# Taller: Modelos de distribución

Kevin Garcia - Cesar Saavedra

6 de marzo de 2018

# Distribución Poisson

Esta distribución es una de las más importantes distribuciones de variable discreta. Sus principales aplicaciones hacen referencia a la modelización de situaciones en las que nos interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio, bajo presupuestos de aleatoriedad. Su función de densidad esta dada por:

$$f(x, \lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}; x \in \{0, 1, 2, 3, ...\}$$

donde:

- x es el número de ocurrencias del evento o fenómeno (la función nos da la probabilidad de que el evento suceda precisamente x veces).

# Distribución Poisson

- $\lambda$ es un parámetro positivo que representa el número de veces que se espera que ocurra el fenómeno durante un intervalo dado.
- $F(x) = \sum_{i=1}^{x} \frac{\lambda_i \cdot e^{-\lambda}}{i!}$
- $f.g.m = m_x(t) = E[e^{tx}] = \sum_x e^{tx} p(x) = e^{[\lambda(e^t - 1)]}$
- $Media = E[x] = \lambda$
- $Varianza = \lambda$
- Coeficiente de asimetría$=\frac{1}{\sqrt{\lambda}}$
- curtosis$=3 + \frac{1}{\lambda}$

# Aplicaciones de la distribución Poisson

La distribución de Poisson se emplea para describir procesos como los siguientes:

- El número de autos que pasan a través de un cierto punto en una ruta (suficientemente distantes de los semáforos) durante un periodo definido de tiempo.
- El número de errores de ortografía que uno comete al escribir una única página.
- El número de llamadas telefónicas en una central telefónica por minuto.
- El número de servidores web accedidos por minuto.
- El número de defectos en una longitud específica de una cinta magnética.
- El número de defectos por metro cuadrado de tela.
- El número de estrellas en un determinado volumen de espacio.

# Distribución Logística

La función de distribución de la logística es una distribución de probabilidad continua que se usa como modelo de crecimiento. Por ejemplo, con un nuevo producto, a menudo encontramos que el crecimiento es inicialmente lento, luego gana impulso, y finalmente se ralentiza cuando el mercado está saturado o hay alguna forma de equilibrio alcanzado
Su función de densidad está dada por:

$$f(x; a, b) = \frac{e^{-(x-a)/b}}{b(1 + e^{-(x-a)/b})^2} = \frac{1}{4b} sech^2\left(\frac{x-a}{2b}\right)$$

# Distribución Logística

Su función de distribución está dada por:

$$F(x; a, b) = \frac{1}{1 + e^{-(x-a)/b}} = \frac{1}{2} + \frac{1}{2} tanh\left(\frac{x-a}{2b}\right)$$

- $f.g.m = e^{at}\Gamma(1 - bt)\Gamma(1 + bt) = \pi bt \frac{e^{at}}{sin(\pi bt)}$
- $Media = E[x] = a$
- $Mediana = a$
- $Moda = a$
- $Varianza = \frac{\pi^2 b^2}{3}$

# Aplicaciones de la distribución Logística

La distribución logística ha sido muy utilizada en áreas como:

- Biología: para describir cómo se comportan las especies en entornos competitivos.
- Epidemiología: para describir la propagación de epidemias.
- Psicología: para describir el proceso de aprendizaje.
- Tecnología: para describir cómo las tecnologías se popularizan y compiten entre sí.
- Márketing: para estudiar la difusión de nuevos productos.
- Energía: para estudiar la difusión y sustitución de unas fuentes de energía primarias por otras.

# Comportamiento distribución Poisson

# Comportamiento distribución Logística

# Variable 'Edad mediana de las viviendas'

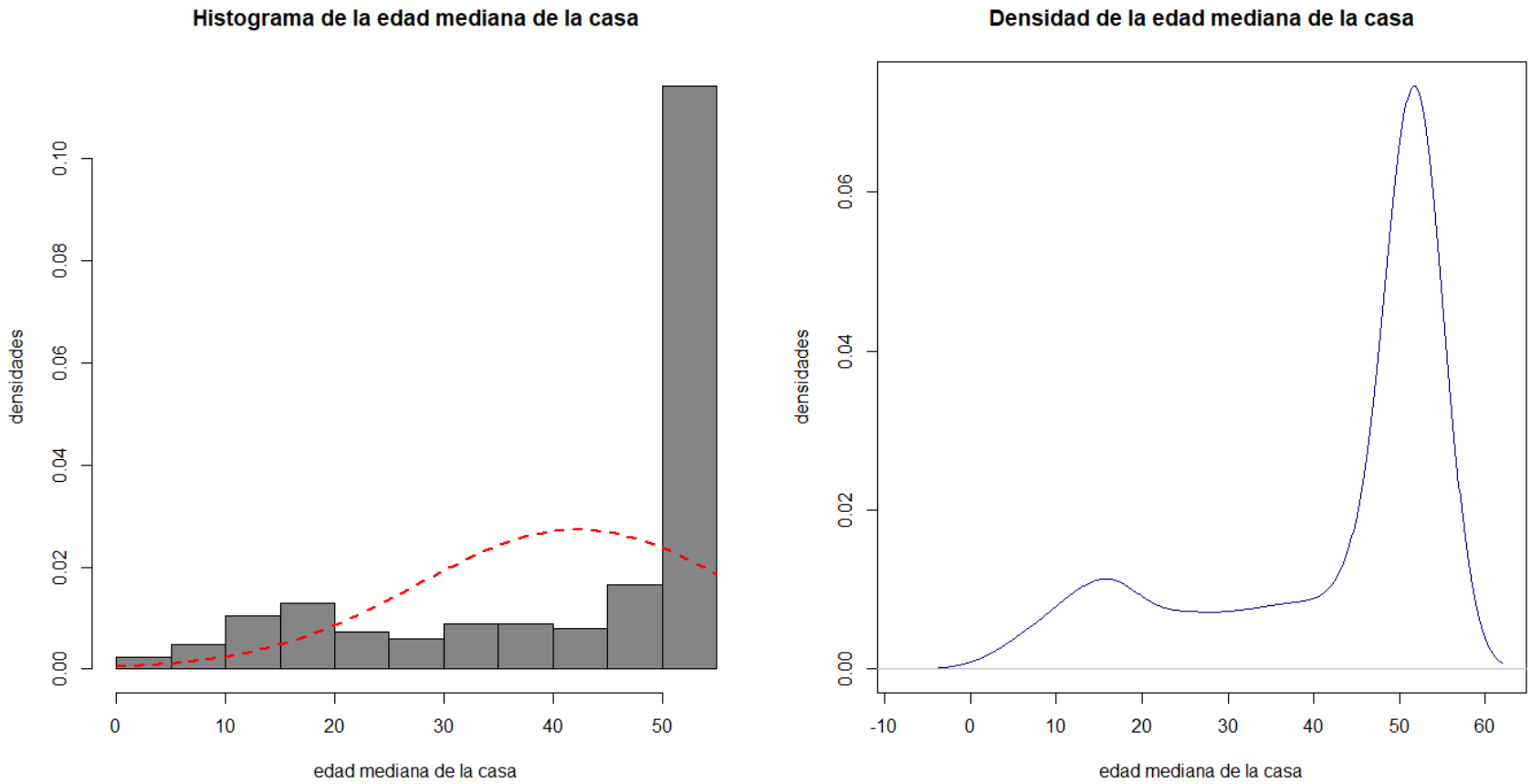


Figura: Histograma y densidad de la variable 'Edad mediana de las viviendas'

# Variable 'Edad mediana de las viviendas'

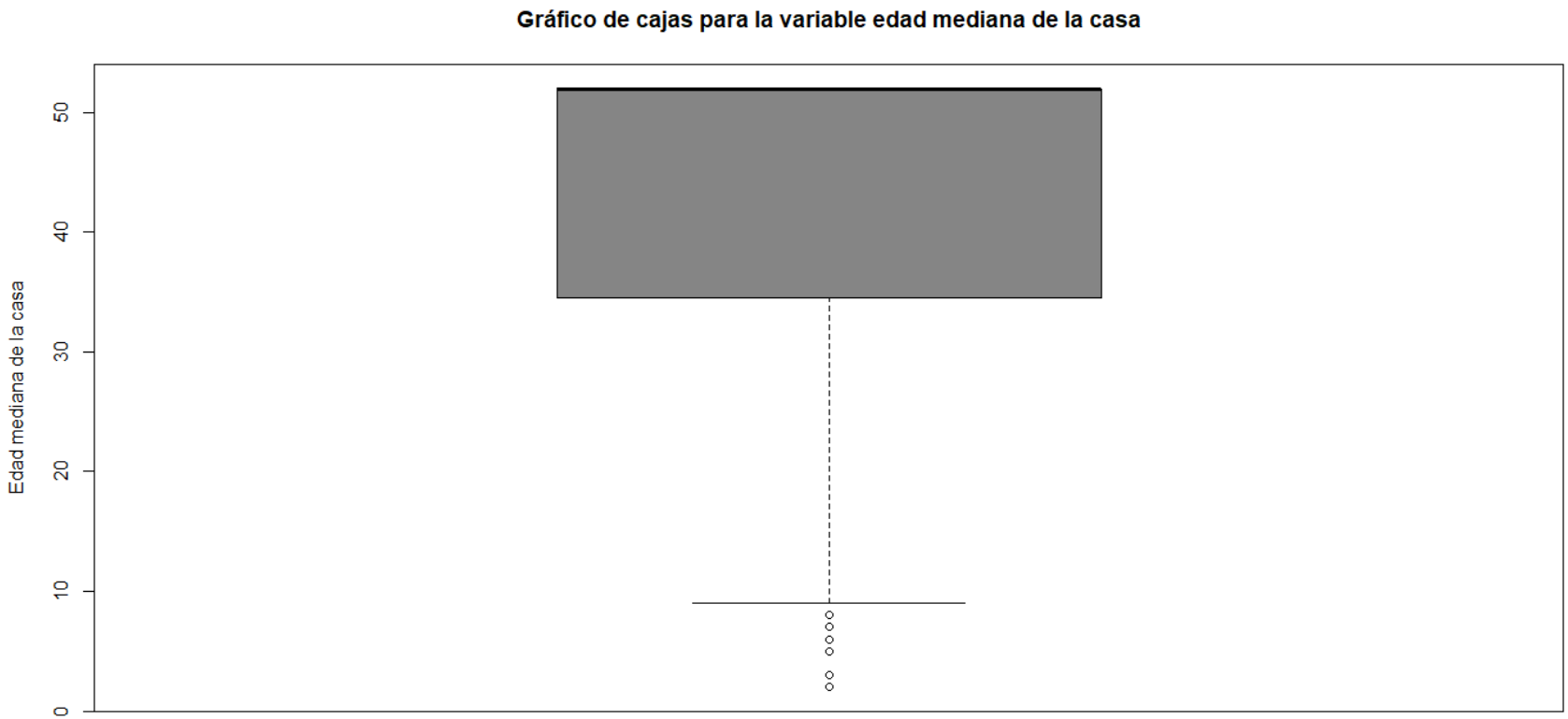


Figura: Gráfico de cajas para la variable 'Edad mediana de las viviendas'

# Variable 'Total de habitaciones'

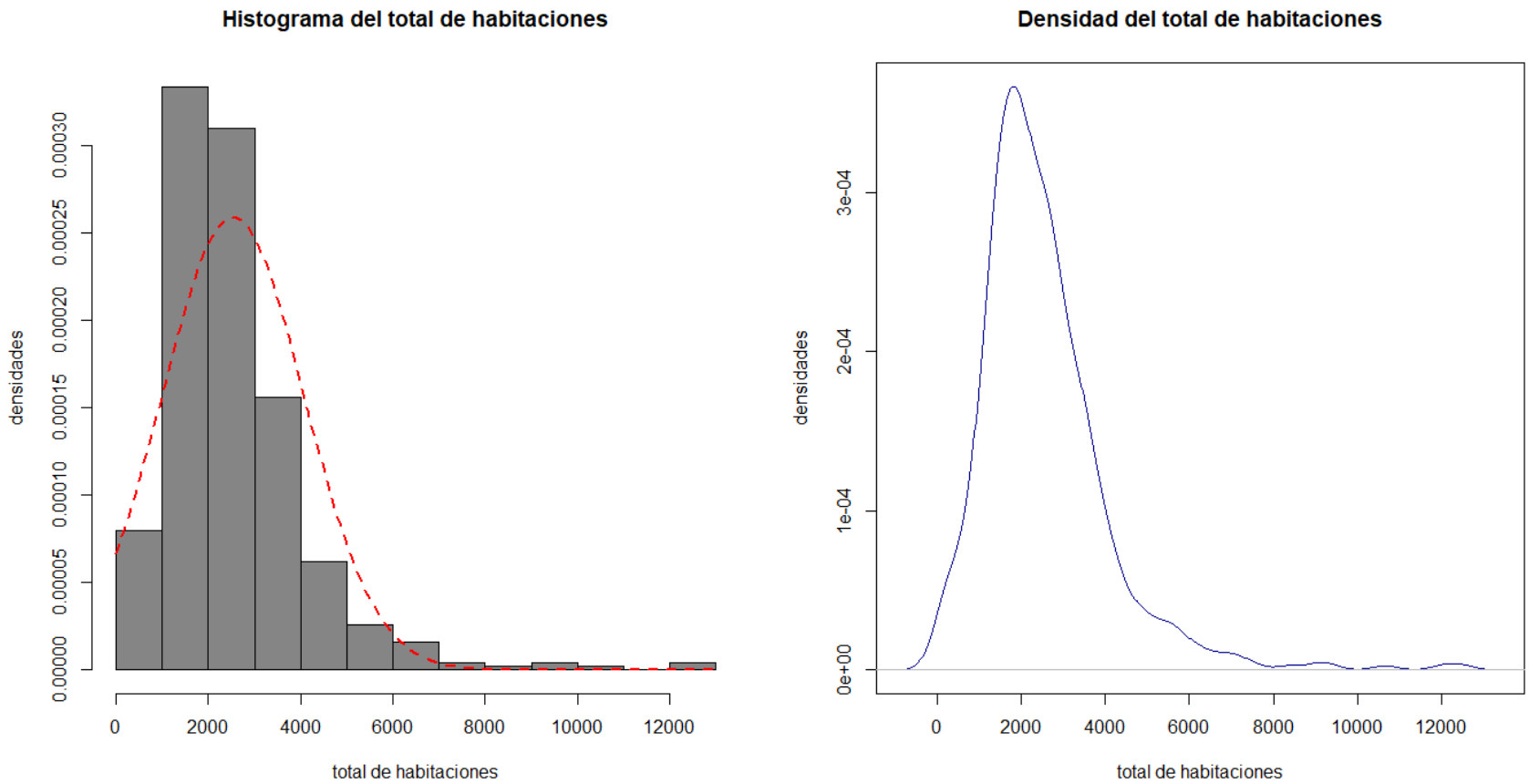


Figura: Histograma y densidad de la variable 'Total de habitaciones'

# Variable 'Total de habitaciones'

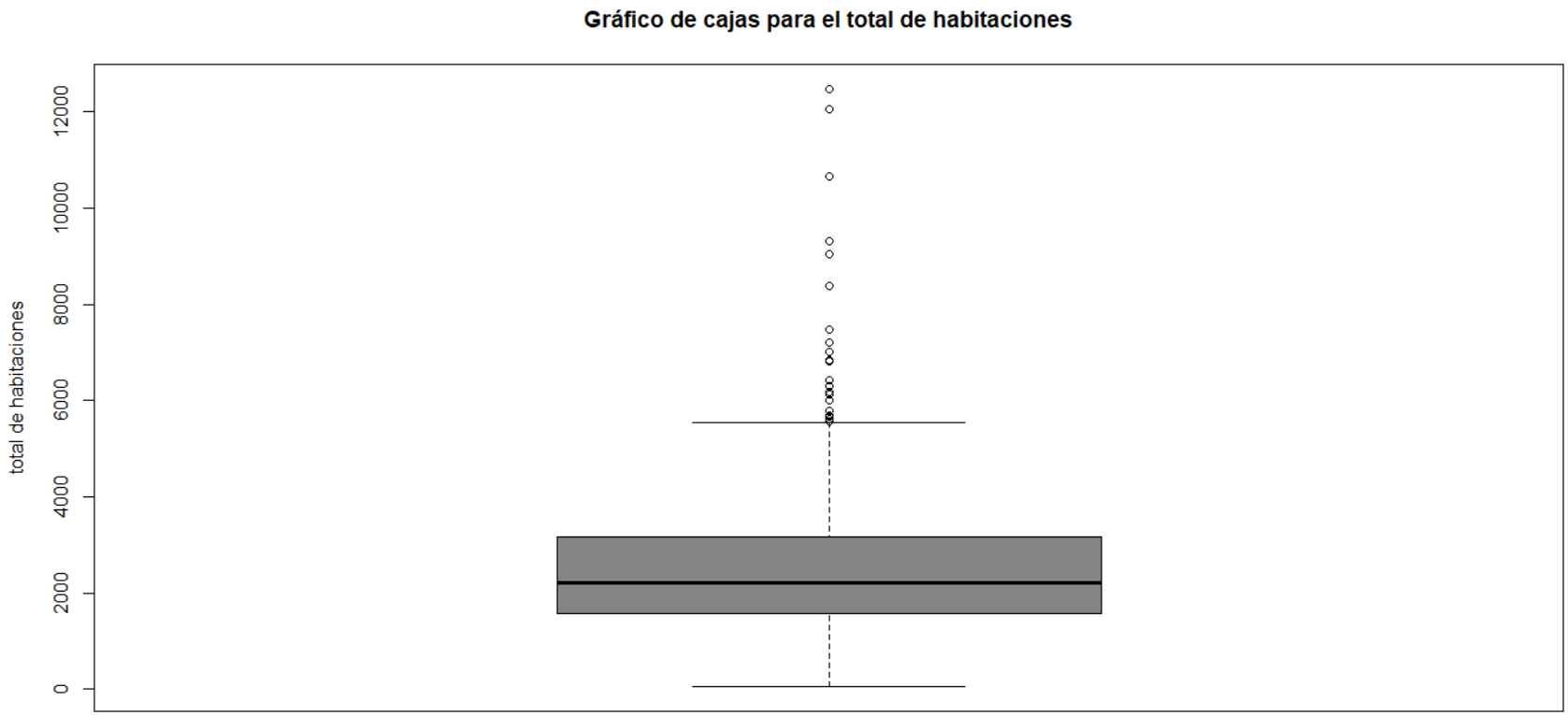


Figura: Gráfico de cajas para la variable 'Total de habitaciones'

# Variable 'Total de dormitorios'

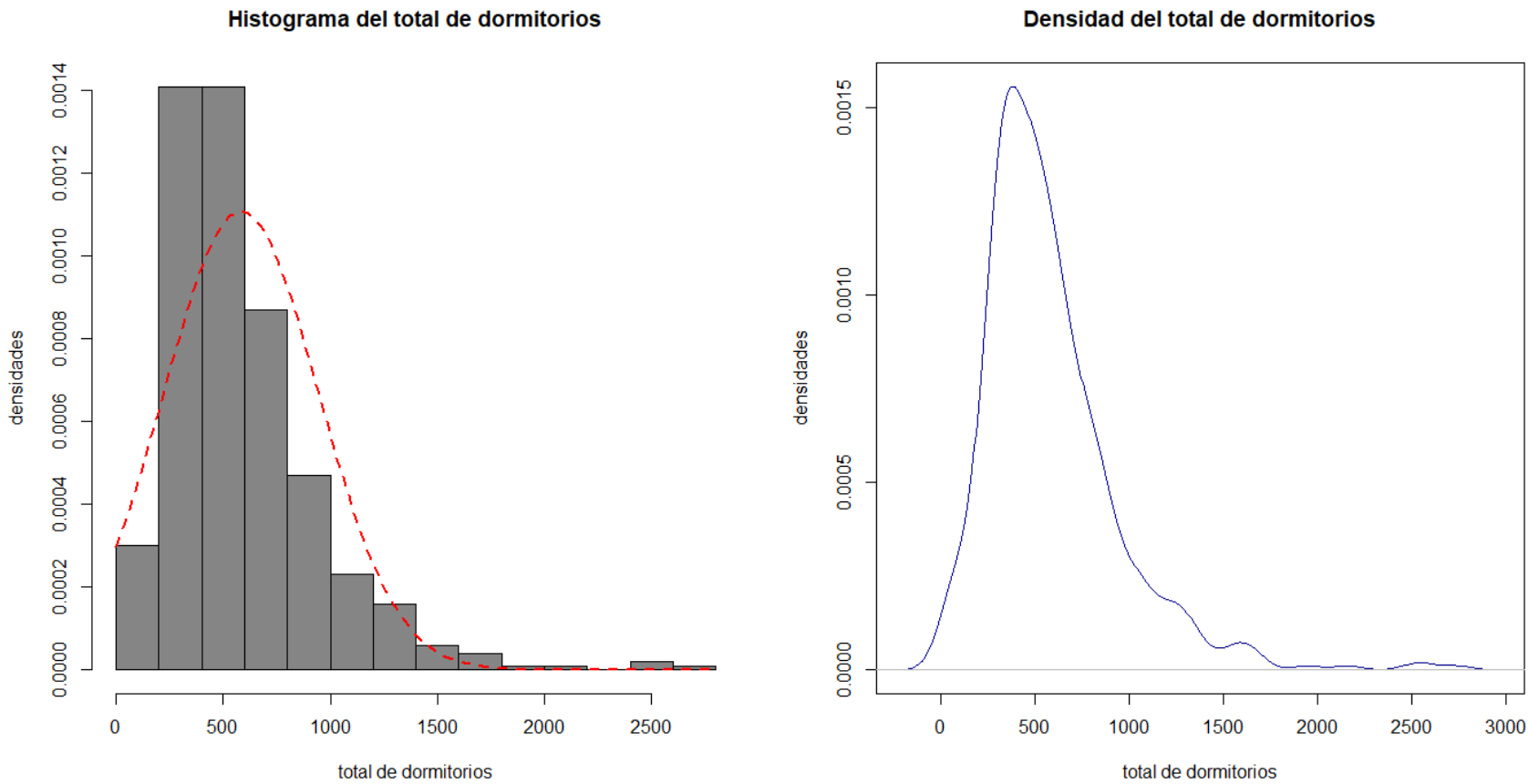


Figura: Histograma y densidad de la variable 'Total de dormitorios'

# Variable 'Total de dormitorios'

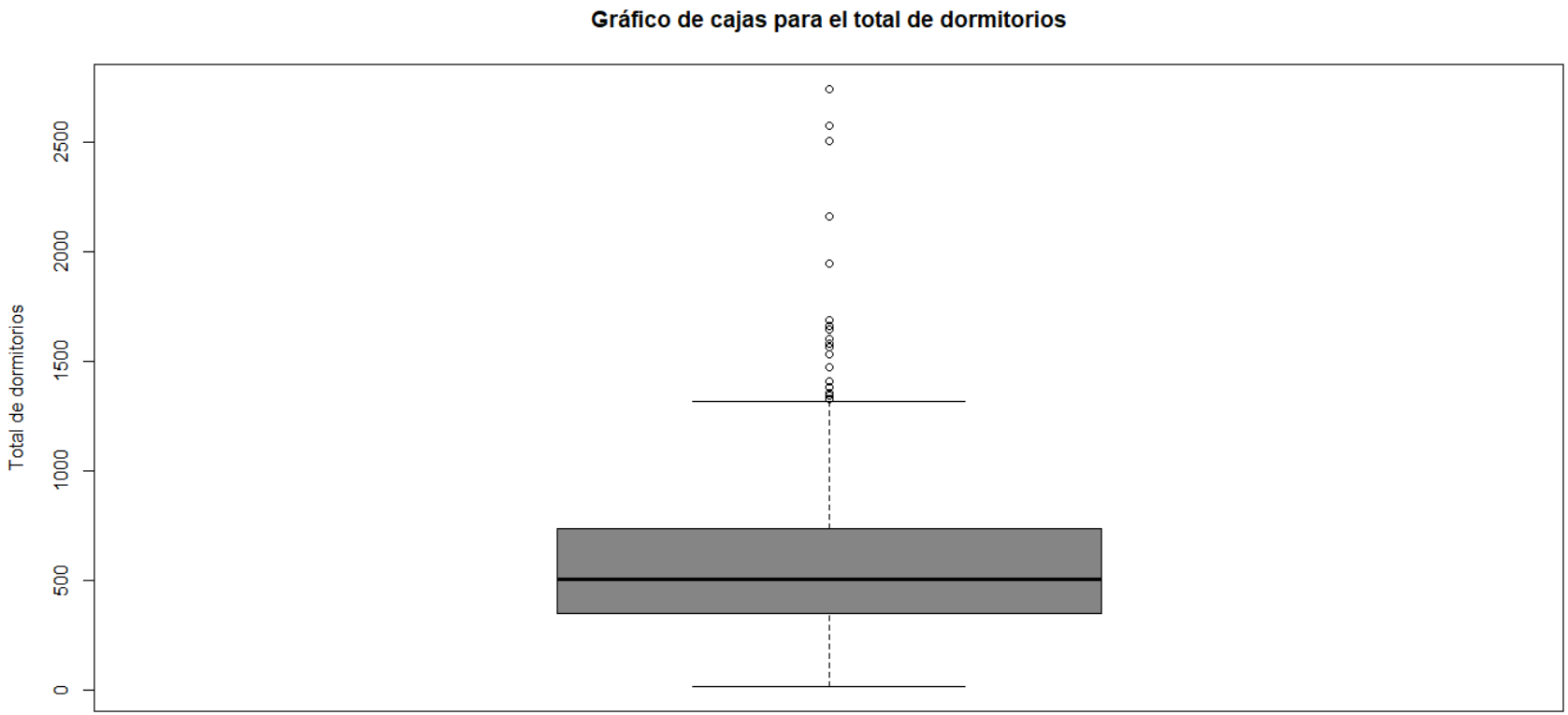


Figura: Gráfico de cajas para la variable 'Total de dormitorios'

# Variable 'Población'

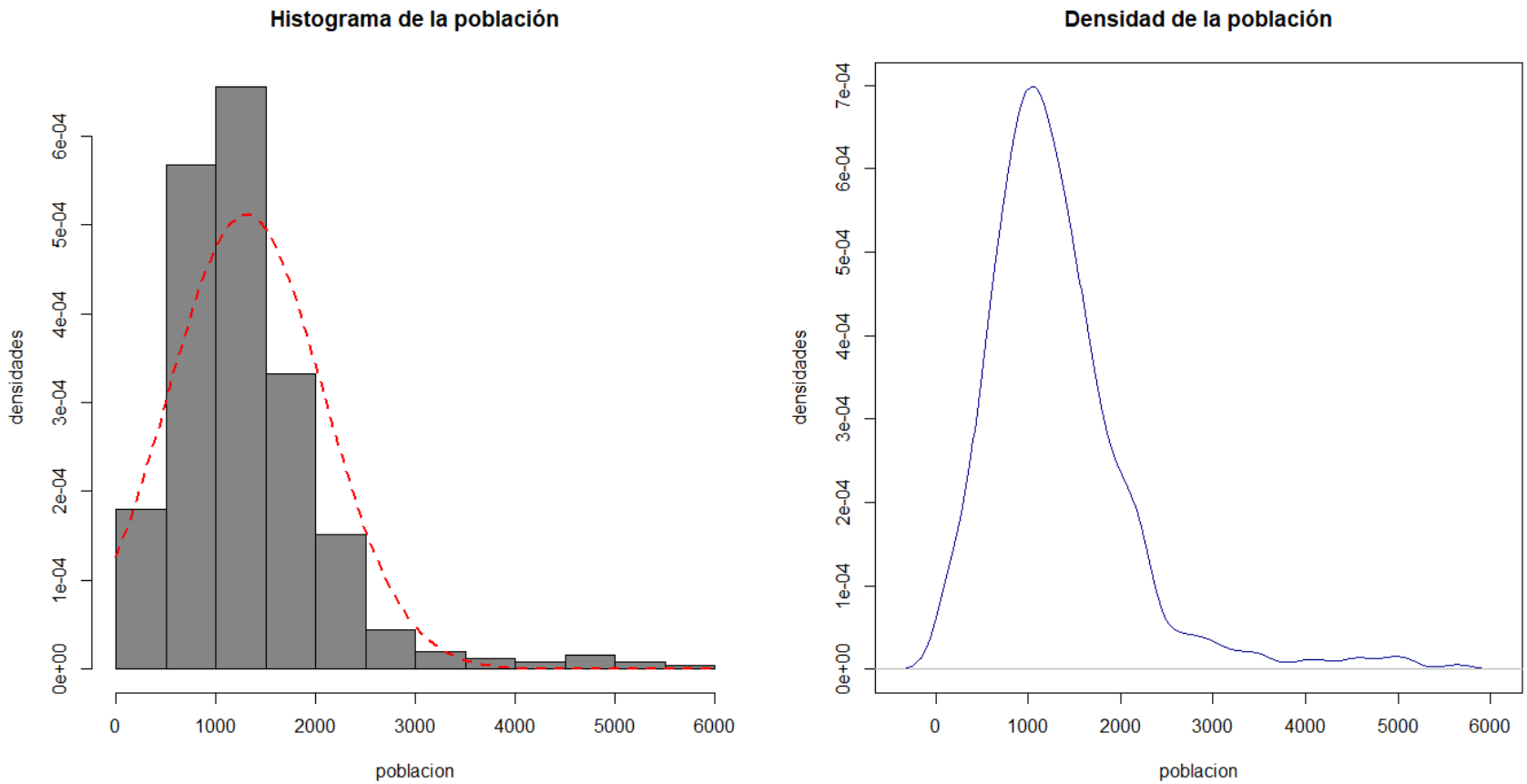


Figura: Histograma y densidad de la variable 'Población'

# Variable 'Población'

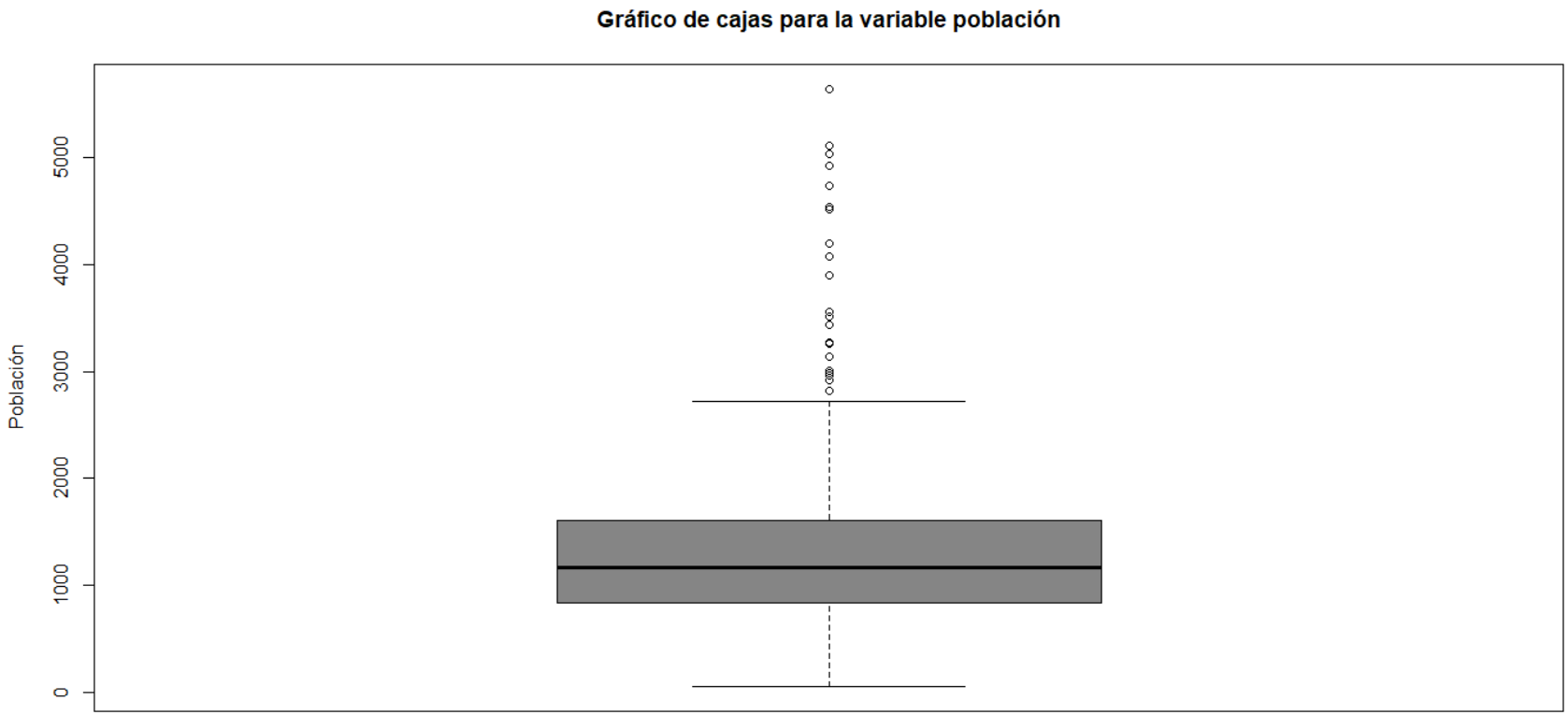


Figura: Gráfico de cajas para la variable 'Población'

# Variable 'Hogares'

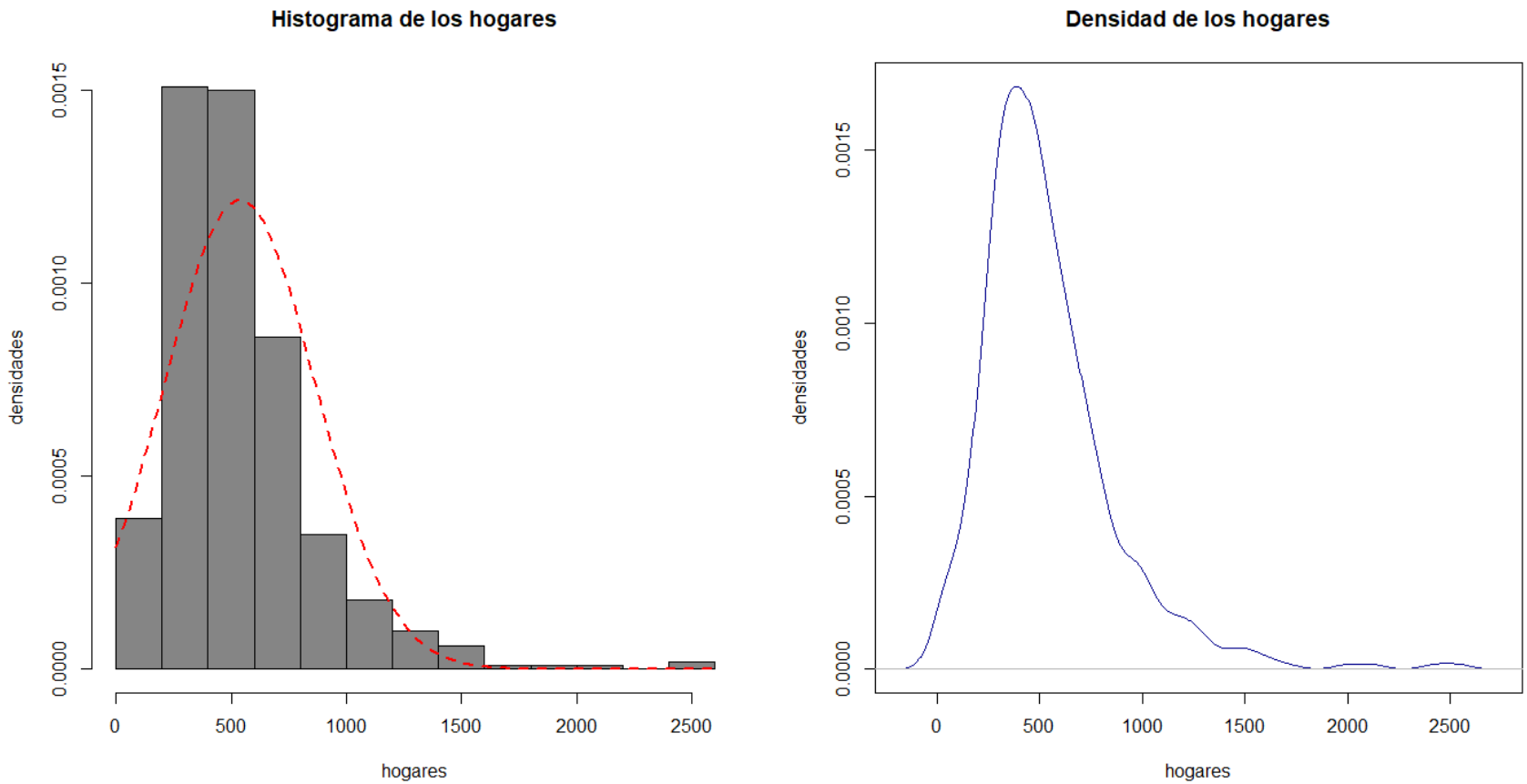


Figura: Histograma y densidad de la variable 'Hogares'

# Variable 'Hogares'

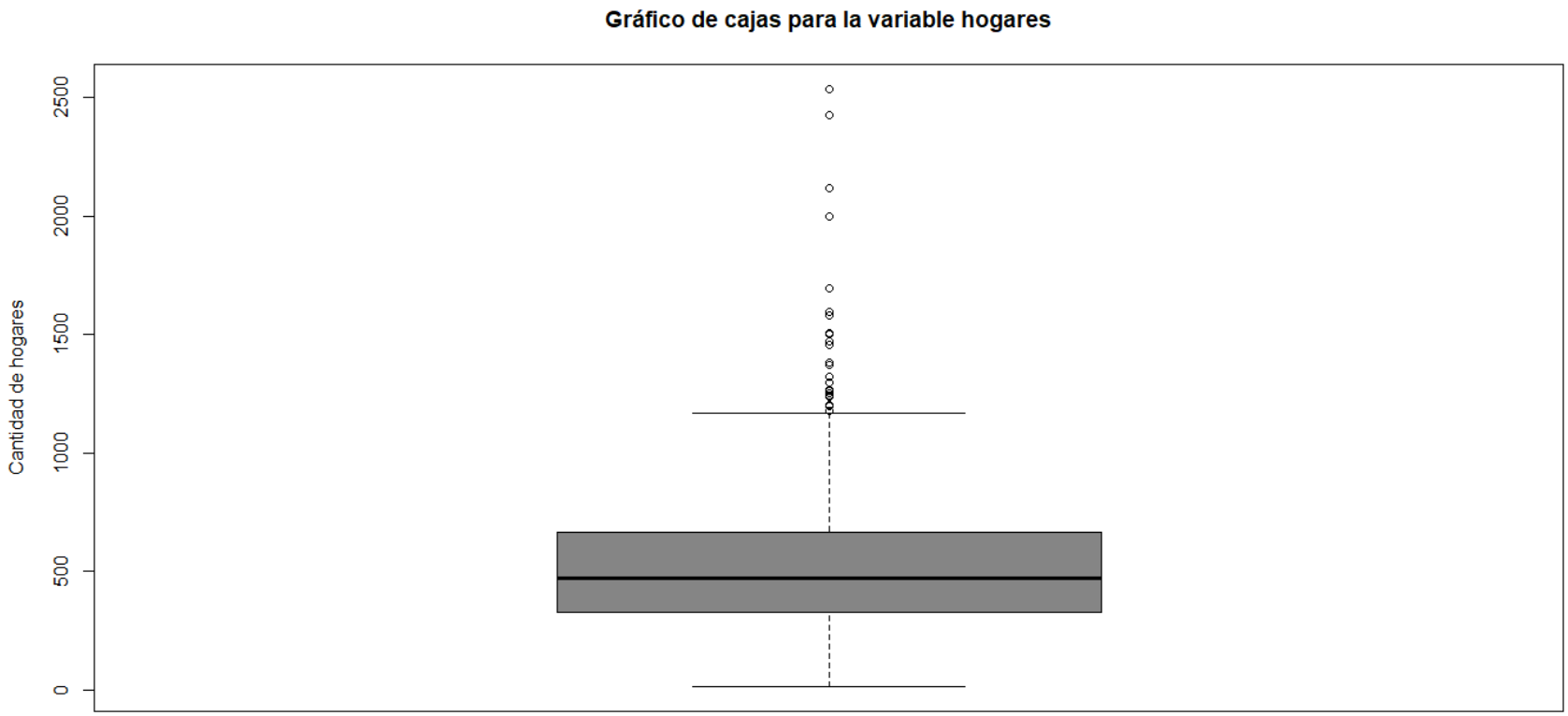


Figura: Gráfico de cajas para la variable 'Hogares'

# California


Figura: Lugar de donde provienen los datos

# San Francisco

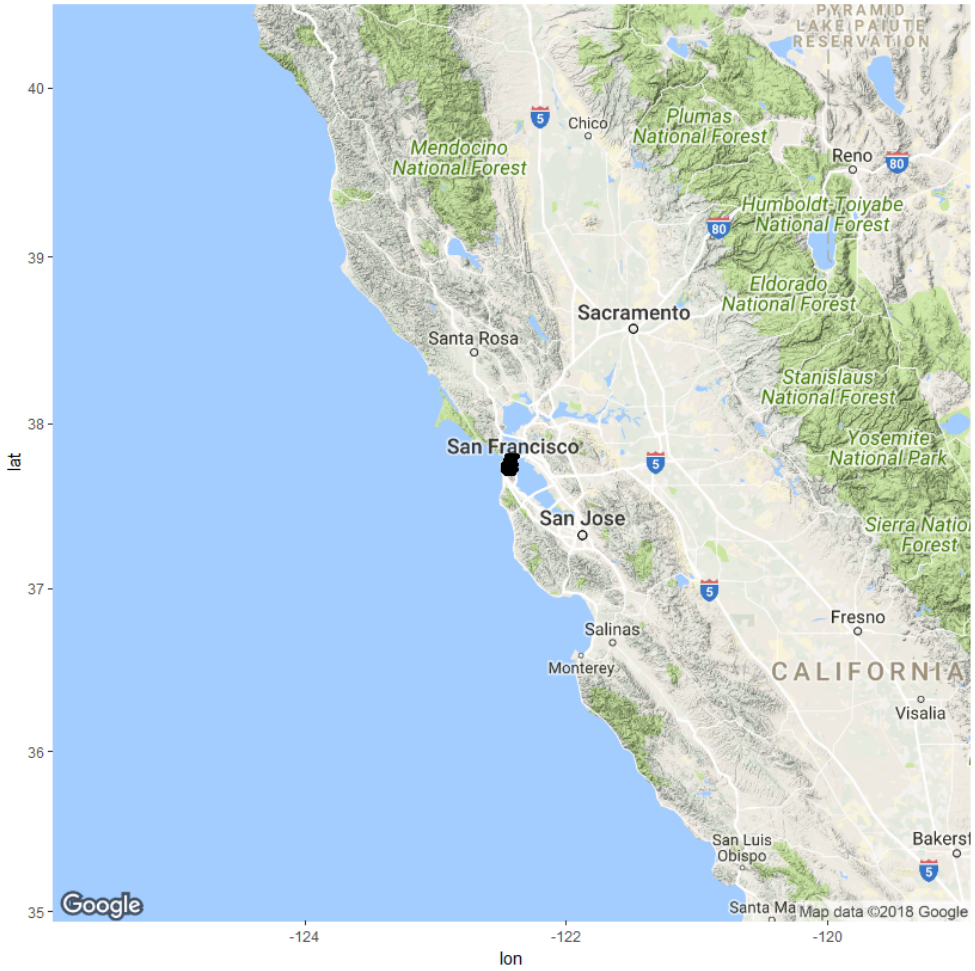


Figura: Mapa con los puntos dados de longitud y latitud

# San Diego

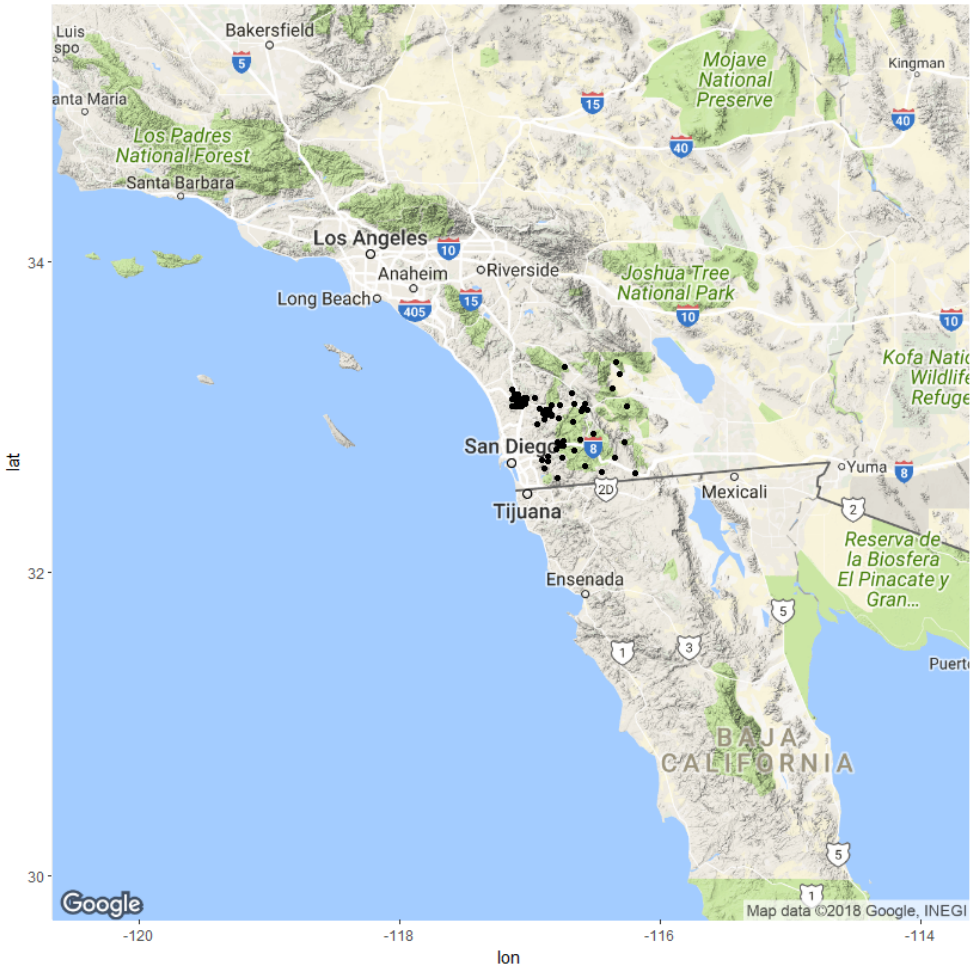


Figura: Mapa con los puntos dados de longitud y latitud

# Posibles relaciones entre las variables explicativas

- Matriz de correlaciones:

| | Ingreso | Edad | Habitaciones | Dormitorios | Población | Hogares |
|---|---|---|---|---|---|---|
| **Ingreso** | 1.00000000 | 0.01725444 | 0.07902878 | -0.1813679 | -0.1625779 | -0.1604979 |
| **Edad** | 0.01725444 | 1.00000000 | -0.31169821 | -0.1702448 | -0.2651765 | -0.1445663 |
| **Habitaciones** | 0.07902878 | -0.31169821 | 1.00000000 | 0.8620727 | 0.8576863 | 0.8652765 |
| **Dormitorios** | -0.18136789 | -0.17024480 | 0.86207267 | 1.0000000 | 0.8340611 | 0.9887048 |
| **Población** | -0.16257787 | -0.26517653 | 0.85768626 | 0.8340611 | 1.0000000 | 0.8539180 |
| **Hogares** | -0.16049785 | -0.14456632 | 0.86527649 | 0.9887048 | 0.8539180 | 1.0000000 |

Figura: Matriz de correlaciones entre covariables

# Posibles relaciones entre las variables explicativas

- Variables 'Total de dormitorios' y 'Hogares':

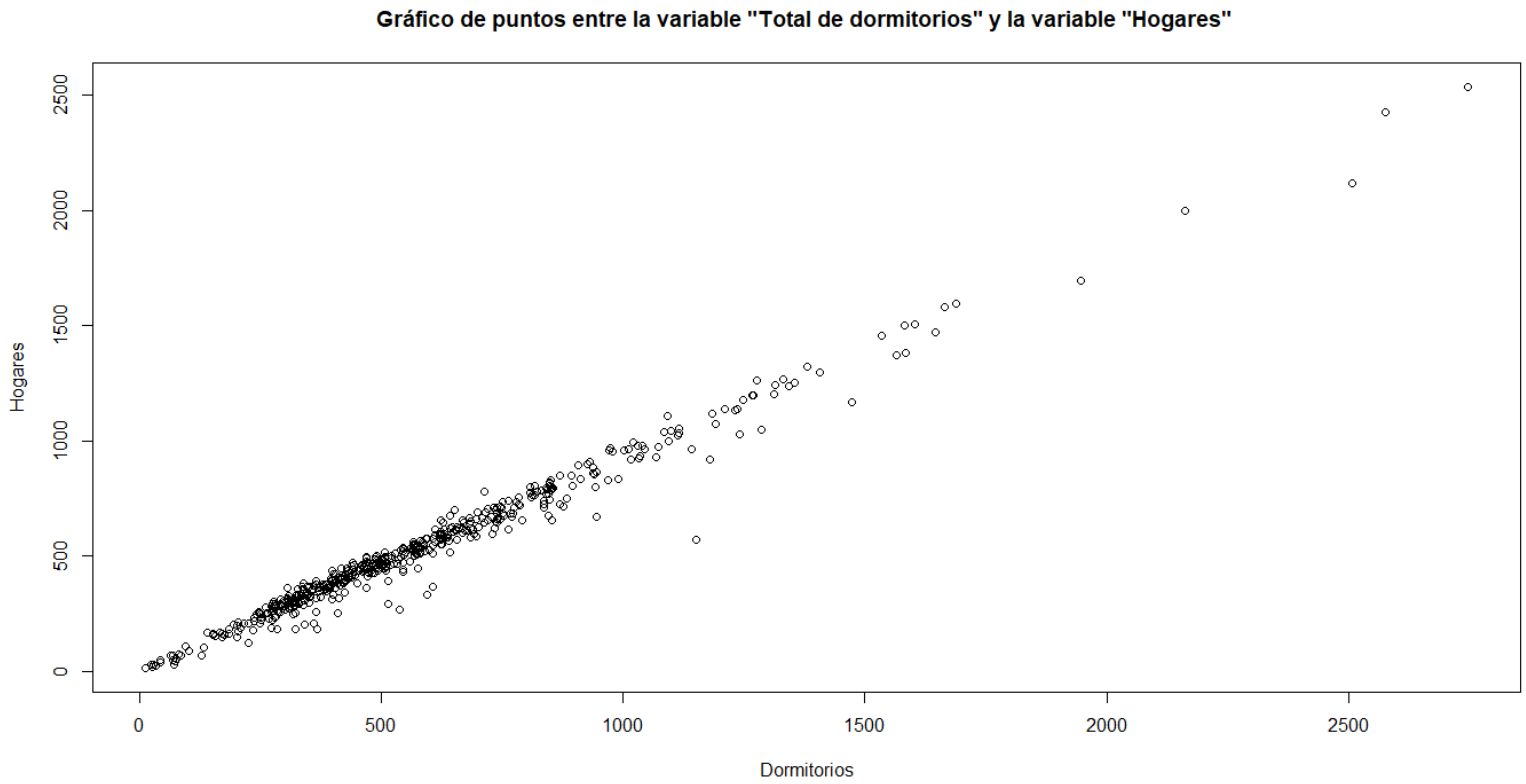


Figura: Gráfico de puntos entre las variables 'Total de dormitorios' y 'Hogares'

# Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}9887048$
- correlación de Spearman: $\rho = 0{,}9819133$

# Posibles relaciones entre las variables explicativas

- Variables 'Población' y 'Hogares':

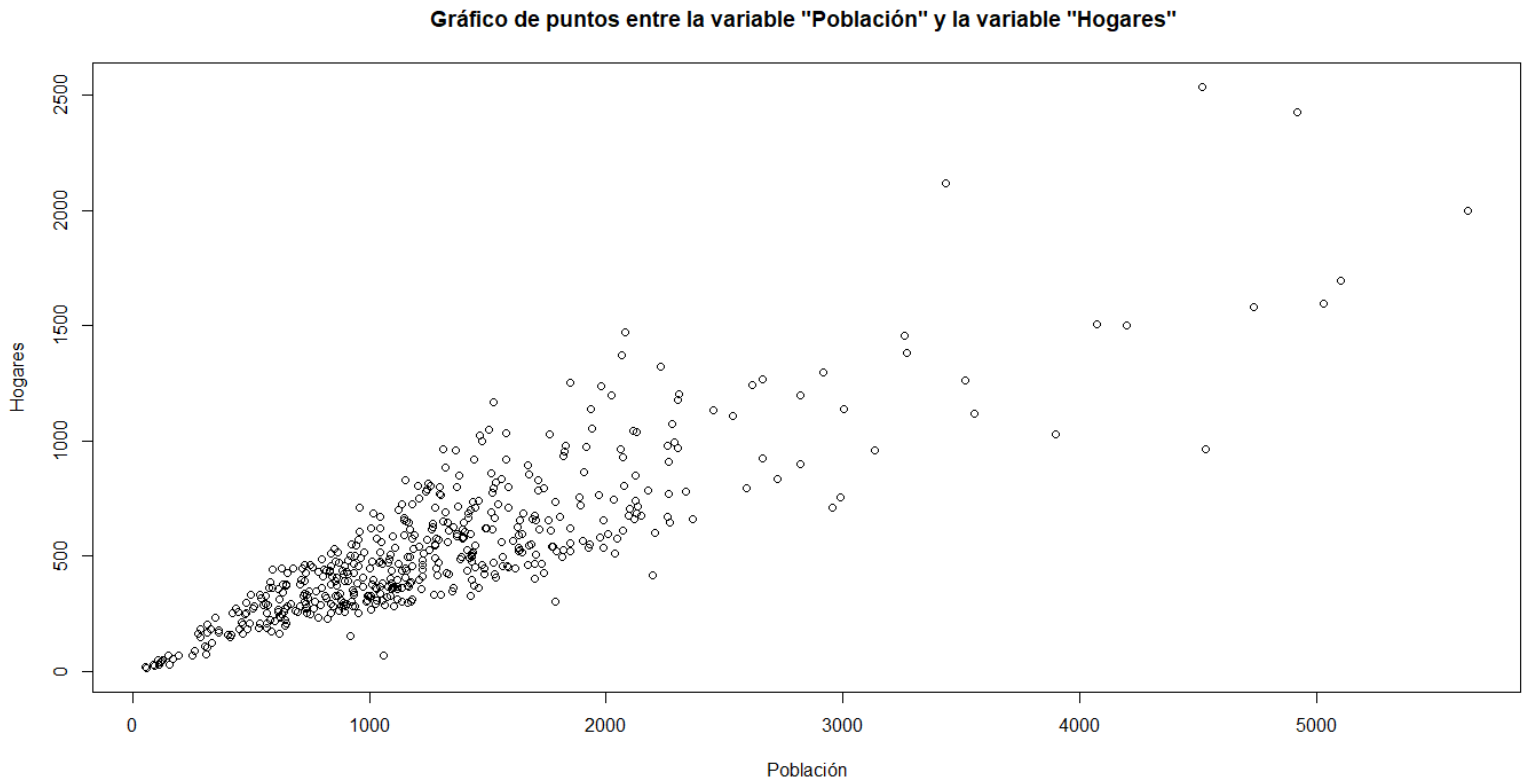


Figura: Gráfico de puntos entre las variables 'Población' y 'Hogares'

# Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}853918$
- correlación de Spearman: $\rho = 0{,}8395669$

# Posibles relaciones entre las variables explicativas

- Variables 'total de habitaciones' y 'total de dormitorios':

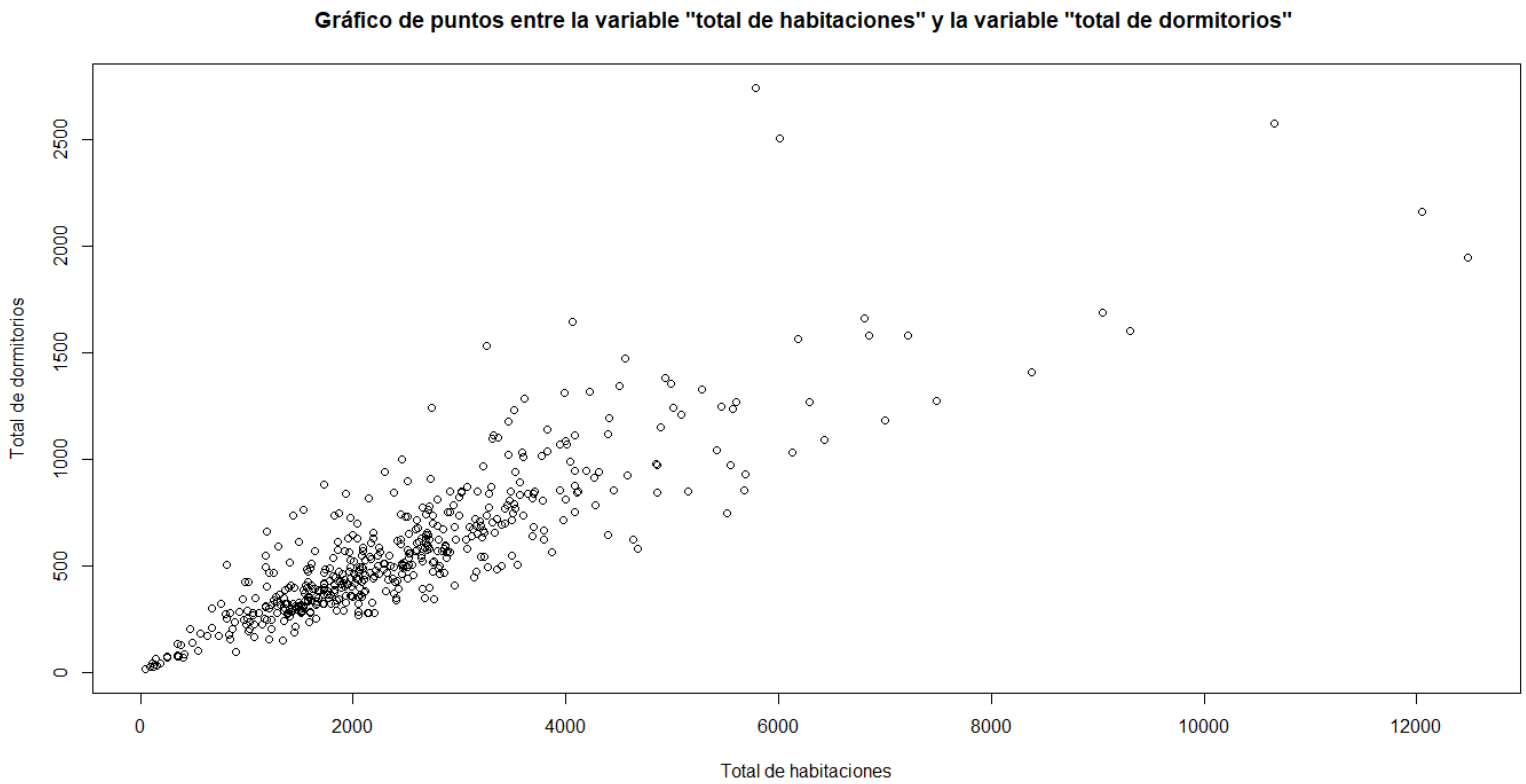


Figura: Gráfico de puntos entre las variables 'total de habitaciones' y 'total de dormitorios'

# Posibles relaciones entre las variables explicativas

- correlación de Pearson: $r = 0{,}862072$
- correlación de Spearman: $\rho = 0{,}8716658$

# Modelo ajustado e interpretación

El modelo ajustado incluyendo todas las variables sin transformación y sin selección de variables es:
$Y = 57720{,}52 + 24261{,}20X_1 + 3443{,}94X_2 + 19{,}09X_3 - 67{,}72X_4 - 121{,}66X_5 + 315{,}92X_6$

Donde: Y=Valor mediano de la casa, $X_1$ =Ingreso mediano, $X_2$ =Edad mediana de la vivienda, $X_3$ =Total de habitaciones, $X_4$ =Total de dormitorios, $X_5$ =Población, $X_6$ =Hogares.

- $R^2 = 0{,}5425$
- $R^2_{ajustado} = 0{,}537$
- $CME = \hat{\sigma^2} = 6724487923$

# Modelo ajustado e interpretación

- $\beta_0$:Representa el valor del intercepto con el eje y, esto quiere decir que Y va a partir de un valor de 57720.52 sin importar en cuantas unidades aumente mis otras variables.
- $\beta_1$:por cada unidad que aumente de ingreso mediano se aumentan 24261.20 unidades del valor mediano de la casa.
- $\beta_2$:por cada unidad que aumente de edad mediana de la vivienda se aumentan 3443.94 unidades del valor mediano de la casa.
- $\beta_3$:por cada unidad que aumente de total de habitaciones se aumentan 19.09 unidades del valor mediano de la casa.
- $\beta_4$:por cada unidad que aumente de total de dormitorios se disminuye 67.72 unidades del valor mediano de la casa.

- $\beta_5$:por cada unidad que aumente de población se disminuye 121.66 unidades del valor mediano de la casa.
- $\beta_6$:por cada unidad que aumente de hogares se aumenta 315.92 unidades del valor mediano de la casa.

# Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método forward para seleccionar variables es exactamente el mismo modelo completo, es decir, el método no me elimino ninguna variable.

# Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método backward para seleccionar variables es:
$Y = 52921{,}688 + 24923{,}244X_1 + 3484{,}164X_2 + 17{,}588X_3 - 118{,}652X_5 + 243{,}265X_6$

Donde: Y=Valor mediano de la casa, $X_1$ =Ingreso mediano, $X_2$ =Edad mediana de la vivienda, $X_3$ =Total de habitaciones, $X_5$ =Población, $X_6$ =Hogares.

- $R^2 = 0{,}5417$
- $R^2_{ajustado} = 0{,}5371$
- $CME = \hat{\sigma^2} = 6722361839$

# Modelo ajustado con selección de variables

El modelo ajustado, utilizando el método stepwise para seleccionar variables es exactamente el mismo modelo que ajustamos por el método anterior (backward), es decir, ambos métodos nos eliminan la variable $X_4$ =Total de dormitorios del modelo completo.

# Comparación

| | Completo | Forward | Backward | Stepwise |
|---|---|---|---|---|
| R^2 ajust. | 0.537 | 0.537 | 0.5371 | 0.5371 |
| CME | 6724487923 | 6724487923 | 6722361839 | 6722361839 |

Figura: Comparación de los modelos generados con selección de variables

# Conclusión

Al finalizar nuestra selección de variables, con el fin de ajustar el mejor modelo posible para la variable valor mediano de la casa, comparamos cada uno de los 4 modelos obtenidos (completo, forward, backward, stepwise) con respecto al $R^2_{ajustado}$ y el $CME = \hat{\sigma^2}$. Podemos concluir que el mejor modelo que logramos obtener para nuestros 500 datos sin hacer transformación de variables, fue el generado por el método de selección 'Backward' y 'Stepwise', los cuales nos eliminaron la variable explicativa $X_4$ :Total de dormitorios.