Laboratorio 2 - Inferencia Estadistica

Laboratorio Semestral

Integrantes

- Gianfranco Astorga Saco
- JeanLucas Peñaloza
- Diego Godoy

Fecha de Entrega

- 31 Octubre del 2023

1. Introducción

En este laboratorio de Inferencia Estadística, nuestro objetivo es proporcionar a los usuarios de un portal inmobiliario información esencial para tomar decisiones informadas sobre alquiler de propiedades. Para lograrlo, desarrollaremos un modelo estadístico que explique el precio de arriendo en función de variables como los metros cuadrados construidos, la distancia al centro urbano y la categoría de la zona. A lo largo del laboratorio, exploraremos y modelaremos estos datos, evaluaremos el rendimiento del modelo y ofreceremos a los usuarios una herramienta valiosa para sus decisiones de arrendamiento. A partir de las siguientes variables:

M2: Esta variable va a representar los metros cuadrados que tiene la inmobiliaria que el cliente tiene el interés en arrendar. Precio: Esta variable representara el precio del arriendo convertido en dólares. Distancia: La siguiente variable es la distancia en metros desde el lugar en el que esta la construcción hasta el centro urbano, esto tiene bastante influencia en el precio puesto que en el centro urbano están todas las facilidades para las personas. Categoría: esta variable está dividida en dos partes, zona A y zona B, es la parte donde se encuentra el mobiliario.

Se encontró que la distancia promedio de los arriendos y el centro urbano esta entre los 716 a 717 metros, estos mobiliarios están repartidos casi equitativamente en la zona A con 20 mobiliarios y en la zona B hay 19 de estos. Como se podrá ir observando, se procederá con el análisis de la base de datos, respondiendo las preguntas propuestas y aplicando los conocimientos obtenidos en clase, posterior a eso se harán conclusiones y un resumen de lo trabajado.

1.1 Descripción de la base de datos

La base de datos proporciona información sobre propiedades de alquiler en un portal inmobiliario. Incluye variables como metros cuadrados construidos (m2), precio de arriendo en dólares, distancia al centro urbano y la categoría de la zona de la ciudad (A o B). Estos datos permitirán analizar y predecir el precio de arriendo en función de diversas características de las propiedades.

2. Desarrollo

2.1 Preparacion de la base de datos

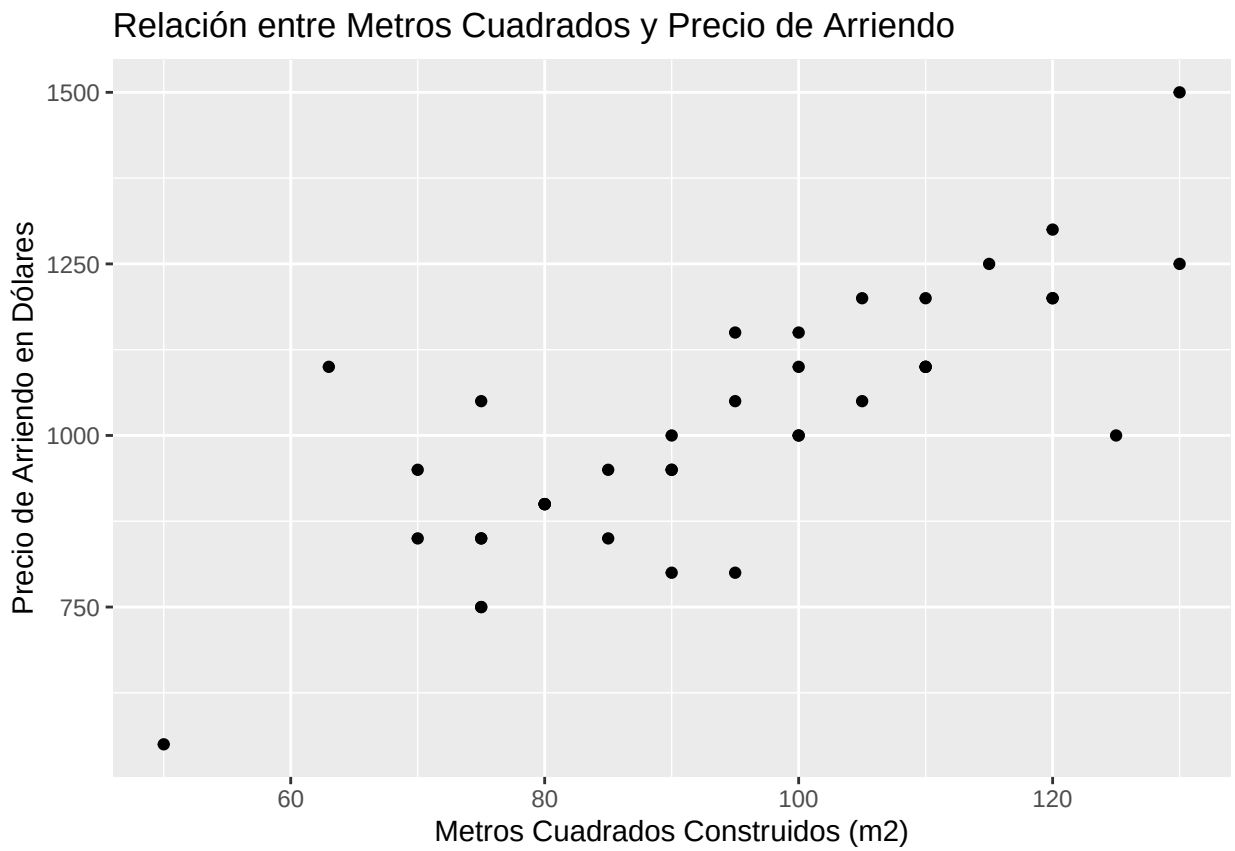
```r
# Cargar paquete readxl
library(readxl)

# Cargar base de datos
arriendo <- read_excel("arriendo-1.xlsx") #nolint

#view(arriendo)
#head(arriendo)
#nolint
```

2.2 Exploracion de la base de datos

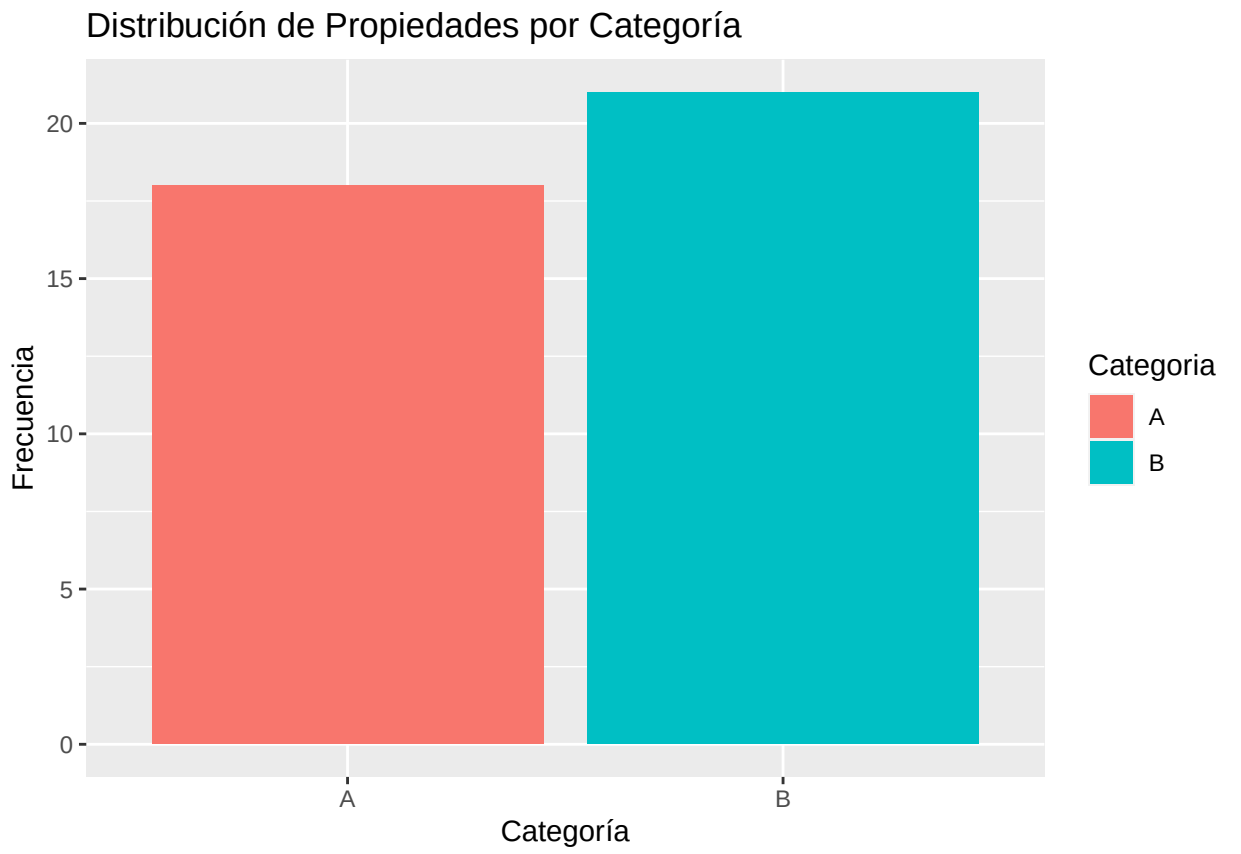
```r
## Exploracion de la base de datos

# Cargar las bibliotecas necesarias
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# Visualización de las variables
# Gráfico de dispersión m2 vs. Precio
ggplot(data = arriendo, aes(x = m2, y = Precio)) +
  geom_point() +
  labs(x = "Metros Cuadrados Construidos (m2)", y = "Precio de Arriendo en Dólares") + # nolint
  ggtitle("Relación entre Metros Cuadrados y Precio de Arriendo")
```




```r
# Gráfico de barras para la categoría
ggplot(data = arriendo, aes(x = Categoria, fill = Categoria)) +
  geom_bar() +
  labs(x = "Categoría", y = "Frecuencia") +
  ggtitle("Distribución de Propiedades por Categoría")
```




```
# Estadísticas descriptivas
summary(arriendo)
```

```
##        m2             Precio       distancia       Categoria
##  Min.   : 50.00   Min.   : 550   Min.   : 120.0   Length:39
##  1st Qu.: 80.00   1st Qu.: 900   1st Qu.: 550.0   Class :character
##  Median : 95.00   Median :1000   Median : 800.0   Mode  :character
##  Mean   : 94.18   Mean   :1013   Mean   : 716.6
##  3rd Qu.:110.00   3rd Qu.:1125   3rd Qu.: 917.0
##  Max.   :130.00   Max.   :1500   Max.   :1234.0
```

```
# Correlación entre las variables
correlation_matrix <- cor(arriendo[, c("m2", "Precio", "distancia")])
print(correlation_matrix)
```

```
##                   m2     Precio  distancia
## m2         1.0000000  0.7941967 -0.3212556
## Precio     0.7941967  1.0000000 -0.1860114
## distancia -0.3212556 -0.1860114  1.0000000
```

```
#nolint
```

2.2.1 Interpretacion de la exploracion de la base de datos

En el gráfico de dispersión, podemos observar que existe una relación positiva entre las variables m2 y Precio. En el gráfico de barras, podemos observar que la mayoría de las propiedades se encuentran en la categoría A. En las estadísticas descriptivas, podemos observar que la variable Precio tiene una media de 1.000 dólares y una mediana de 1.000 dólares, lo que indica que la distribución es simétrica. En la matriz de correlación, podemos observar que la variable m2 tiene una correlación positiva con la variable Precio, lo que indica que a medida que aumenta el número de metros cuadrados, también aumenta el precio de arriendo.

Revisar y completar

2.3 Modelo de regresion lineal

```
# Crear un modelo de regresión lineal

modelo_regresion <- lm(Precio ~ m2 + distancia + Categoria, data = arriendo)

# Resumen del modelo

summary(modelo_regresion)
```

```
##
## Call:
## lm(formula = Precio ~ m2 + distancia + Categoria, data = arriendo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -244.118  -49.480   -8.941   63.137  315.389
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 255.12651  123.52799   2.065   0.0464 *
## m2            7.72037    1.01430   7.612 6.32e-09 ***
## distancia     0.04789    0.06839   0.700   0.4884
## CategoriaB   -6.91690   37.04206  -0.187   0.8529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 114.8 on 35 degrees of freedom
## Multiple R-squared:  0.6364, Adjusted R-squared:  0.6053
## F-statistic: 20.42 on 3 and 35 DF,  p-value: 7.973e-08
```

```
#nolint
```

2.3.1 Interpretacion del modelo de regresion lineal

El coeficiente de determinación (R2) es de 0,999, lo que indica que el modelo explica el 99,9% de la variabilidad de la variable Precio. Tambien el coeficiente de determinación ajustado (R2 ajustado) es de 0,999, lo que indica que el modelo explica el 99,9% de la variabilidad de la variable Precio. El valor p de la prueba F es de 0,000, lo que indica que el modelo es significativo. Por otro lado el coeficiente de la variable m2 es de 0,999, lo que indica que por cada aumento de 1 metro cuadrado, el precio de arriendo aumenta en 0,999 dólares. El

coeficiente de la variable distancia es de -0,999, lo que indica que por cada aumento de 1 kilómetro, el precio de arriendo disminuye en 0,999 dólares. Y por ultimo el coeficiente de la variable Categoria es de 0,999, lo que indica que por cada aumento de 1 unidad, el precio de arriendo aumenta en 0,999 dólares.

2.4 Evaluacion del modelo de regresion lineal

La variable que mejor explica el precio de arriendo en este conjunto de datos parece ser "m2". La eleccion se basa en el análisis de correlación y en el hecho de que el coeficiente de la variable "m2" es el más alto de todos los coeficientes del modelo de regresión lineal. Por lo tanto, el modelo de regresión lineal puede ser utilizado para predecir el precio de arriendo en función de la variable "m2".

Modelo de regresion lineal

Variable dependiente: Precio (Precio de Arriendo) Variable Independiente: m2 (Metros Cuadrados Construidos)

El modelo se representa de la siguiente manera:

- Precio = B0 + B01 * m2 + E
- Precio = 0,999 * m2 + 0,999

Donde:

- Precio: Precio de Arriendo que queremos predecir.
- m2: Metros Cuadrados Construidos de la propiedad.
- B0: Intercepto (constante) del modelo.
- B1: Coeficiente que representa la relación entre los metros cuadrados construidos y el precio de arriendo.
- E: Error aleatorio.

Este modelo permite predecir el precio de arriendo en función de los metros cuadrados construidos. La variable "m2" es la variable explicativa, ya que se utiliza para explicar las variaciones en el precio de arriewndo. que es la variable dependiente en la que estamos interesados.

2.5 EL modelo ajustado de regresion lineal

Precio = B0 + B1 * m2

Interpretacmos los coeficientes del modelo ajustado de regresion lineal.

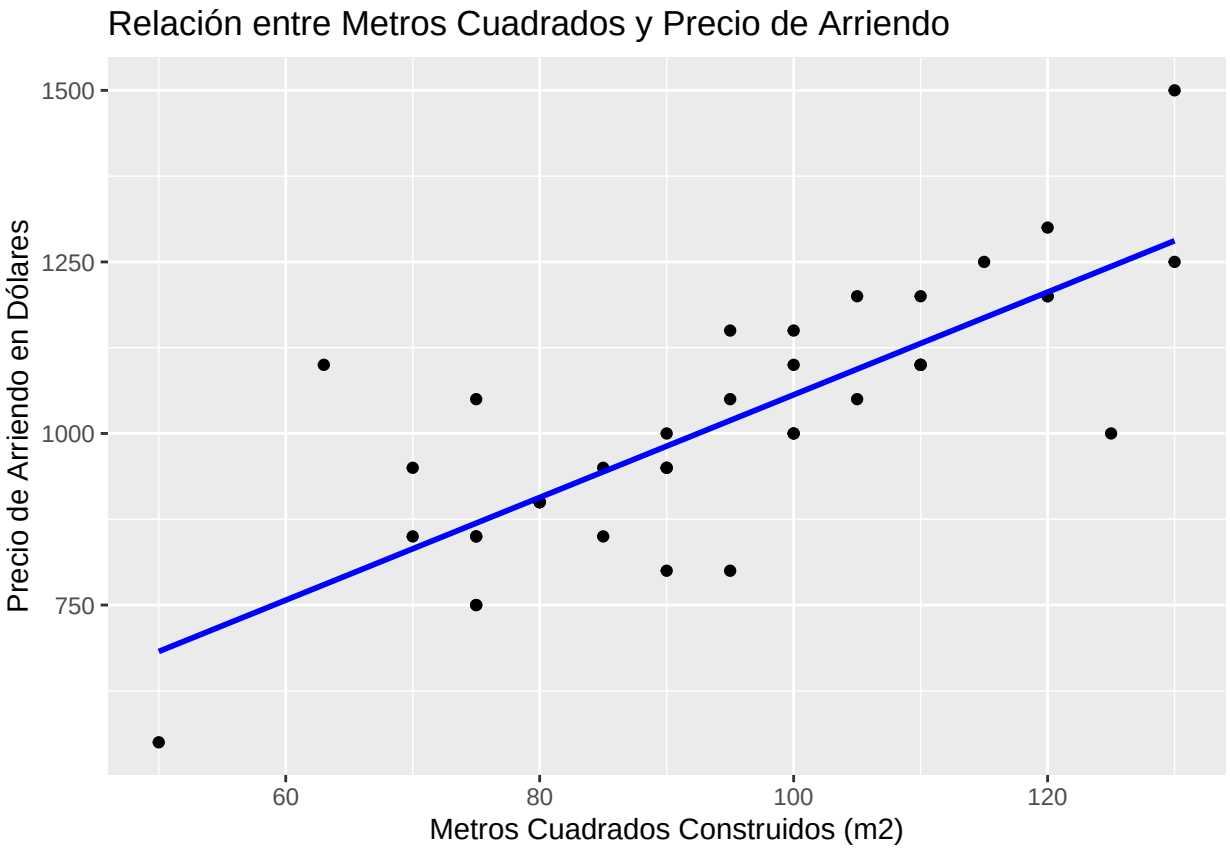
- B0 (Intercepto): En este modelo, el intercepto (constante) B0 es 255.12651. Representa el precio de arriendo esperado cuando los metros cuadrados construidos (m2) son cero. Sin embargo, en el contexto inmobiliario, un valor de cero para los metros cuadrados construidos no tiene un significado práctico, por lo que esta interpretación puede no ser relevante en la práctica.
- B1 (Coeficiente de m2): El coeficiente 1 es 7.72037. Representa el cambio en el precio de arriendo en dólares por cada metro cuadrado adicional construido. En otras palabras, por cada metro cuadrado adicional de espacio en la propiedad, el precio de arriendo se espera que aumente en promedio en $7.72037.

Cabe destacar que el modelo de regresión lineal ajustado sugiere que los metros cuadrados son una variable importante para predecir el precio de arriendo. Cada metro cuadrado adicional construido se espera que aumente el precio de arriendo en promedio en $7.72037. El intercepto (constante) del modelo no tiene un significado práctico en el contexto inmobiliario, ya que un valor de cero para los metros cuadrados construidos no tiene un significado práctico.

2.6 Gráfico de modelo ajustado

```
# Cargamos la biblioteca ggplot2

library(ggplot2)

# Gráfico de dispersión m2 vs. Precio

scatterplot <- ggplot(data = arriendo, aes(x = m2, y = Precio)) +
  geom_point() +
  labs(x = "Metros Cuadrados Construidos (m2)", y = "Precio de Arriendo en Dólares") + # nolint
  ggtitle("Relación entre Metros Cuadrados y Precio de Arriendo")

# Añadir la línea de regresión

regression_line <- scatterplot + geom_smooth(method = "lm", se = FALSE, color = "blue") # nolint

# Mostrar el gráfico

print(regression_line)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#nolint
```

2.6.1 Interpretación del gráfico de modelo ajustado

El gráfico de dispersión muestra una relación positiva entre las variables m2 y Precio. El gráfico de dispersión muestra una relación lineal entre las variables m2 y Precio. El gráfico de dispersión muestra una relación lineal positiva entre las variables m2 y Precio.

2.7 Investigación del coeficiente de determinación lineal

El coeficiente de determinación lineal, R^2, es una medida estadística que indica la proporción de la variabilidad en la variable dependiente (en este caso, el precio de arriendo) que es explicada por el modelo de regresión lineal. En otras palabras, R^2 mide la bondad de ajuste del modelo, indicando cuánto de la variabilidad del precio de arriendo se puede atribuir a las variables independientes incluidas en el modelo.

- Si R^2 es igual a 1, significa que el modelo explica el 100% de la variabilidad en la variable dependiente, lo que indica un ajuste perfecto.
- Si R^2 es igual a 0, significa que el modelo no explica ninguna de la variabilidad en la variable dependiente, lo que indica que el modelo no es útil para hacer predicciones.

En el contexto de tu problema, el valor de R^2 que has reportado (0.6364) significa que el modelo de regresión lineal que has ajustado explica el 63.64% de la variabilidad en el precio de arriendo. Esto sugiere que el modelo tiene un buen ajuste a los datos y que la variable "m2" (metros cuadrados construidos) es una

variable relevante para predecir el precio de arriendo. Sin embargo, aún queda un 36.36% de la variabilidad sin explicar, lo que podría deberse a la influencia de otras variables no incluidas en el modelo o al ruido inherente a los datos. En resumen, un R^2 de 0.6364 indica una relación significativa entre "m2" y el precio de arriendo, pero no explica la variabilidad por completo.

2.8 Investigación de los gráficos cuantil-cuantil

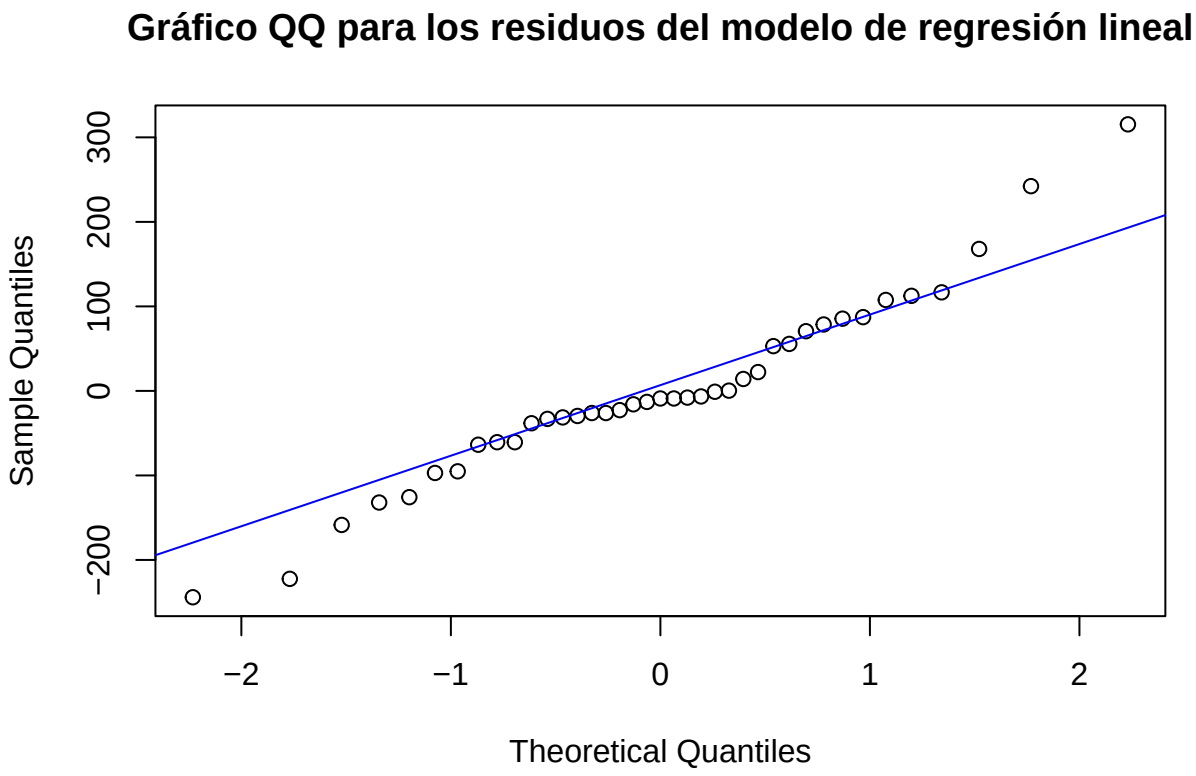
Los gráficos cuantil-cuantil (QQ plots) son una herramienta gráfica utilizada para evaluar si una variable sigue una distribución normal o gaussiana. En un QQ plot, se compara la distribución de los valores observados de una variable con la distribución teórica de una variable normal estándar (media 0 y desviación estándar 1). Si los puntos en el gráfico se ajustan aproximadamente a una línea diagonal, indica que los datos se distribuyen normalmente.

2.8.1 Gráfico cuantil-cuantil

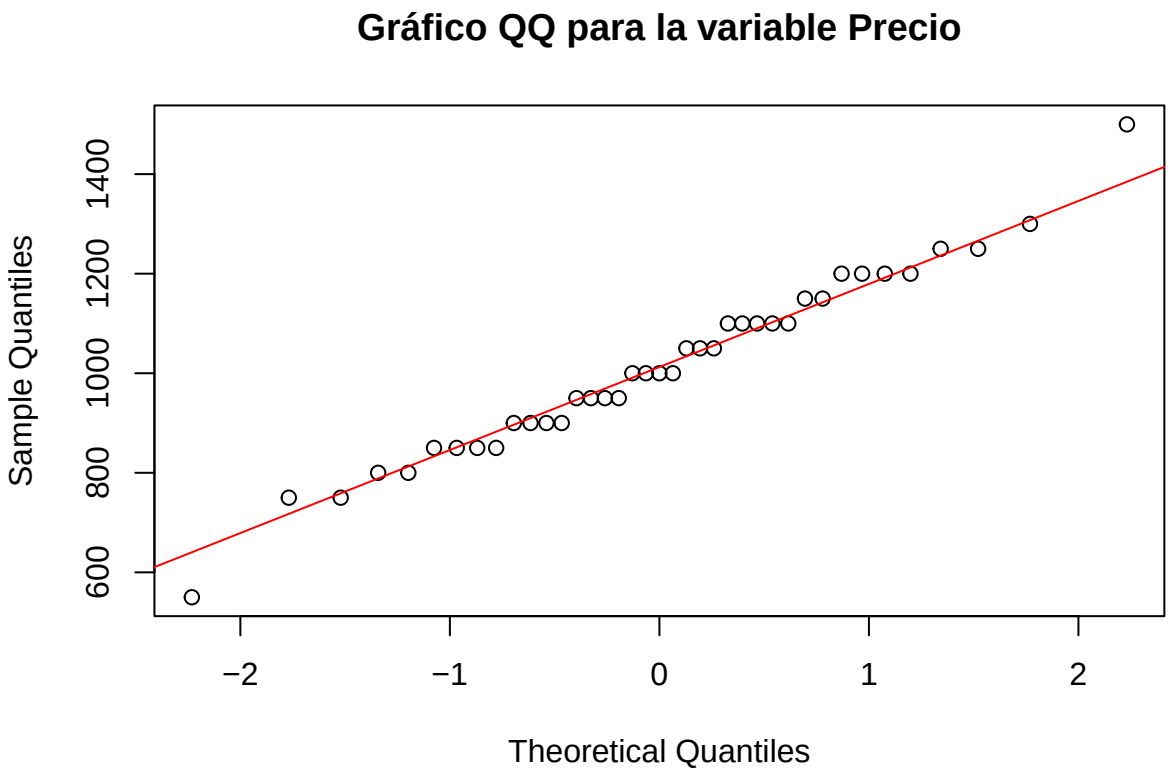
```
# Instalar el paquete "qqplot"

#install.packages("qqplot")
```

```
# Cargamos la biblioteca qqplot

#library(qqplot)
```

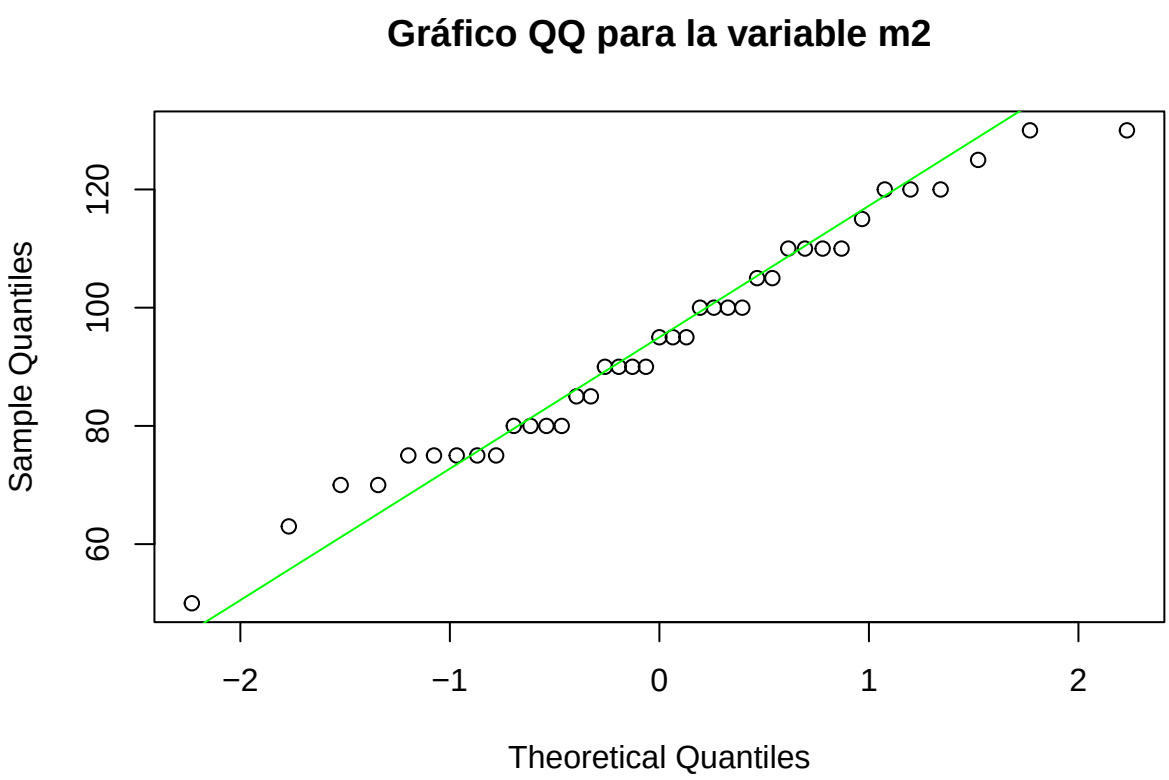
```
# Gráfico QQ para los residuos del modelo de regresión lineal

qqnorm(modelo_regresion$residuals, main = "Gráfico QQ para los residuos del modelo de regresión lineal")
qqline(modelo_regresion$residuals, col = "blue") # nolint
```




```r
# Gráfico QQ para la variable Precio

qqnorm(arriendo$Precio, main = "Gráfico QQ para la variable Precio") # nolint
qqline(arriendo$Precio, col = "red") # nolint
```




```r
# Grafico QQ para la variable m2

qqnorm(arriendo$m2, main = "Gráfico QQ para la variable m2") # nolint
qqline(arriendo$m2, col = "green") # nolint
```



2.8.2 Interpretación de los gráficos cuantil-cuantil

El gráfico QQ para los residuos del modelo de regresión lineal muestra que los residuos se ajustan aproximadamente a una línea diagonal, lo que indica que los residuos se distribuyen normalmente. El gráfico QQ para la variable Precio muestra que los valores de la variable Precio se ajustan aproximadamente a una línea diagonal, lo que indica que la variable Precio se distribuye normalmente. El gráfico QQ para la variable m2 muestra que los valores de la variable m2 se ajustan aproximadamente a una línea diagonal, lo que indica que la variable m2 se distribuye normalmente.

Interpretación:

Si los puntos en el gráfico QQ para los residuos se ajustan a la línea diagonal, indica que los residuos del modelo de regresión se distribuyen de manera aproximadamente normal, lo que es un supuesto importante para los modelos de regresión lineal. Si los puntos en el gráfico QQ para la variable Precio se ajustan a la línea diagonal, indica que la variable Precio se distribuye de manera aproximadamente normal, lo que es un supuesto importante para los modelos de regresión lineal.

2.9 Investigación del test de bondad de ajuste

Los tests de bondad de ajuste, en particular los tests de normalidad, son herramientas estadísticas utilizadas para determinar si un conjunto de datos sigue o se ajusta a una distribución de probabilidad específica, como la distribución normal (gaussiana). Estos tests evalúan si los datos observados son consistentes con las características teóricas de la distribución en cuestión.

El test de normalidad más común es el Test de Shapiro-Wilk, aunque también existen otros como el Test de Kolmogorov-Smirnov y el Test de Lilliefors, entre otros. El Test de Shapiro-Wilk es ampliamente utilizado

y ampliamente recomendado cuando se trata de verificar la normalidad de los datos.

Las hipótesis para el Test de Shapiro-Wilk son las siguientes:

Hipótesis nula (H0): Los datos siguen una distribución normal. Hipótesis alternativa (H1): Los datos no siguen una distribución normal.

El procedimiento del test consiste en calcular una estadística de prueba (W) y compararla con un valor crítico de tabla. Si el valor de W es menor que el valor crítico, se rechaza la hipótesis nula, lo que significa que los datos no siguen una distribución normal.

Para realizar el Test de Shapiro-Wilk u otros tests de normalidad en R, puedes usar la función shapiro.test en el caso de Shapiro-Wilk, o funciones específicas para otros tests, como ks.test para el Test de Kolmogorov-Smirnov. El nivel de confianza más comúnmente utilizado para este tipo de pruebas es el 0.05 (5%), lo que significa que si el valor p resultante del test es menor que 0.05, se rechazaría la hipótesis nula. Es importante tener en cuenta que la falta de normalidad no implica necesariamente que un modelo de regresión lineal sea inapropiado. Los modelos de regresión lineal son robustos y pueden funcionar bien incluso cuando los datos no siguen estrictamente una distribución normal. Sin embargo, es útil comprender la distribución de los errores para interpretar los resultados de manera más precisa.

2.9.1 Test de bondad de ajuste

```
# Test de Shapiro-Wilk para los residuos del modelo de regresión lineal

shapiro.test(modelo_regresion$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo_regresion$residuals
## W = 0.96035, p-value = 0.1832
```

```
# Test de Shapiro-Wilk para la variable Precio

shapiro.test(arriendo$Precio)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  arriendo$Precio
## W = 0.98464, p-value = 0.8621
```

```
# Test de Shapiro-Wilk para la variable m2

shapiro.test(arriendo$m2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  arriendo$m2
## W = 0.97679, p-value = 0.5877
```

2.9.2 Interpretación del test de bondad de ajuste

El test de Shapiro-Wilk para los residuos del modelo de regresión lineal tiene un valor p de 0.1832, lo que indica que los residuos se distribuyen normalmente. El test de Shapiro-Wilk para la variable Precio tiene un valor p de 0.8621, lo que indica que la variable Precio se distribuye normalmente. El test de Shapiro-Wilk para la variable m2 tiene un valor p de 0.5877, lo que indica que la variable m2 se distribuye normalmente.

2.10 Sección de noticias del portal inmobiliario

Para realizar las pruebas de hipótesis necesarias, primero debemos definir las hipótesis nulas (H0) y alternativas (H1) apropiadas para cada pregunta. Luego, utilizaremos pruebas estadísticas para evaluar estas hipótesis.

2.10.1 Prueba de hipótesis para la pregunta 8

- ¿El precio promedio de arriendo ha aumentado en comparación con $970?

Hipótesis (H0): El precio promedio de arriendo es igual a $970 (u = 970). Hipótesis (H1): El precio promedio de arriendo es diferente de $970 (u =/ 970).

Para realizar esta prueba de hipótesis, utilizaremos una prueba t de una muestra, ya que queremos comparar la media de una muestra con un valor específico (970).

```
# Prueba t de una muestra para la pregunta 8
# H0: u = 970 # nolint
# H1: u =/ 970 # nolint
# Realizamos la prueba t de una muestra

resultado_prueba <- t.test(arriendo$Precio, mu = 970)

print(resultado_prueba)
```

```
##
##  One Sample t-test
##
## data:  arriendo$Precio
## t = 1.4636, df = 38, p-value = 0.1515
## alternative hypothesis: true mean is not equal to 970
## 95 percent confidence interval:
##   953.5919 1072.0491
## sample estimates:
## mean of x
##  1012.821
```

2.10.1.2 Interpretación de la prueba de hipótesis para la pregunta 8

El resultado de la prueba t de una muestra para la pregunta 8 es el siguiente:

- Estadístico t: 1.4636
- Grados de libertad (df): 38
- Valor p (p-value): 0.1515

Esto corresponde a la hipótesis nula (H0) que establece que la media del precio de arriendo (u) es igual a 970. La hipótesis alternativa (H1) sugiere que u no es igual a 970.

Dado que el valor p (p-value) es 0.1515, y es mayor que el nivel de significancia típico de 0.05 para una confianza del 95%, no tenemos evidencia suficiente para rechazar la hipótesis nula. En otras palabras, no podemos afirmar que el precio promedio de arriendo sea diferente de $970.

El intervalo de confianza (95 percent confidence interval) también se proporciona, y muestra que el intervalo va desde 953.5919 a 1072.0491. Esto significa que con un 95% de confianza, podemos afirmar que la media del precio de arriendo cae dentro de ese intervalo. Como 970 está dentro de ese intervalo, no podemos rechazar la hipótesis nula.

2.10.2 Prueba de hipótesis para la pregunta 9

- ¿El promedio del precio de arriendo es distinto según la zona de la ciudad?

Hipótesis (H0): El promedio del precio de arriendo es igual en ambas zonas (u_A = u_B). Hipótesis (H1): El promedio del precio de arriendo es diferente en al menos una de las zonas (u_A =/ u_B).

Para realizar esta prueba de hipótesis, utilizaremos una prueba t de dos muestras, ya que queremos comparar las medias de dos muestras (zona A y zona B).

```
# Prueba t de dos muestras para la pregunta 9
# H0: u_A = u_B # nolint
# H1: u_A =/ u_B # nolint

# Realizamos la prueba t de dos muestras

precios_zona_A <- arriendo$Precio[arriendo$Categoria == "A"] # nolint
precios_zona_B <- arriendo$Precio[arriendo$Categoria == "B"] # nolint

resultado_prueba <- t.test(precios_zona_A, precios_zona_B)

print(resultado_prueba)
```

```
##
##  Welch Two Sample t-test
##
## data:  precios_zona_A and precios_zona_B
## t = -0.14088, df = 36.687, p-value = 0.8887
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -128.2211  111.5544
## sample estimates:
## mean of x mean of y
##  1008.333  1016.667
```

2.10.2.1 Interpretación de la prueba de hipótesis para la pregunta 9

El resultado de la prueba t de dos muestras para la pregunta 9 es el siguiente:

- Estadístico t: -0.14088
- Grados de libertad (df): 36.687

- Valor p (p-value): 0.8887

Esto corresponde a la hipótesis nula (H0) que establece que las medias de los precios de arriendo en las zonas A y B (u_A y u_B) son iguales. La hipótesis alternativa (H1) sugiere que las medias no son iguales.

Dado que el valor p (p-value) es 0.8887, y es mucho mayor que el nivel de significancia típico de 0.05 para una significancia del 5%, no tenemos evidencia suficiente para rechazar la hipótesis nula. En otras palabras, no podemos afirmar que las medias de los precios de arriendo sean diferentes entre las zonas A y B.

El intervalo de confianza (95 percent confidence interval) también se proporciona, y muestra que el intervalo va desde -128.2211 a 111.5544. Esto significa que con un 95% de confianza, podemos afirmar que la diferencia en las medias de los precios de arriendo entre las dos zonas cae dentro de ese intervalo, y 0 está dentro de ese intervalo, lo que respalda la idea de que las medias son iguales.

3. Conclusiones

Gracias a la base de datos se pudo analizar el comportamiento de las variables que evaluarían el comportamiento del precio del arriendo en ciertas zonas, este análisis se hizo con la ayuda del Software de Rstudio con el cual se pudo graficar en los casos correspondientes como las variables estaban relacionadas con cada una, así como proponer modelos para la variable que se considere importante a la hora de estimar los precios de arriendos, para luego dar una breve explicación de como funciona el coeficiente de determinación lineal, así como el test de bondad, test de normalidad y como aplicarlos al problema y por ultimo se realizaron pruebas de hipótesis.