

Taller RLM 2

Sofía Cuartas García

Simón Cuartas Rendón

Julián Alejandro Úsuga Ortiz

Deivid Zhang Figueroa

Enero de 2022

Punto uno. Descripción de la base de datos.

La base de datos consta de 205 observaciones de las cuales solo tendremos en cuenta 100, esta contiene variables continuas y categóricas; las observaciones corresponden a características de automóviles como lo pueden ser sus características físicas que en este caso serían variables continuas por nombrar algunas tenemos: largo, ancho y altura del automóvil; también se le caracterizan según: su fabricante, que tan susceptible es a pérdidas y/o daños, número de puertas, entre otras categorías. Lo que nos interesa es buscar si el *precio* del auto puede ser explicado por la variable continua *City-mpg* y la variable categórica *body-style*, de las cuales haremos una breve descripción a continuación:

City-mpg. Las millas por galón es una medida de la distancia promedio recorrida por unidad de energía consumida, puede ser utilizada para comparar el consumo de combustible entre automóviles. Los valores observados de esta variable están entre 13 mi/gal.us y 49 mi/gal.us

Body-style. Esta variable clasifica el automóvil según la forma de su carrocería, tendrá tres niveles de medición que se definen como:

- *Hatchback* En términos de estilo automotriz son vehículos diseñados en dos pilares, en comparación con la mayoría de los sedan estos tienen más capacidad de almacenamiento y tienen un menor consumo de combustible.
- *Sedan* Son vehículos diseñados en tres pilares, en comparación con los hatchback son más estables a altas velocidades y el espacio para los pasajeros es más generoso.
- *Otro* En la base de datos se habla también de las carrocerías hardtop, wagon y convertibles pero por facilidad el nivel “Otro” estará constituido por estas tres.

Price Precio de los autos, será nuestra variable de respuesta, los valores observados de esta variable toma valores entre 5118 y 45400 dólares estadounidenses.

Punto dos. Análisis descriptivo.

Parámetro	Valor
Media	13443
Desviación estándar	9163.297
Mínimo	5151
Primer cuantil (Q1)	7254
Mediana (Q2)	9754
Tercer cuantil (Q3)	16500

Parámetro	Valor
Máximo (Q4)	45400
Rango intercuartílico	9246.5
Coefficiente de asimetría	1.735634
Curtosis	5.394072

Como se observa en la tabla anterior el precio promedio de los automóviles de la muestra es de 13443, con una desviación estándar de 9163.297, también podemos ver que el auto de menor precio de la muestra tiene un precio de 5151 y el de mayor precio es de 45400. Por otro lado, el rango intercuartílico es de 9246.5 por lo que podemos anticipar que los datos centrales tienen una fuerte dispersión. Destaca que la curtosis es de 5.394072 lo que nos indica que la distribución de nuestra variable tiene colas más pesadas que la distribución normal, coherente con esto tenemos un coeficiente de asimetría de 1.735634 un indicio de que la distribución de la variable es asimétrica hacia la derecha, para ver esto mejor realizaremos el respectivo histograma

Histograma para para el precio del automovil

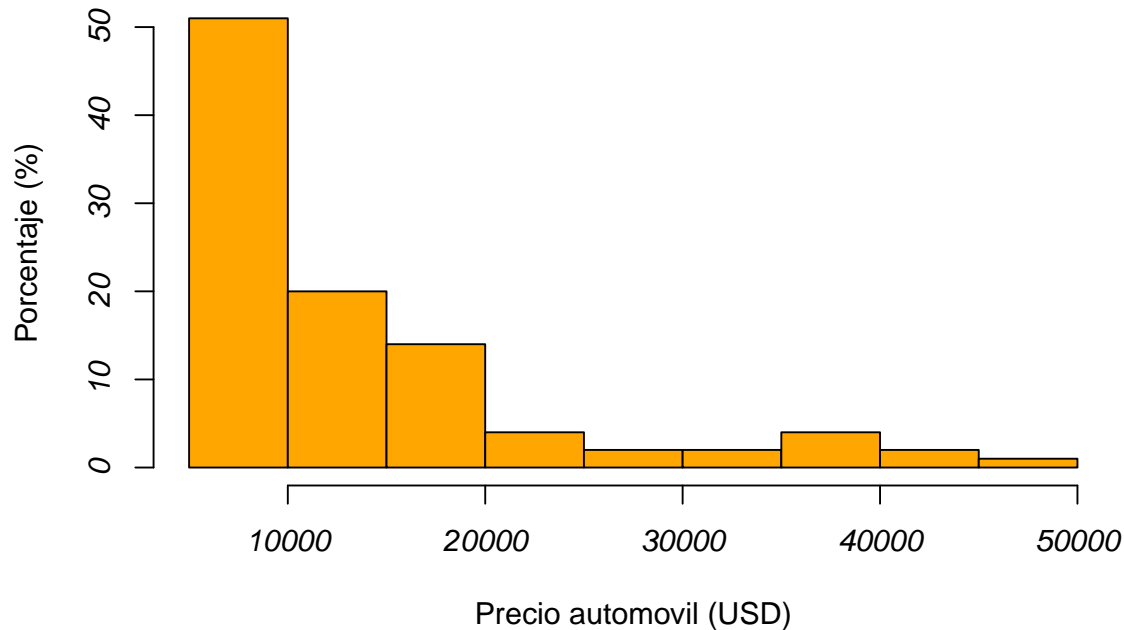


Gráfico de dispersión de millas recorridas por galón en área urbana contra precio

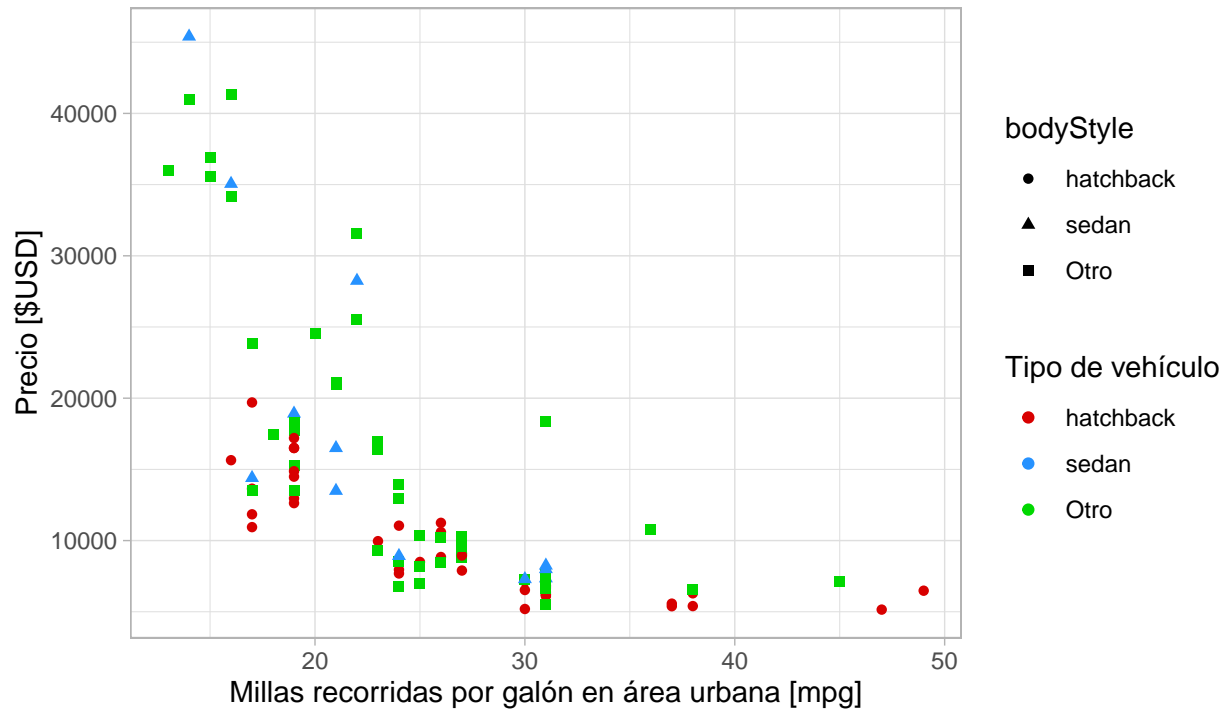


Figura XXXX.

El gráfico evidencia que entre el precio y las millas por galón parece existir un relación lineal negativa para todos los tipos de vehículo, lo que era de esperarse ya que los coches pequeños suelen ser los más baratos y estos tienden a tener mejores Mpg; ahora, si discriminamos por el tipo de vehículo podemos ver que las observaciones pertenecientes a vehículos hard top, wagon y convertibles que fueron las que agrupamos en la categoría “otro” tienen la pendiente con inclinación más elevada, de estas observaciones también podemos comentar que son bastante dispersas esto puede deberse a que en esta categoría están mezclados distintos tipos de vehículos; pasando a las observaciones pertenecientes a la categoría “hatchback” podemos comentar que su pendiente es por mucho la menos inclinada además estas observaciones tienen poca dispersión; la categoría “sedan” parecería tener una pendiente ni tan plana como la de “hatchback” ni tan pronunciada como la de “otro” de no ser porque algunas observaciones alcanzan precios bastante altos, debido a esto último podemos decir que el intercepto con el “Eje Y” para “sedan” y para “otro” podrían ser iguales y de este podríamos esperar que sea un valor alto en comparación con el intercepto de las observación tipo “hatchback”.

Punto tres. Planteamiento del modelo de regresión lineal.

Como se vio en el punto uno, la base de datos asignada está compuesta por tres variables: `price`, que corresponde a la variable respuesta del modelo a plantear y que será denotada como Y ; `cityMpg`, que es una variable cuantitativa y que será una variable regresora que será denotada como X_1 y, finalmente, `bodyStyle`, denotada como X_2 , es la otra variable regresora, y de la cual se debe tener en cuenta que es cuantitativa, de manera que para la realización del modelo de regresión se deben tener en cuenta sus tres niveles de medición: `hatchback`, `sedan` y `Otro`, y para ello se define al nivel `Otros` como el de referencia. Para definir cómo se va a plantear el modelo de regresión, esto es, si deben ser incluidos términos de interacción en el modelo, vale la pena observar la figura XXXX, en la cual se muestra la relación que existe entre `price` y `cityMpg` diferenciando los niveles de `bodyStyle` con diferentes figuras.

Como se observa en el gráfico anterior, existen tendencias diferentes para cada uno de los niveles de medición.

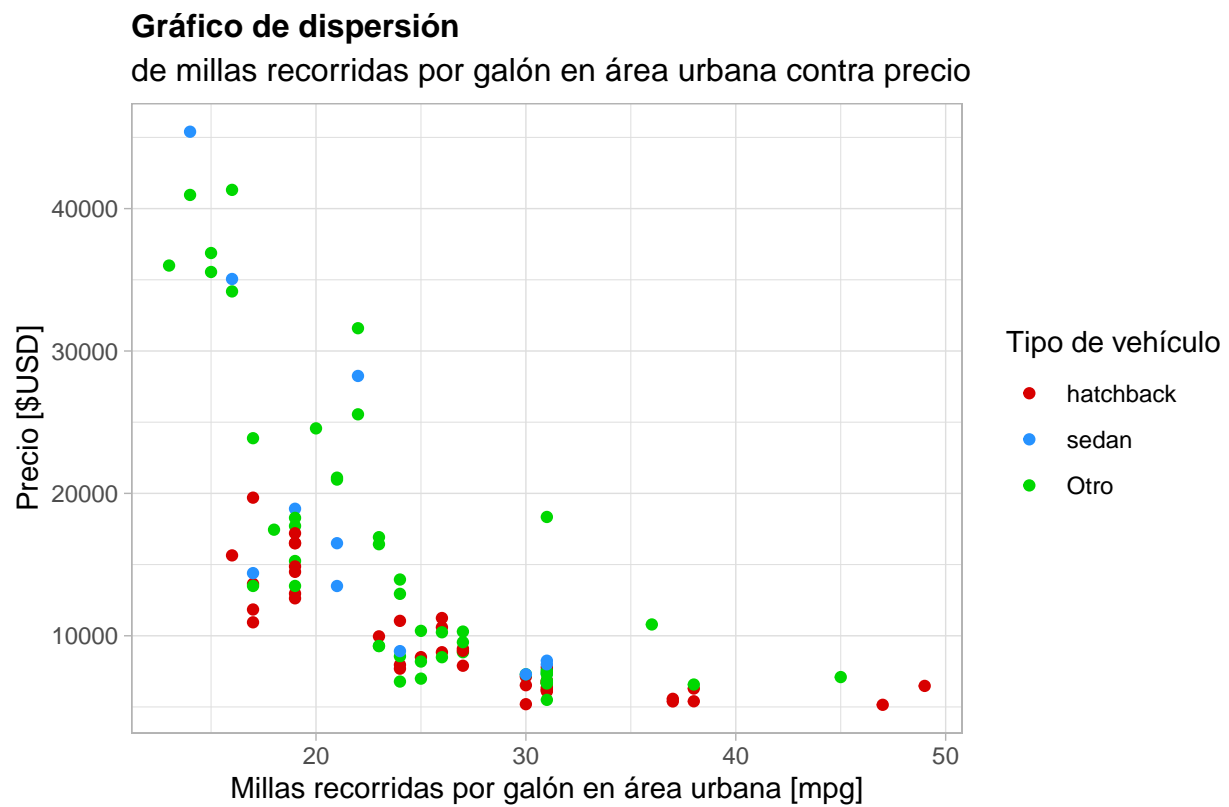


Figura XXXX.

Figure 1: Dispersión de mpg vs. precio.

Para empezar, las observaciones asociadas a vehículos catalogados como **Otro** parece tener una pendiente bastante pronunciada, lo cual se refleja en el hecho de que existen varias observaciones que toman altos precios, lo cual contrasta con lo que sucede con los otros dos niveles: **hatchback** y **sedan**, que no toman valores tan altos de precios, salvo algunas pocas excepciones para el nivel **sedan**. Además, debe tenerse en cuenta que hay algunas observaciones del nivel **Otro** que tiene un comportamiento semejante al de **hatchback** y **sedan**, lo cual se puede deber al hecho de que esta variable recoge distintos tipos de vehículos, por lo que es posible que existan algunos de estos tipos que tengan precios bajos.

En todo caso, teniendo en cuenta este análisis, es necesario incluir variables indicadoras al modelo. Así, tomando al nivel **Otro** como el de referencia para **bodyStyle**, I_1 la variable indicadora asociada a los *hatchbacks* e I_2 la variable indicadora que referencia a los sedanes, entonces el modelo a desarrollar sería como sigue:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} X_{i1} I_{i1} + \beta_{1,2} X_{i1} I_{i2} + E_i, \quad [1]$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

De manera que pueden ser planteados tres modelos diferentes en función del nivel de la variable categórica **bodyShape** considerado, a saber:

- Para los automóviles catalogados como **Otro** ($I_1 = I_2 = 0$):

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i \quad [2]$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

- Para los automóviles catalogados como **hatchback** ($I_1 = 1 \wedge I_2 = 0$):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 + \beta_{1,1} X_{i1} + E_i$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

Lo cual equivale a:

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1}) X_{i1} + E_i \quad [3]$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

- Para los automóviles catalogados como **sedan** ($I_1 = 0 \wedge I_2 = 1$):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_3 + \beta_{1,2} X_{i1} + E_i$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

Lo cual equivale a:

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2}) X_{i1} + E_i \quad [4]$$

$$E_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2), i = 1, 2, \dots, 100$$

Definido esto, se puede proceder a obtener el modelo con ayuda de [R](#) y el resultado es el que se visualiza enseguida:

```
##
## Call:
## lm(formula = price ~ cityMpg * bodyStyle, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12713.3  -2919.2   -907.8   2025.3  15997.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      43564.3      3169.3   13.746 < 2e-16 ***
## cityMpg          -1140.4       124.6   -9.150 1.18e-14 ***
## bodyStylehatchback -23252.4      4525.6   -5.138 1.50e-06 ***
## bodyStylessedan    9983.4       7305.3    1.367  0.175
## cityMpg:bodyStylehatchback  744.5       169.5    4.393 2.93e-05 ***
## cityMpg:bodyStylessedan   -414.6       302.8   -1.369  0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5742 on 94 degrees of freedom
## Multiple R-squared:  0.6272, Adjusted R-squared:  0.6074
## F-statistic: 31.63 on 5 and 94 DF,  p-value: < 2.2e-16
```

A partir de la tabla anterior se tiene que el vector de coeficientes está dado por:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \widehat{\beta_{1,1}} \\ \widehat{\beta_{1,2}} \end{pmatrix} = \begin{pmatrix} 43,564.3 \\ -1,140.4 \\ -23,252.4 \\ 9,983.4 \\ 744.5 \\ -414.6 \end{pmatrix}$$

pueden obtener entonces las rectas ajustadas para cada uno de los niveles de **bodyShape** como sigue:

- Para los vehículos catalogados como **Other**, reemplazando en [2]:

$$\hat{Y}_i = 43564.3 - 1140.4X_{i1}, \quad i = 1, \dots, 100 \quad [\hat{2}]$$

- Para los vehículos catalogados como **hatchback**, reemplazando en [3]:

$$\hat{Y}_i = 20311.9 - 395.9X_{i1}, \quad i = 1, \dots, 100 \quad [\hat{3}]$$

* Para los vehículos etiquetados como **sedan**, reemplazando en [4]:

$$\hat{Y}_i = 53547.7 - 1555X_{i1}, \quad i = 1, \dots, 100 \quad [\hat{4}]$$

Además, debe notarse que el $R^2 = 0.6272$, lo que significa que el 62.72 % de la variabilidad del precio es explicada por el tipo de vehículo empleado y la cantidad de millas recorridas por galón de combustible consumido en una área urbana.

Punto cuatro. Supuesto de normalidad y varianza constante.

Uno de los supuestos asumidos en el punto anterior para poder plantear los modelos presentados es que los errores tienen la misma distribución, siendo esta normal con media cero y varianza σ^2 , por lo que es necesario validar este último aspecto para comprobar los modelos propuestos, para lo cual vale la pena revisar los siguientes dos gráficos, los cuales contrastan a los residuos internamente estudentizados contra el número de millas recorridas por galón consumida en recorridos urbanos y contra los valores ajustados.

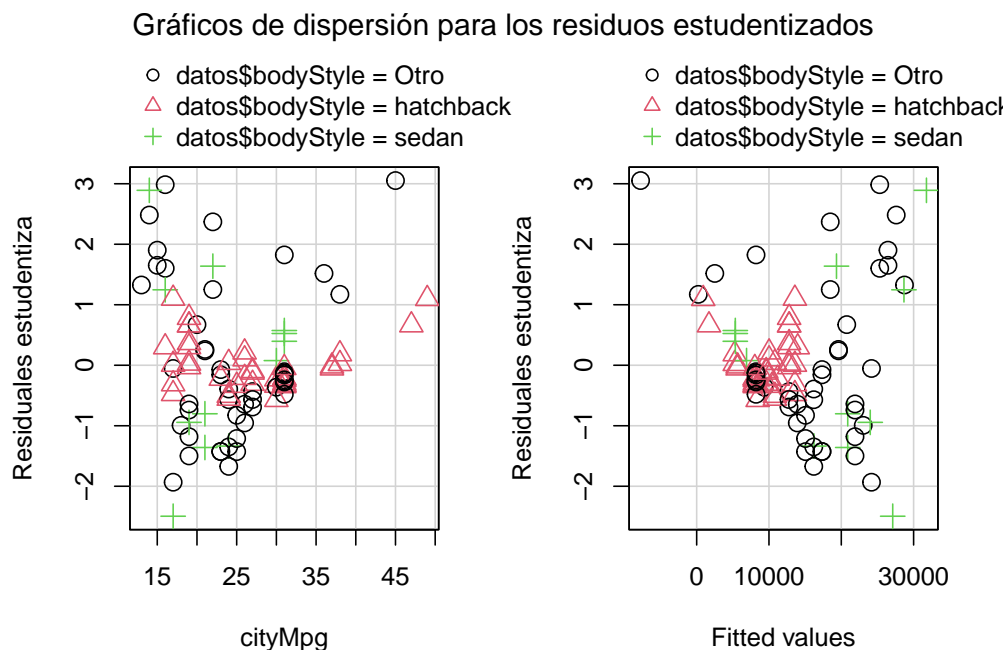


Figure 2: Residuos estudentizados.

Comenzando con el gráfico de la izquierda, que grafica a las millas recorridas por galón en la ciudad contra los residuos estudentizados, se puede ver que, en general, no parece haber alguna distribución particular en cada uno de los puntos, salvo que tienen una acumulación hacia bajos valores de millas recorridas por galón, lo cual se anticipaba del gráfico de la figura XXXX, donde la mayor concentración de puntos se da en valores bajos de dicha variable. No obstante, al revisar para cada tipo de vehículo, resaltan los triángulos rojos asociados a los `hatchback`, que parecen tener una tendencia parabólica y cóncava hacia arriba, lo cual muestra que el mejor modelo pudo haber sido uno de orden cuadrático, si bien este patrón no se repite con los vehículos etiquetados como `Otro` ni como `sedan`. Cabe destacarse también que pueden existir valores atípicos, ya que en el gráfico se ven algunas observaciones de `hatchbacks` en el margen derecho.

Después, al pasar al gráfico de la derecha, donde se esquematiza un gráfico de dispersse observa que tampoco hay una distribución particular de los diferentes puntos, aunque se puede diferenciar que los vehículos que son clasificados como `sedan` o como `Otro` tienen una mayor variabilidad de valores ajustados y residuos internamente estudentizados, mientras que los `hatchback` están fuertemente concentrados en precios ajustados alrededor de los diez mil dólares estadounidenses (USD\$ 10,000.00) y sus residuales estudentizados se agrupan alrededor de valores negativos cercanos al cero. Además, también llama la atención que existe un valor ajustado asociado a un vehículo tipo `Otro` negativo, lo cual en términos prácticos no tiene sentido (¿se debe pagar al comprador para que se quede con el vehículo?).

PREGUNTA. ¿Qué se hace con el valor ajustado negativo?

A partir de lo anterior se puede concluir que no hay heterocedasticidad, por lo que se cumple el primer supuesto que se quería verificar, si bien se debe detallar que existe una observación potencialmente atípicas,

dado que su residual internamente estudentizado es mayor que tres. Asimismo, se tienen varias observaciones que tienen residuales internamente estudentizados mayores a dos y menores a menos dos, lo cual da cuenta de que existen varias observaciones potencialmente influenciadas, siendo cuatro de ellas asociadas a vehículos tipo **Otro** y dos a **sedan**. Además, debe tenerse en cuenta que el modelo propuesto presenta un problema de ajuste, como se explicó antes.

Por otro lado, es útil analizar descriptivamente a los residuos internamente estudentizados, para lo cual vale la pena analizar la figura XXXX que contraste tres boxplots para estos valores según el tipo de vehículo:

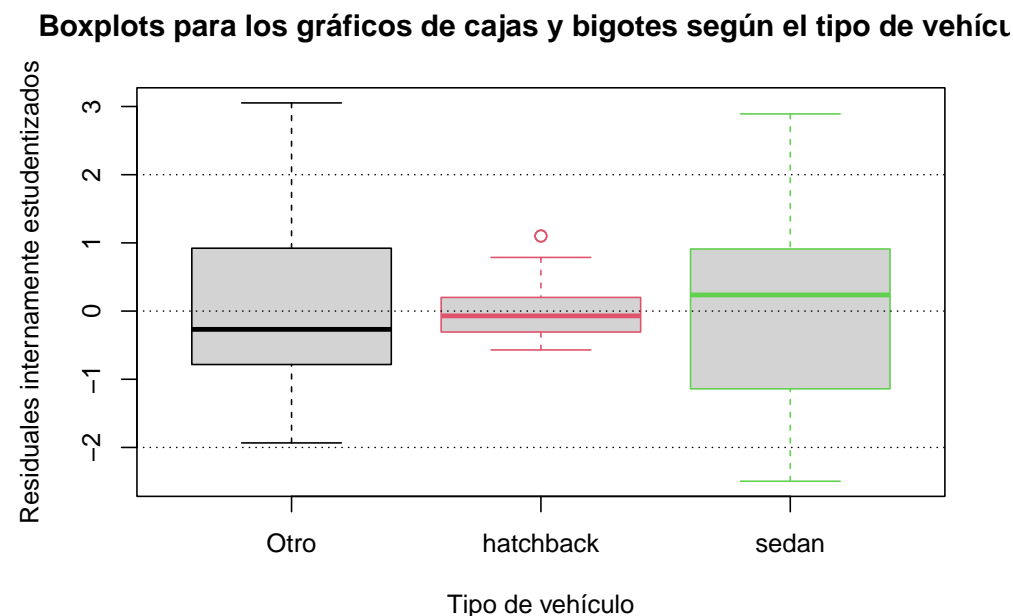


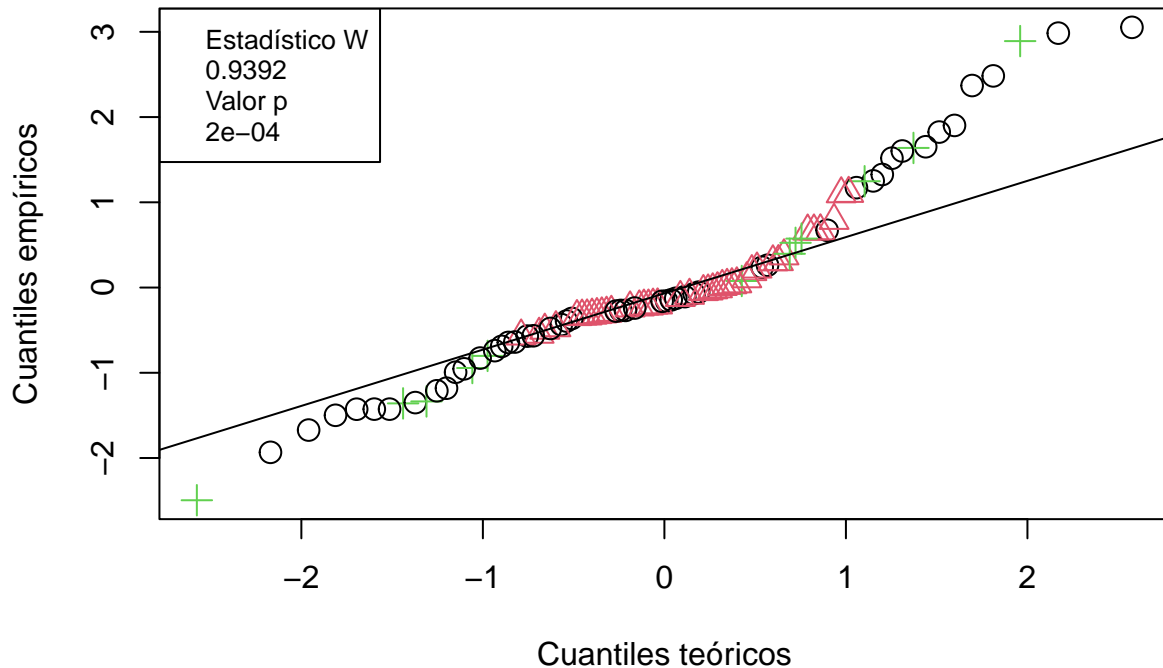
Figure 3: Gráficos de cajas y bigotes para los residuales estudentizados según el tipo de vehículo.

De la figura XXXX lo primero que llama la atención es que los vehículos tipo **hatchback** poseen una caja muy angosta, lo cual está asociado con la concentración de los residuales estudentizados que se analizó en el gráfico derecho de la figura XXXX, además debe notarse que su mediana es negativa, lo cual se anticipaba al ver que los puntos se concentraban debajo del cero. Nótese, además, que los bigotes no superan los umbrales demarcados en dos y dos negativo, y que existe un residual estudentizado atípico en el contexto de este tipo de vehículos.

Por otro lado, se tiene que la situación con los vehículos tipo **Otro** o **sedan** es diferente, ya que hay una mayor dispersión en los valores de sus residuales internamente estudentizados, lo que se refleja en sus cajas más anchas, y que hay bigotes que superan el umbral de dos positivo y dos negativo. En el caso de los vehículos tipo **Otro**, se ve que hay un bigote que supera el valor de dos y se aproxima al de tres, mientras que con los sedanas se tienen a los dos bigotes superando los umbrales demarcados, pero sin llegar a tres positivo o negativo, lo que se corresponde con el análisis realizado a partir de la figura XXXX con diagramas de dispersión.

Luego, se debe verificar la normalidad de los residuales, ya que como se mencionó, se partió del supuesto de que los errores son normales, para lo cual vale la pena observar el siguiente QQ plot:

QQ pplot para normalidad



Y como se puede observar, existe una amplia concentración de valores alrededor de los cuantiles teóricos cuando se está cerca a la media, es decir, a cero, pero este no es el caso en las colas, ya que los cuantiles empíricos se alejan de los teóricos, sobre todo con la cola superior, y esto sucede fundamentalmente con los vehículos tipo **Otro** y con algunos tipo **sedan**, por lo que es razonable pensar que los vehículos **hatchback**, cuyo modelo ajustado está dado por la ecuación $\hat{3}$ sí cumplen con el supuesto de normalidad, mientras que los demás presentan dificultades al tener distribuciones con colas más pesadas. Además, se realizó un test de Shapiro-Wilk para comprar la normalidad para las cien observaciones, con las siguientes hipótesis:

- H_0 : Los residuales tienen una distribución normal.
- H_1 : Los residuales **no** tienen una distribución normal.

Y sus resultados se pueden ver en el margen superior izquierdo de la gráfica, donde se ve que el valor p es $V_p = 2 \times 10^{-4}$, lo cual es menor que cualquier α , y en particular, que un $\alpha = 0.05 > V_p$, de manera que se rechaza la hipótesis nula y se concluye que no hay evidencia muestral suficiente para sugerir que los errores tienen una distribución normal.

PREGUNTA. ¿Concluyo sobre errores o residuales? ¿Y es correcto decir que uno de los modelos cumple y los otros no?