Modelo de regresión lineal múltiple

del precio de vehículos de 1985 en Estados Unidos a partir del número de millas recorridas por galón en zonas urbanas y del tipo de vehículo

Sofía Cuartas García   Simón Cuartas Rendón
Julian Alejandro Úsuga Ortiz   Deivid Zhang Figueroa

Universidad Nacional de Colombia

Febrero 2 de 2022

Contenidos

1. Contexto
   - Objetivo
2. 1. Descripción de las variables
3. 2. Análisis descriptivo
   - Resúmenes numéricos
   - Gráficos
4. 3. Modelo de regresión lineal
5. 4. Verificación de supuestos
6. 5. Diferencias entre los interceptos para cada tipo de vehículo
7. 6. Diferencia entre las pendientes para cada tipo de vehículo
8. 7. Prueba de cuadrados extra con test lineal
9. 8. ¿Es la recta de regresión diferente para cada tipo de vehículo?

Base de datos

Se tiene una base de datos que contiene **veintiséis características** de 205 vehículos importados a los Estados Unidos en 1985.

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R*

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.
Algunas de las variables de esta base de datos son:

- Largo

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.

Algunas de las variables de esta base de datos son:

- Largo
- Ancho

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.
Algunas de las variables de esta base de datos son:

- Largo
- Ancho
- Altura

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.

Algunas de las variables de esta base de datos son:

- Largo
- Ancho
- Altura
- Fabricante

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.

Algunas de las variables de esta base de datos son:

- Largo
- Ancho
- Altura
- Fabricante
- Susceptibilidad a daños y/o pérdidas

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.
Algunas de las variables de esta base de datos son:

- Largo
- Ancho
- Altura
- Fabricante
- Susceptibilidad a daños y/o pérdidas
- Tipo de combustible

Contexto

Esta base de datos se encuentra en el paquete `randomForest` de *R* y fue publicada por la Universidad de California en Irvine (UCI) en su repositorio de bases de datos para técnicas de aprendizaje automatizado.

Algunas de las variables de esta base de datos son:

- Largo
- Ancho
- Altura
- Fabricante
- Susceptibilidad a daños y/o pérdidas
- Tipo de combustible
- Etcétera

Objetivo

¿Qué se va a hacer?

Se quiere plantear un modelo de regresión lineal múltiple que permita obtener el **precio** de un vehículo como función de solo dos características y empleando solo las primeras cien observaciones de la base de datos.

Variables a emplear

Las variables a usar son:

Precio

Denominada `price`. **Continua.** Es el valor en dólares estadounidenses ($USD) de un vehículo.

Variables a emplear

Las variables a usar son:

Precio

Denominada `price`. **Continua.** Es el valor en dólares estadounidenses ($USD) de un vehículo.

Millas recorridas por galón en zona urbana

Denominada `cityMpg`. **Continua.** Es la cantidad de millas que un vehículo puede recorrer en una zona urbana de Estados Unidos consumiendo un galón de combustible.

Variables a emplear

Las variables a usar son:

Precio

Denominada `price`. **Continua.** Es el valor en dólares estadounidenses ($USD) de un vehículo.

Millas recorridas por galón en zona urbana

Denominada `cityMpg`. **Continua.** Es la cantidad de millas que un vehículo puede recorrer en una zona urbana de Estados Unidos consumiendo un galón de combustible.

Tipo de vehículo

Denominada `bodyStyle`. **Categórica.** Describe el tipo de carrocería del vehículo.

Estadísticos de resumen

Estadístico	**Valor**
Media	13443
Desviación estándar	9163.297
Mínimo	5151
Primer cuantil (Q1)	7254
Mediana (Q2)	9754
Tercer cuantil (Q3)	16500
Máximo (Q4)	45400
Rango intercuartílico	9246.5
Coeficiente de asimetría	1.736
Curtosis	5.394

Histograma para la variable 'precio'

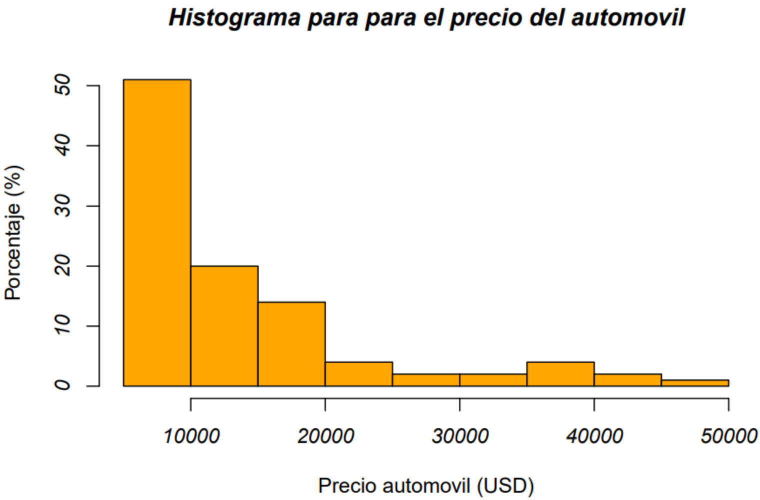
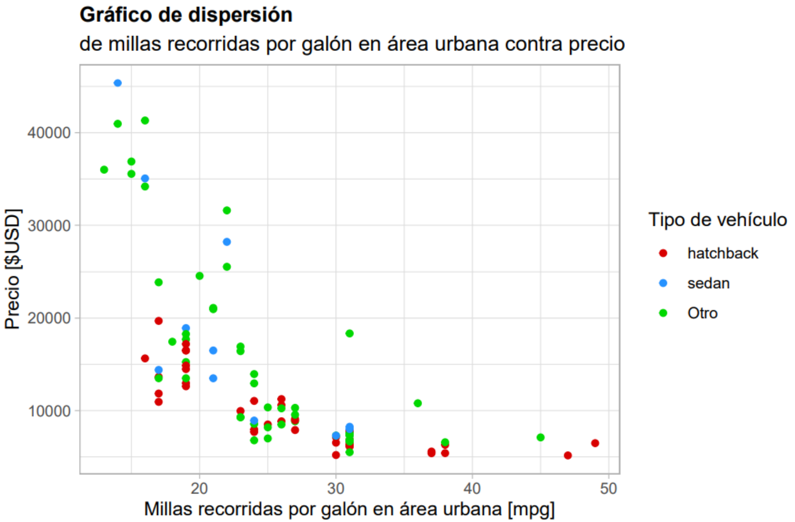


Gráfico de dispersión



Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**. Para ello, se va a tomar como nivel de referencia a los vehículos tipo **Otro**.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**. Para ello, se va a tomar como nivel de referencia a los vehículos tipo **Otro**.

- I_{i1}. Vehículo tipo **hatchback**.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**. Para ello, se va a tomar como nivel de referencia a los vehículos tipo **Otro**.

- I_{i1}. Vehículo tipo **hatchback**.
- I_{i2}. Vehículo tipo **sedán**.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**. Para ello, se va a tomar como nivel de referencia a los vehículos tipo **Otro**.

- I_{i1}. Vehículo tipo **hatchback**.
- I_{i2}. Vehículo tipo **sedán**.
- E_i. Error aleatorio.

Notación

Para desarrollar el modelo de RLM, se va a emplear la siguiente notación para las variables cuantitativas:

- Y_i. Precio del i-ésimo vehículo.
- X_{i1}. Número de millas recorridas por galón de combustible del i-ésimo vehículo en zona urbana.

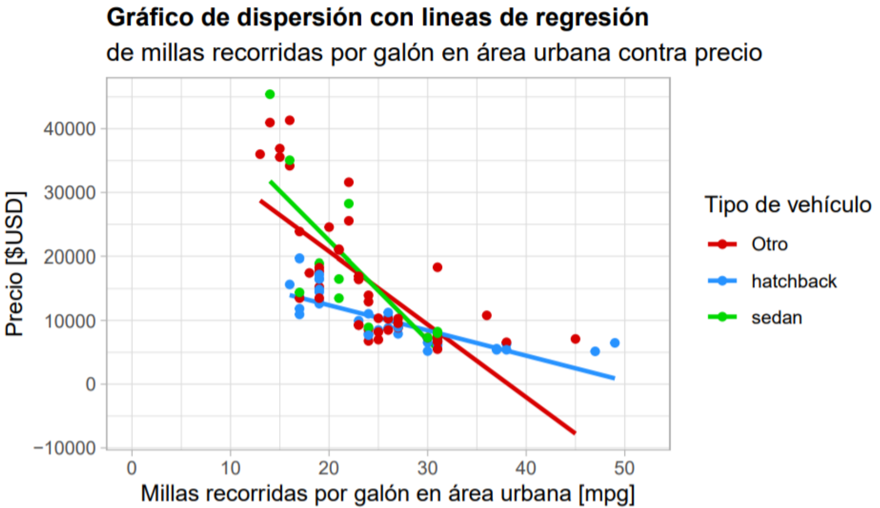
Variable categórica

Se tiene una variable categórica: **tipo de vehículo**. Para poder trabajar con ella, es necesario emplear **variables indicadoras**. Para ello, se va a tomar como nivel de referencia a los vehículos tipo **Otro**.

- I_{i1}. Vehículo tipo **hatchback**.
- I_{i2}. Vehículo tipo **sedán**.
- E_i. Error aleatorio.

$$i = 1,\ 2,\ \ldots,\ 100$$

Tipo de modelo a plantear



Modelo

Así, el modelo a ajustar es:

Modelo

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

Modelo

Así, el modelo a ajustar es:

Modelo

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

Modelo

Así, el modelo a ajustar es:

Modelo

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

No obstante, se debe tener en cuenta que este modelo variará en función del **tipo de vehículo**.

Modelo según el tipo de vehículo

Otro

$I_1 = I_2 = 0$

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i$$

Modelo según el tipo de vehículo

Otro

$I_1 = I_2 = 0$

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i$$

Hatchback

$I_1 = 1,\ \ I_2 = 0$

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

Modelo según el tipo de vehículo

Otro

$I_1 = I_2 = 0$

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i$$

Hatchback

$I_1 = 1,\ I_2 = 0$

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

Sedan

$I_1 = 0,\ I_2 = 1$

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

$$E_i \overset{i.i.d}{\sim} Normal(0,\ \sigma^2)$$

Modelo ajustado

Coe.	Término	Estim.	Error std.	Est. T	Valor P
β_0	Intercepto	43564.3	3163.3	13.746	$<$2e-16
β_1	MPG	-1140.4	124.6	-9.150	1.18e-14
β_2	Hatchback	-23252.4	4525.6	-5.138	1.50e-6
β_3	Sedán	9983.4	7305.3	1.367	0.175
$\beta_{1,1}$	MPG - Hatchback	744.5	169.5	4.393	2.93e-5
$\beta_{1,2}$	MPG - Sedán	-414.6	302.8	-1.369	0.174

Modelo ajustado

Coe.	Término	Estim.	Error std.	Est. T	Valor P
β_0	Intercepto	43564.3	3163.3	13.746	$<$2e-16
β_1	MPG	-1140.4	124.6	-9.150	1.18e-14
β_2	Hatchback	-23252.4	4525.6	-5.138	1.50e-6
β_3	Sedán	9983.4	7305.3	1.367	0.175
$\beta_{1,1}$	MPG - Hatchback	744.5	169.5	4.393	2.93e-5
$\beta_{1,2}$	MPG - Sedán	-414.6	302.8	-1.369	0.174

$$R^2 = 0.6272$$

Modelo ajustado

Coe.	Término	Estim.	Error std.	Est. T	Valor P
β_0	Intercepto	43564.3	3163.3	13.746	$<$2e-16
β_1	MPG	-1140.4	124.6	-9.150	1.18e-14
β_2	Hatchback	-23252.4	4525.6	-5.138	1.50e-6
β_3	Sedán	9983.4	7305.3	1.367	0.175
$\beta_{1,1}$	MPG - Hatchback	744.5	169.5	4.393	2.93e-5
$\beta_{1,2}$	MPG - Sedán	-414.6	302.8	-1.369	0.174

$$R^2 = 0.6272$$
$$R^2_{adj} = 0.6074$$

Modelo ajustado

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \widehat{\beta_{1,1}} \\ \widehat{\beta_{1,2}} \end{pmatrix} = \begin{pmatrix} 43,564.3 \\ -1,140.4 \\ -23.252.4 \\ 9,983.4 \\ 744.5 \\ -414.6 \end{pmatrix}$$

Modelo ajustado

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \widehat{\beta_{1,1}} \\ \widehat{\beta_{1,2}} \end{pmatrix} = \begin{pmatrix} 43,564.3 \\ -1,140.4 \\ -23.252.4 \\ 9,983.4 \\ 744.5 \\ -414.6 \end{pmatrix}$$

Modelo para vehículos tipo *otro*

$$\hat{Y}_i = 43564.3 - 1140.4X_{i1}, \quad i = 1, \ ..., \ 100$$

Modelo ajustado

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \widehat{\beta_{1,1}} \\ \widehat{\beta_{1,2}} \end{pmatrix} = \begin{pmatrix} 43,564.3 \\ -1,140.4 \\ -23.252.4 \\ 9,983.4 \\ 744.5 \\ -414.6 \end{pmatrix}$$

Modelo para vehículos tipo *otro*

$$\hat{Y}_i = 43564.3 - 1140.4X_{i1}, \qquad i = 1,\ \dots,\ 100$$

Modelo para vehículos tipo *hatchback*

$$\hat{Y}_i = 20311.9 - 395.9X_{i1}, \qquad i = 1,\ \dots,\ 100$$

Modelo ajustado

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \widehat{\beta_{1,1}} \\ \widehat{\beta_{1,2}} \end{pmatrix} = \begin{pmatrix} 43,564.3 \\ -1,140.4 \\ -23.252.4 \\ 9,983.4 \\ 744.5 \\ -414.6 \end{pmatrix}$$

Modelo para vehículos tipo *otro*

$$\hat{Y}_i = 43564.3 - 1140.4X_{i1}, \qquad i = 1, \; ..., \; 100$$

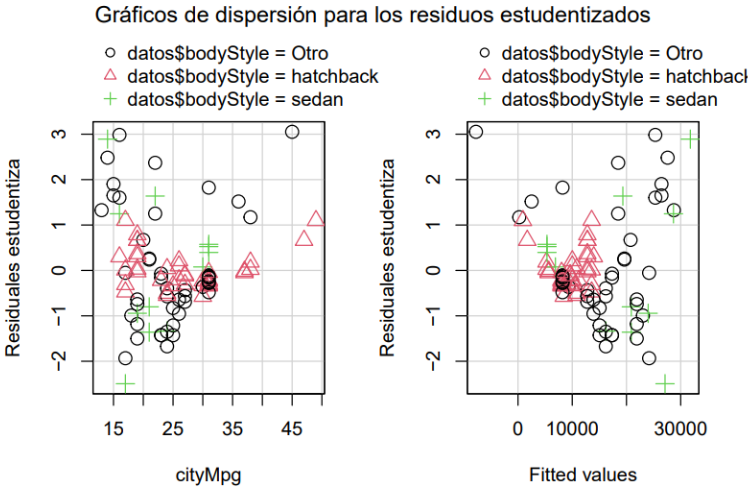
Modelo para vehículos tipo *hatchback*

$$\hat{Y}_i = 20311.9 - 395.9X_{i1}, \qquad i = 1, \; ..., \; 100$$

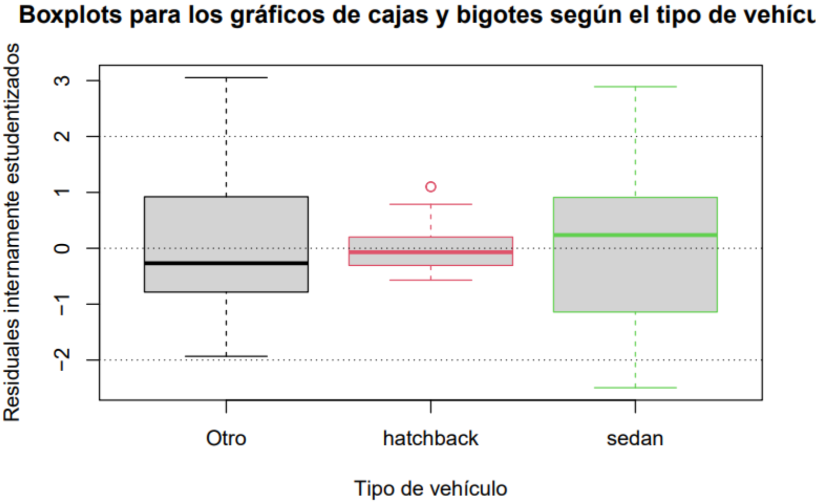
Modelo para vehículos tipo *sedán*

$$\hat{Y}_i = 53547.7 - 1555X_{i1}, \qquad i = 1, \; ..., \; 100$$

Residuales internamente estudentizados



Residuales internamente estudentizados



Test de normalidad

Uno de los supuestos es que los errores siguen una distribución **normal**.

Test de normalidad

Uno de los supuestos es que los errores siguen una distribución **normal**. Para verificar que el modelo realmente cumpla esto, se va a realizar una prueba de hipótesis sobre los errores aleatorios del modelo propuesto.

Test de normalidad

Uno de los supuestos es que los errores siguen una distribución **normal**. Para verificar que el modelo realmente cumpla esto, se va a realizar una prueba de hipótesis sobre los errores aleatorios del modelo propuesto. Así, las hipótesis son:

Hipótesis

H_0 : Los errores siguen una distribución normal.
H_1 : Los errores **no** siguen una distribución normal.

Test de normalidad

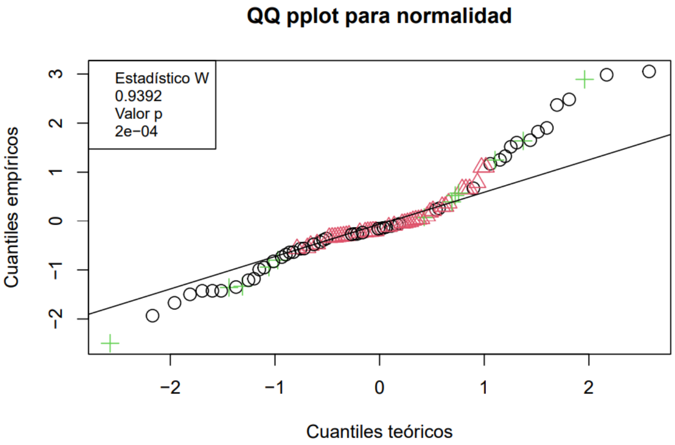
Uno de los supuestos es que los errores siguen una distribución **normal**. Para verificar que el modelo realmente cumpla esto, se va a realizar una prueba de hipótesis sobre los errores aleatorios del modelo propuesto. Así, las hipótesis son:

Hipótesis

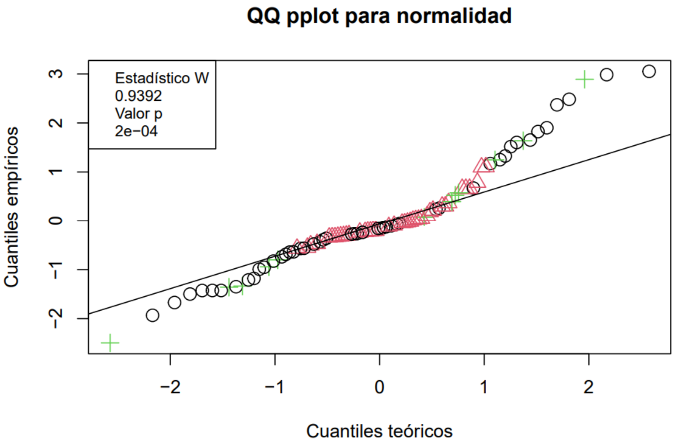
H_0 : Los errores siguen una distribución normal.
H_1 : Los errores **no** siguen una distribución normal.

Para evaluar este test, vale la pena observar un **QQ plot** de normalidad y el resultado de un test de **Shapiro-Wilk** con un nivel de significancia de $\alpha = 0.05$.

Test de normalidad



Test de normalidad



Se rechaza la hipótesis nula y por tanto, se concluye que no hay evidencia muestral suficiente para sugerir que los errores siguen una distribución normal con una significancia de $\alpha = 0.05$

Recordar

Modelo completo (MF)

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

Otro

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i$$

Hatchback

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

Sedan

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

$$E_i \overset{i.i.d}{\sim} Normal(0,\ \sigma^2)$$

¿Qué se va a evaluar?

Quiere verificarse si las ordenadas en el origen para las rectas ajustadas de cada tipo de vehículo son significativamente diferentes.

¿Qué se va a evaluar?

Quiere verificarse si las ordenadas en el origen para las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_0 = (\beta_0 + \beta_2) = (\beta_0 + \beta_3)$,

¿Qué se va a evaluar?

Quiere verificarse si las ordenadas en el origen para las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_0 = (\beta_0 + \beta_2) = (\beta_0 + \beta_3)$, lo cual equivale a $\beta_2 = \beta_3 = 0$.

¿Qué se va a evaluar?

Quiere verificarse si las ordenadas en el origen para las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_0 = (\beta_0 + \beta_2) = (\beta_0 + \beta_3)$, lo cual equivale a $\beta_2 = \beta_3 = 0$. Para esto, se plantean las siguientes hipótesis:

Hipótesis

$H_0: \ \beta_2 = \beta_3 = 0$
$H_1: \ \exists\, j: \ \beta_j \neq 0, \ j = 2, \ 3$

¿Qué se va a evaluar?

Quiere verificarse si las ordenadas en el origen para las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_0 = (\beta_0 + \beta_2) = (\beta_0 + \beta_3)$, lo cual equivale a $\beta_2 = \beta_3 = 0$. Para esto, se plantean las siguientes hipótesis:

Hipótesis

$H_0: \ \beta_2 = \beta_3 = 0$
$H_1: \ \exists \, j: \ \beta_j \neq 0, \ j = 2, \ 3$

Y el estadístico de prueba es:

$$F_0 = \frac{[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}]/v}{\text{MSE}_{(\text{MF})}}$$

donde MR hace referencia al modelo reducido y MF está asociado al modelo completo.

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{\mathrm{SSR}(I_1, I_2 | X_1, X_1 * I_1, X_1 * I_2)/(c-1)}{\mathrm{MSE}(X_1, I_1, I_2, X_1 * I_1, X_1 * I_2)} = \frac{(4280394376 - 3098866721)/(3-1)}{32966667} = 17.92$$

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{\text{SSR}(I_1, I_2 | X_1, X_1 * I_1, X_1 * I_2)/(c-1)}{\text{MSE}(X_1, I_1, I_2, X_1 * I_1, X_1 * I_2)} = \frac{(4280394376 - 3098866721)/(3-1)}{32966667} = 17.92$$

Luego, se calcula el valor p:

$$V_p = P(f_{3-1,100-3*2} > F_0) = P(f_{2,94} > 17.92) = 2.55 \times 10^{-7}$$

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{SSR(I_1, I_2 | X_1, X_1 * I_1, X_1 * I_2)/(c-1)}{MSE(X_1, I_1, I_2, X_1 * I_1, X_1 * I_2)} = \frac{(4280394376 - 3098866721)/(3-1)}{32966667} = 17.92$$

Luego, se calcula el valor p:

$$V_p = P(f_{3-1,100-3*2} > F_0) = P(f_{2,94} > 17.92) = 2.55 \times 10^{-7}$$

Y tomando un nivel de significancia de $\alpha = 0.05$, se tiene que $V_p = 2.55 \times 10^{-7} < \alpha = 0.05$, por lo que **se rechaza la hipótesis nula**.

Conclusión

Hay evidencia muestral suficiente para sugerir que las ordenadas de los orígenes de las rectas de regresión ajustadas **son diferentes**, con una significancia de $\alpha = 0.05$.

Gráfico

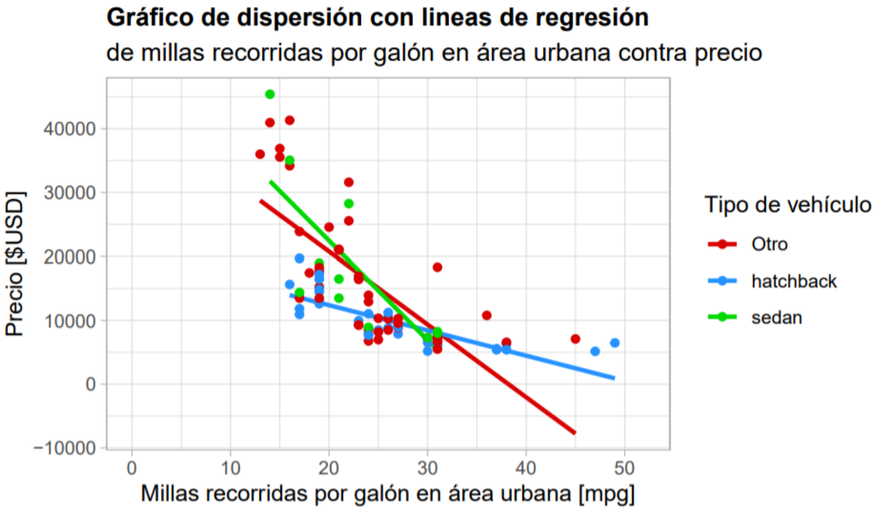


Gráfico de dispersión con lineas de regresión
de millas recorridas por galón en área urbana contra precio
Precio [$USD]
40000
30000
20000
10000
0
−10000
0
10
20
30
40
50
Millas recorridas por galón en área urbana [mpg]
Tipo de vehículo
Otro
hatchback
sedan

Recordar

Modelo completo (MF)

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

Otro

$$Y_i = \beta_0 + \beta_1 \; X_{i1} + E_i$$

Hatchback

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

Sedan

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

$$E_i \overset{i.i.d}{\sim} Normal(0, \; \sigma^2)$$

¿Qué se va a evaluar?

Quiere verificarse si las **pendientes** las rectas ajustadas de cada tipo de vehículo son significativamente diferentes.

¿Qué se va a evaluar?

Quiere verificarse si las **pendientes** las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_1 = (\beta_1 + \beta_{1,1}) = (\beta_0 + \beta_{1,2})$,

¿Qué se va a evaluar?

Quiere verificarse si las **pendientes** las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_1 = (\beta_1 + \beta_{1,1}) = (\beta_0 + \beta_{1,2})$, lo cual equivale a $\beta_{1,1} = \beta_{1,2} = 0$.

¿Qué se va a evaluar?

Quiere verificarse si las **pendientes** las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_1 = (\beta_1 + \beta_{1,1}) = (\beta_0 + \beta_{1,2})$, lo cual equivale a $\beta_{1,1} = \beta_{1,2} = 0$.
Para esto, se plantean las siguientes hipótesis:

Hipótesis

$$H_0 : \ \beta_{1,1} = \beta_{1,2} = 0$$
$$H_1 : \ \exists \ k : \ \beta_{1,k} \neq 0, \ k = 1, \ 2$$

¿Qué se va a evaluar?

Quiere verificarse si las **pendientes** las rectas ajustadas de cada tipo de vehículo son significativamente diferentes. Si esto no sucede es porque $\beta_1 = (\beta_1 + \beta_{1,1}) = (\beta_0 + \beta_{1,2})$, lo cual equivale a $\beta_{1,1} = \beta_{1,2} = 0$. Para esto, se plantean las siguientes hipótesis:

Hipótesis

$H_0: \; \beta_{1,1} = \beta_{1,2} = 0$
$H_1: \; \exists \; k: \; \beta_{1,k} \neq 0, \; k = 1, \; 2$

Y el estadístico de prueba es:

$$F_0 = \frac{[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}]/v}{\text{MSE}_{(\text{MF})}}$$

donde *MR* hace referencia al modelo reducido y *MF* está asociado al modelo completo.

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{SSR(X_1 * I_1, X_1 * I_2 | X_1, I_1, I_2)/(c-1)}{MSE(X_1, I_1, I_2, X_1 * I_1, X_1 * I_2)} = \frac{(3999435253-3098866721)/(3-1)}{32966667} = 13.65877$$

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{\text{SSR}(X_1 * I_1, X_1 * I_2 | X_1, I_1, I_2)/(c-1)}{\text{MSE}(X_1, I_1, I_2, X_1 * I_1, X_1 * I_2)} = \frac{(3999435253 - 3098866721)/(3-1)}{32966667} = 13.65877$$

Luego, se calcula el valor p:

$$V_p = P(f_{3-1,100-3*2} > F_0) = P(f_{2,94} > 13.65877) = 6.203 \times 10^{-6}$$

Evaluación

Así, sabiendo que se tienen $c = 3$ tipos de vehículos, el estadístico de prueba calculado está dado por:

$$\frac{SSR(X_1{*}I_1, X_1{*}I_2 | X_1, I_1, I_2)/(c-1)}{MSE(X_1, I_1, I_2, X_1{*}I_1, X_1{*}I_2)} = \frac{(3999435253-3098866721)/(3-1)}{32966667} = 13.65877$$

Luego, se calcula el valor p:

$$V_p = P(f_{3-1,100-3*2} > F_0) = P(f_{2,94} > 13.65877) = 6.203 \times 10^{-6}$$

Y tomando un nivel de significancia de $\alpha = 0.05$, se tiene que $V_p = 6.203 \times 10^{-6} < \alpha = 0.05$, por lo que **se rechaza la hipótesis nula**.

Conclusión

Hay evidencia muestral suficiente para sugerir que las pendientes de las rectas de regresión ajustadas **son diferentes**, con una significancia de $\alpha = 0.05$.

Gráfico

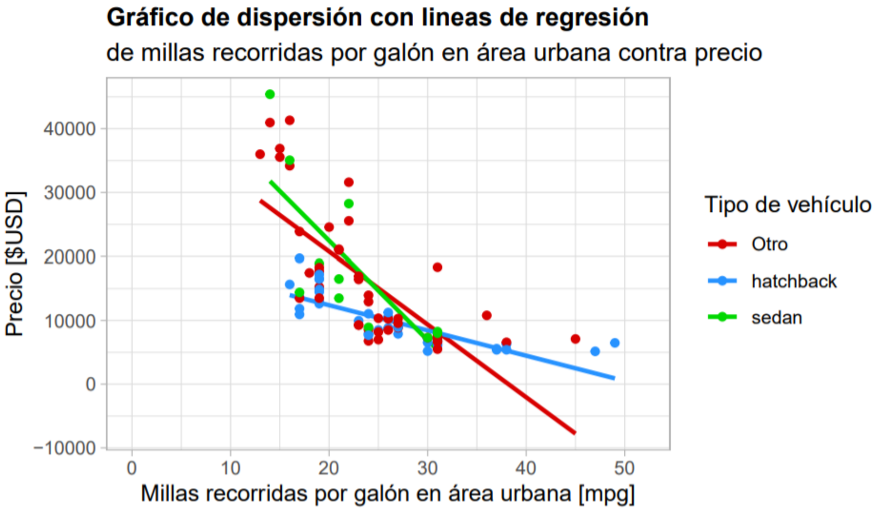


Gráfico de dispersión con lineas de regresión
de millas recorridas por galón en área urbana contra precio
Precio [$USD]
40000
30000
20000
10000
0
−10000
0
10
20
30
40
50
Millas recorridas por galón en área urbana [mpg]
Tipo de vehículo
Otro
hatchback
sedan

Recordar

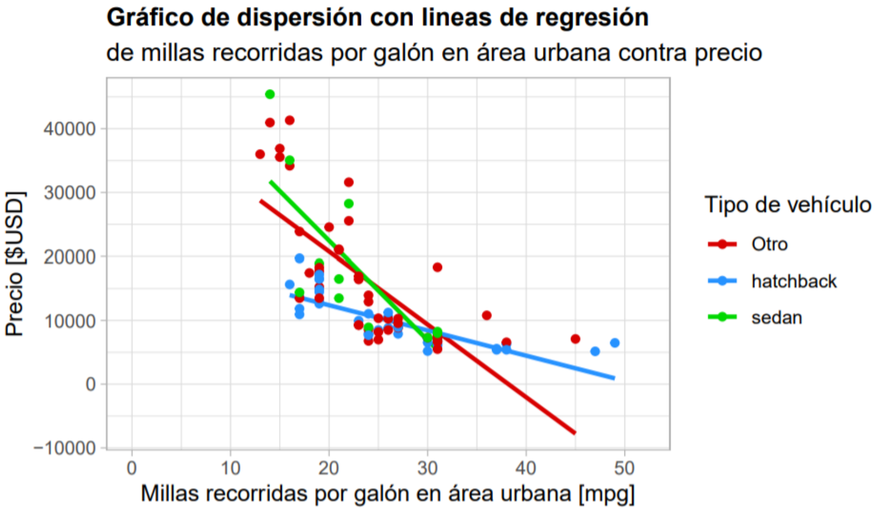


Gráfico de dispersión con lineas de regresión
de millas recorridas por galón en área urbana contra precio
Precio [$USD]
40000
30000
20000
10000
0
−10000
0
10
20
30
40
50
Millas recorridas por galón en área urbana [mpg]
Tipo de vehículo
Otro
hatchback
sedan

¿Qué se va a evaluar?

En el gráfico de dispersión se observa que las rectas ajustadas para los vehículos tipo **otro** y **sedán**, por lo que vale la pena chequear si sus pendientes son estadísticamente iguales con una significancia de $\alpha = 0.05$.

¿Qué se va a evaluar?

En el gráfico de dispersión se observa que las rectas ajustadas para los vehículos tipo **otro** y **sedán**, por lo que vale la pena chequear si sus pendientes son estadísticamente iguales con una significancia de $\alpha = 0.05$. Para ello vale la pena recordar sus dos ecuaciones:

Otro

$$Y_i = \beta_0 + \beta_1\ X_{i1} + E_i$$

Sedan

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

¿Qué se va a evaluar?

En el gráfico de dispersión se observa que las rectas ajustadas para los vehículos tipo **otro** y **sedán**, por lo que vale la pena chequear si sus pendientes son estadísticamente iguales con una significancia de $\alpha = 0.05$. Para ello vale la pena recordar sus dos ecuaciones:

Otro

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

Sedan

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

Así, lo que se quiere verificar es si $\beta_0 = \beta_0 + \beta_3$ y si $\beta_1 = \beta_1 + \beta_{1,2}$,

¿Qué se va a evaluar?

En el gráfico de dispersión se observa que las rectas ajustadas para los vehículos tipo **otro** y **sedán**, por lo que vale la pena chequear si sus pendientes son estadísticamente iguales con una significancia de $\alpha = 0.05$. Para ello vale la pena recordar sus dos ecuaciones:

Otro

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i$$

Sedan

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

Así, lo que se quiere verificar es si $\beta_0 = \beta_0 + \beta_3$ y si $\beta_1 = \beta_1 + \beta_{1,2}$, que es lo mismo que probar que $\beta_3 = 0$ y que $\beta_{1,2} = 0$.

Prueba de hipótesis

Así, se va realizar el siguiente test lineal:

Hipótesis

$H_0 : \ \beta_3 = \beta_{1,2} = 0$
$H_1 : \ \beta_3 \neq 0 \vee \beta_{1,2} \neq 0$

Prueba de hipótesis

Así, se va realizar el siguiente test lineal:

Hipótesis

$H_0: \ \beta_3 = \beta_{1,2} = 0$
$H_1: \ \beta_3 \neq 0 \vee \beta_{1,2} \neq 0$

Cuyo estadístico de prueba es:

$$F_0 = \frac{[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}]/v}{\text{MSE}_{(\text{MF})}}$$

donde MR hace referencia al modelo reducido y MF está asociado al modelo completo.

Prueba de hipótesis

Así, se va realizar el siguiente test lineal:

Hipótesis

$H_0: \ \beta_3 = \beta_{1,2} = 0$
$H_1: \ \beta_3 \neq 0 \vee \beta_{1,2} \neq 0$

Cuyo estadístico de prueba es:

$$F_0 = \frac{[\mathrm{SSE}_{(\mathrm{MR})} - \mathrm{SSE}_{(\mathrm{MF})}]/v}{\mathrm{MSE}_{(\mathrm{MF})}}$$

donde MR hace referencia al modelo reducido y MF está asociado al modelo completo.
Por lo que, al calcular el valor del estadístico de prueba, se obtiene que este es:

$$F_0 = \frac{62720421/2}{32966667} = 0.9513$$

Prueba de hipótesis

También es posible calcular el valor p:

$$V_p = 0.3899$$

Y resulta pues evidente que $V_p = 0.3899 > 0.05 = \alpha$, por lo que **no** se rechaza la hipótesis nula.

Prueba de hipótesis

También es posible calcular el valor p:

$$V_p = 0.3899$$

Y resulta pues evidente que $V_p = 0.3899 > 0.05 = \alpha$, por lo que **no** se rechaza la hipótesis nula.

Conclusión

No hay evidencia muestral suficiente para sugerir que la recta de regresión ajustada para los vehículos tipo **otro** son **diferentes** a la asociada a los vehículos tipo **sedán**, por lo que se asume que las rectas de regresión de ambos tipos de vehículo son estadísticamente iguales con un nivel de significancia de $\alpha = 0.05$.

¿Qué se va a evaluar?

Se quiere determinar si la recta de regresión que ajusta el precio de un vehículo en función del número del millas que recorra por galón de combustible consumido en áreas urbanas es la misma para los tres tipos de vehículos considerados.

¿Qué se va a evaluar?

Se quiere determinar si la recta de regresión que ajusta el precio de un vehículo en función del número del millas que recorra por galón de combustible consumido en áreas urbanas es la misma para los tres tipos de vehículos considerados. Así, vale la pena recordar sus ecuaciones:

Otro

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

Hatchback

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

Sedan

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

¿Qué se va a evaluar?

Se quiere determinar si la recta de regresión que ajusta el precio de un vehículo en función del número del millas que recorra por galón de combustible consumido en áreas urbanas es la misma para los tres tipos de vehículos considerados. Así, vale la pena recordar sus ecuaciones:

Otro

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

Hatchback

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

Sedan

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

¿Qué se va a evaluar?

Así, se quiere probar si se cumple que $\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ y que $\beta_1 = \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2}$,

¿Qué se va a evaluar?

Así, se quiere probar si se cumple que $\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ y que $\beta_1 = \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2}$, lo que resulta equivalente a verificar si $\beta_2 = \beta_3 = 0$ y que $\beta_{1,1} = \beta_{1,2} = 0$.

¿Qué se va a evaluar?

Así, se quiere probar si se cumple que $\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ y que $\beta_1 = \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2}$, lo que resulta equivalente a verificar si $\beta_2 = \beta_3 = 0$ y que $\beta_{1,1} = \beta_{1,2} = 0$.
Por tanto, las hipótesis a evaluar son:

Hipótesis

$$H_0 : \ \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0$$
$$H_1 : \ \beta_2 \neq \beta_3 \vee \beta_{1,1} \neq \beta_{1,2}$$

¿Qué se va a evaluar?

Así, se quiere probar si se cumple que $\beta_0 = \beta_0 + \beta_2 = \beta_0 + \beta_3$ y que $\beta_1 = \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2}$, lo que resulta equivalente a verificar si $\beta_2 = \beta_3 = 0$ y que $\beta_{1,1} = \beta_{1,2} = 0$.
Por tanto, las hipótesis a evaluar son:

Hipótesis

$$
\begin{aligned}
H_0 &: \ \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0 \\
H_1 &: \ \beta_2 \neq \beta_3 \vee \beta_{1,1} \neq \beta_{1,2}
\end{aligned}
$$

De manera que, empleando un nivel de significancia de $\alpha = 0.05$, se tiene que el estadístico de prueba a emplear está dado por:

$$F_0 = \frac{[\text{SSE}_{(\text{MR})} - \text{SSE}_{(\text{MF})}]/v}{\text{MSE}_{(\text{MF})}}$$

Evaluación

Así, realizando los cálculos, se tiene que el valor del estadístico de prueba es:

$$F_0 = \frac{1305775049/4}{32966667} = 9.9022$$

.

Evaluación

Así, realizando los cálculos, se tiene que el valor del estadístico de prueba es:

$$F_0 = \frac{1305775049/4}{32966667} = 9.9022$$

.

Asimismo es posible calcular el valor p:

$$V_p = 9.929 \times 10^{-7}$$

Evaluación

Así, realizando los cálculos, se tiene que el valor del estadístico de prueba es:

$$F_0 = \frac{1305775049/4}{32966667} = 9.9022$$

.

Asimismo es posible calcular el valor p:

$$V_p = 9.929 \times 10^{-7}$$

Así pues, dado que $V_p = 9.929 \times 10^{-7} < 0.05 = \alpha$, se **rechaza** la hipótesis nula.

Conclusión

Hay evidencia muestral suficiente para sugerir que al menos una de las rectas de regresión es diferente a las otras dos con una significancia de $\alpha = 0.05$

Preguntas

Gracias