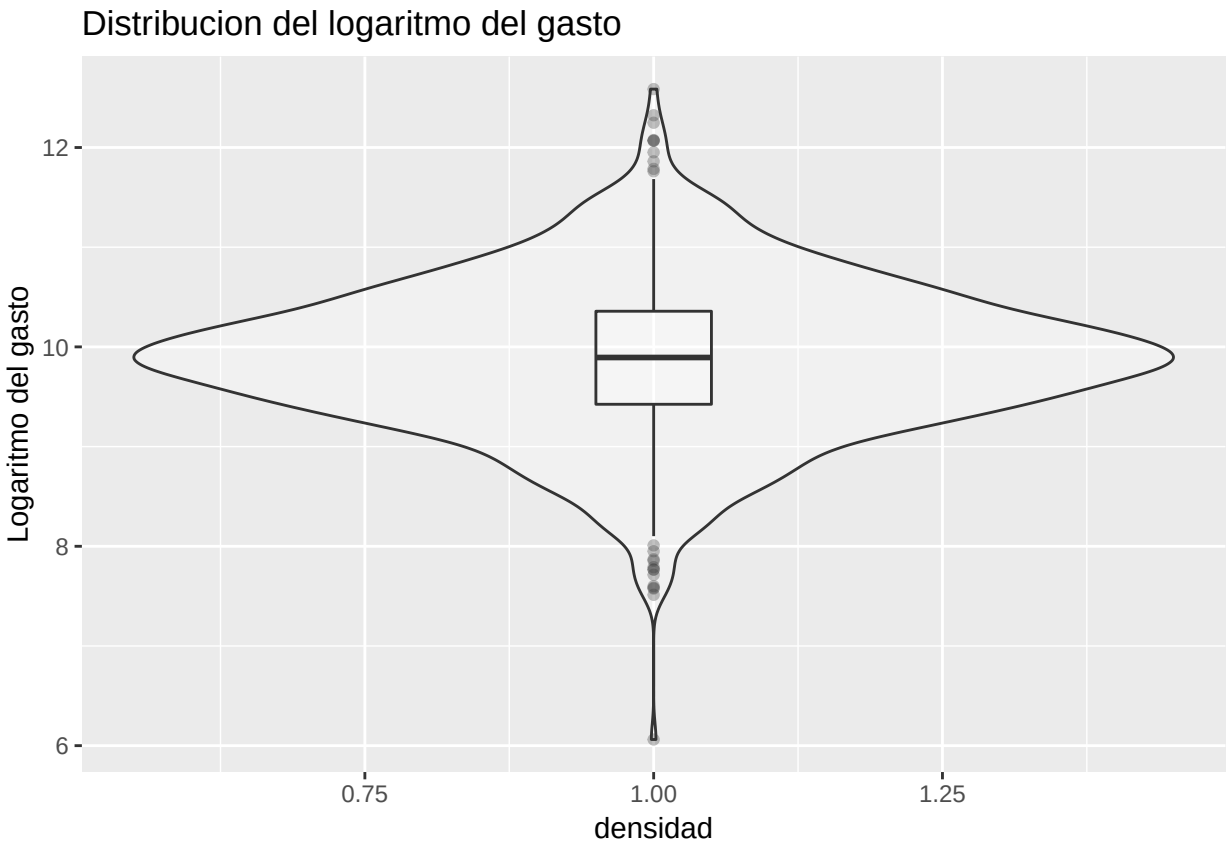
El analisis exploratorio de los datos consiste en descubrir las relaciones entre las variables propuestas para el modelo, para que asi se puedan presentar de forma correcta. Tambien es util para conocer la estructura de los datos y conocer su consistencia. La forma en la que presento esta seccion es primero, con estadisticas descrptivas de las variables numericas, entre las cuales se incluyen media, moda, desviacion estandard, cuartiles, minimo y maximo. A continuacion prosigo con graficas de estadistica univariada, entre los que se incluyen, diagramas de caja y brazos, histogramas y graficos de densidad esto con el proposito de conocer la forma, el sesgo y los parametros de localizacion de la distribucion de estas variables. Despues exploro algunas relaciones multivariadas entre distintos parametros categoricos y numericos, esto con el fin de conocer las caracteristicas de la poblacion.

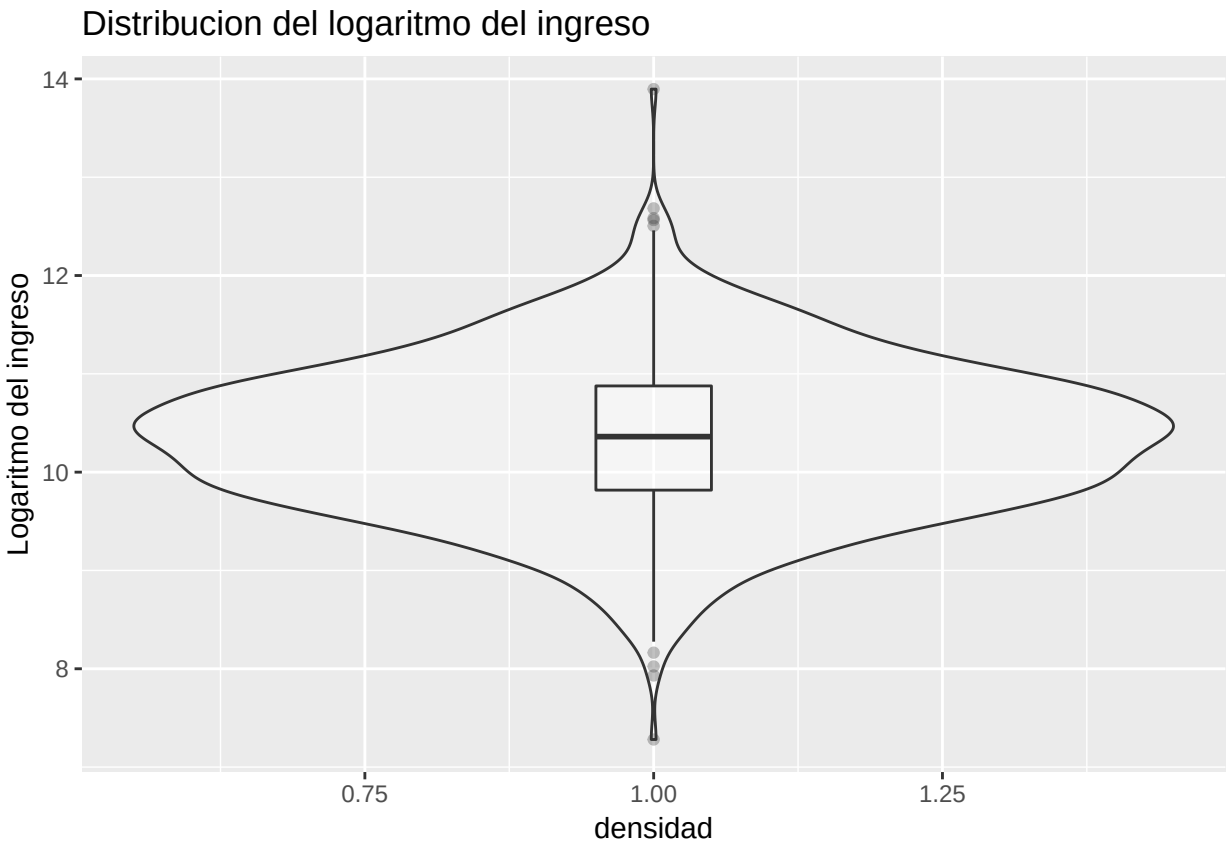
## 4.1 Estadistica Descriptiva

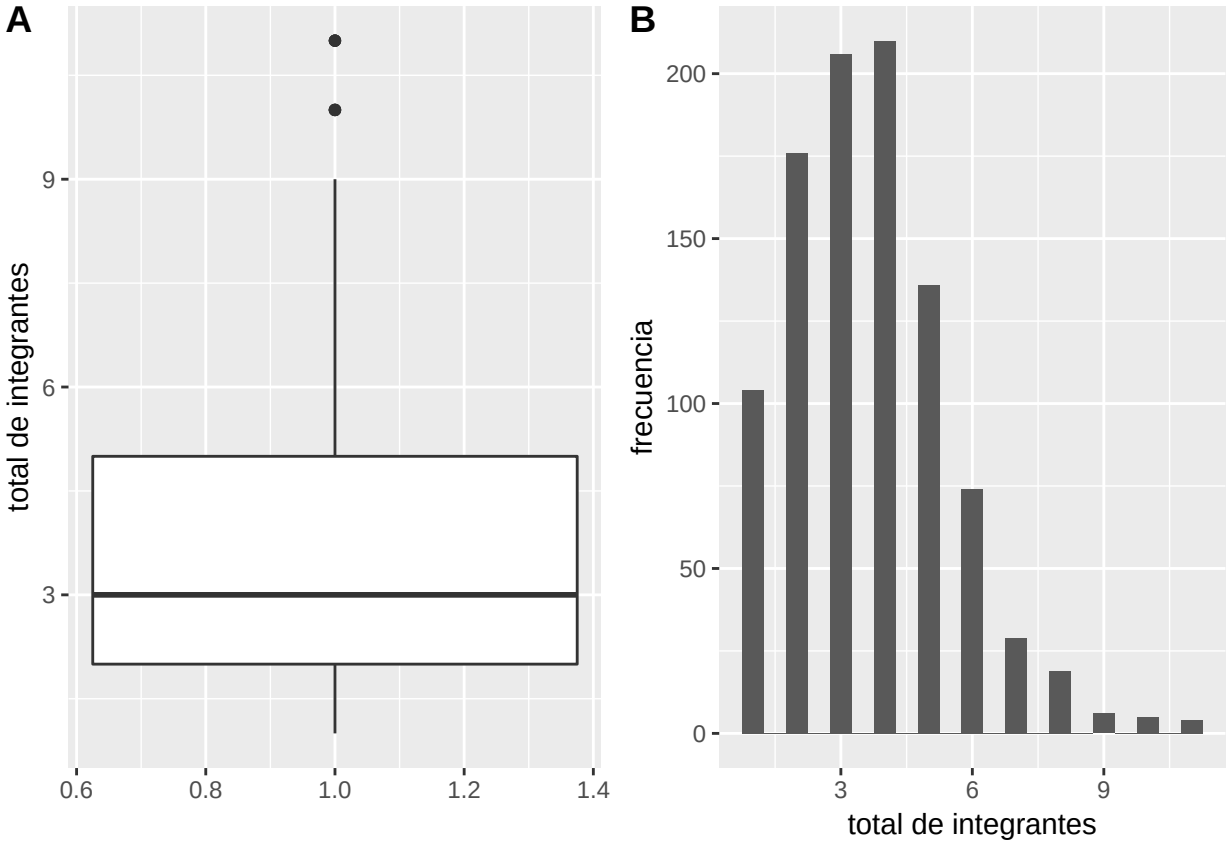
| | datos2 (N = 969) |
|---|---|
| **Edad del jefe de familia** | |
| min | 17 |
| median (IQR) | 47 (37.00, 61.00) |
| mean (sd) | 49.28 ± 16.41 |
| max | 104 |
| **Total de integrantes del hogar** | |
| min | 1 |
| median (IQR) | 3 (2.00, 5.00) |
| mean (sd) | 3.65 ± 1.83 |
| max | 11 |
| **Gasto general(log)** | |
| min | 6.061783 |
| median (IQR) | 9.89 (9.42, 10.36) |
| mean (sd) | 9.89 ± 0.79 |
| max | 12.5858 |
| **Ingreso(log)** | |
| min | 7.280429 |
| median (IQR) | 10.36 (9.82, 10.88) |
| mean (sd) | 10.35 ± 0.80 |
| max | 13.8962 |

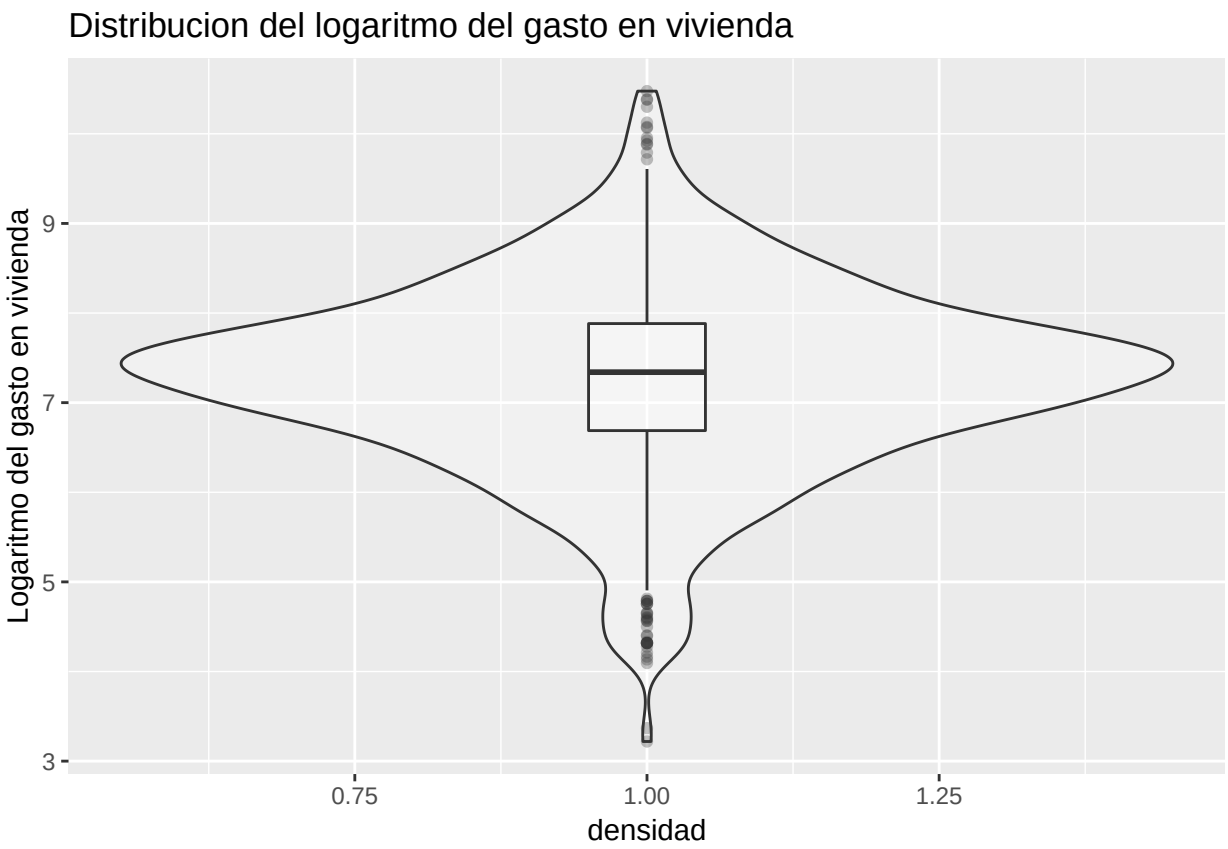
| | datos2 (N = 969) |
|---|---|
| **Gasto en vivienda(log)** | |
| min | 3.218876 |
| median (IQR) | 7.34 (6.69, 7.88) |
| mean (sd) | 7.27 ± 1.08 |
| max | 10.47658 |

## 4.2 Analisis Univariado


Distribucion del logaritmo del gasto
Logaritmo del gasto
12
10
8
6
0.75
1.00
1.25
densidad


Distribucion del logaritmo del ingreso
Logaritmo del ingreso
14
12
10
8
0.75
1.00
1.25
densidad


A
total de integrantes
9
6
3
0.6
0.8
1.0
1.2
1.4
B
frecuencia
200
150
100
50
0
3
6
9
total de integrantes


Distribucion del logaritmo del gasto en vivienda
Logaritmo del gasto en vivienda
9
7
5
3
0.75
1.00
1.25
densidad

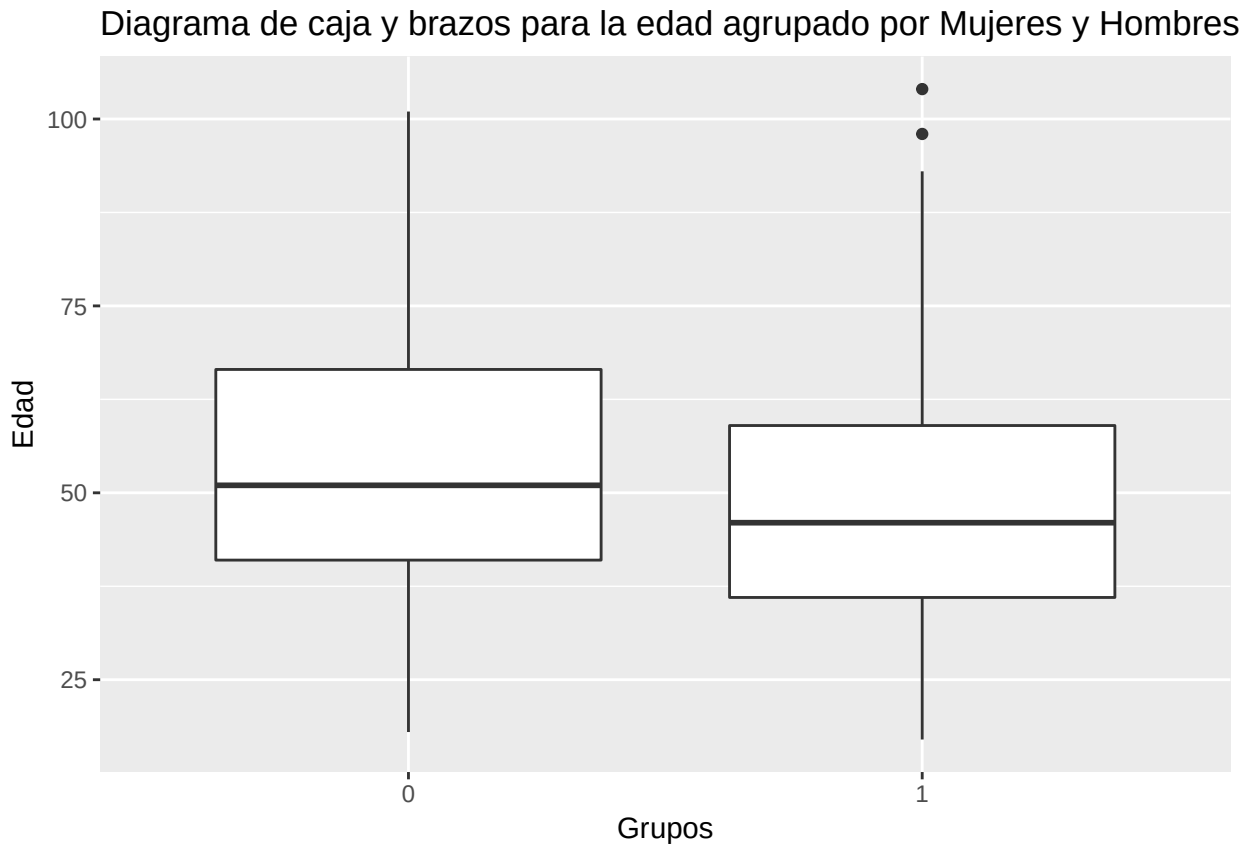
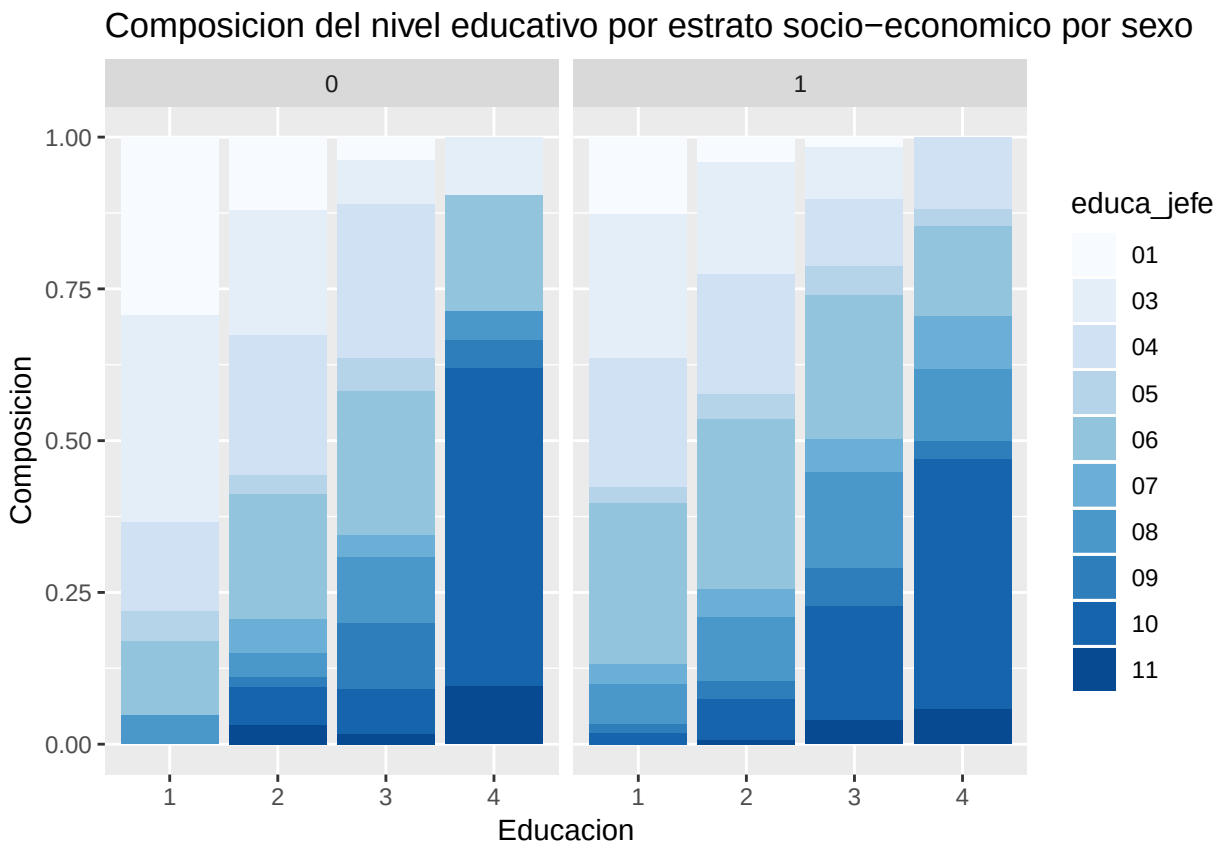
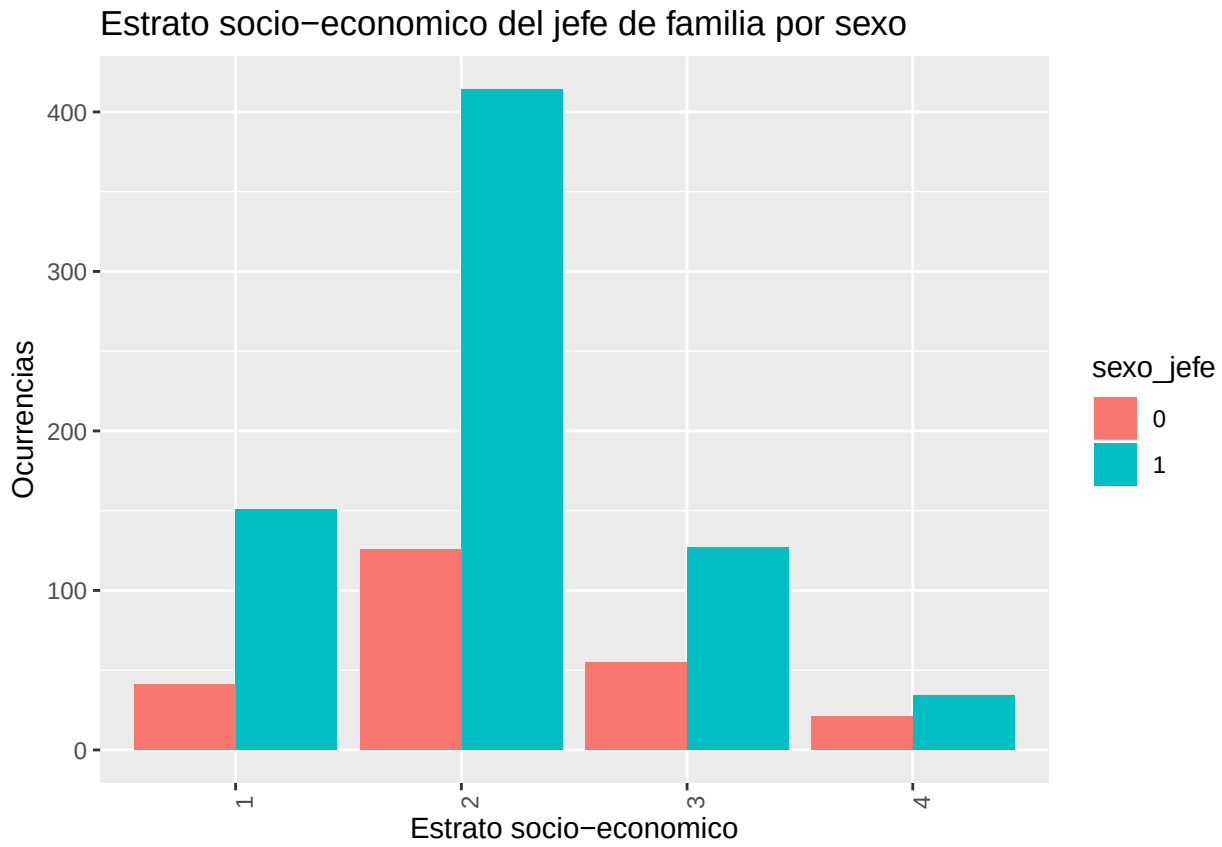
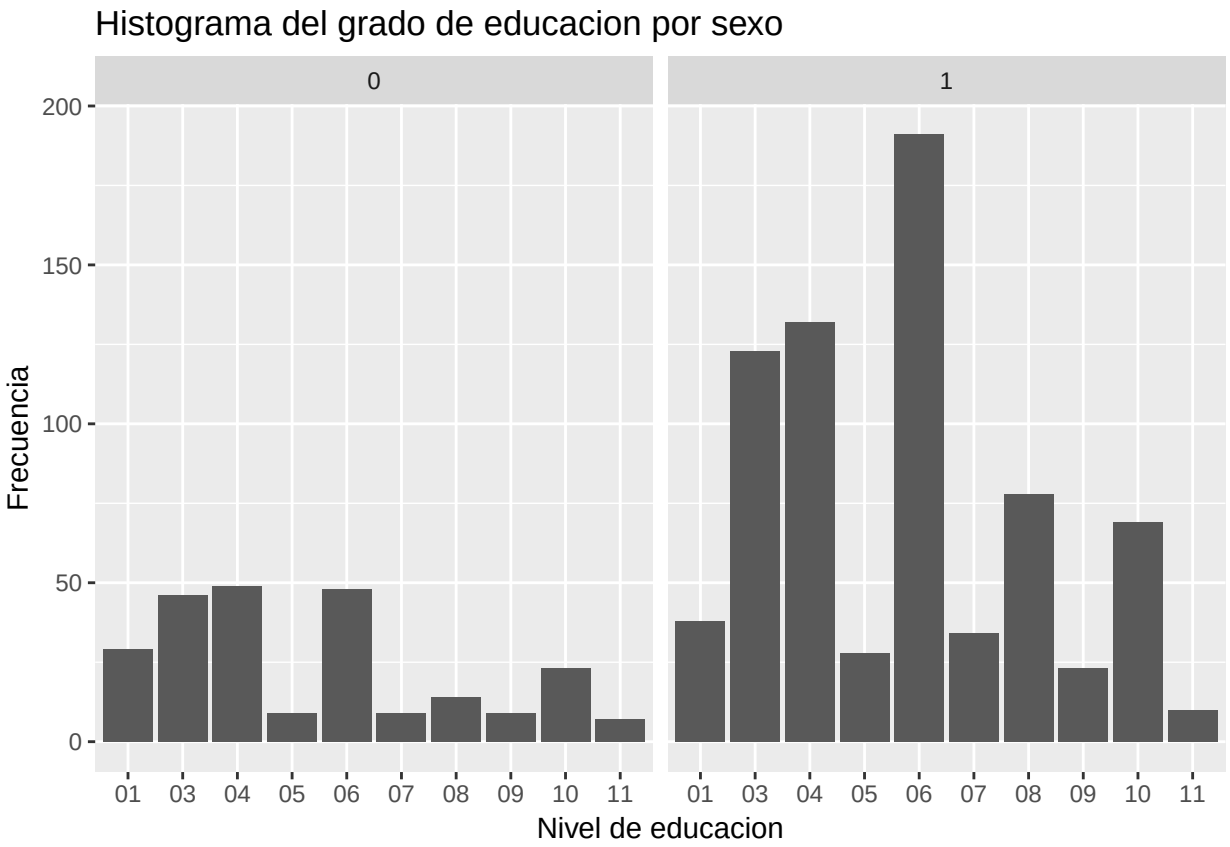


Diagrama de caja y brazos para la edad agrupado por Mujeres y Hombres
Edad
100
75
50
25
0
1
Grupos

## 4.3 Analisis Multivariado

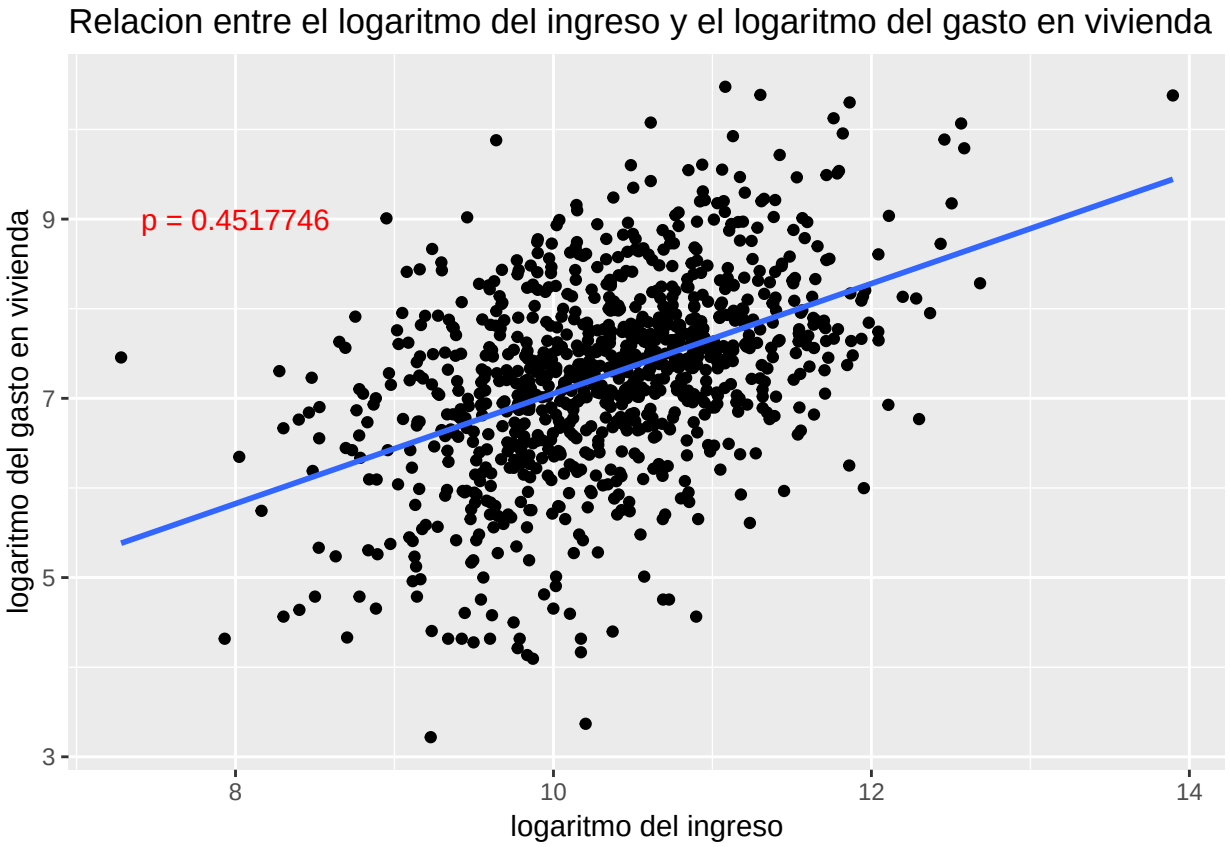


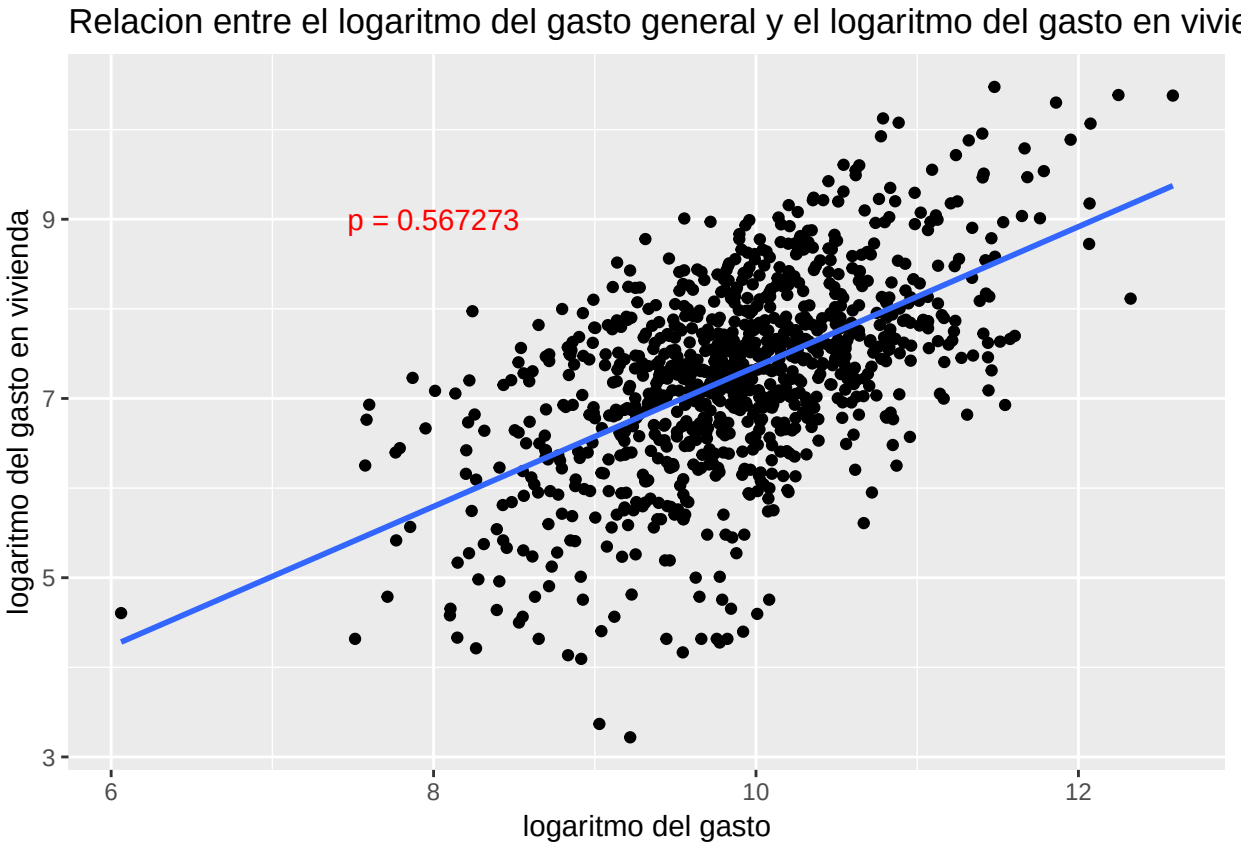

Estrato socio−economico del jefe de familia por sexo
Ocurrencias
400
300
200
100
0
1
2
3
4
Estrato socio−economico
sexo_jefe
0
1

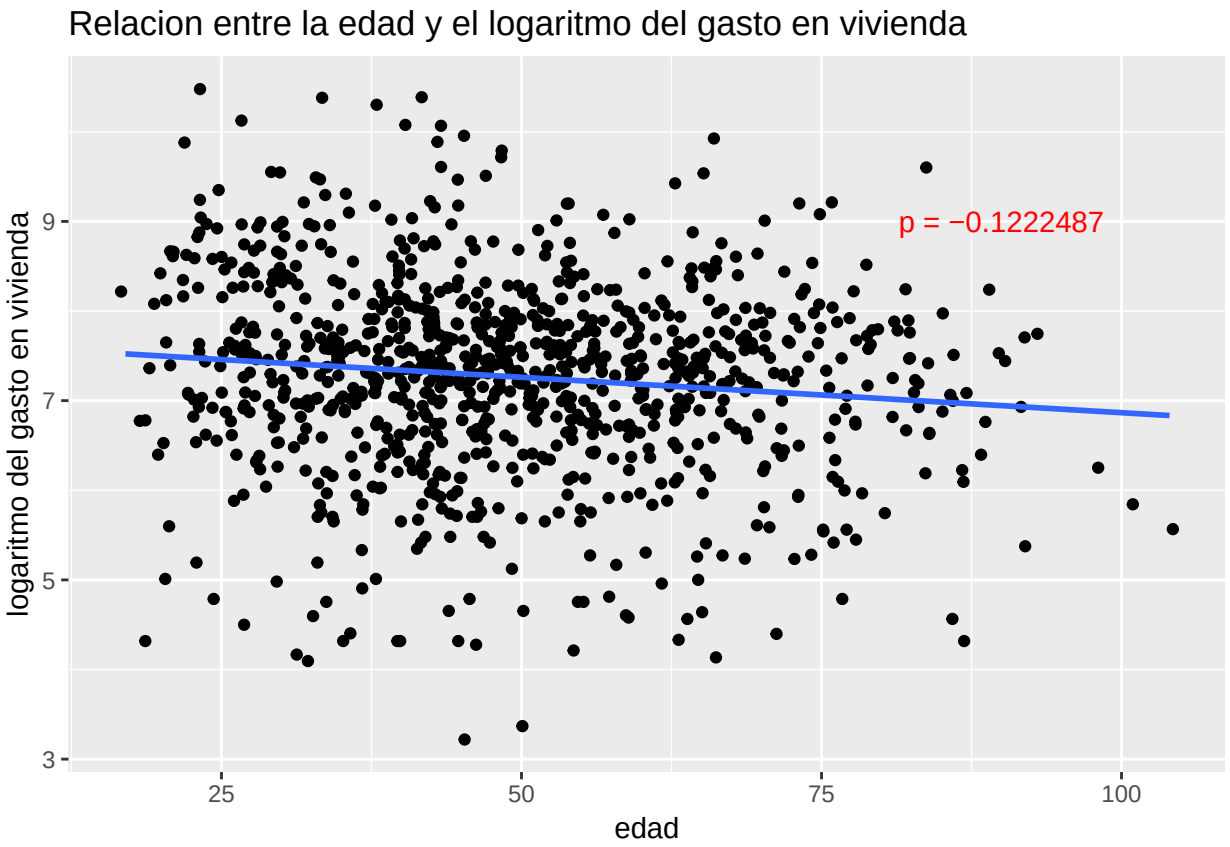

Histograma del grado de educacion por sexo
0
1
200
150
100
50
0
Frecuencia
01
03
04
05
06
07
08
09
10
11
Nivel de educacion

### 4.3.1 Relaciones entre regresores y variable explicada

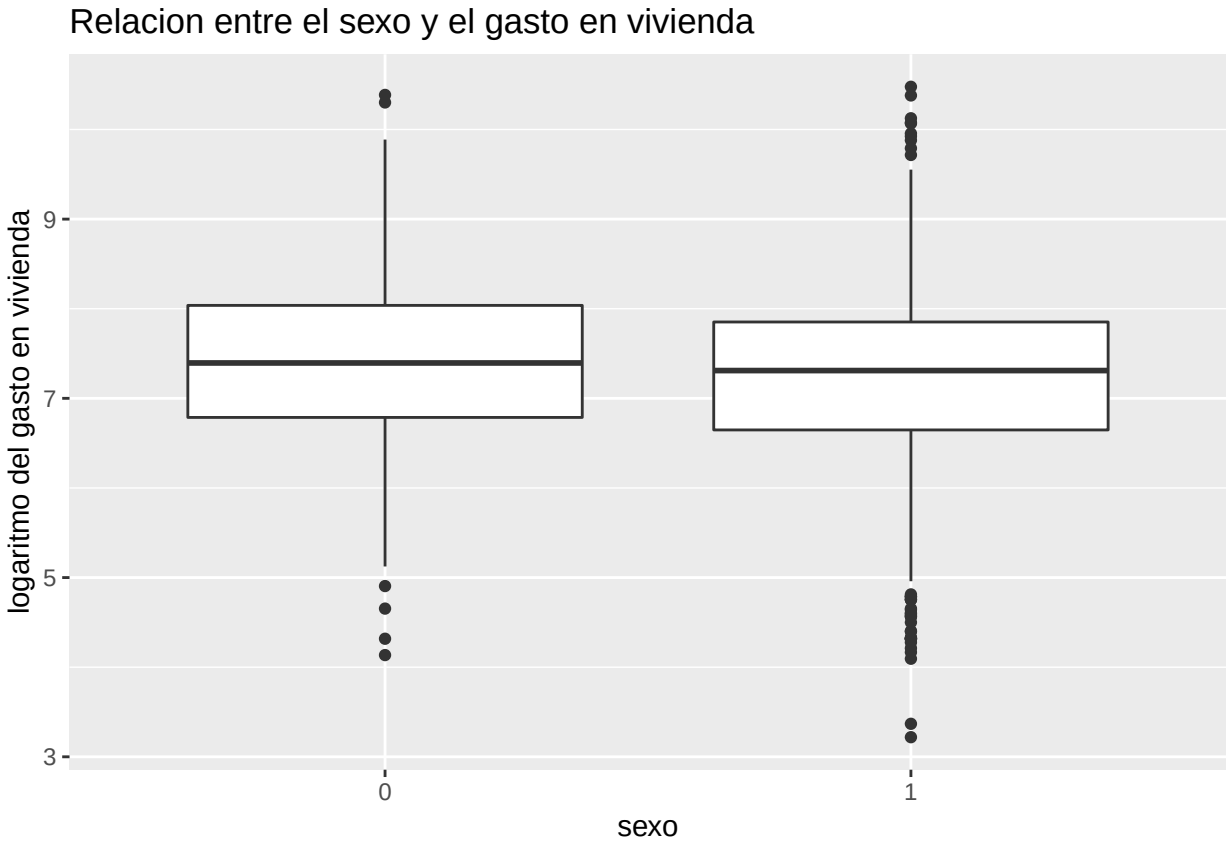
#### 4.3.1.1 Diagramas de relacion lineal

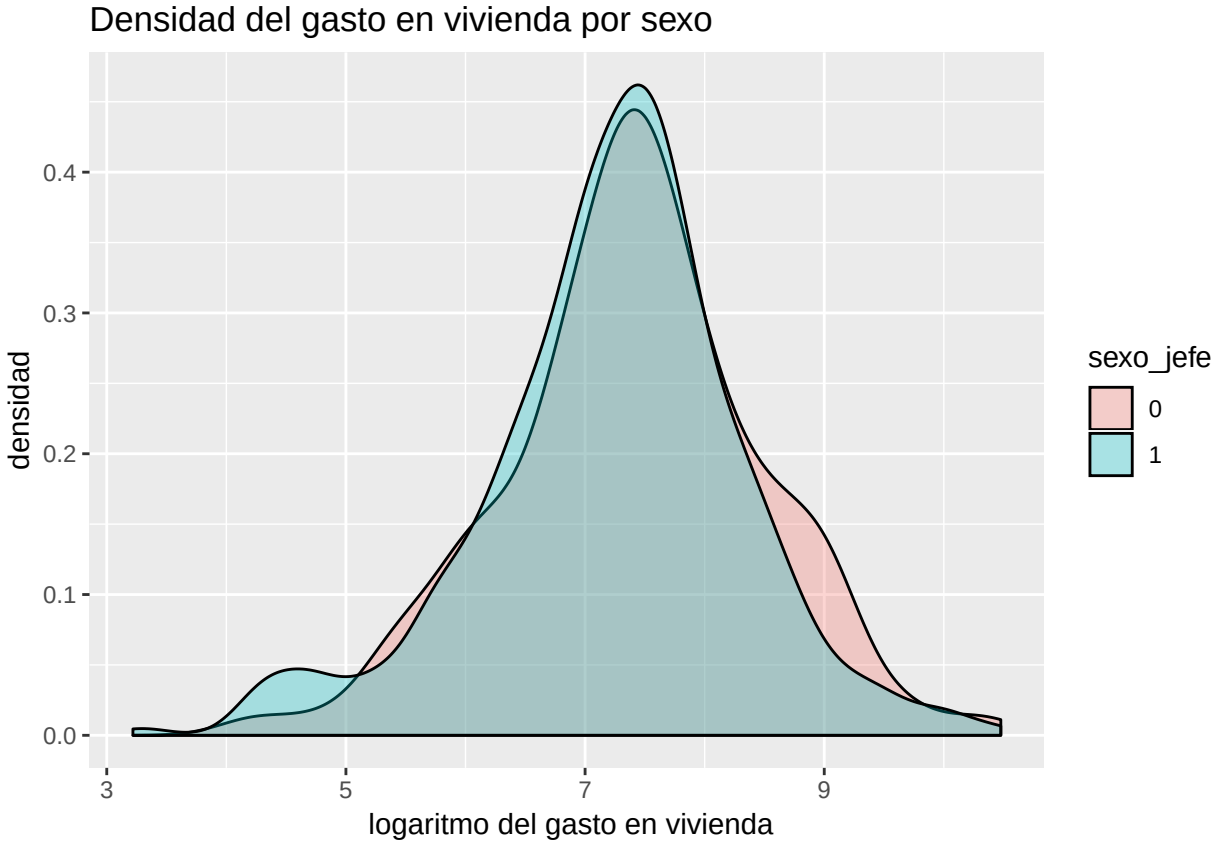


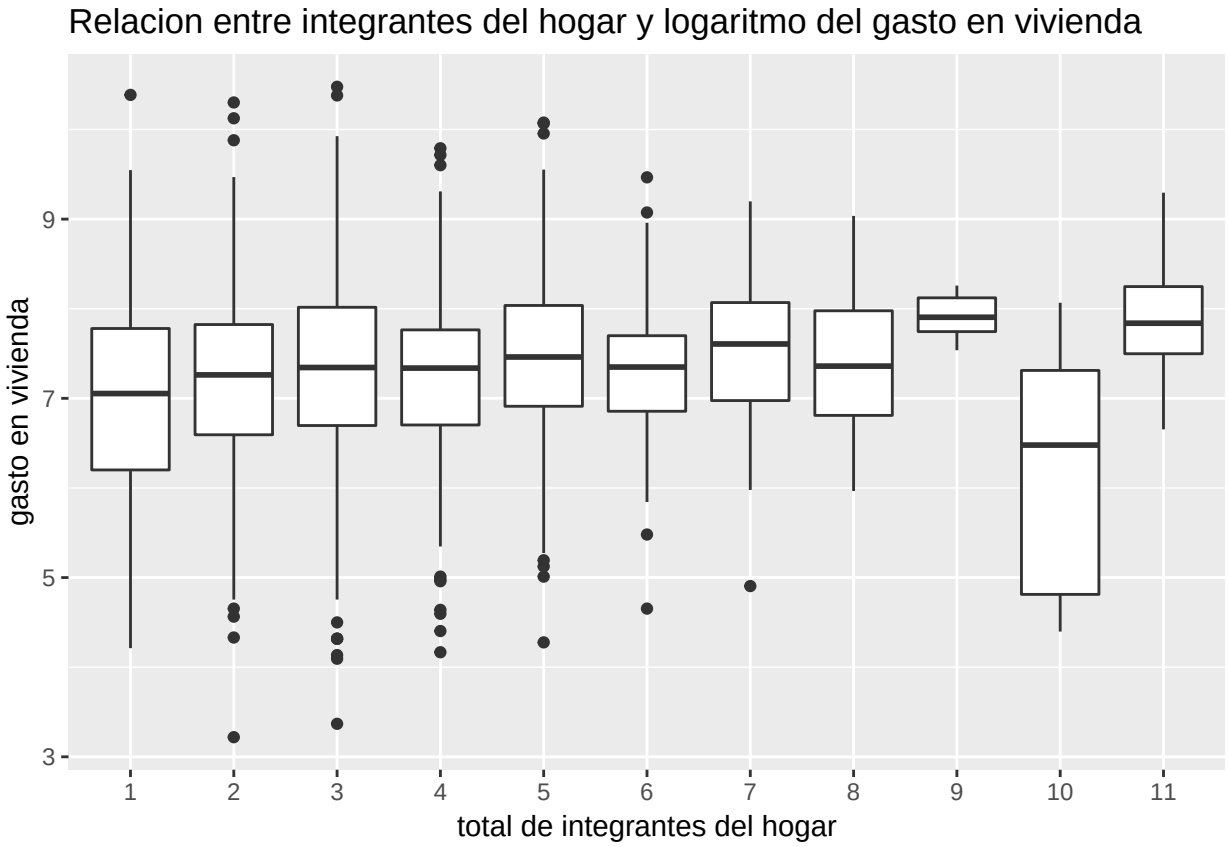

Relacion entre la edad y el logaritmo del gasto en vivienda
p = −0.1222487
logaritmo del gasto en vivienda
9
7
5
3
25
50
75
100
edad

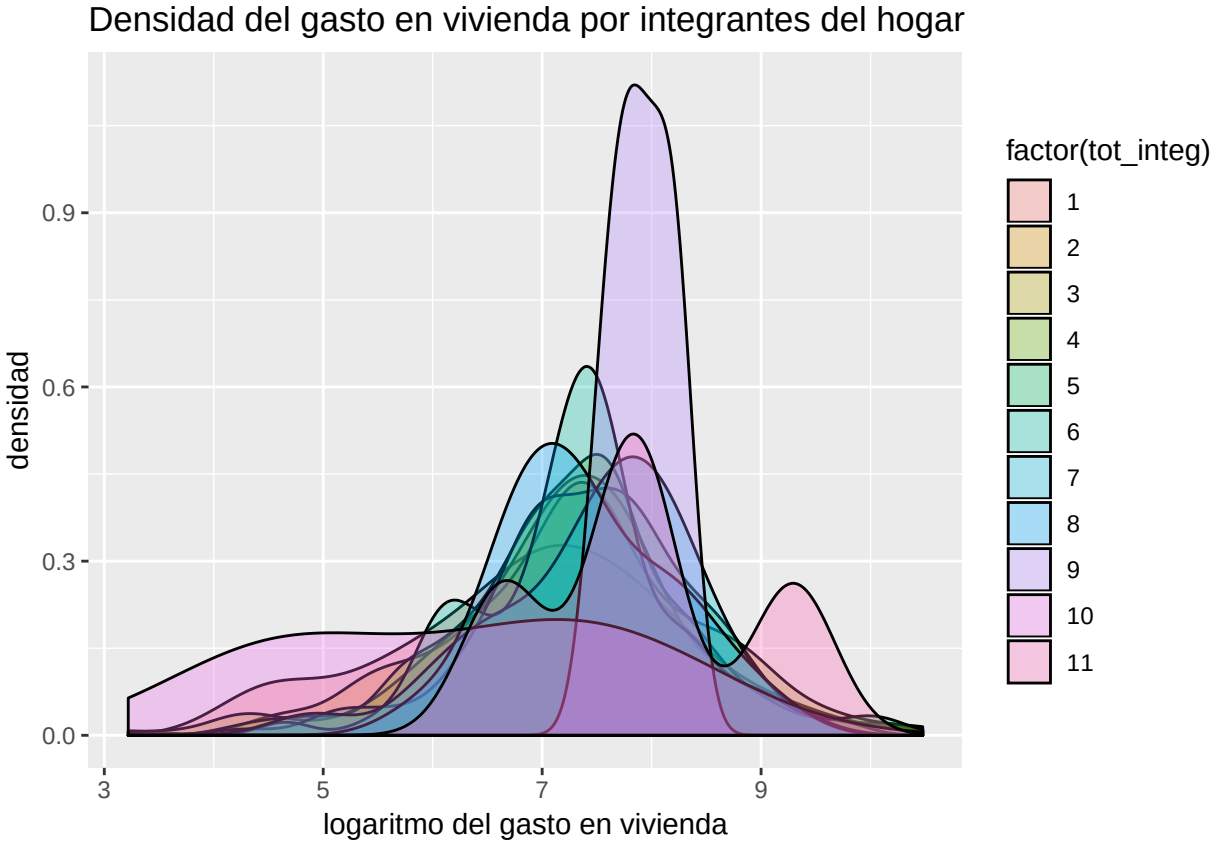
#### 4.3.1.2 Distribucion del gasto en vivienda agrupado por sexo



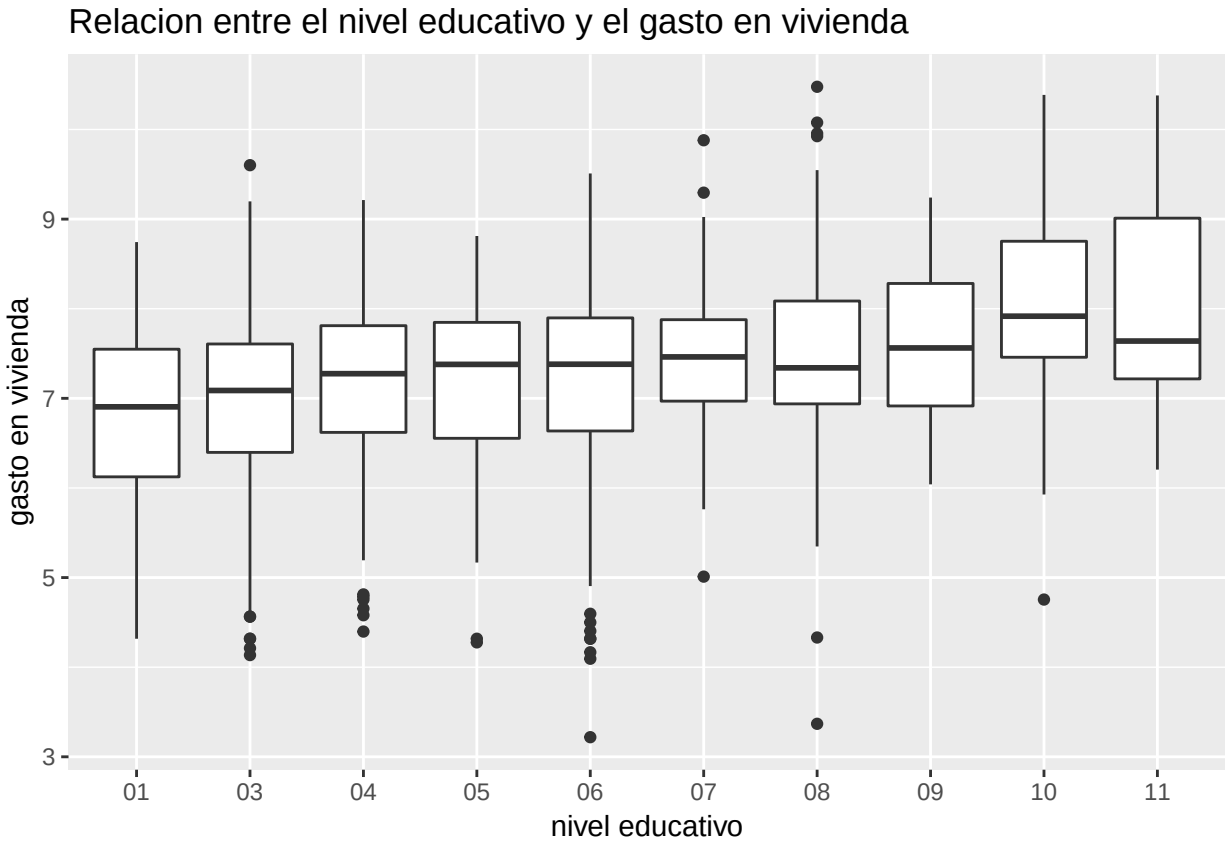



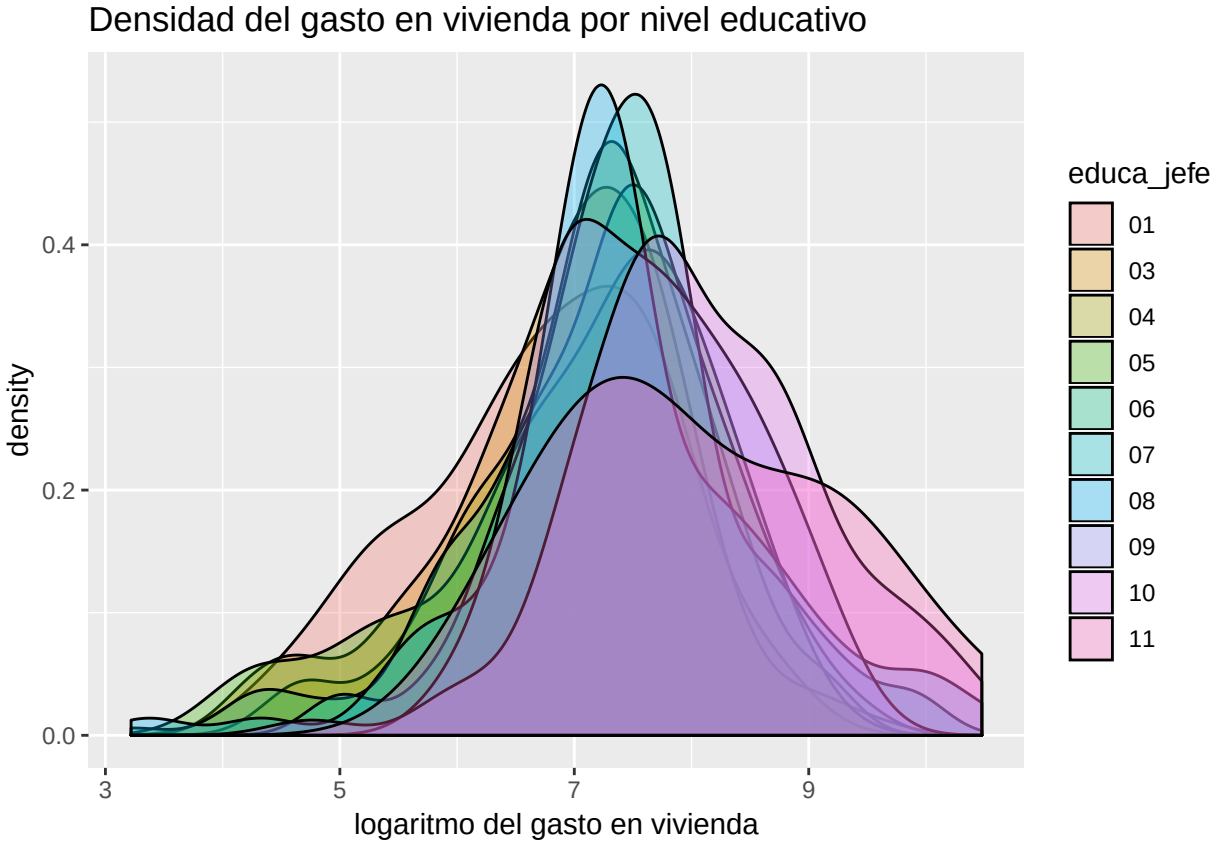
#### 4.3.1.3 Distribucion del gasto en vivienda agrupado por total de integrantes del hogar



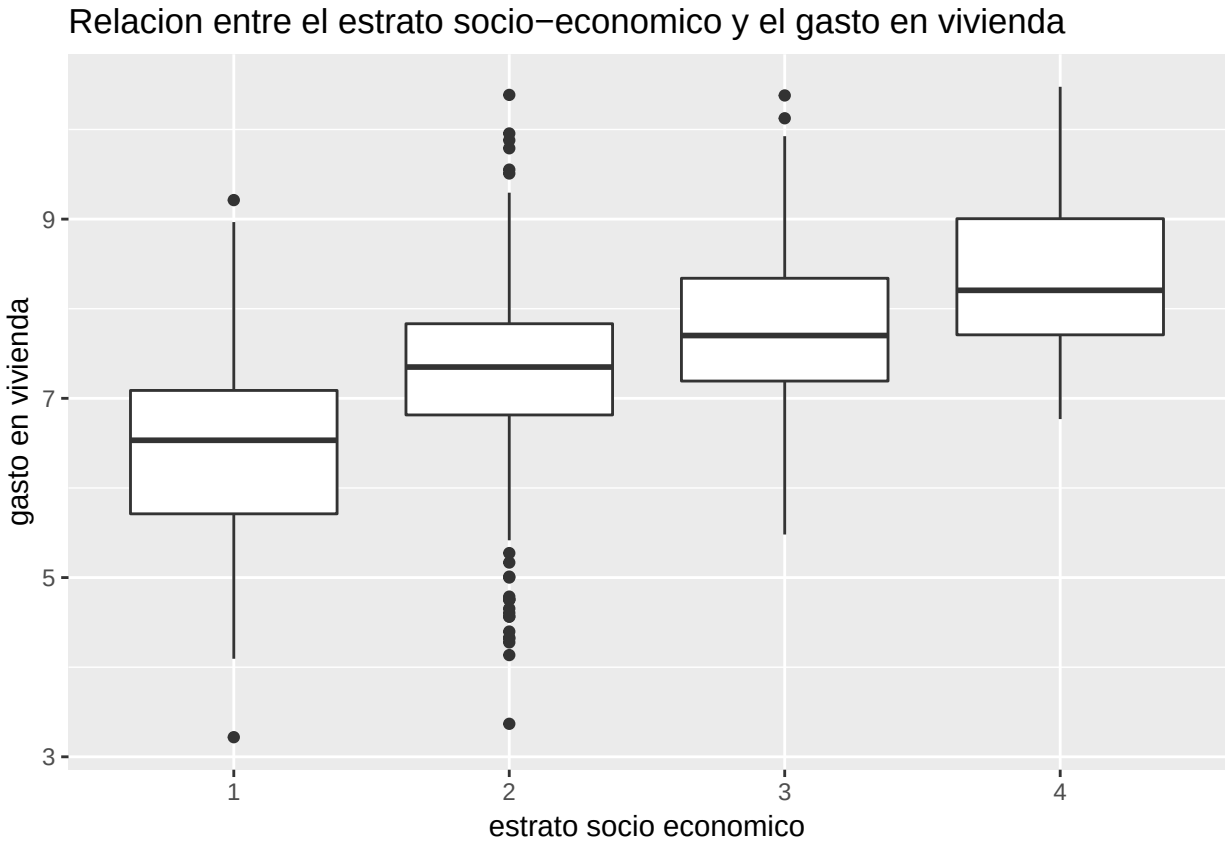



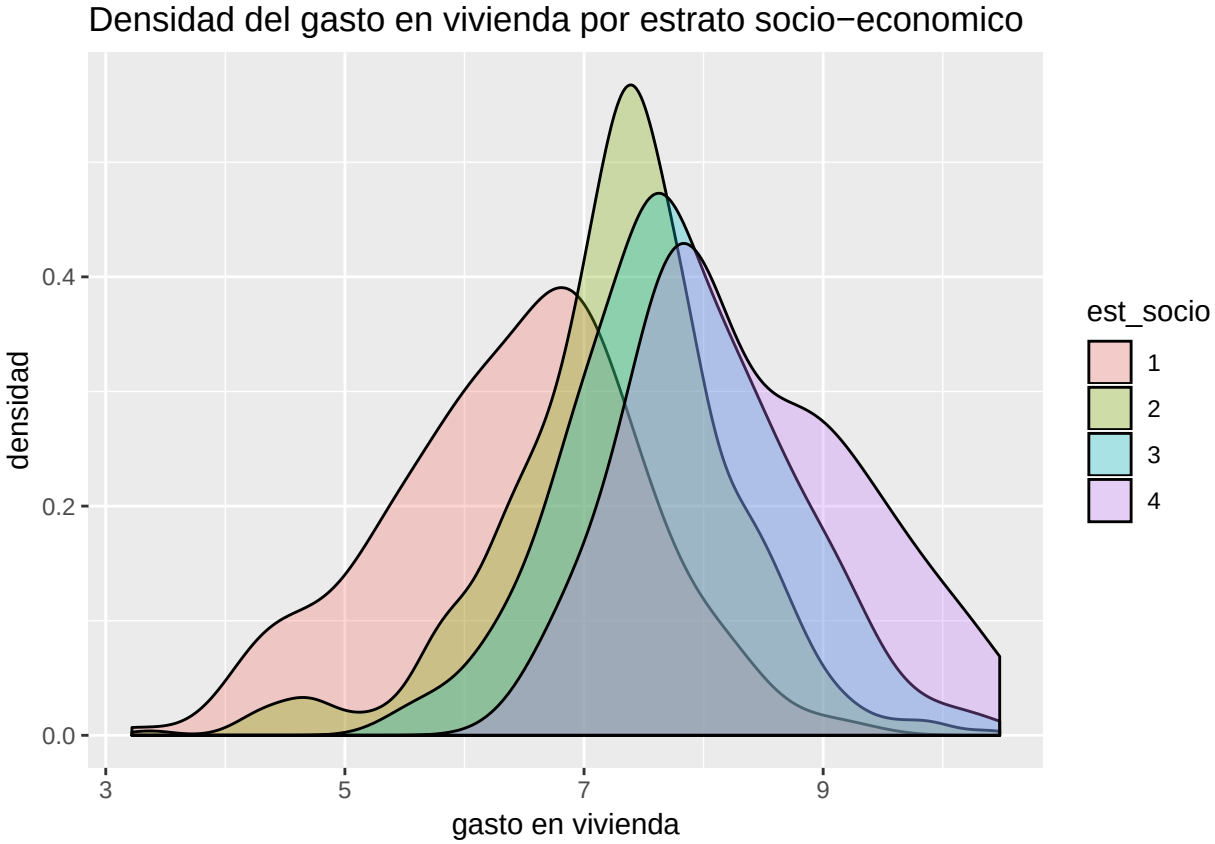
#### 4.3.1.4 Distribucion del gasto en vivienda agrupado por nivel educativo del jefe de familia



#### 4.3.1.5 Distribucion del gasto en vivienda agrupado por estrato socio-economico



## 4.4 Conclusiones del analisis exploratorio de de datos

- Todas las variables numericas en escala logaratmica que componen el estudio presentan evidencia de ser distribuidas de forma unimodal y simetrica.
- Existen mas hombres jefes de familia que mujeres
- Los hombres asumen la jefatura de la familia a edades menores
- El estrato socio economico mas comun al que pertenecen las familias mexicanas es el de Medio-Bajo sin distincion en el sexo del jefe de familia.
- El nivel de instruccion formal mas comun de un jefe de familia perteneciente al sexo masculino es el de secundaria completa. En caso de que el jefe de familia pertenezca al sexo femenino el grado de instruccion mas comun es el de primaria completa.
- Mas del 50% de las personas en el estrato socio-economico bajo tienen instruccion formal entre “sin instruccion” y “primaria completa”.
- Mas del 50% de la personas en el estrato socio economico alto tienen estudios de entre profesional completa y posgrado
- La relacion entre el logaritmo del ingreso y el logaritmo del gasto en vivienda es lineal, positiva y moderadamente correlacionada.
- La relacion entre el logaritmo del gasto general y el logaritmo del gasto en vivienda es lineal, positivo y moderadamente correlacionado.
- La relacion entre la edad y el logaritmo del gasto en vivienda es lineal, negativa y debilmente correlacionada.
- No hay evidencia de existir relacion alguna entre el gasto en vivienda y los integrantes del hogar.
- No hay evidencia de la existencia de relacion alguna entre el gasto en vivienda y el nivel educativo del jefe de familia
- Existe evidencia suficiente que apoye la relacion positiva entre el gasto en vivienda y el estrato socio economico al que pertenece una familia mexicana.

# Chapter 5

# Modelo econometrico

El modelo econometrico que propongo es un modelo lineal generalizado, cuya variable respuesta se encuentra en escala logaritmica y que puede ser descrito con la siguiente exucacion :

$$\ln(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

Donde:
$x_1$ Es esl gasto general trimestral reportado por la unidad de observaci?n medido en escala logaritmica.
$x_2$ Es la edad del jefe de familia.
$x_3$ Es la edad del jefe de familia al cuadrado.
$x_4$ Es que el sexo del jefe de familia es masculino.
$x_5$ La unidad de an?lisis pertenece al estrato socio economico 2.
$x_6$ La unidad de analisis pertenece al estrato socio economico 3.
$x_7$ La unidad de analisis pertenece al estrato socio economico 4.

## 5.1 Resultados del modelo de regresion

```
##
## Call:
## lm(formula = log(vivienda) ~ log(gasto_mon) + edad_jefe + I(edad_jefe^2) +
##     sexo_jefe + est_socio, data = datos2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2057 -0.5016  0.0397  0.5480  2.2927
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.336e+00  4.186e-01   3.191  0.00146 **
## log(gasto_mon)  6.602e-01  3.904e-02  16.912  < 2e-16 ***
## edad_jefe      -4.155e-02  8.802e-03  -4.720 2.70e-06 ***
## I(edad_jefe^2)  3.897e-04  8.278e-05   4.708 2.87e-06 ***
## sexo_jefe1     -2.055e-01  6.247e-02  -3.289  0.00104 **
## est_socio2      5.864e-01  7.197e-02   8.147 1.15e-15 ***
## est_socio3      8.707e-01  9.115e-02   9.553  < 2e-16 ***
## est_socio4      1.133e+00  1.361e-01   8.320 3.00e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8268 on 961 degrees of freedom
## Multiple R-squared:  0.4188, Adjusted R-squared:  0.4146
## F-statistic: 98.94 on 7 and 961 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: log(vivienda)
##                 Df Sum Sq Mean Sq  F value    Pr(>F)
## log(gasto_mon)   1 363.78  363.78 532.1165 < 2.2e-16 ***
## edad_jefe        1   0.64    0.64   0.9384    0.3329
## I(edad_jefe^2)   1  15.10   15.10  22.0924 2.978e-06 ***
## sexo_jefe        1  14.37   14.37  21.0195 5.146e-06 ***
## est_socio        3  79.58   26.53  38.8010 < 2.2e-16 ***
## Residuals      961 656.99    0.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##                        2.5 %        97.5 %
## (Intercept)     0.5143712362  2.1572688403
## log(gasto_mon)  0.5835896271  0.7368019585
## edad_jefe      -0.0588241483 -0.0242766405
## I(edad_jefe^2)  0.0002272565  0.0005521679
## sexo_jefe1     -0.3280454049 -0.0828596406
## est_socio2      0.4451288375  0.7276122742
## est_socio3      0.6918135486  1.0495494610
## est_socio4      0.8654571212  1.3997876225
```

### 5.1.1 Interpretacion de los resultados de regresion

Despues de ajustar el modelo con los regresores propuestos, podemos escribir la ecuacion de regresion como:

$$\hat{ln(y)} = 1.336 + 0.6602x_2 - 0.04155x_2 + 0.0003897x_3 - 0.255x_4 + 0.5864x_5 + 0.8707x_6 + 1.133x_7$$

El P-value de la prueba F de significancia global del modelo esta por debajo del $\alpha = 0.05$ (numero que, generalmente se utiliza para evaluar la significancia de pruebas estad?sticas), recordemos que la hipotesis a contrastar en la prueba de significancia global son

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0 \quad vs. \quad H_1 : \exists \;\; \beta_i \neq 0 \quad p.a \quad i \;\; \epsilon \;\; \{1, 2, \ldots, 7\}$$

Del resultado de la prueba, rechazao la hipotesis nula, por lo que alguno de los coeficientes de mi modelo es distinto de cero, y por lo tanto el modelo es globalmente significativo.

El P-value de las pruebas t de significancia individual de todos los parametros esta por debajo del $\alpha = 0.05$ ,por lo que rechazo la hipotesis nula, recordemos que las hipotesis a contrastar de la prueba t de significancia indivual es :

$$H_0 : \frac{\hat{\beta}_j}{S(\beta_j)} = 0 \quad vs \quad H_1 : \frac{\hat{\beta}_j}{S(\hat{\beta}_j)} \neq 0$$

Por lo que todas las variables incluidas en el modelo tienen algún (**Caeteris Paribus**) sobre el gasto en vivienda que no es debido solamente al azar de tal forma que son estadisticamente significativas.

Puedo decir que, de acuerdo a la medida de bondad de ajuste $R^2$ ajustado que 41% de la desviacion del modelo base es directamente imputable a la existencia de correlacion de la variable explicada con los regresores.

El modelo base es aquel donde solo se tienen en cuenta los efectos capturados por el intercepto al origen ($\hat{\beta_0}$), es decir, cuando el resto de las $ x_{i} $ se mantienen en 0.

Las unidades observacionales que no cumplen explicitamenente alguna de las caracteristicas de las variables categericas( por ejemplo, que el sexo del jefe de familia sea femenino, o que la familia pertenezca al estrato socio-economico 1) son efectos capturados en el modelo base

## 5.2 Analisis de residuales

El analisis de residuales es una herramienta que me ayudara a comprobar los supuestos que todo modelo de regresion lineal multiple (**RLM**) debe cumplir, esto para saber que la inferencia sobre los parametros del modelo es correcta y confiable. Los supuestos de **RLM** son:

**Independencia de los errores**

$$F_{\epsilon_1,\epsilon_2\ldots,\epsilon_n}(\epsilon_1,\epsilon_2,\ldots,\epsilon_n) = F_{\epsilon_1}(\epsilon_1)F_{\epsilon_2}(\epsilon_2)\ldots F_{\epsilon_n}(\epsilon_n)$$

donde F es la funcion de distribucion de las perturbaciones.

1. **Linealidad en los parametros** : para cualquier combinacion de los valores de $x_i$ se tiene que:

   $$E(\hat{\epsilon}|X) = 0$$

   Esto es para que los estimadores de los efectos ceteris paribus sean insesgados

2. **Homocedasticidad condicional** : La varianza de los residuos, dados los parametros es constante, i.e:

   $$Var(\hat{\epsilon_i}|X) = \sigma_\epsilon^2$$

3. **Normalidad multivariada** : Los residuos se distribuyen normal con media 0 y varianza constante, i.e:

   $$\hat{\epsilon} \sim N(0, \sigma_\epsilon^2)$$

Esto es para que los estimadores de los coeficientes de regresion sean eficientes (de minima varianza) y que los intervalos de confianza sean exactos.

4. **No existencia de multicolinealidad perfecta** : Ninguna de las columnas de la matriz $X$ (la matriz de diseno) es combinacion lineal del resto de las columnas, esto es:

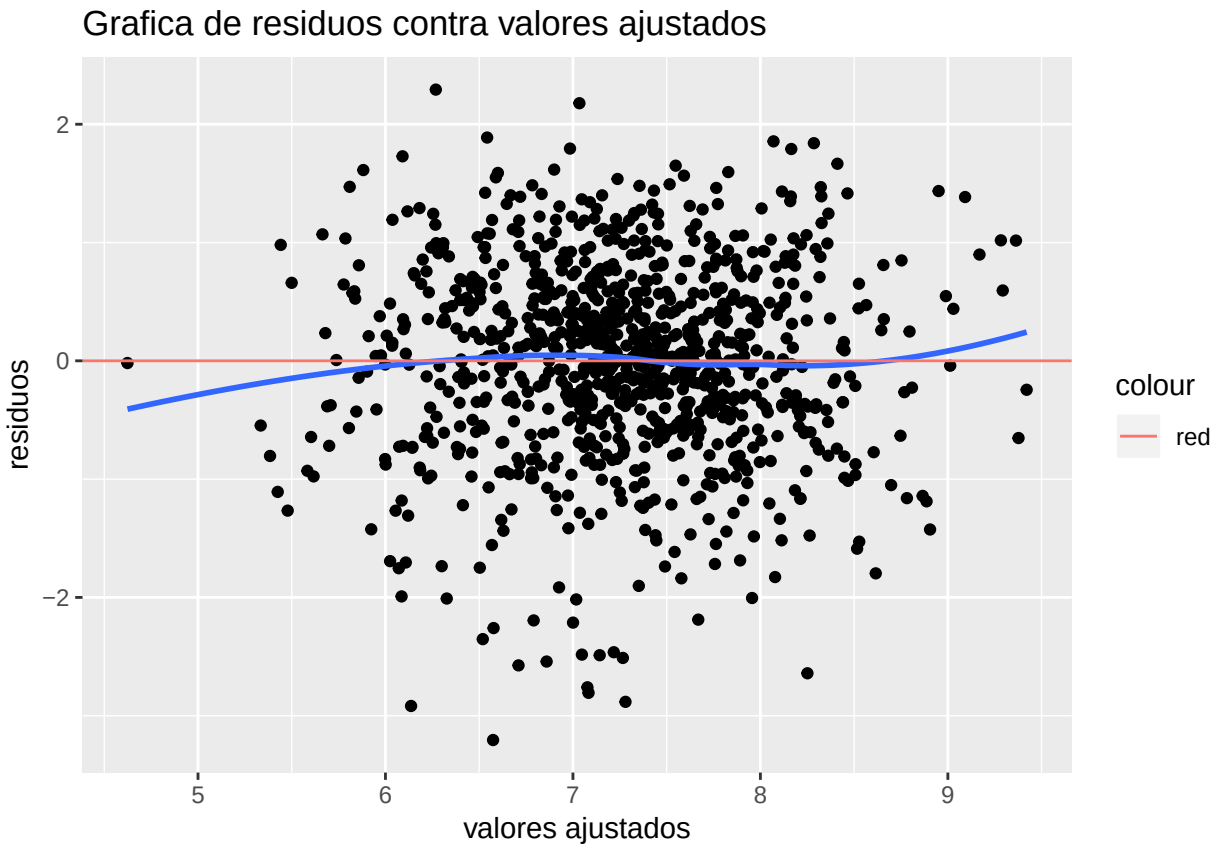
   $$|(X^T X)^{-1}| > 0$$

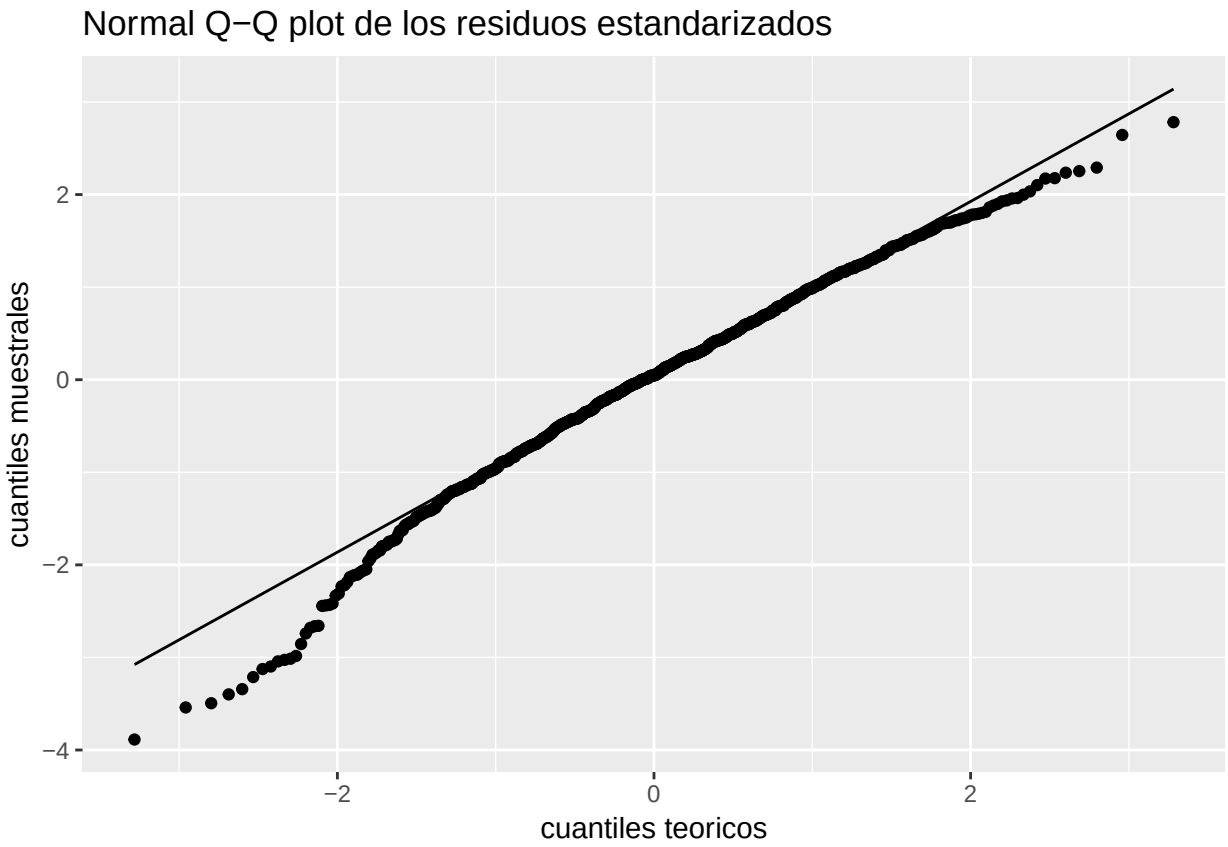
   Esto se pide para que sea posible calcular los estimadores de los coeficientes de regresion, sin empargo, la **Multicolinealidad imperfecta** que es:

   $$|(X^T X)^{-1}| \approx 0$$

   que se presenta cuando existe una alta correlacion entre variables tambien representa un grave error en un modelo, ya que se “inflan” los errores estandar de los estimadores, lo cual genera una impresicion y “ensanchamiento” de los intervalos de confianza.

### 5.2.1 Tests graficos para comprobar los supuestos de regresion lineal multiple




Normal Q−Q plot de los residuos estandarizados
cuantiles muestrales
cuantiles teoricos
2
0
−2
−4
−2
0
2

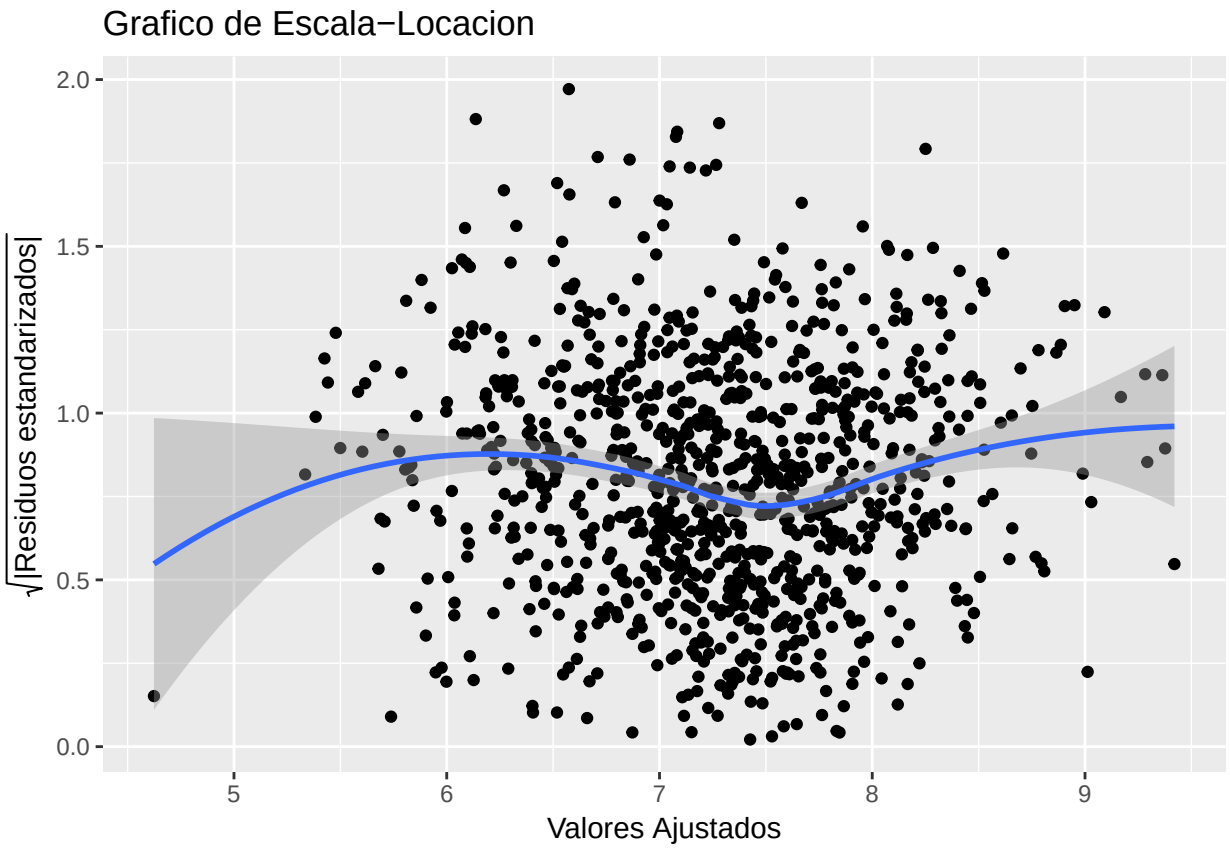
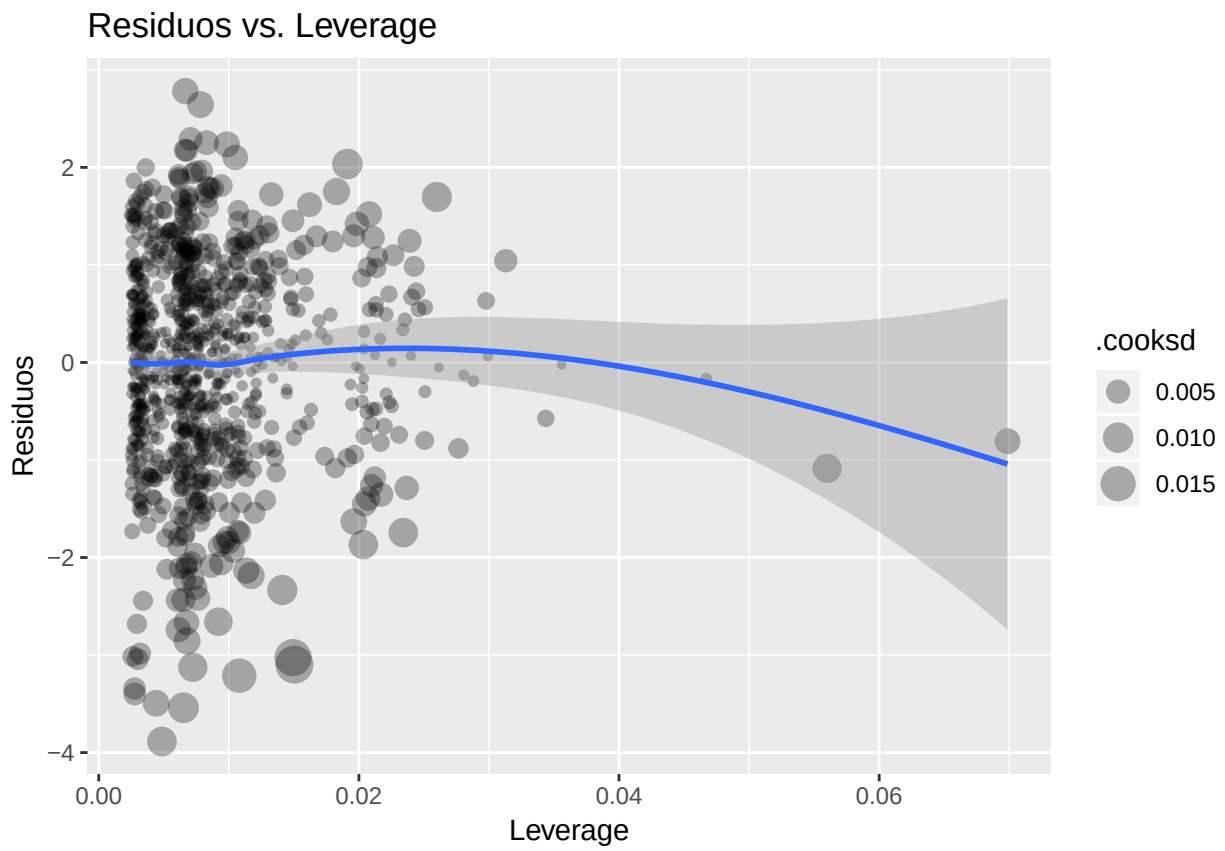
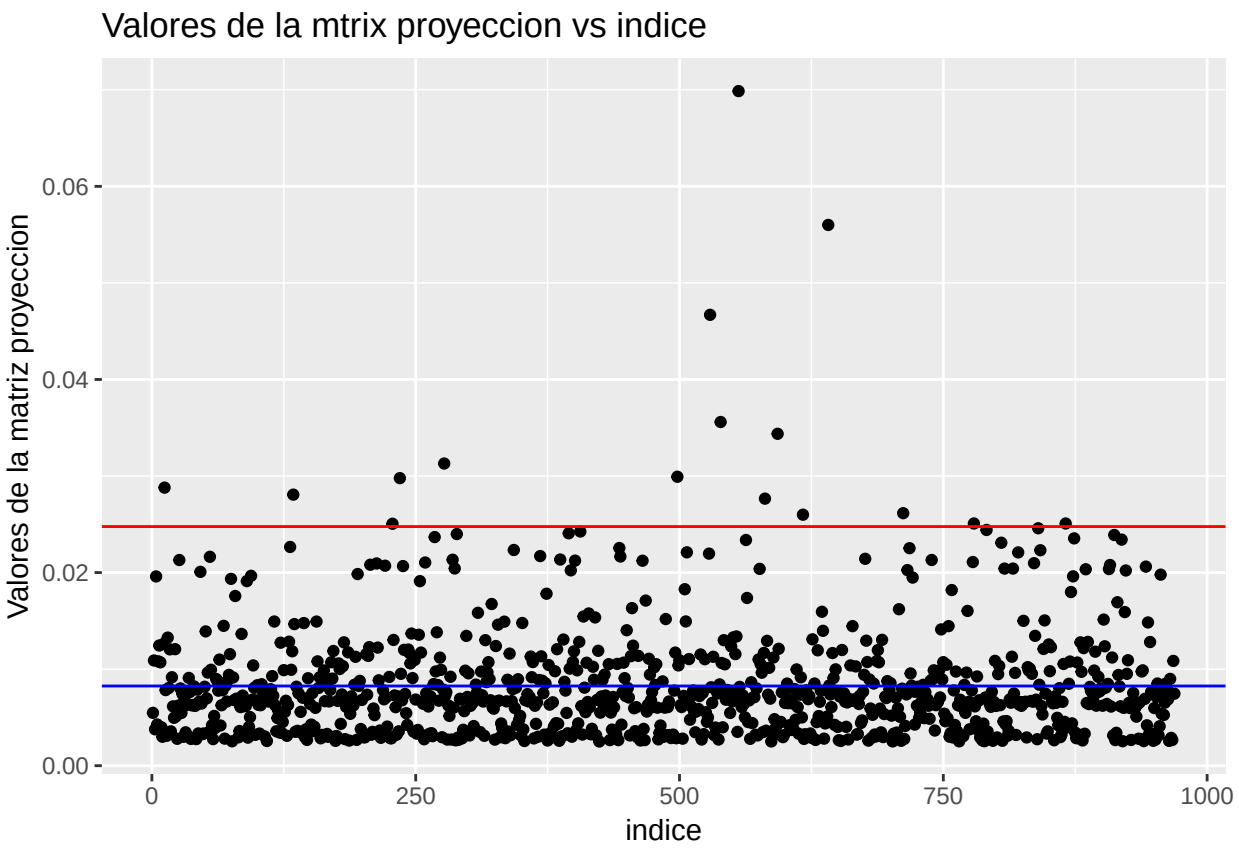
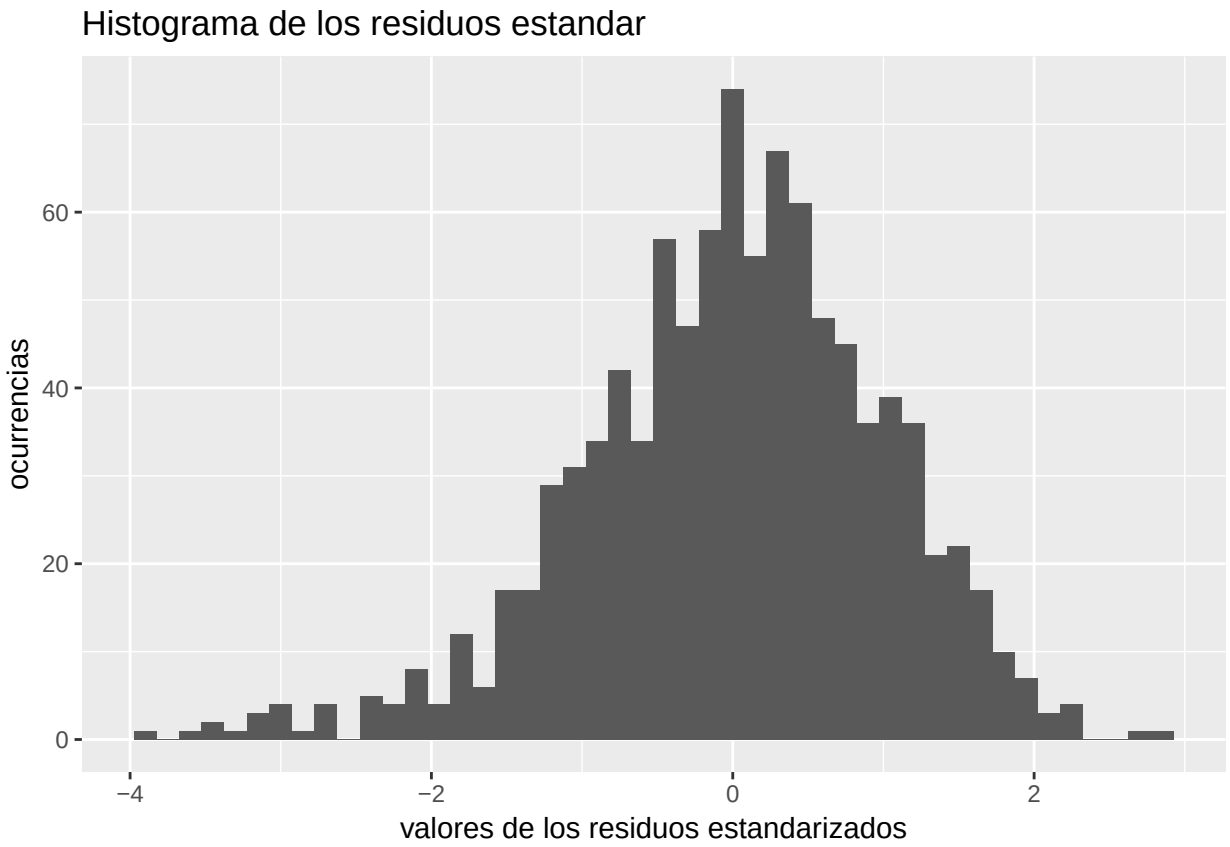


Grafico de Escala−Locacion
2.0
1.5
1.0
0.5
0.0
√|Residuos estandarizados|
5
6
7
8
9
Valores Ajustados


Residuos vs. Leverage
Residuos
2
0
−2
−4
0.00
0.02
0.04
0.06
Leverage
.cooksd
0.005
0.010
0.015


Valores de la mtrix proyeccion vs indice
Valores de la matriz proyeccion
0.06
0.04
0.02
0.00
0
250
500
750
1000
indice


Histograma de los residuos estandar
ocurrencias
0
20
40
60
−4
−2
0
2
valores de los residuos estandarizados

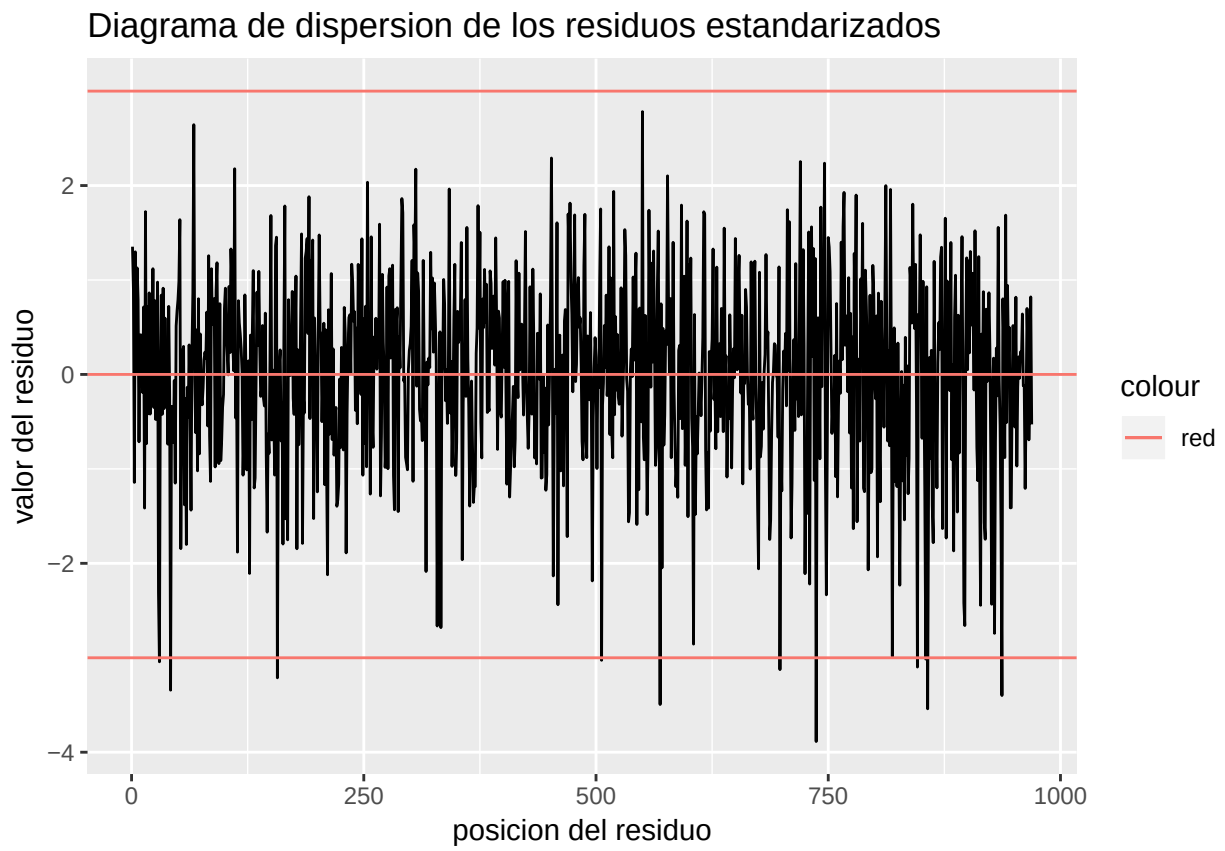
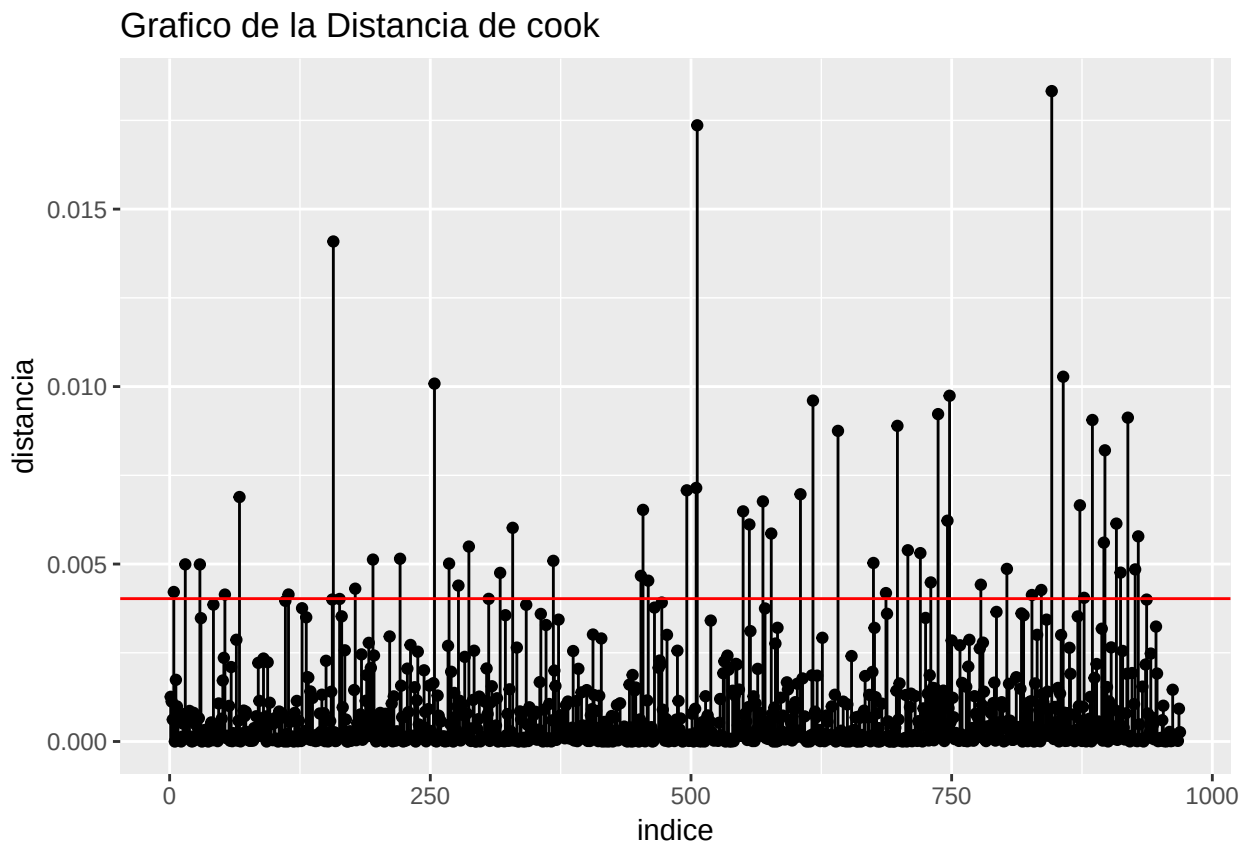


Diagrama de dispersion de los residuos estandarizados
valor del residuo
2
0
−2
−4
0
250
500
750
1000
posicion del residuo
colour
red




### 5.2.2 Tests estadisticos para comprobar las hipotesis de regresion lineal multiple

Quiero comprobar que la esperanza de los residuos sea 0 , por lo que hare el siguiente contraste de hipotesis con una prueba t

$$H_0 : E(\epsilon|X) = 0 \quad vs \quad H_1 : E(\epsilon|X) \neq 0$$

```
## [1] 1
```

Como el P-value es mayor que el valor de significancia $\alpha = 0.05$ acepto la hipotesis nula de que los residuos tienen valor esperado 0.

Aplico la prueba de Durbin-Watson para detectar si el coeficiente de correlacion es 0 o distinto de cero, esto para verificar la independencia de los residuos, sus hipotesis son :

$$H_0 : \rho(i, i+1) = 0 \quad vs \quad H_1 : \rho(i, i+1) \neq 0$$

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.01480526      1.968207    0.66
##  Alternative hypothesis: rho != 0
```

Como el P-value es mayor que el valor $\alpha = 0.05$ entonces acepto la hipotesis nula de que el coeficiente de correlacion es 0.

Aplico la prueba de Breusch-Pagan, cuyas hipotesis a contrastar son : La varianza de los residuos es constante vs. la varianza de los residuos es una funcion de los valores ajustados del modelo. Esto es para verificar la homocedasticidad.

$$H_0 : Var(\hat{\epsilon}|X) = \sigma_\epsilon^2 \quad vs \quad H_1 : Var(\hat{\epsilon}|X) = \sigma_\epsilon^2(\hat{Y})$$

```
##
##  studentized Breusch-Pagan test
##
## data:  fitb
## BP = 5.9653, df = 7, p-value = 0.5438
```

Como el P-value es mayor que $\alpha = 0.05$ , por lo tanto acepto la hipotesis nula de una varianza constante.

Tambien proveo el resultado de la prueba de Goldfeld-Quandt, cuyas hipotesis son: La varianza es igual en un primer grupo de residuos que en un segundo grupo de residuos vs, la varianza en el primer grupo de residuo es es menor que la varianza en un segundo grupo (es decir, la varianza aumenta conforme crecen los valores ajustados del modelo) h

$$H_0 : \sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} \quad vs \quad H_1 : \sigma^2_{\epsilon_1} < \sigma^2_{\epsilon_2}$$

```
##
##  Goldfeld-Quandt test
##
## data:  fitb
## GQ = 0.58906, df1 = 477, df2 = 476, p-value = 1
## alternative hypothesis: variance increases from segment 1 to 2
```

Como el P-value es mayor que el valor de significancia $\alpha = 0.05$ entonces acepto la hipotesis nula de que la varianza es igual en ambos segmentos

Proveo tambien el resultado de la prueba de Kolmogorov-Smirnov, cuyas hipotesis a contrastar son :

$$H_0 : \frac{\hat{\epsilon}}{\sqrt{S^2(1-h_i)}} \sim N(0,1) \quad vs \quad \frac{\hat{\epsilon}}{\sqrt{S^2(1-h_i)}} \nsim N(0,1)$$

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  resultados_fit$.std.resid
## D = 0.037732, p-value = 0.1267
## alternative hypothesis: two-sided
```

Como el P-value es mayor que el valor de significancia $\alpha = 0.05$ entonces acepto la hipotesis nula de que los residuos estandarizados se distribuyan normal con parametros de media 0 y varianza 1.

La identificacion de la multicolinealidad no se puede hacer de forma tradicional en este modelo, ya que se incluye un termino cuadratico (la edad del jefe de familia al cuadrado) como regresor, por lo cual, para proveer evidencia de que no existe multicolinealidad entre los regresores, cree un modelo de regresion auxiliar donde excluyo el termino cuadratico. A continuacion presento los factores de inflacion de varianza (FIV) de esa regresion auxiliar.

Recordemos que los FIV se calculan en dos pasos, primero se crean i distintas regresiones por el m?todo de m?nimos cuadrados cuya variable explicada es $x_i$ y los regresores son el resto de las variables, es decir

$$x_i = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k + \epsilon$$

Despues se calcula el FIV para cada coeficiente $\hat{\beta}_i$ del modelo de regresion original (En nuestro caso es el modelo que no incluye el termino cuadratico)

$$FIV_i = \frac{1}{1 - R^2_i}$$

Donde $R^2_i$ es el coeficiente de determinacion de la regresion cuya variable explicada es $x_i$ y sus regresores son el resto de las k variables explicativas.

Table 5.1: Tabla de los valores influyentes

| count | .std.resid | .cooksd |
|---|---|---|
| 157 | -3.212012 | 0.0140883 |
| 506 | -3.026296 | 0.0173619 |
| 569 | -3.494389 | 0.0067647 |
| 698 | -3.125113 | 0.0088935 |
| 737 | -3.886520 | 0.0092244 |
| 846 | -3.098029 | 0.0183249 |
| 857 | -3.540299 | 0.0102776 |

Por lo general se dice que una variable aporta colinealidad al modelo si $FIV(\hat{\beta}_i) > 10$ lo cual claramente no ocurre entre los regresores, por lo que podemos descartar la existencia de multicolinealidad ya que no existe evidencia suficiente en pro de esta.

```
##                    GVIF Df GVIF^(1/(2*Df))
## log(gasto_mon) 1.296543  1        1.138658
## edad_jefe      1.090427  1        1.044235
## sexo_jefe      1.037531  1        1.018593
## est_socio      1.226484  3        1.034611
```

### 5.2.3 Identificacion de los valores extremos e influyentes

A continuacion presento una tabla cuyos metricas de residuo estandar y distancia de cook son aparentemente mas grandes que el resto (residuo estandar $> 3$ & distancia de cook $> 4$ veces el promedio de la distancia de cook de los residuos), esto por que son potenciales observaciones extremas.

# Chapter 6

# Interpretacion economica del modelo

Segun los resultados del estudio, y manteniendo todo lo demas constante (**Ceteris Paribus**) :

- El gasto en vivienda parte de los $3.8073 pesos.
- Un incremento de un punto porcentual en el gasto, repercute en un incremento porcentual del 0.6602% en el gasto en vivienda, por la especificacion del modelo, este tiene una interpretaci?n de elasticidad con respecto a esta variable.
- Un incremento en una unidad de edad del jefe de familia representa un detrimento del 4.155% en el gasto en vivienda, pero ya que anadi el termino cuadratico a la forma funcional del modelo tambien significa un incremento del .03987% en el gasto en vivienda, esta relacion proviene de los rendimientos marginales decrecientes proporcionados por la edad.
- El hecho de que el jefe de familia pertenezca al sexo masculino se puede traducir en un detrimento del gasto en vivienda de 25.54%.
- Al pertenecer al estrato socio economico Medio-Bajo el hogar gasta 58.64% en vivienda mas en vivienda que las familias que pertenecen al estrato socio economico Bajo.
- Al pertenecer al estrato socio-economico Medio-Alto la unidad de observacion gasta un 87.07% mas en vivienda que las familias que pertenecen al estrato socio economico Bajo.
- Al pertenecer al estrato socio-economico Alto, la unidad la familia gasta un 113.3% mas en vivienda que las familias que pertenecen al estrato socio economico Bajo.
- Para realizar predicciones sobre el gasto en vivienda de una familia mexicana, es conveniente regresar a la unidad original, es decir expresar la esperanza de Y en pesos en vez de en logaritmos, esto lo logramos exponenciando la ecuacion del modelo de ambos lados, de lo cual resulta la siguiente ecuaci?n.

$$E[Y|X] = e^{1.336+0.6602x_2-0.04155x_2+0.0003897x_3-0.255x_4+0.5864x_5+0.8707x_6+1.133x_7}$$

# Chapter 7

# Evaluacion predictiva del modelo

Para esta etapa de la investigacion seleccione otra muestra aleatoria de tamano 200, elimine las observaciones que ya se habian incluido en la muestra aleatoria original (los datos de entrenamiento) todo esto con la finalidad conformar un conjunto de datos de prueba.

Una medida simple para evaluar el poder predictivo del modelo es el coeficiente del correlacion lineal de Pearson, definido como:

$$r_{xy}(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

El cual es de:

```
## [1] 0.5184405
```

Este resultado entre mas cercano a 1 es mejor.

Una segunda medida que proveo es la exactitud Min-Max, que se calcula como:

$$media(\frac{min(actuales, predichos)}{max(actuales, predichos)})$$

y cuyo resultado para el modelo es:

```
## [1] 0.9040671
```

Table 7.1: Tabla que recoje algunos de los valores exactos vs. los valores predichos por el modelo

| actuals | predicted |
|---|---|
| 7.373374 | 6.996638 |
| 5.010635 | 6.661774 |
| 7.326985 | 8.096949 |
| 7.928406 | 7.781373 |
| 8.144752 | 7.732080 |
| 6.452270 | 7.455156 |
| 5.010635 | 6.781590 |
| 7.272398 | 6.164553 |
| 4.890800 | 6.310569 |
| 5.228592 | 7.351888 |

Este numero, que se encuentra entre 0 y 1, entre mas alto significa mayor precision del modelo, en este caso la precision max-min es del 90%.

Tambien exploro una medida conocida como la media del porcentaje de error absoluto calculado como

$$MPEA = media(\frac{abs(predichos - actuales)}{actuales})$$

```
## [1] 0.1096576
```

Este numero,que se encuentra entre 0 y 1, nos dice que el el MAPE del modelo es de aproximadamente 10%. Este numero entre mas bajo mejor, significa la desviacion absoluta promedio entre valores predichos y actuales.

(?) (?) (?) (?) (?)

# Bibliography

Champernowne, D. G. and Theil, H. (1972). Principles of Econometrics. *The Economic Journal*, 82(325):222.

Editors, S., Gentleman, R., Hornik, K., and Parmigiani, G. G. (2009). *Use R !*

Farnsworth, G. V. (2008). Econometrics in R. *I Can*, page 50.

Learning, C., Reserved, A. R., and Learning, C. *Jeffrey M. Wooldridge/Introductory Econometrics A Modern Approach*.

Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. *Booksgooglecom*, 58(2):752.