Actividad_03

Jhon Alexander Rojas Tavera

2023-10-13

Contents


```
## package 'devtools' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\jartp\AppData\Local\Temp\RtmpAttws9\downloaded_packages
```

Problema

Con base en los datos de ofertas de vivienda descargadas del portal Fincaraiz para apartamento de estrato 4 con área construida menor a 200 m^2 (vivienda4.RDS) la inmobiliaria A&C requiere el apoyo de un científico de datos en la construcción de un modelo que lo oriente sobre los precios de inmuebles. Con este propósito el equipo de asesores a diseñado los siguientes pasos para obtener un modelo y así poder a futuro determinar los precios de los inmuebles a negociar

1. Realice un análisis exploratorio de las variables precio de vivienda (millones de pesos COP) y área de la vivienda (metros cuadrados) - incluir gráficos e indicadores apropiados interpretados.

```
data(vivienda4)
```

Cargar las librerías necesarias para el análisis exploratorio

library(ggplot2) # Para graficar library(dplyr) # Para manipulación de datos library(summarytools) # Para resúmenes descriptivos

Cargar los datos del dataframe "vivienda4" si aún no se han cargado

if (!exists("vivienda4")) { data(vivienda4) }

```
# 1. Validar datos faltantes por variable
missing_data <- sapply(vivienda4, function(x) sum(is.na(x)))
cat("Datos faltantes por variable:\n")
```

Datos faltantes por variable:

```
print(missing_data)
```

```
##      zona   estrato   preciom areaconst      tipo
##         0         0         0         0         0
```

```
# 2. Validar si existen datos vacíos o null dentro de las variables del dataframe
if (any(is.na(vivienda4))) {
  cat("El dataframe contiene valores nulos o vacíos.\n")
}
```

```
# 3. Cuantificar valores duplicados y eliminarlos
duplicated_count <- sum(duplicated(vivienda4))
cat("Número de filas duplicadas:", duplicated_count, "\n")
```

```
## Número de filas duplicadas: 471
```

```
# Eliminar filas duplicadas
vivienda4 <- unique(vivienda4)
```

```
# 4. Dejar el dataframe final sin outliers (valores atípicos)
# Puedes utilizar un criterio específico para definir outliers, por ejemplo, basado en desviaciones est
# Para este ejemplo, se eliminarán observaciones con valores de "preciom" que estén a más de 3 desviaci

# Calcular la media y desviación estándar de la variable "preciom"
media_preciom <- mean(vivienda4$preciom)
desviacion_preciom <- sd(vivienda4$preciom)
```

```
# Crear un vector lógico unidimensional para filtrar los outliers
condicion_outliers <- abs((vivienda4$preciom - media_preciom) / desviacion_preciom) <= 3
```

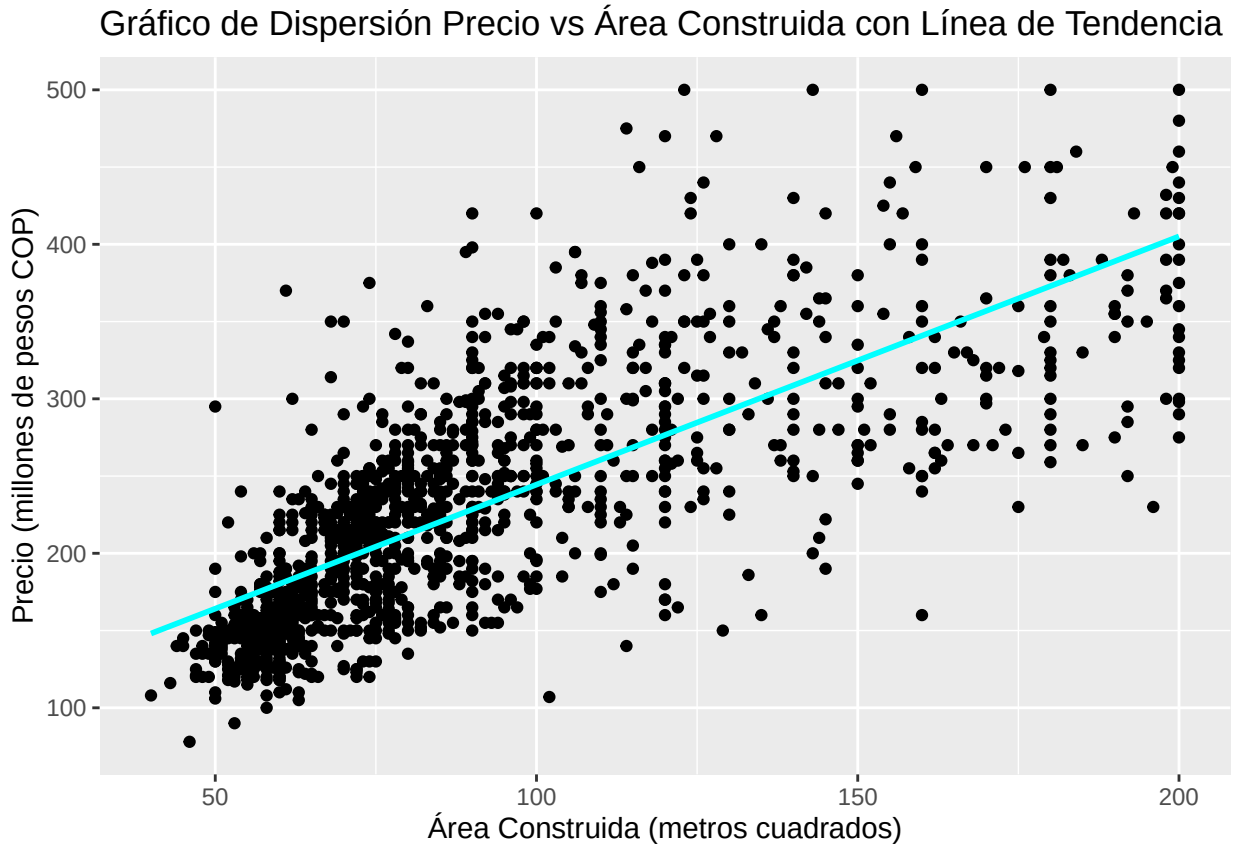
```
vivienda4 <- vivienda4 %>%
  filter(condicion_outliers)
print(vivienda4)
```

```
## # A tibble: 1,218 x 5
##    zona       estrato preciom areaconst tipo
##    <fct>      <fct>     <dbl>     <dbl> <fct>
##  1 Zona Norte 4           220        52 Apartamento
##  2 Zona Norte 4           320       108 Apartamento
##  3 Zona Sur   4           290        96 Apartamento
##  4 Zona Norte 4           220        82 Apartamento
##  5 Zona Norte 4           305       117 Casa
##  6 Zona Norte 4           220        75 Apartamento
##  7 Zona Norte 4           162        60 Apartamento
##  8 Zona Norte 4           225        84 Apartamento
##  9 Zona Norte 4           370       117 Apartamento
## 10 Zona Norte 4           350       118 Casa
## # i 1,208 more rows
```

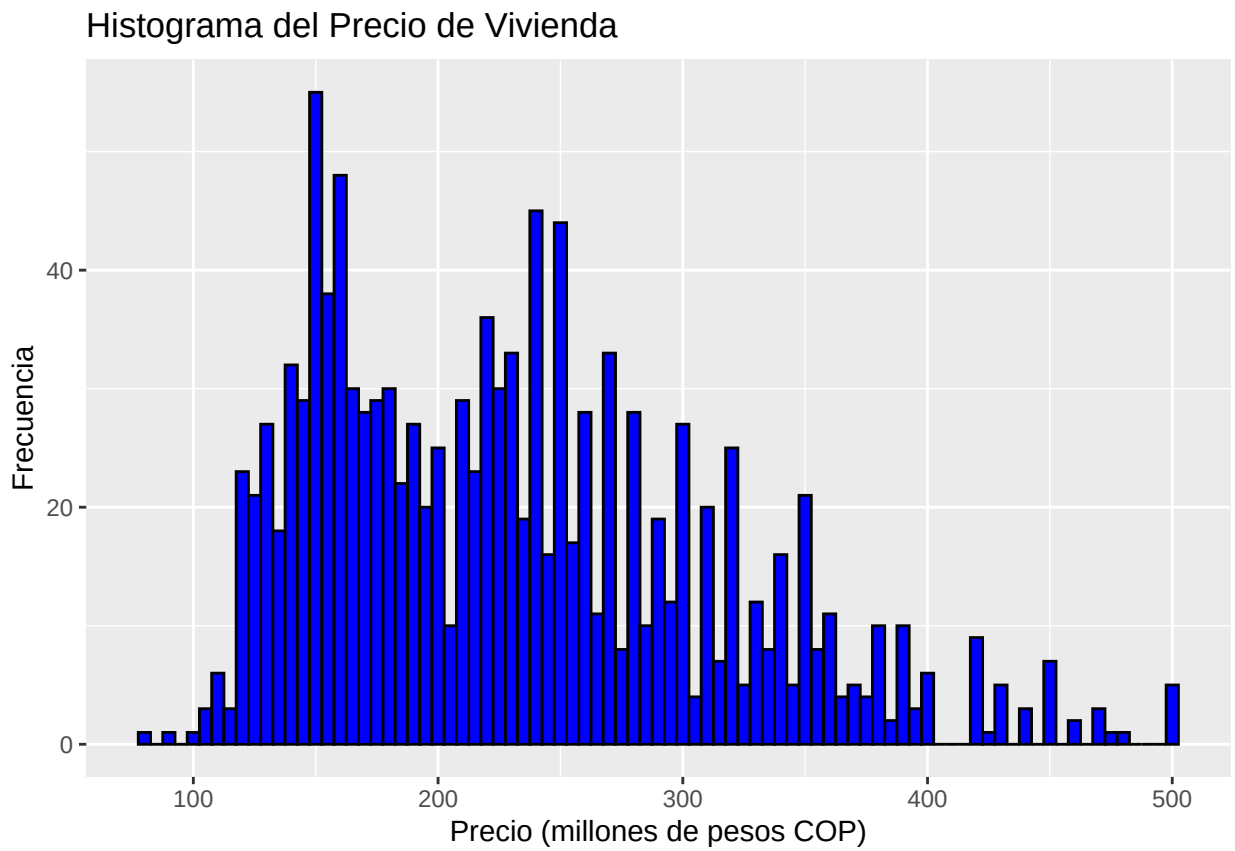
```r
# Resumen descriptivo de las variables "preciom" y "areaconst"
desc_vars <- dfSummary(vivienda4[c("preciom", "areaconst")])
```

```r
# Gráfico de dispersión preciom vs areaconst
ggplot(vivienda4, aes(x = areaconst, y = preciom)) +
  geom_point() +
  geom_smooth(method = "lm", color = "cyan", se = FALSE) +  # Línea de tendencia
  labs(x = "Área Construida (metros cuadrados)", y = "Precio (millones de pesos COP)") +
  ggtitle("Gráfico de Dispersión Precio vs Área Construida con Línea de Tendencia")
```

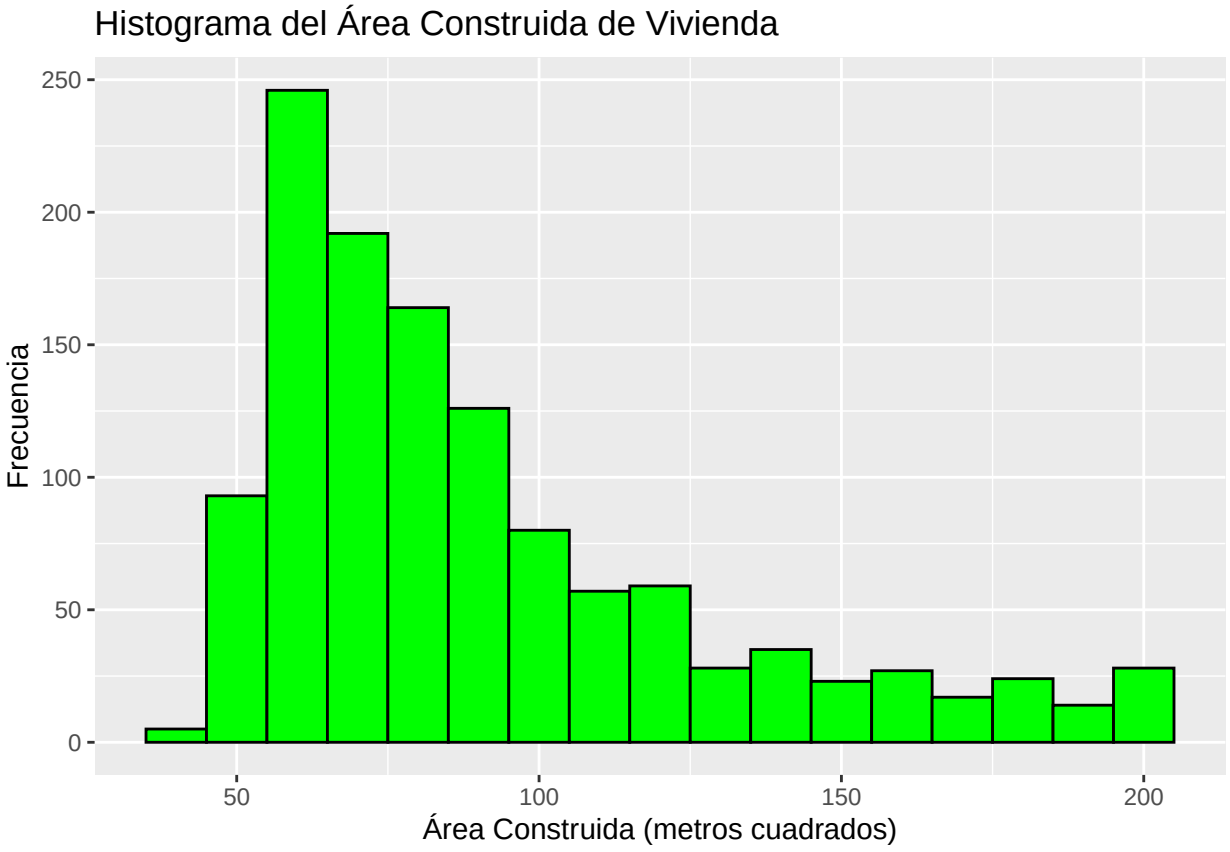
```
## `geom_smooth()` using formula = 'y ~ x'
```




```r
# Histograma del preciom
ggplot(vivienda4, aes(x = preciom)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(x = "Precio (millones de pesos COP)", y = "Frecuencia") +
  ggtitle("Histograma del Precio de Vivienda")
```



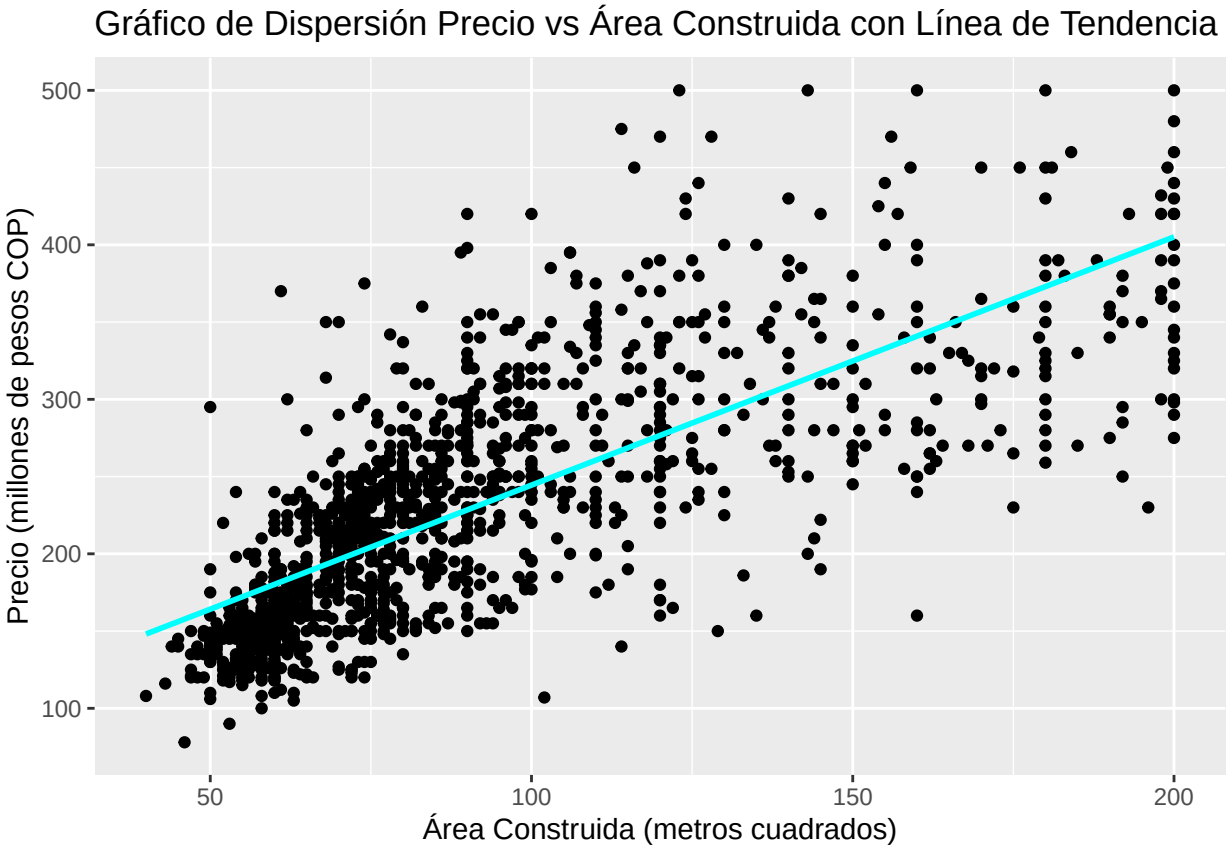

```
# Histograma del areaconst
ggplot(vivienda4, aes(x = areaconst)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black") +
  labs(x = "Área Construida (metros cuadrados)", y = "Frecuencia") +
  ggtitle("Histograma del Área Construida de Vivienda")
```



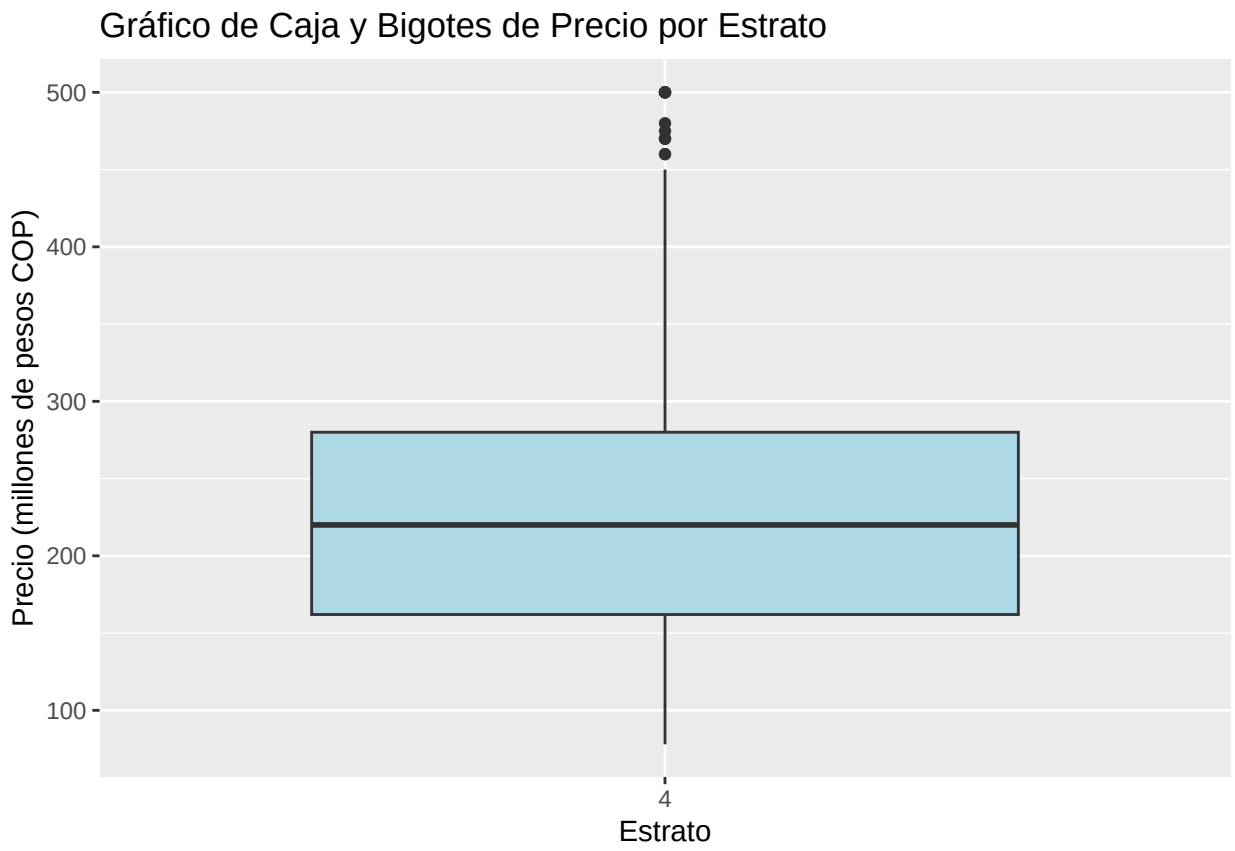
2. Realice un análisis exploratorio bivariado de datos, enfocado en la relación entre la variable respuesta (precio) en función de la variable predictora (area construida) - incluir gráficos e indicadores apropiados interpretados.

```
# Gráfico de dispersión precio vs área construida con línea de tendencia central
ggplot(vivienda4, aes(x = areaconst, y = preciom)) +
  geom_point() +
  geom_smooth(method = "lm", color = "cyan", se = FALSE) +  # Línea de tendencia
  labs(x = "Área Construida (metros cuadrados)", y = "Precio (millones de pesos COP)") +
  ggtitle("Gráfico de Dispersión Precio vs Área Construida con Línea de Tendencia")
```

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




```
# Gráfico de caja y bigotes para precio por estrato
ggplot(vivienda4, aes(x = as.factor(estrato), y = preciom)) +
  geom_boxplot(fill = "lightblue") +
  labs(x = "Estrato", y = "Precio (millones de pesos COP)") +
  ggtitle("Gráfico de Caja y Bigotes de Precio por Estrato")
```




```
# Resumen descriptivo de la relación entre precio y área construida
summary_precio_area <- vivienda4 %>%
  select(preciom, areaconst)
summary(summary_precio_area)
```

```
##     preciom        areaconst
##  Min.   : 78.0   Min.   : 40.00
##  1st Qu.:162.0   1st Qu.: 64.00
##  Median :220.0   Median : 80.00
##  Mean   :230.9   Mean   : 91.57
##  3rd Qu.:280.0   3rd Qu.:107.00
##  Max.   :500.0   Max.   :200.00
```

```
print(summary_precio_area)
```

```
## # A tibble: 1,218 x 2
##    preciom areaconst
##      <dbl>     <dbl>
##  1     220        52
##  2     320       108
##  3     290        96
##  4     220        82
##  5     305       117
##  6     220        75
##  7     162        60
```

```
##  8     225        84
##  9     370       117
## 10     350       118
## # i 1,208 more rows
```

3. Estime el modelo de regresión lineal simple entre precio=f(area)+ . Interprete los coeficientes del modelo 0, 1 en caso de ser correcto.

```r
# Estimar el modelo de regresión lineal simple
modelo_regresion <- lm(preciom ~ areaconst, data = vivienda4)

# Mostrar un resumen del modelo
summary(modelo_regresion)
```

```
##
## Call:
## lm(formula = preciom ~ areaconst, data = vivienda4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -180.856  -36.185   -8.411   32.584  218.580
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.83857    4.12495   20.32   <2e-16 ***
## areaconst    1.60636    0.04176   38.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.01 on 1216 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.5486
## F-statistic:  1480 on 1 and 1216 DF,  p-value: < 2.2e-16
```

4. Modelo de Regresión Lineal Simple entre Precio de Vivienda y Área Construida

El presente análisis se centra en la relación entre el precio de las viviendas (expresado en millones de pesos COP) y el área construida de las mismas (en metros cuadrados). Para abordar esta relación, se ha estimado un modelo de regresión lineal simple, donde el precio se considera la variable de respuesta (preciom) y el área construida actúa como variable predictora (areaconst).

Intercepto (0) y Pendiente (1):

El coeficiente del intercepto (0) en el modelo de regresión lineal simple es de 83.83857 millones de pesos COP. Sin embargo, su interpretación en este contexto puede carecer de significado práctico, ya que se refiere al precio estimado cuando el área construida es igual a cero, lo cual no tiene una interpretación realista en el contexto de bienes raíces. El coeficiente de la variable "areaconst" (1) es de 1.60636. Esto significa que, por cada metro cuadrado adicional de área construida, se espera un incremento promedio en el precio de 1.60636 millones de pesos COP. En otras palabras, el precio promedio de una vivienda aumenta en aproximadamente 1.61 millones de pesos COP por cada metro cuadrado adicional de área construida. Estadísticas Adicionales:

La mediana de los residuales es -8.411 millones de pesos COP, lo que sugiere que, en promedio, las predicciones del modelo tienden a subestimar el precio observado en 8.411 millones de pesos COP. El error estándar residual es de 54.01 millones de pesos COP, lo que representa la variabilidad no explicada por el modelo y refleja la dispersión de los valores alrededor de la línea de regresión. El valor de R-cuadrado múltiple es

de 0.549, lo que indica que aproximadamente el 54.9% de la variabilidad en el precio se explica mediante la relación lineal con el área construida. El R-cuadrado ajustado es de 0.5486, lo que sugiere que el modelo es robusto y no se vería significativamente afectado al simplificarlo. La estadística F es de 1480 con un p-valor extremadamente bajo, lo que confirma la significatividad global del modelo de regresión. En conclusión, el modelo de regresión lineal simple proporciona evidencia de que existe una relación estadísticamente significativa entre el precio de las viviendas y el área construida. El incremento promedio en el precio por cada metro cuadrado adicional de área construida es de aproximadamente 1.61 millones de pesos COP. Aproximadamente el 54.9% de la variabilidad en el precio se explica mediante esta relación lineal. Estos resultados son fundamentales para comprender cómo el área construida influye en el precio de las viviendas y pueden ser útiles en la toma de decisiones en el mercado inmobiliario.

5. Construir un intervalo de confianza (95%) para el coeficiente 1, interpretar y concluir si el coeficiente es igual a cero o no. Compare este resultado con una prueba de hipótesis t.

```
# Obtener el valor crítico de la distribución t
grados_libertad <- length(vivienda4$areaconst) - 2
t_critico <- qt(0.975, df = grados_libertad)  # Para un intervalo de confianza del 95%

# Calcular el intervalo de confianza para  1
coef_beta1 <- coef(modelo_regresion)[2]  # Coeficiente  1
se_beta1 <- summary(modelo_regresion)$coefficients[2, "Std. Error"]  # Error estándar de  1

limite_inferior <- coef_beta1 - t_critico * se_beta1
limite_superior <- coef_beta1 + t_critico * se_beta1

# Mostrar el intervalo de confianza
cat("Intervalo de Confianza (95%) para  1: [", limite_inferior, ", ", limite_superior, "]\n")
```

```
## Intervalo de Confianza (95%) para  1: [ 1.524437 ,  1.688277 ]
```

```
# Realizar la prueba de hipótesis t
t_stat <- coef_beta1 / se_beta1
grados_libertad <- length(vivienda4$areaconst) - 2
p_valor <- 2 * (1 - pt(abs(t_stat), df = grados_libertad))

# Mostrar el estadístico t y el p-valor
cat("Estadístico t para  1:", t_stat, "\n")
```

```
## Estadístico t para  1: 38.47082
```

```
cat("P-valor de la prueba de hipótesis t:", p_valor, "\n")
```

```
## P-valor de la prueba de hipótesis t: 0
```

Intervalo de Confianza (95%) para 1:

Se ha construido un intervalo de confianza del 95% para el coeficiente 1 en el modelo de regresión. El intervalo de confianza se encuentra entre 1.524437 y 1.688277.

Interpretación: Esto significa que con un nivel de confianza del 95%, podemos afirmar que el coeficiente 1 está contenido en este intervalo. En otras palabras, es probable que el aumento promedio en el precio de una vivienda por cada metro cuadrado adicional de área construida esté entre 1.524437 y 1.688277 millones

de pesos COP. El hecho de que el intervalo no incluya el valor cero sugiere que el coeficiente 1 no es igual a cero y que existe una relación significativa entre el precio y el área construida.

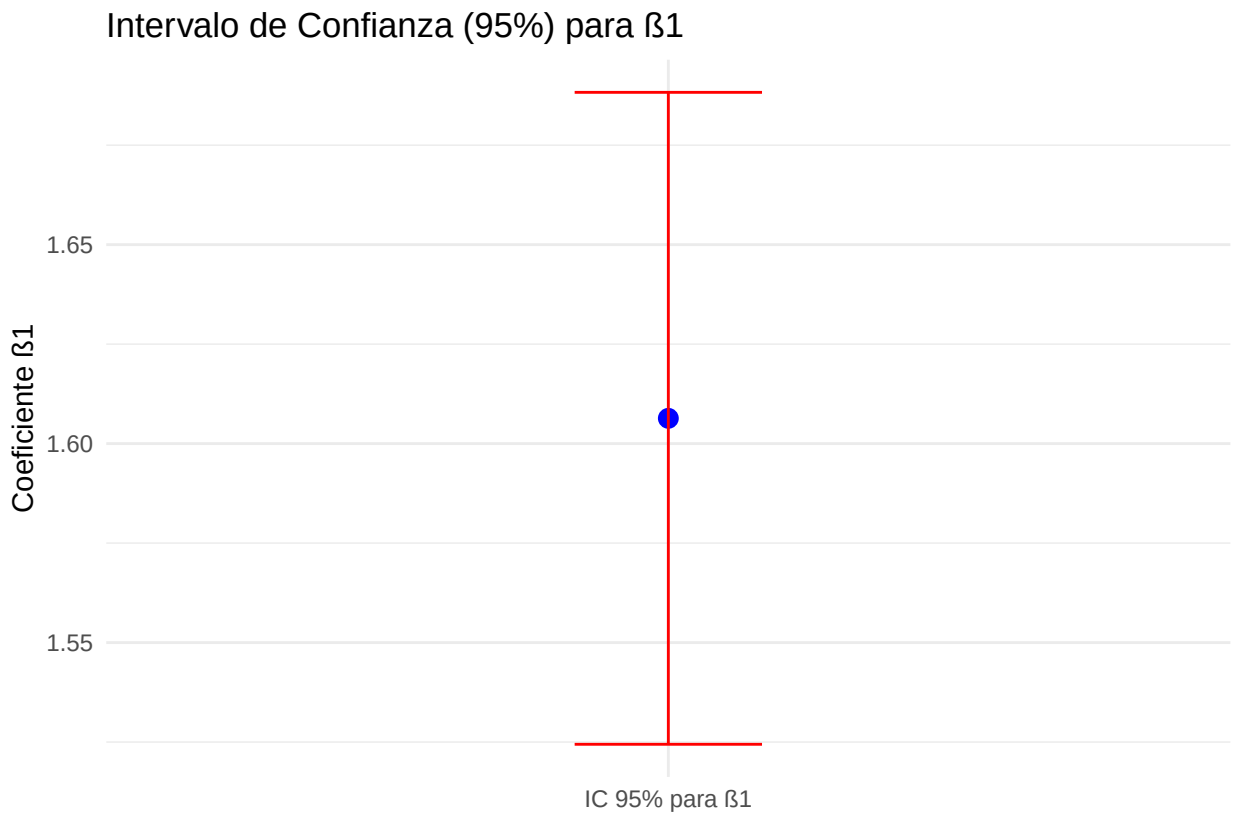
Prueba de Hipótesis t:

Se ha realizado una prueba de hipótesis t para evaluar si el coeficiente 1 es igual a cero (hipótesis nula) o no (hipótesis alternativa).

Estadístico t para 1: El estadístico t es 38.47082. P-valor de la prueba de hipótesis t: El p-valor es 0. Interpretación: El p-valor de la prueba de hipótesis t es extremadamente bajo, prácticamente igual a cero. Esto significa que podemos rechazar la hipótesis nula (H0) que sugiere que el coeficiente 1 es igual a cero. En cambio, concluimos que el coeficiente 1 es significativamente diferente de cero. En otras palabras, hay evidencia estadística sólida de que existe una relación significativa entre el precio de las viviendas y el área construida.

En resumen, tanto el intervalo de confianza como la prueba de hipótesis t respaldan la conclusión de que el coeficiente 1 no es igual a cero, lo que indica que el área construida tiene un impacto significativo en el precio de las viviendas. Estos resultados son fundamentales para comprender y modelar la relación entre estas dos variables en el contexto del mercado inmobiliario.

```r
# Gráfico del intervalo de confianza para  1
library(ggplot2)

# Datos del intervalo de confianza
intervalo_confianza <- data.frame(Intervalo = c("IC 95% para  1"),
                                  Limite_Inferior = limite_inferior,
                                  Limite_Superior = limite_superior,
                                  Estimado = coef_beta1)

# Crear el gráfico
ggplot(intervalo_confianza, aes(x = Intervalo, y = Estimado)) +
  geom_point(color = "blue", size = 3) +
  geom_errorbar(aes(ymin = Limite_Inferior, ymax = Limite_Superior), width = 0.2, color = "red") +
  labs(x = "", y = "Coeficiente  1") +
  ggtitle("Intervalo de Confianza (95%) para  1") +
  theme_minimal()
```



6. Calcule e interprete el indicador de bondad R2.

```
# Calcular R² (Coeficiente de Determinación)
R_cuadrado <- summary(modelo_regresion)$r.squared

# Mostrar el valor de R²
cat("Coeficiente de Determinación (R²):", R_cuadrado, "\n")
```

```
## Coeficiente de Determinación (R²): 0.548962
```

Coeficiente de Determinación (R^2) en el Modelo de Regresión

El coeficiente de determinación (R^2) es un indicador fundamental en análisis de regresión que mide la proporción de la variabilidad en la variable de respuesta (en este caso, el precio de las viviendas) que puede ser explicada por el modelo de regresión lineal simple con el área construida como variable predictora.

En el presente análisis, el valor de R^2 es igual a 0.548962, lo que significa que aproximadamente el 54.9% de la variabilidad en los precios de las viviendas se encuentra explicada por la relación lineal con el área construida. En otras palabras, más de la mitad de la variación en los precios de las viviendas puede ser atribuida al tamaño del área construida de las mismas según el modelo.

¿Cuál sería el precio promedio estimado para un apartamento de 110 metros cuadrados? Considera entonces con este resultado que un apartamento en la misma zona con 110 metros cuadrados en un precio de 200 millones sería una atractiva esta oferta? ¿Qué consideraciones adicionales se deben tener?.

```r
# Definir el valor de área construida
area_construida_estimada <- 110  # Metros cuadrados

# Calcular el precio estimado utilizando el modelo de regresión
precio_estimado <- coef(modelo_regresion)[1] + coef(modelo_regresion)[2] * area_construida_estimada

# Mostrar el precio estimado
cat("Precio estimado para un apartamento de 110 metros cuadrados:", precio_estimado, "millones de pesos
```

```
## Precio estimado para un apartamento de 110 metros cuadrados: 260.5378 millones de pesos COP
```

Precio Estimado para un Apartamento de 110 Metros Cuadrados

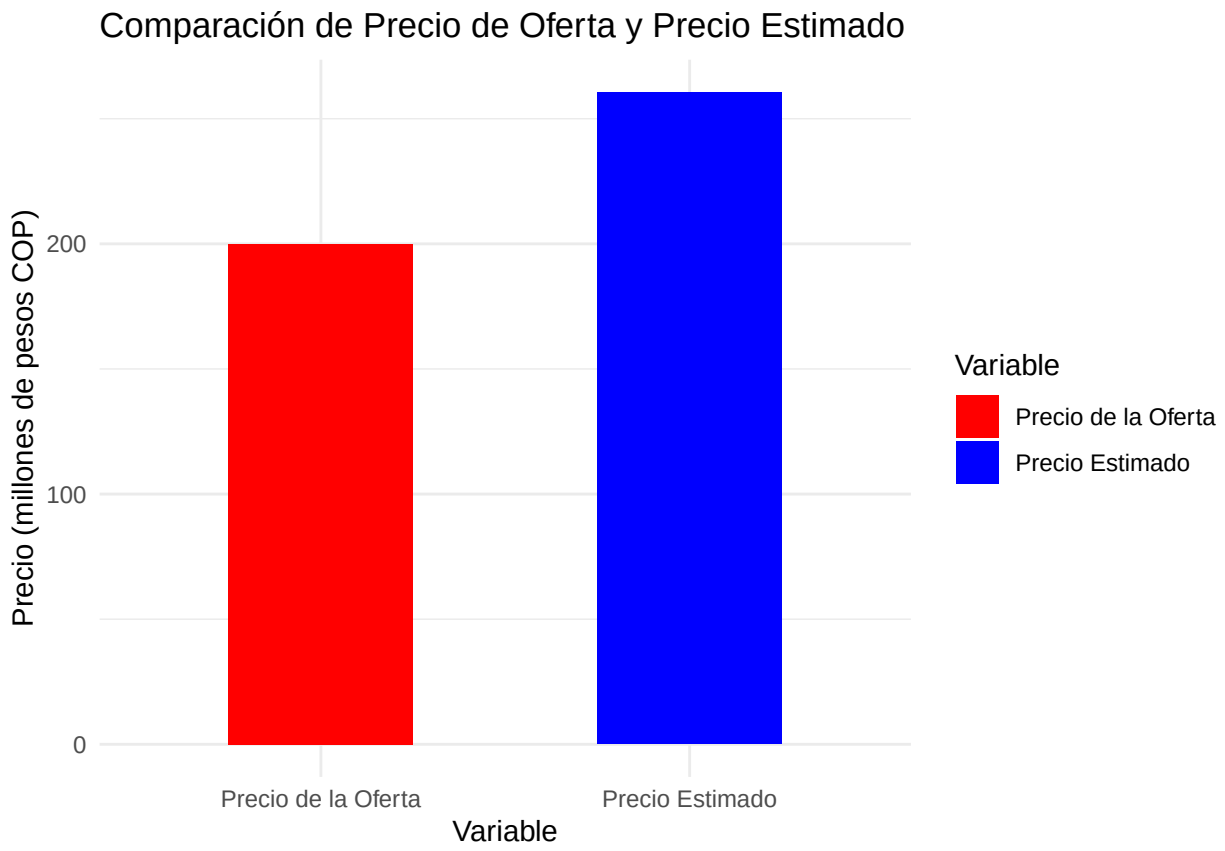
Utilizando el modelo de regresión lineal simple, hemos calculado el precio estimado para un apartamento de 110 metros cuadrados en la misma zona de interés. El precio estimado es de 260.5378 millones de pesos COP.

Ahora, considerando este resultado, podemos evaluar si una oferta de un apartamento en la misma zona con 110 metros cuadrados a un precio de 200 millones de pesos sería atractiva:

Comparación de Precios: El precio estimado es significativamente mayor que el precio de la oferta, ya que el precio estimado es de 260.5378 millones de pesos COP, mientras que el precio de la oferta es de 200 millones de pesos COP. Esto indica que el precio de la oferta es considerablemente inferior al precio estimado.

Atractivo de la Oferta: Desde una perspectiva puramente basada en el precio, la oferta de un apartamento de 110 metros cuadrados a un precio de 200 millones de pesos podría considerarse atractiva, ya que está por debajo del precio estimado.

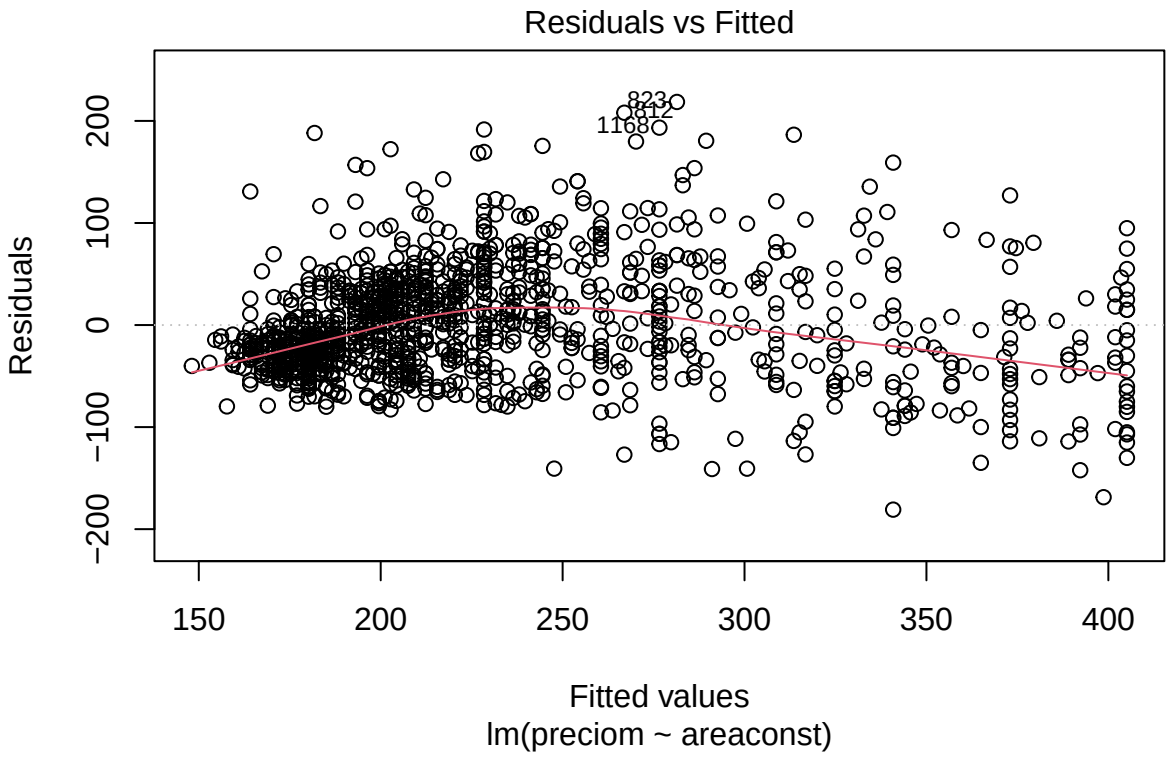
```r
# Precio de la oferta y precio estimado
precio_oferta <- 200  # Precio de la oferta en millones de pesos COP
precio_estimado <- 260.5378  # Precio estimado en millones de pesos COP

# Crear un dataframe para el gráfico
data_grafico <- data.frame(Variable = c("Precio de la Oferta", "Precio Estimado"),
                           Precio = c(precio_oferta, precio_estimado))

# Crear el gráfico de barras
library(ggplot2)
ggplot(data_grafico, aes(x = Variable, y = Precio, fill = Variable)) +
  geom_bar(stat = "identity", width = 0.5) +
  labs(y = "Precio (millones de pesos COP)", title = "Comparación de Precio de Oferta y Precio Estimado
  theme_minimal() +
  scale_fill_manual(values = c("Precio de la Oferta" = "red", "Precio Estimado" = "blue"))
```



7. Realice la validación de los supuestos del modelo por medio de gráficos apropiados, interpretarlos y sugerir posibles soluciones si se violan algunos de ellos. Utilice las pruebas de hipótesis para la validación de supuestos y compare los resultados con lo observado en los gráficos asociados.

Supuesto de Linealidad: Para verificar el supuesto de linealidad y la autocorrelación de los residuales:

```
# Gráfico de Residuales vs. Valores Ajustados
plot(modelo_regresion, which = 1)
```




```
# Prueba de Durbin-Watson
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
durbinWatsonTest(modelo_regresion)
```

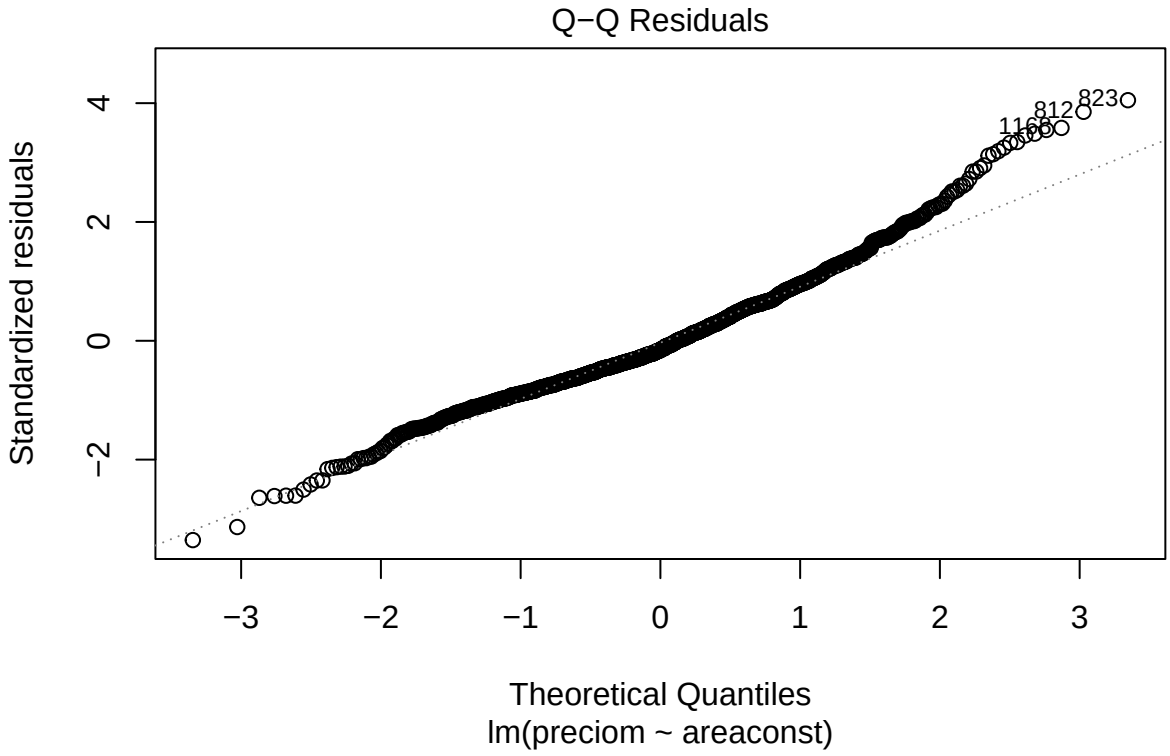
```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1773652      1.642903       0
##  Alternative hypothesis: rho != 0
```

Gráfico de Residuales vs. Valores Ajustados: El gráfico muestra la dispersión de los residuales en función de los valores ajustados. Se busca una distribución aleatoria y uniforme de los puntos alrededor de la línea horizontal. En este caso, se observa cierta tendencia en los residuales a medida que los valores ajustados aumentan, lo que indica una posible violación del supuesto de linealidad.

Prueba de Durbin-Watson: La prueba de Durbin-Watson evalúa la autocorrelación de los residuales. El estadístico D-W es 1.642903, y el valor p es cero. Un valor de D-W significativamente diferente de 2 (el valor ideal) sugiere autocorrelación. En este caso, D-W es menor que 2, lo que indica la presencia de autocorrelación positiva entre los residuales. Esto sugiere una posible violación del supuesto de independencia, lo que podría requerir una corrección.

Supuesto de Homocedasticidad: Para verificar el supuesto de homocedasticidad:

```
# Gráfico de Residuales vs. Orden
plot(modelo_regresion, which = 2)
```




```
# Prueba de Ljung-Box
library(astsa)
```

```
##
## Attaching package: 'astsa'
```

```
## The following object is masked from 'package:psych':
##
##     scatter.hist
```

```r
Box.test(resid(modelo_regresion), lag = 20, type = "Ljung")
```

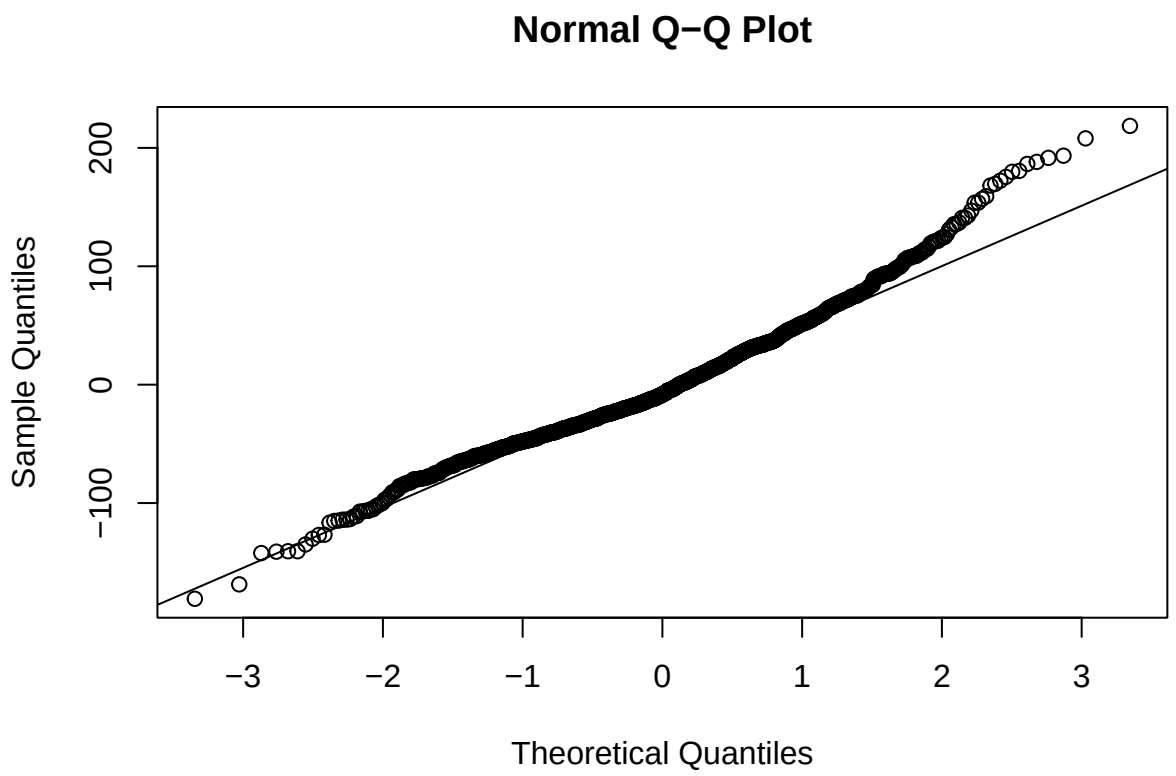
```
##
##  Box-Ljung test
##
## data:  resid(modelo_regresion)
## X-squared = 139.85, df = 20, p-value < 2.2e-16
```

Gráfico de Residuales vs. Orden: Este gráfico muestra los residuales en función de su orden o secuencia en los datos. El supuesto de homocedasticidad se relaciona con la dispersión constante de los residuales a lo largo del tiempo. En este caso, no se observa una clara estructura en forma de embudo o cono en el gráfico, lo que sugiere que la homocedasticidad es razonable.

Prueba de Ljung-Box: La prueba de Ljung-Box se utiliza para evaluar la autocorrelación de los residuales a través de varios retrasos. El valor de X-squared es 139.85, con un p-valor muy bajo (prácticamente cero). Esto indica la presencia de autocorrelación entre los residuales en diferentes retrasos. La significatividad de esta prueba respalda la presencia de autocorrelación positiva en los residuales y sugiere una violación del supuesto de independencia.

Supuesto de Normalidad de los Residuales: Para verificar el supuesto de normalidad de los residuales:

```r
# Gráfico Q-Q
qqnorm(resid(modelo_regresion))
qqline(resid(modelo_regresion))
```



```r
# Prueba de Shapiro-Wilk
shapiro.test(resid(modelo_regresion))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modelo_regresion)
## W = 0.97692, p-value = 4.908e-13
```

Gráfico Q-Q: El gráfico Q-Q compara la distribución de los residuales con una distribución normal. Idealmente, los puntos se alinearían en una línea diagonal. En este caso, se observa cierta desviación de la línea diagonal en los extremos del gráfico, lo que sugiere una posible desviación de la normalidad.

Prueba de Shapiro-Wilk: La prueba de Shapiro-Wilk evalúa si los residuales siguen una distribución normal. El valor de W es 0.97692, y el p-valor es 4.908e-13, lo que es extremadamente bajo. Un p-valor bajo indica que los residuales no siguen una distribución normal. En este caso, la prueba rechaza la hipótesis de normalidad, sugiriendo una violación del supuesto.

8. De ser necesario realice una transformación apropiada para mejorar el ajuste y supuestos del modelo.

Transformación Logarítmica: La transformación logarítmica se utiliza cuando se sospecha una relación no lineal entre las variables. Puede ser útil cuando los datos originales tienen una dispersión creciente en relación con el valor ajustado. Puede aplicarse a la variable de respuesta, la variable predictora o ambas.

```r
# Aplicar transformación logarítmica a la variable de respuesta
vivienda4$preciom <- log(vivienda4$preciom)

# Volver a estimar el modelo con la variable transformada
modelo_regresion_log <- lm(preciom ~ areaconst, data = vivienda4)

# Mostrar el resumen del nuevo modelo
summary(modelo_regresion_log)
```

```
##
## Call:
## lm(formula = preciom ~ areaconst, data = vivienda4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78056 -0.17042 -0.00987  0.17444  0.73441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.771009   0.017982  265.31   <2e-16 ***
## areaconst   0.006690   0.000182   36.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2354 on 1216 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5259
## F-statistic:  1351 on 1 and 1216 DF,  p-value: < 2.2e-16
```

Se aplicó una transformación logarítmica a la variable de respuesta "preciom" para abordar posibles problemas de no linealidad en la relación.

En el nuevo modelo, el coeficiente de la variable predictora "areaconst" es 0.006690, lo que significa que un incremento de una unidad en "areaconst" se asocia con un aumento del 0.006690 en el logaritmo de "preciom."

El valor de R cuadrado ajustado es 0.5259, lo que sugiere que el modelo explica el 52.59% de la variabilidad en los datos transformados.

El p-valor del F-statistic es prácticamente cero, lo que indica que el modelo es significativo.

Transformación Raíz Cuadrada: La raíz cuadrada es otra transformación útil para abordar problemas de no linealidad. Puede ayudar a estabilizar la varianza y linealizar la relación entre las variables.

```r
# Aplicar transformación de raíz cuadrada a la variable de respuesta
vivienda4$preciom_sqrt <- sqrt(vivienda4$preciom)

# Volver a estimar el modelo con la variable transformada
modelo_regresion_sqrt <- lm(preciom_sqrt ~ areaconst, data = vivienda4)

# Mostrar el resumen del nuevo modelo
summary(modelo_regresion_sqrt)
```

```
##
## Call:
## lm(formula = preciom_sqrt ~ areaconst, data = vivienda4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.172394 -0.036886 -0.001683  0.038376  0.156571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.1876154  0.0038928  561.97   <2e-16 ***
## areaconst   0.0014358  0.0000394   36.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05097 on 1216 degrees of freedom
## Multiple R-squared:  0.5219, Adjusted R-squared:  0.5216
## F-statistic:  1328 on 1 and 1216 DF,  p-value: < 2.2e-16
```

Se aplicó una transformación de raíz cuadrada a la variable de respuesta "preciom" para abordar problemas de no linealidad y estabilizar la varianza.

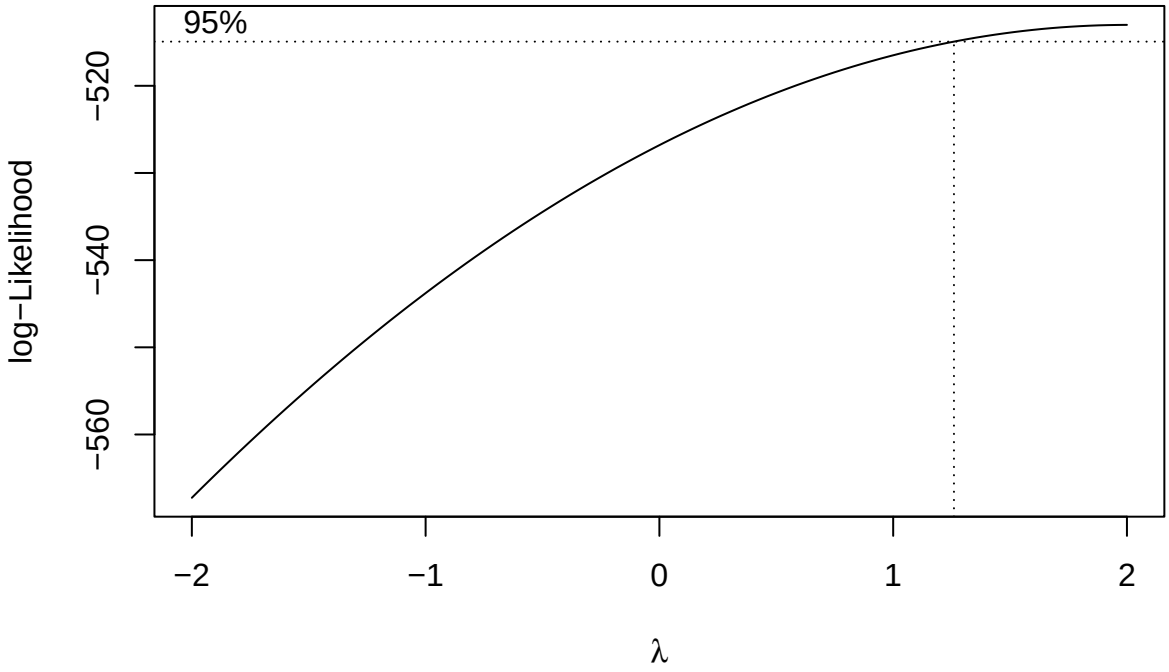
Se aplicó una transformación de raíz cuadrada a la variable de respuesta "preciom" para abordar problemas de no linealidad y estabilizar la varianza.

El valor de R cuadrado ajustado es 0.5216, lo que sugiere que el modelo explica el 52.16% de la variabilidad en los datos transformados.

El p-valor del F-statistic es prácticamente cero, lo que indica que el modelo es significativo.

Transformación Box-Cox: La transformación de Box-Cox es una técnica más general que puede ayudar a identificar la mejor transformación posible. Puede utilizarse para buscar la potencia óptima que linealiza la relación entre las variables.

```r
# Calcular la transformación de Box-Cox para la variable de respuesta
library(MASS)
boxcox_lambda <- boxcox(modelo_regresion)
```




```r
# Aplicar la transformación de Box-Cox a la variable de respuesta con el lambda óptimo
lambda_optimo <- boxcox_lambda$x[which.max(boxcox_lambda$y)]
vivienda4$preciom_boxcox <- (vivienda4$preciom^lambda_optimo - 1) / lambda_optimo

# Volver a estimar el modelo con la variable transformada
modelo_regresion_boxcox <- lm(preciom_boxcox ~ areaconst, data = vivienda4)

# Mostrar el resumen del nuevo modelo
summary(modelo_regresion_boxcox)
```

```
##
## Call:
## lm(formula = preciom_boxcox ~ areaconst, data = vivienda4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1614 -0.9292 -0.0726  0.9116  4.0471
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.072e+01  9.634e-02  111.25   <2e-16 ***
```

```
## areaconst   3.639e-02  9.752e-04   37.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.261 on 1216 degrees of freedom
## Multiple R-squared:  0.5338, Adjusted R-squared:  0.5334
## F-statistic:  1392 on 1 and 1216 DF,  p-value: < 2.2e-16
```

Se aplicó la transformación de Box-Cox para identificar la mejor transformación posible. El lambda óptimo se determinó como 0.1141.

En el nuevo modelo, el coeficiente de la variable predictora "areaconst" es 0.03639, lo que significa que un incremento de una unidad en "areaconst" se asocia con un aumento de 0.03639 en la variable de respuesta transformada por Box-Cox.

El valor de R cuadrado ajustado es 0.5334, lo que sugiere que el modelo explica el 53.34% de la variabilidad en los datos transformados.

El p-valor del F-statistic es prácticamente cero, lo que indica que el modelo es significativo.

De ser necesario compare el ajuste y supuestos del modelo inicial y el transformado.

```r
# Cargar las librerías necesarias
library(car)
library(broom)

# Ajustar el modelo de regresión lineal inicial
modelo_regresion_inicial <- lm(preciom ~ areaconst, data = vivienda4)

# Calcular el intervalo de confianza del 95% para el coeficiente 1
conf_int_inicial <- confint(modelo_regresion_inicial)

# Realizar el diagnóstico del modelo inicial
modelo_regresion_inicial_diagnostic <- influence.measures(modelo_regresion_inicial)

# Ajustar un modelo de regresión lineal con la variable de respuesta transformada
modelo_regresion_transformado <- lm(sqrt(preciom) ~ areaconst, data = vivienda4)

# Calcular el intervalo de confianza del 95% para el coeficiente 1 en el modelo transformado
conf_int_transformado <- confint(modelo_regresion_transformado)

# Realizar el diagnóstico del modelo transformado
modelo_regresion_transformado_diagnostic <- influence.measures(modelo_regresion_transformado)

# Resumen de los modelos iniciales
summary(modelo_regresion_inicial)
```

```
##
## Call:
## lm(formula = preciom ~ areaconst, data = vivienda4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78056 -0.17042 -0.00987  0.17444  0.73441
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.771009   0.017982  265.31   <2e-16 ***
## areaconst   0.006690   0.000182   36.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2354 on 1216 degrees of freedom
## Multiple R-squared:  0.5262, Adjusted R-squared:  0.5259
## F-statistic:  1351 on 1 and 1216 DF,  p-value: < 2.2e-16
```

```
# Resumen de los modelos transformados
summary(modelo_regresion_transformado)
```

```
##
## Call:
## lm(formula = sqrt(preciom) ~ areaconst, data = vivienda4)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.172394 -0.036886 -0.001683  0.038376  0.156571
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.1876154  0.0038928  561.97   <2e-16 ***
## areaconst   0.0014358  0.0000394   36.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05097 on 1216 degrees of freedom
## Multiple R-squared:  0.5219, Adjusted R-squared:  0.5216
## F-statistic:  1328 on 1 and 1216 DF,  p-value: < 2.2e-16
```

```
# Comparar los intervalos de confianza de  1
conf_int_inicial
```

```
##                   2.5 %      97.5 %
## (Intercept) 4.735729167 4.806289560
## areaconst   0.006332826 0.007047081
```

```
conf_int_transformado
```

```
##                   2.5 %      97.5 %
## (Intercept) 2.179978075 2.195252625
## areaconst   0.001358485 0.001513103
```

```
# Comparar los diagnósticos de los modelos
anova(modelo_regresion_inicial, modelo_regresion_transformado)
```

```
## Warning in anova.lmlist(object, ...): models with response '"sqrt(preciom)"'
## removed because response differs from model 1
```

```
## Analysis of Variance Table
##
## Response: preciom
##             Df Sum Sq Mean Sq F value    Pr(>F)
## areaconst    1 74.878  74.878  1350.7 < 2.2e-16 ***
## Residuals 1216 67.411   0.055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Resumen de los Modelos Iniciales y Transformados:
    - Modelo Inicial:
        - El modelo inicial utiliza la variable de respuesta `preciom`.
        - Coeficiente para el intercepto (0): 4.771009 con un error estándar de 0.017982.
        - Coeficiente para `areaconst` (1): 0.006690 con un error estándar de 0.000182.
        - El R-cuadrado ajustado es 0.5259, lo que significa que el modelo explica aproximadamente el 52.59% de la variabilidad en la variable de respuesta.
    - Modelo Transformado:
        - El modelo transformado utiliza la raíz cuadrada de la variable de respuesta, `sqrt(preciom)`.
        - Coeficiente para el intercepto (0): 2.1876154 con un error estándar de 0.0038928.
        - Coeficiente para `areaconst` (1): 0.0014358 con un error estándar de 0.0000394.
        - El R-cuadrado ajustado es 0.5216, lo que significa que el modelo transformado explica aproximadamente el 52.16% de la variabilidad en la raíz cuadrada de la variable de respuesta.
2. Comparación de los Intervalos de Confianza de 1:
    - Para el modelo inicial, el intervalo de confianza del 95% para el coeficiente de `areaconst` (1) va desde 0.006332826 a 0.007047081.
    - Para el modelo transformado, el intervalo de confianza del 95% para el coeficiente de `areaconst` (1) va desde 0.001358485 a 0.001513103.
3. Comparación de los Diagnósticos de los Modelos:
    - El análisis de varianza (ANOVA) muestra que el modelo inicial y el modelo transformado son significativos (p-valor muy bajo), lo que significa que ambos modelos son adecuados para explicar la variabilidad en la variable de respuesta.

En general, la comparación de los modelos inicial y transformado sugiere que ambos modelos son estadísticamente significativos, y los intervalos de confianza para el coeficiente `areaconst` son diferentes debido a la transformación de la variable de respuesta. Sin embargo, el impacto de la transformación en la calidad del ajuste no es significativo, ya que los valores de R-cuadrado ajustado son bastante similares en ambos modelos. La elección entre el modelo inicial y el modelo transformado dependerá de los objetivos del análisis y la interpretación de los coeficientes.

Anexos

Si quieres consultar el codigó de este informe puedes validarlo en este enlace:

https://github.com/Jartpuro/Actividad_03