Taller uno

Simón Cuartas Rendón

Marzo de 2022

```
# Borrado de memoria previa
rm(list = ls())

# Lectura de paquetes
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Taller uno

En cada una de tres zonas residenciales de cierta ciudad fueron seleccionadas al azar cinco casas recientemente vendidas, con el fin es estudiar la relación entre el precio de venta de la propiedad (Y) y el valor catastral de la propiedad (X). Los datos se pueden ver en el enunciado completo del taller.

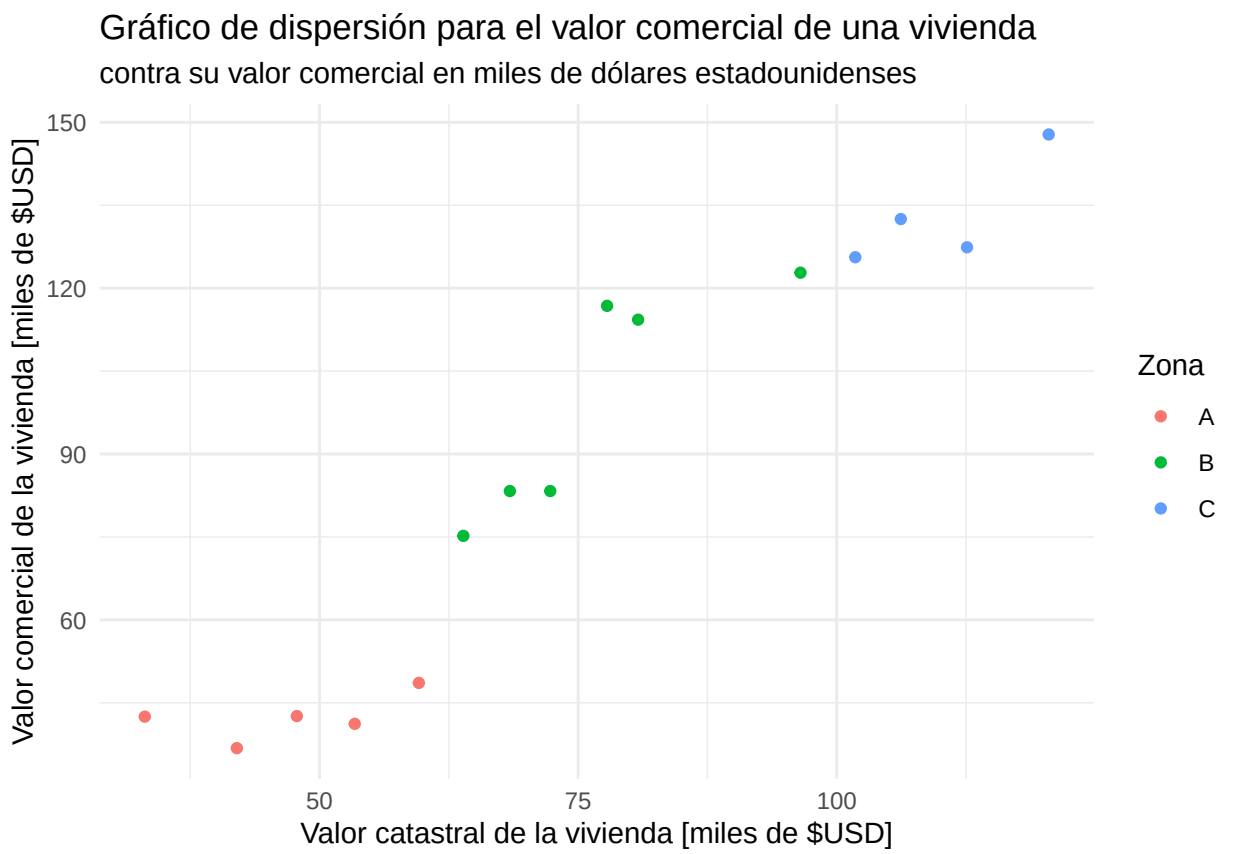
```
datos = read.csv('DatosVivienda.csv', header = TRUE, dec = ".", sep = ",")
names(datos)[1] <- "Propiedad"
```

Punto uno

Analice gráficamente la relación entre el precio de venta y el valor catastral teniendo en cuenta la información de la zona. ¿Qué se puede concluir acerca de la relación entre el precio de venta y el valor catastral? ¿difiere según la zona? ¿tendrá sentido realizar una regresión lineal con las 15 observaciones ignorando la información de la zona?

Sea X el valor catastral de una vivienda y Y su valor comercial, ambos en miles de dólares estadounidenses (USD). Para poder resolver este punto se va a realizar un gráfico de dispersión en el cual se muestra el valor catastral en el eje horizontal y el comercial en el eje vertical, y se diferencia cada una de las zonas analizadas con un color y una figura particular.

```r
p1 = ggplot(data = datos, mapping = aes(x = Valor, y = Precio)) +
     geom_point(mapping = aes(color = Zona)) +
     xlab("Valor catastral de la vivienda [miles de $USD]") +
     ylab("Valor comercial de la vivienda [miles de $USD]") +
     ggtitle("Gráfico de dispersión para el valor comercial de una vivienda",
             subtitle = "contra su valor comercial en miles de dólares estadounidenses") +
     theme_minimal()
p1
```



Y como se puede observar en el gráfico, para cada color, representando una zona diferente de la ciudad en la que se está abordando el estudio se tiene un efecto promedio del valor catastral de la vivienda sobre su valor comercial, al igual que la media del valor comercial para cada zona, por lo que es razonable plantear un modelo de regresión que considere la influencia de la zona en el valor comercial de la vivienda.

Punto dos

Si se considera que la relación lineal entre el precio de venta vs. el valor catastral puede diferir según la zona residencial, postule el modelo de regresión indicado con variables indicadoras para modelar la situación planteada, tome como zona de referencia.

Literal A. La zona A es el nivel de referencia.

```
attach(datos)
Zona = as.factor(Zona)
Zona = relevel(Zona, ref = "A")
```

Considerando las definiciones que se dieron previamente para el valor comercial y catastral de la vivienda, y tomandoteniendo en cuenta las siguientes variables indicadoras.

- Sea I_1 la variable indicadora que toma el valor de uno si una vivienda está ubicada en la zona A de la ciudad y cero en cualquier otro caso.
- Sea I_2 la variable indicadora que toma el valor de uno si una vivienda está ubicada en la zona B de la ciudad y cero en cualquier otro caso.
- Sea I_3 la variable indicadora que toma el valor de uno si una vivienda está ubicada en la zona C de la ciudad y cero en cualquier otro caso.

Teniendo en cuenta que para este literal se está tomando al nivel asociado a la zona A como el referencia, entonces para plantear el modelo de regresión lineal se van a emplear las variables indicadoras I_2 e I_3, de tal suerte que el modelo general es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i2} + \beta_3 I_{i3} + \beta_{1,2} X_1 I_{i2} + \beta_{1,3} X_{i1} I_{i3} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Para $i = 1,\ ...,\ n = 15$, ya que es el número de observaciones con el que se está construyendo el modelo. Y de este modo, la ecuación para cada una de las zonas son:

- **Ecuación para la zona A.**

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona B.**

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,2}) X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona C.**

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,3}) X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Literal B. La zona de referencia es la B.

En este caso será posible prescindir de la variable indicadora asociada a la zona B, ya que será implícito que se está tratando a esta zona cuando las variables indicadoras relacionadas con las zonas A y C sean ambas nulas, esto es, $I_1 = I_3 = 0$. Así, teniendo esto presente, se tiene que el modelo general va a estar dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i3} + \beta_{1,1} X_1 I_{i1} + \beta_{1,3} X_{i1} I_{i3} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Para $i = 1,\ ...,\ n = 15$, ya que es el número de observaciones con el que se está construyendo el modelo. Y de este modo, la ecuación para cada una de las zonas son:

- **Ecuación para la zona A.**

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona B.**

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona C.**

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,3})X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Literal C. La zona C es la de referencia.

En este caso será posible prescindir de la variable indicadora asociada a la zona C, ya que será implícito que se está tratando a esta zona cuando las variables indicadoras relacionadas con las zonas A y B sean ambas nulas, esto es, $I_1 = I_2 = 0$. Así, teniendo esto presente, se tiene que el modelo general va a estar dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_2 I_{i2} + \beta_{1,1} X_1 I_{i1} + \beta_{1,2} X_{i1} I_{i2} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Para $i = 1,\ ...,\ n = 15$, ya que es el número de observaciones con el que se está construyendo el modelo. Y de este modo, la ecuación para cada una de las zonas son:

- **Ecuación para la zona A.**

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona B.**

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

- **Ecuación para la zona C.**

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$$

Punto tres.

Tome los resultados del modelo ajustado en 2-a. Escriba la ecuación general ajustada así como las ecuaciones ajustadas para cada una de las tres zonas.

Se va a usar a la zona A como la zona de referencia, por lo que las ecuaciones generales y para cada zona serán como las mostradas en el literal ***a*** del segundo punto de este taller. Luego, con ayuda de R es posible calcular los coeficientes estimados para poder construir las ecuaciones ajustadas como sigue:

```r
modelo = lm(Precio~Valor*Zona)
summary(modelo)
```

```
##
## Call:
## lm(formula = Precio ~ Valor * Zona)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0874 -4.0521  0.1159  3.4029 15.6263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.3724    19.6945   1.593   0.1456
## Valor         0.2325     0.4098   0.567   0.5844
## ZonaB       -54.4846    31.9773  -1.704   0.1226
## ZonaC       -10.9211    68.6479  -0.159   0.8771
## Valor:ZonaB   1.3650     0.5235   2.607   0.0284 *
## Valor:ZonaC   0.7911     0.7226   1.095   0.3020
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.382 on 9 degrees of freedom
## Multiple R-squared:  0.9712, Adjusted R-squared:  0.9551
## F-statistic: 60.62 on 5 and 9 DF,  p-value: 1.172e-06
```

Y con esto, la ecuación general ajustada para el modelo es:

$$\hat{Y}_i = 31.3724 + 0.2352X_{i1} - 54.4846I_{i2} - 10.9211I_{i3} + 1.365X_{i1}I_{i2} + 0.7911X_{i1}I_{i3}$$

Luego, para cada una de las zonas, sus ecuaciones ajustadas son:

- ***Zona A.*** $\hat{Y} = 31.724 + 0.2352X_{i1}$
- ***Zona B.*** $\hat{Y} = -23.1122 + 1.6002X_{i1}$
- ***Zona C.*** $\hat{Y} = 20.4513 + 1.0263X_{i1}$

Punto cuatro.

De nuevo con base en los resultados del modelo ajustado en 2-a, ¿existe alguna razón para creer que las pendientes en las zonas A y B son diferentes? Realice el test de hipótesis para concluir.

Para determinar si las pendientes de las zonas A y B son diferentes, es decir, saber si el efecto promedio del valor catastral de las viviendas es diferente sobre sus valores comerciales son diferentes en las zonas A y B, se debe determinar si se cumple que $\beta_1 \neq \beta_1 + \beta_{1,2}$, lo cual, al restar β_1 a ambos lados de la inequidad, resulta equivalente a probar si se cumple que $\beta_{1,2} \neq 0$, lo cual se verifica a través de la siguiente prueba de hipótesis:

- $H_0: \ \beta_{1,2} = 0$. El efecto medio del valor catastral sobre el precio comercial es el mismo en las zonas A y B.

- $H_0: \ \beta_{1,2} \neq 0$. El efecto medio del valor catastral sobre el precio comercial es diferente en las zonas A y B.

Así, tomando un nivel de significancia de $\alpha = 0.05$, se tiene que el estadístico de prueba bajo H_0 está dado por:

$$T_0 = \frac{\widehat{\beta_{1,2}}}{s.e.[\widehat{\beta_{1,2}}]} \sim t(n-p-1)$$

Donde n y p se refieren al número de observaciones consideradas y al número de parámetros estimados respectivamente, por lo que en este caso se tiene que $n = 15$ y que $p = 5$, de manera que los grados de libertad del estadístico de prueba son $n - p - 1 = 15 - 5 - 1 = 9$. De esta forma, a partir de la información obtenida en la tabla resumen anterior, se tiene que:

$$T_0 = \frac{1.3650}{2.60745} = 0.5235$$

Ahora bien, para tomar una decisión se va a apelar al valor p, V_p, el cual puede ser calculado como sigue:

```
t0 = 1.3650 / 0.5235
vp = 1 - pt(t0, df = 9, lower.tail = TRUE) + pt(-t0, df = 9, lower.tail = TRUE)
```

$$V_p = P(|t(n-p-1)| > |T_0|) = 0.0284 < 0.05 = \alpha$$

Es decir, es muy poco probable haber obtenido los resultados presentados bajo la idea de el efecto medio del valor catastral sobre el precio comercial de las viviendas es el mismo para las zonas A y B, de forma tal que hay evidencia muestral suficiente para rechazar la hipótesis nula y, en consecuencia, se concluye que el efecto medio del valor catrastral sobre el precio comercial de las viviendas difiere entre las zonas A y B, lo cual se evidencia geométricamente en dos rectas con pendientes distintas para los modelos ajustados para cada una de estas zonas.

Punto cinco

Ajuste el modelo donde se considera que el efecto del valor de catastro sobre el precio de venta no difiere según la zona aunque el precio medio de venta es diferente en cada zona. Analice los residuales de este modelo.

Planteamiento del modelo

Bajo la suposición de este punto, será posible prescindir de los términos de interacción. De esta forma, conservando a la zona A como la de referencia, se tiene que la ecuación del modelo a ajustar está dada por:

$$Y_i = \beta_0 + \beta 1 X_{i1} + \beta_2 I_{i2} + \beta_3 I_{i3} + E_i, \quad E \overset{i.i.d.}{\sim} N(0, \ \sigma^2)$$

De forma tal que las ecuaciones para cada zona son como sigue:

- ***Zona A.*** $Y_i = \beta_0 + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0, \ \sigma^2)$
- ***Zona B.*** $Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0, \ \sigma^2)$

- ***Zona C.*** $Y_i = (\beta_0 + \beta_3) + \beta_1 X_{i1} + E_i, \quad E \overset{i.i.d.}{\sim} N(0,\ \sigma^2)$

Así, con ayuda de R es posible ajustar este modelo como sigue:

```
modelo2 = lm(Precio~Valor+Zona)
summary(modelo2)
```

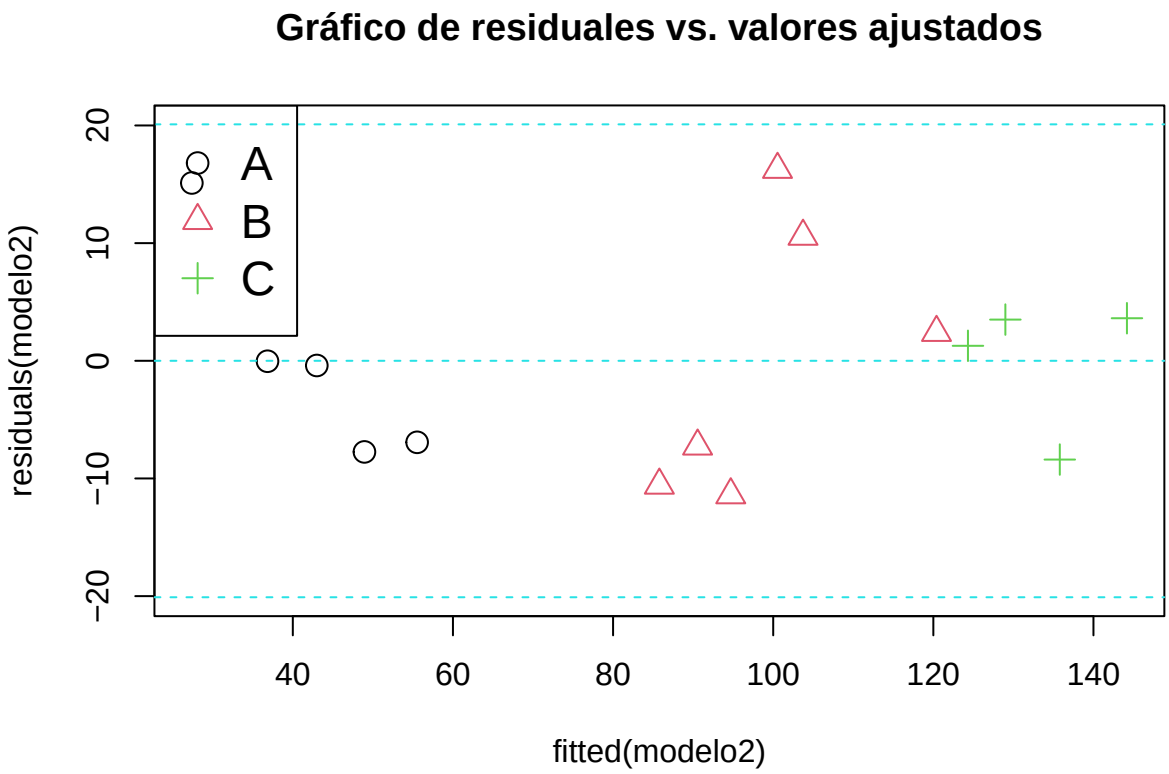
```
##
## Call:
## lm(formula = Precio ~ Valor + Zona)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3995  -7.5016  -0.0394   3.5597  16.2601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -7.760     13.996  -0.554  0.59039
## Valor          1.062      0.281   3.780  0.00305 **
## ZonaB         25.685     10.267   2.502  0.02942 *
## ZonaC         23.985     18.965   1.265  0.23211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.05 on 11 degrees of freedom
## Multiple R-squared:  0.9494, Adjusted R-squared:  0.9356
## F-statistic: 68.74 on 3 and 11 DF,  p-value: 2.069e-07
```

De manera que el modelo general ajustado está dado por:

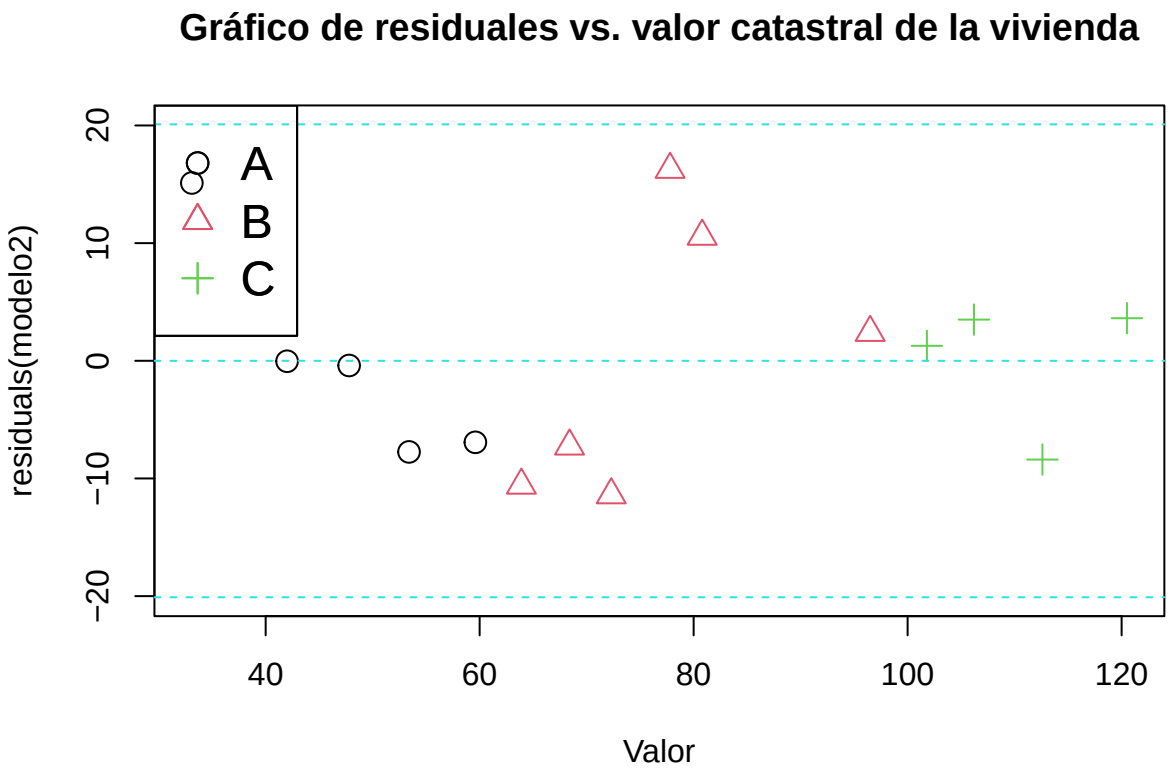
$$\hat{Y}_i = -7.76 + 1.062 X_i + 25.685 I_{i2} + 23.985 I_{i3}$$

Evaluación de los supuestos de homocedasticidad y ajuste

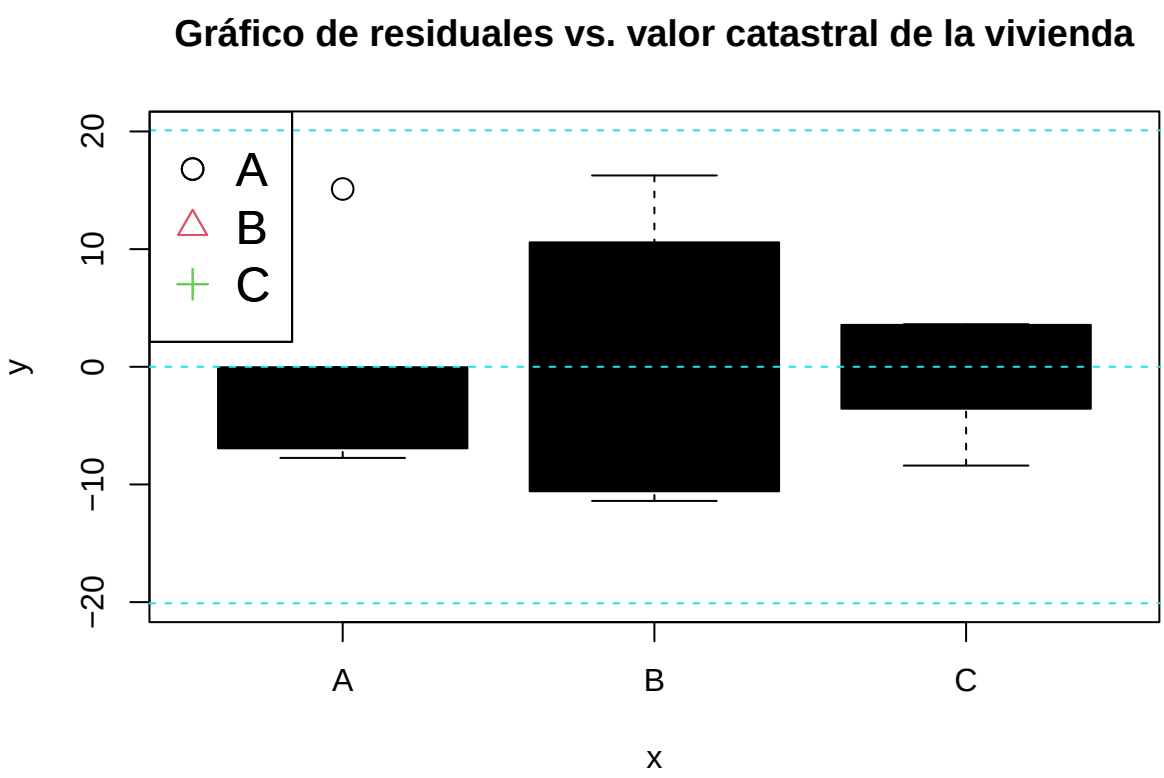
```
#Residuales vs. valores ajustados, con representación de las secciones
  plot(fitted(modelo2),residuals(modelo2),pch=as.numeric(Zona),
       col=as.numeric(Zona),cex=1.5,
       ylim=c(min(residuals(modelo2),-2*summary(modelo2)$sigma),
       max(residuals(modelo2),2*summary(modelo2)$sigma)))
abline(h=c(-2*summary(modelo2)$sigma,0,2*summary(modelo2)$sigma),lty=2,col=5)
legend("topleft",legend=c("A","B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
title("Gráfico de residuales vs. valores ajustados")
```




```
#Residuales vs. predictor cuantitativo X1, con representación de las secciones
plot(Valor,residuals(modelo2),pch=as.numeric(Zona),
     col=as.numeric(Zona),cex=1.5,
       ylim=c(min(residuals(modelo2),-2*summary(modelo2)$sigma),
       max(residuals(modelo2),2*summary(modelo2)$sigma)))
abline(h=c(-2*summary(modelo2)$sigma,0,2*summary(modelo2)$sigma),lty=2,col=5)
legend("topleft",legend=c("A","B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
legend("topleft",legend=c("A", "B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
title("Gráfico de residuales vs. valor catastral de la vivienda")
```




```
#Residuales vs. predictor cualitativo X2 (Zonas), con representación de las secciones
plot(Zona,residuals(modelo2),pch=as.numeric(Zona),
     col=as.numeric(Zona),cex=1.5,
       ylim=c(min(residuals(modelo2),-2*summary(modelo2)$sigma),
       max(residuals(modelo2),2*summary(modelo2)$sigma)))
abline(h=c(-2*summary(modelo2)$sigma,0,2*summary(modelo2)$sigma),lty=2,col=5)
legend("topleft",legend=c("A","B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
legend("topleft",legend=c("A", "B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
title("Gráfico de residuales vs. valor catastral de la vivienda")
```



A partir de los tres gráficos anteriores se pueden percibir varios problemas asociados con los residuales. Para comenzar, se puede observar el tercer gráfico el cual contrasta los residuales para las observaciones según su zona, y se destaca el hecho de que el gráfico de cajas y bigotes asociado a la zona A no es simétrico para valores positivos y negativos de los residuales internamente estudentizados, y que hay evidencia según los boxplots para sugerir que el valor de los residuales tiene dispersiones diferentes para cada zona, estando relativamente concentrados para la zona A, muy dispersos para la zona B y nuevamente concetrados para la zona C, lo cual indica que no se cumple el supuesto de homocedasticidad.

Asimismo, vale destacar en la gráfica de residuales internamente estudentizados contra el valor catastral que, para el caso de las viviendas de la zona A, hay una tendencia en los residuales, lo cual contrasta con el aspecto estocástico que se esperaría para este gráfico de residuales, de tal suerte que se puede pensar que hay problema de ajuste en el modelo.

Evaluación del supuesto de normalidad

Ahora, se va a verificar si los residuales de distribuyen siguiendo una normal con media nula y varianza $sigma^2$, para lo cual se va a recurrir al test de Shapiro-Wilks y a un gráfico cuantil-cuantil, lo cual es lícito considerando que en el enunciado se habla de que los datos fueron recolectados en el marco de un muestreo aleatorio, de tal forma que es posible asumir la independencia de los errores. Así, vale la pena comenzar definiendo el siguiente par de hipótesis:

H_0. $e_i \sim Normal, i = 1, 2, ..., 15$ H_1. $e_i \nsim Normal, i = 1, 2, ..., 15$

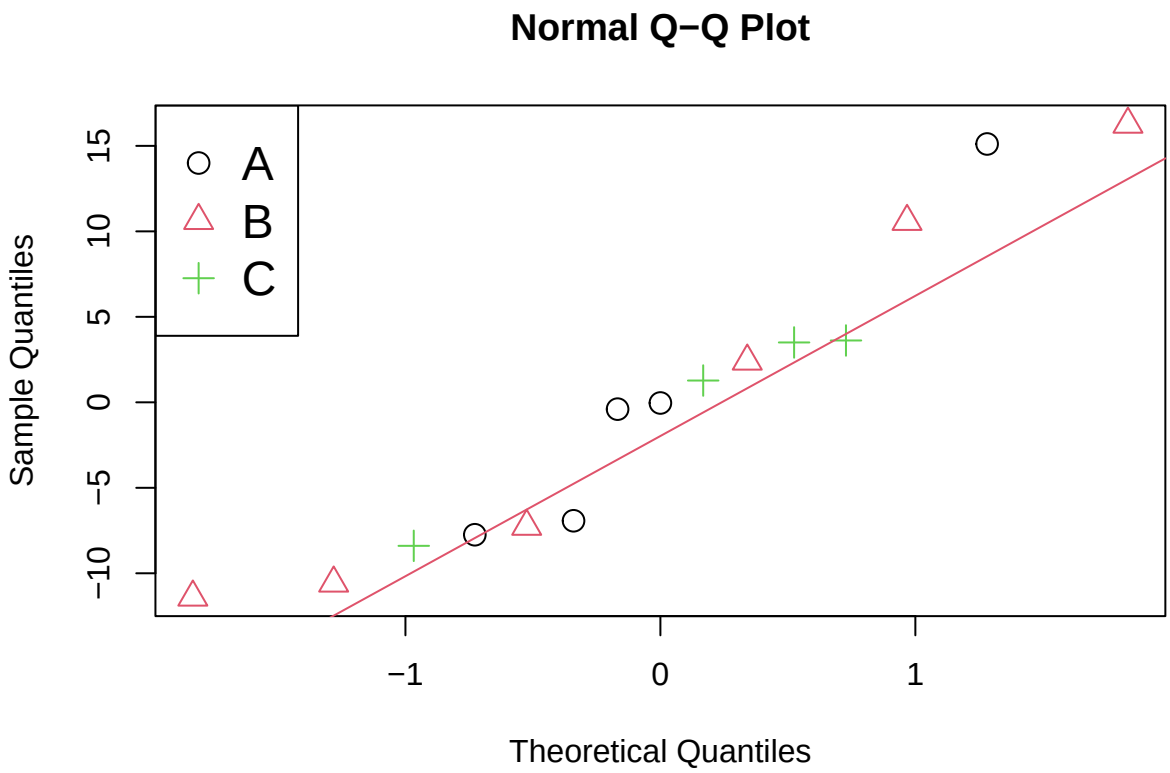
```
shapiro.test(residuals(modelo2))
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo2)
## W = 0.91959, p-value = 0.1899
```

Y como se observa, se tiene que el valor p asociado a este test de normalidad es $V_p = 0.1899 > 0.05 = \alpha$, por lo que no se rechaza la hipótesis nula, lo que quiere decir que hay evidencia muestral suficiente para sugerir que los residuales del modelo ajustado siguen una distribución normal. No obstante, hay que recordar que la prueba de Shapiro-Wilk pierde potencia conforme se tienen menos datos y que de hecho esta fue construida para evaluar normalidad cuando se tienen treinta o más datos, de tal forma que es buena idea apelar al gráfico de cuantil-cuantil

```
qqnorm(residuals(modelo2),pch=as.numeric(Zona),
       col=as.numeric(Zona),cex=1.5)
qqline(residuals(modelo2),col=2)
legend("topleft",legend=c("A","B","C"),pch=c(1:3),col=c(1:3),cex=1.5)
```



Y como se puede observar, hay un pobre ajuste entre los cuantiles teóricos y los empíricos, lo cual se evidencia en el hecho de que los diferentes puntos que aparecen graficados en el esquema anterior no se ubican sobre la recta de cuantiles teóricos de una distribución normal, por lo que finalmente se concluye que los residuales no siguen una distribución normal y, en consecuencia, se invalida este supuesto.