Universidad Nacional De Colombia

Sede Medellín

Facultad de Ciencias

Departamento de Estadística

Proyecto 2 - Taller 3 - Entrega

Daniel Felipe Villa Rengifo

Luis David Hernández Pérez

Juan Gabriel Carvajal Negrete

Modelos de Regresión

Febrero, 2025

Contexto de datos

Planteamiento del problema

Una empresa automotriz china, Geely Auto, tiene la intención de ingresar al mercado de Estados Unidos estableciendo una planta de fabricación allí y produciendo vehículos localmente para competir con sus contrapartes estadounidenses y europeas.

Para ello, han contratado una consultora automotriz para comprender los factores que influyen en la fijación de precios de los vehículos. Específicamente, desean entender los factores que afectan el precio de los autos en el mercado estadounidense, ya que estos pueden ser muy diferentes a los del mercado chino. La empresa desea conocer:

- Qué variables son significativas para predecir el precio de un automóvil.
- Qué tan bien estas variables describen el precio de un automóvil.

A partir de diversas encuestas de mercado, la consultora ha recopilado un conjunto de datos https://www.kaggle.com/datasets/hellbuoy/car-price-prediction/data sobre diferentes tipos de vehículos en el mercado estadounidense.

Objetivo Empresarial

Se requiere modelar el precio de los automóviles utilizando las variables independientes disponibles. Este modelo será utilizado por la gerencia para comprender cómo varían exactamente los precios en función de las variables independientes. De esta manera, podrán ajustar el diseño de los vehículos, la estrategia empresarial, entre otros, para alcanzar ciertos niveles de precio. Además, el modelo será una herramienta útil para que la gerencia entienda la dinámica de precios de un nuevo mercado.

Descripción de las variables

- `Car_ID`: ID único de cada observación (entero).
- `Symboling`: Clasificación del riesgo de seguro; un valor de +3 indica que el automóvil tiene alto riesgo y un valor de -3 indica que probablemente es seguro (categórico).
- `fueltype`: Tipo de combustible del automóvil, por ejemplo, gasolina o diésel (categórico).
- `aspiration`: Tipo de aspiración utilizado en el automóvil (categórico).
- `doornumber`: Número de puertas del automóvil (categórico).
- `carbody`: Tipo de carrocería del automóvil (categórico).
- `drivewheel`: Tipo de tracción (ruedas motrices) del automóvil (categórico).
- `enginelocation`: Ubicación del motor del automóvil (categórico).
- `wheelbase`: Distancia entre los ejes del automóvil (numérico).
- `carlength`: Longitud del automóvil (numérico).
- `carwidth`: Ancho del automóvil (numérico).
- `carheight`: Altura del automóvil (numérico).
- `curbweight`: Peso del automóvil sin ocupantes ni equipaje (numérico).
- `enginetype`: Tipo de motor del automóvil (categórico).
- `cylindernumber`: Número de cilindros del motor (categórico).
- `enginesize`: Tamaño del motor del automóvil (numérico).

- `fuelsystem`: Sistema de combustible del automóvil (categórico).
- `boreratio`: Relación de diámetro del cilindro (numérico).
- `stroke`: Carrera o volumen dentro del motor (numérico).
- `compressionratio`: Relación de compresión del motor (numérico).
- `horsepower`: Potencia del motor en caballos de fuerza (numérico).
- `peakrpm`: Revoluciones máximas por minuto (RPM) del motor (numérico).
- `citympg`: Rendimiento de combustible en ciudad, medido en millas por galón (numérico).
- `highwaympg`: Rendimiento de combustible en carretera, medido en millas por galón (numérico).
- `price`: Precio del automóvil, considerado como la variable dependiente (numérico).

El conjunto de datos está formado por 205 registros y 26 variables, sin valores ausentes en las variables.

Limpieza de los datos

En la variable `car_name` podemos observar que los valores almacenan tanto el nombre de la empresa como el nombre del coche por lo cual hay 147 categorías distintas. Por tanto limpiaremos esa variable separando los nombres de las empresas de carros de la variable `car_name`. Por lo tanto crearemos una variable llamada `company_name` la cual tendrá solo el nombre de la empresa o compañía a la cual pertenece el carro.

Vemos que hay algunas categorías de la variable `company_name` están mal escritas como:

- maxda = mazda
- Nissan = nissan
- porsche = porcshce
- toyota = toyouta
- vokswagen = volkswagen = vw

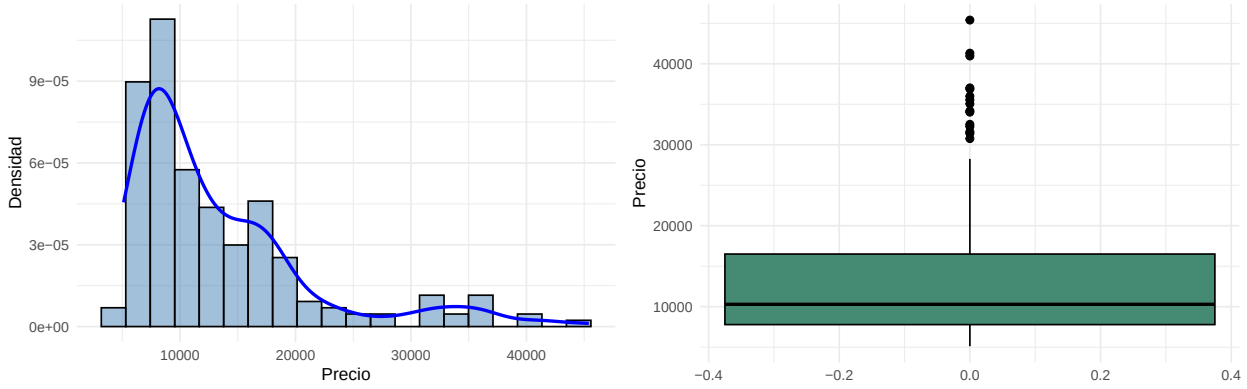
Reemplazaremos los nombres incorrectos con el nombre correcto de la empresa.

Como es sabido venimos trabajando con esta base de datos por lo que se a encontrado información importante y avances interesantes en este proyecto, para esta entrega se van a considerar dos variables de nuestro conjunto de datos de variables predictoras una de tipo continua y otra de tipo categórica con mínimo tres categorías, estas variables son:**carwidth**(Ancho del automóvil) ya que en resultados ya vistos, esta variable es la que mas aporta a nuestro modelo y ademas tiene una alta correlación con la variable respuesta y para la variable categórica **drivewheel** que representa el tipo de tracción de un automóvil y tiene tres categorías **rwd** (Rear-Wheel Drive) tracción trasera, **fwd** (Front-Whill Drive) y **4wd** (Four-Wheel Driive) Tracción en las cuatro ruedas.

Punto 1: Análisis descriptivo de los datos.

Un análisis descriptivo de las variables que se van a tener en cuenta en el modelo. Concluya.

Variable respuesta

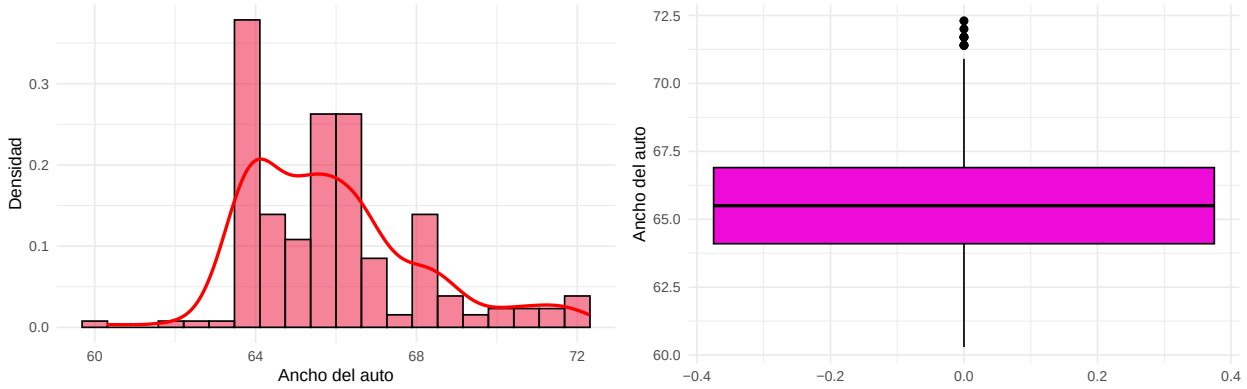


(a) Histograma (b) Boxplot

Figura 1: Distribución de la variable respuesta (Precio)

De la Figura 1 podemos decir que los precios de los carros tienen una distribución asimétrica, concentrándose en valores bajos, pero con una minoría de carros significativamente más caros, debido a la distribución puede existen problemas de normalidad.Los valores atípicos en el rango superior deben considerarse, ya que pueden representar carros de lujo o especiales por tanto analizaremos ahora la variable que corresponde al nombre de la empresa o compañía a la cual pertenece el carro para ver que relación existe con el precio.

Variable continua (carwidth)

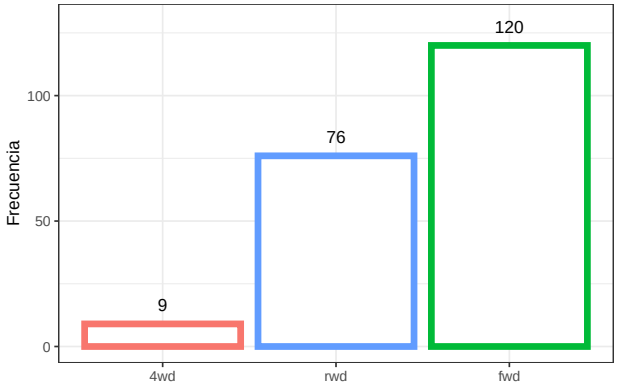


(a) Histograma (b) Boxplot

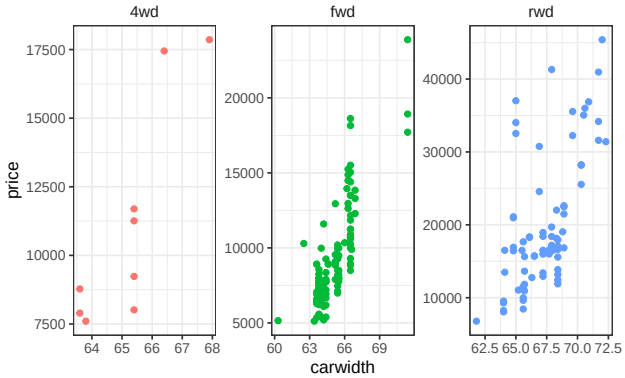
Figura 2: Distribución de la variable respuesta (Carwidht)

De la figura Figura 2 podemos visualizar como se distribuyen los anchos de los autos en el conjunto de datos, mostrando cuales son los anchos mas comunes y cuales son menos frecuentes, la distribución parece ser algo asimétrica, con una mayor concentración de autos en el rango de anchos mas bajos. por parte del boxplot vemos que unos datos con una media aproximada de 66 y algo importante tenemos algunos anchos de autos un poco distantes.

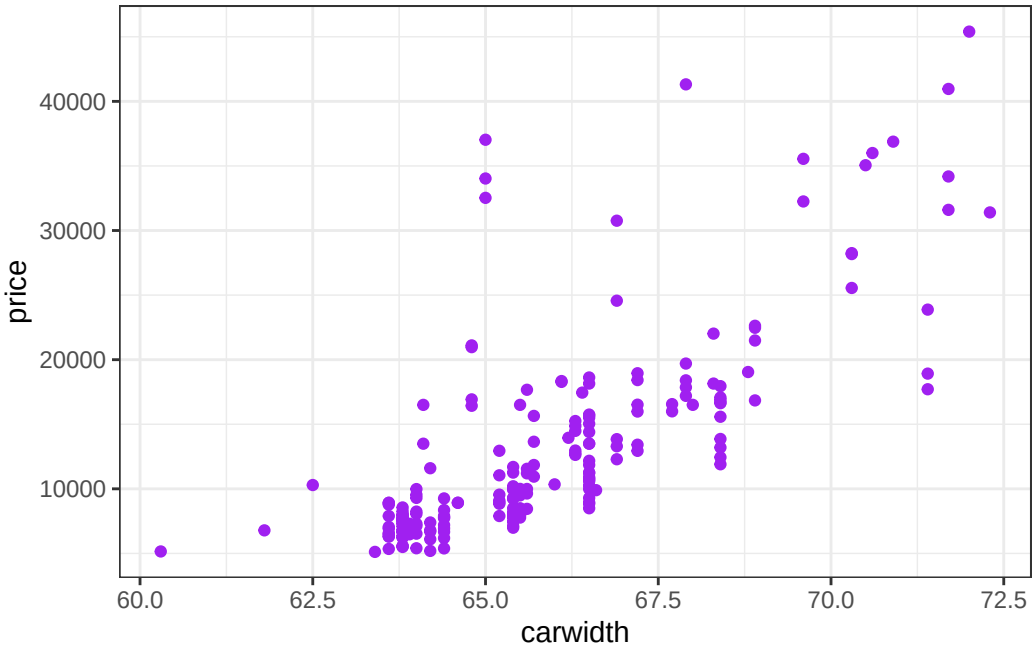
Variable categórica (drivewheel)



(a) Histograma



(b) Dispersion por categoria



(c) Dispersion

Figura 3: influencia variable categorica (drivewheel)

Los gráficos muestran que la mayoría de los vehículos tienen tracción delantera y que existe una relación general entre el ancho del vehículo y su precio, aunque con una dispersión considerable que indica la influencia de otras variables. Los vehículos con tracción delantera son los más comunes y tienden a tener precios y anchos moderados, mientras que los vehículos con tracción trasera

muestran una mayor variación en precio y ancho. Los vehículos con tracción en las cuatro ruedas son los menos frecuentes y no siguen un patrón claro en relación con el precio y el ancho.

Punto 2: Modelo de regresión apropiado.

Plantee el modelo de regresion apropiado si se espera una diferencia entre las rectas de Y vs. X que corresponden a los niveles de Z.

En este punto plantearemos el modelo de regresión correspondiente al caso donde se espera que existan diferencias entre las rectas de Y vs X que corresponden a los niveles de la variable categórica.Así el modelo planteado es el siguiente.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} X_i I_{i1} + \beta_{1,2} X_i I_{i2} + E_i, \quad E_i \ N(0, \sigma^2)$$

Donde:

$$I_{i1} = \begin{cases} 1 & \text{si en la unidad experimental es observada la categoría } fwd \\ 0 & \text{si en la unidad experimental no es observada la categoría } fwd. \end{cases}$$

$$I_{i2} = \begin{cases} 1 & \text{si en la unidad experimental es observada la categoría } rwd \\ 0 & \text{si en la unidad experimental no es observada la categoría } rwd. \end{cases}$$

X_i *y* Y_i Son las variables Carwidth y price respectivamente

Note que la categoría de referencia es `4wd`

Punto 3: Ajuste del modelo planteado.

Realice el ajuste del modelo e interprete las estimaciones de los parametros.

Resumen del modelo planteado

Coeficientes	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	-147214.6606	75620.4391	-1.95	0.0530
carwidth	2427.5330	1159.3832	2.09	0.0375
drivewheelfwd	45184.5715	77625.6922	0.58	0.5612
drivewheelrwd	-3314.6485	77290.3741	-0.04	0.9658
carwidth:drivewheelfwd	-717.2091	1190.2654	-0.60	0.5475
carwidth:drivewheelrwd	103.7724	1183.3976	0.09	0.9302

Por lo tanto el modelo ajustado es:

$$\begin{aligned} \hat{Y}_i = -147214.66 &+ 2427.53\texttt{carwidth} + 45184.57\texttt{drivewheelfwd} - 3314.65\texttt{drivewheelrwd} \\ &- 717.21(\texttt{carwidth} \times \texttt{drivewheelfwd}) + 103.77(\texttt{carwidth} \times \texttt{drivewheelrwd}) + \epsilon \end{aligned}$$

Interpretación de los parámetros:

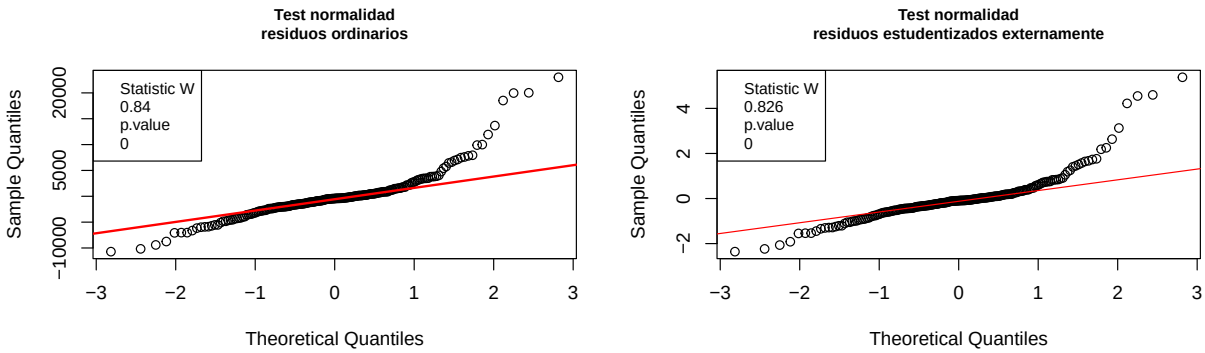
- **Intercepto** ($\beta_0 = -147214.66$): Representa el valor estimado de Y cuando todas las variables predictoras son cero. En este caso, se interpreta como el valor esperado de la variable de respuesta cuando el ancho del carro es 0 y la tracción es `drivewheel=4wd` (categoría de referencia).
- **Coeficiente de carwidth** ($\beta_1 = 2427.53$): Indica que, manteniendo constante la tracción del vehículo `(drivewheel)`, por cada unidad adicional en `carwidth`, se espera un aumento de 2427.53 unidades en la variable respuesta.
- **Coeficiente de drivewheelfwd** ($\beta_2 = 45184.57$): Representa la diferencia en la variable respuesta entre los carros con tracción delantera `(fwd)` y la categoría de referencia `(4wd)`, cuando `carwidth` $= 0$. Su efecto directo no es significativo (p-valor $= 0.5612$).
- **Coeficiente de drivewheelrwd** ($\beta_3 = -3314.65$): Indica la diferencia en la variable respuesta entre los carros con tracción trasera `(rwd)` y la categoría de referencia `(4wd)`, cuando `carwidth` $= 0$. No es significativo (p-valor $= 0.9658$).
- **Interacción carwidth:drivewheelfwd** ($\beta_{1,1} = -717.21$): Representa el cambio en la pendiente de `carwidth` cuando el vehículo tiene tracción delantera `(fwd)`. No es estadísticamente significativo (p-valor $= 0.5475$).
- **Interacción carwidth:drivewheelrwd** ($\beta_{1,2} = 103.77$): Representa el cambio en la pendiente de `carwidth` cuando el vehículo tiene tracción trasera `(rwd)`. Tampoco es significativo (p-valor $= 0.9302$).

Punto 4:

Supuestos del modelo

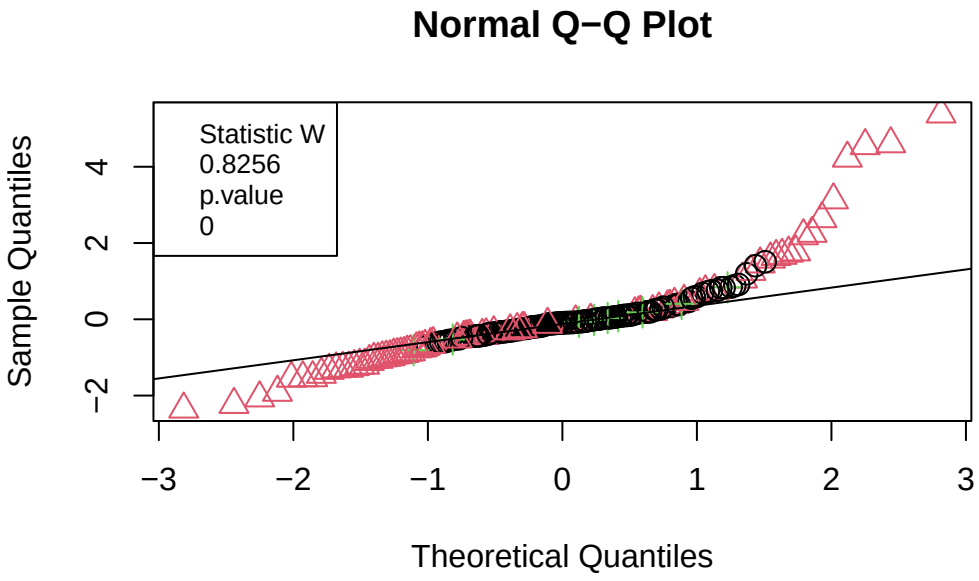
Analice supuestos de normalidad y varianza constante. Identifique en los graficos las observaciones segun la variable Z.

Gráficas de normalidad



(a) Residuales ordinarios

(b) Residuales estudenntizados



(c) residuales estudenntizados identificando categorias.

Figura 4: Gráfico Q-Q

Los gráficos Figura 4 de los residuales ordinarios y estudentizados muestran desviaciones significativas de la normalidad, especialmente en las colas, lo que indica la presencia de valores extremos. La prueba de Shapiro-Wilk confirma este problema, con estadísticos W de 0.84 y 0.826 respectivamente, y p-valores de 0, rechazando la hipótesis de normalidad en ambos casos.

prueba de homocedasticidad

Con ayudad del test Breusch-Pagan

```
ncvTest(modelo,var.formula=~carwidth*drivewheel)

#Version Breusch-Pagan estudentizado
bptest(modelo,studentize=TRUE)
```

```
Non-constant Variance Score Test
Variance formula: ~ carwidth * drivewheel
Chisquare = 119.1784, Df = 5, p = < 2.22e-16

    studentized Breusch-Pagan test

data:  modelo
BP = 26.734, df = 5, p-value = 6.425e-05
```

Dado que ambas pruebas rechazan la hipótesis nula de homocedasticidad, hay evidencia fuerte de heterocedasticidad en el modelo. Esto puede afectar la eficiencia de los coeficientes estimados y la validez de las pruebas de hipótesis.

Punto 5: Pruebas de hipótesis

Determine si existe diferencia entre las ordenadas en el origen de las rectas correspondientes a los diferentes niveles de Z. Plantee la hipótesis a probar, el estadístico de prueba y región critica al nivel de 0.05, realice la prueba y concluya.

Las ordenadas al origen son los interceptos, luego la igualdad de interceptos de las rectas correspondientes a las categorías de la variable `drivewheel` implica que:

$$\text{Se requiere que: } \beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 \iff \beta_2 = \beta_3 = 0$$

Luego se debe probar que:

$$H_0 : \beta_2 = \beta_3 = 0 \quad vs \quad H_1 : \text{Al menos uno no es cero } \beta_2, \beta_3$$

El estadístico de prueba correspondiente a esta prueba es:

$$F_0 = \frac{[SSE_{(MR)-SSE_{(MF)}}]/r}{MSE_{(MF)}}$$

Con

- $r = gl(SSE_{MR}) - gl(SSE_{MF}) = 201 - 199 = 2$
- SSE_{MR} es el SSE de modelo (MR) bajo H_0

- $SSE_{(MF)}$ es el SSE del modelo (MF) con todas las k variables

Se rechaza H_0,

- Si $F_0 > f_{0.05,2,199}$

```
linearHypothesis(modelo,c("drivewheelfwd=0","drivewheelrwd=0"))
```

```
Linear hypothesis test:
drivewheelfwd = 0
drivewheelrwd = 0

Model 1: restricted model
Model 2: price ~ carwidth * drivewheel

  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    201 4362529308
2    199 4271513634  2  91015674 2.1201 0.1227
```

El valor p (0.1227) es mayor que un nivel de significancia típico (0.05), lo que significa que no tenemos suficiente evidencia para rechazar la hipótesis nula. En otras palabras, los coeficientes de `drivewheelfwd` y `drivewheelrwd` no son significativamente diferentes de cero de manera conjunta.Ademas decimos que no existen diferencias en las coordenadas al origen no tienen diferencias significativas.

Punto 6: pruebas de hipótesis

Determine si existe diferencia en las pendientes de las rectas correspondientes a los diferentes niveles de Z. Plantee la hipótesis a probar, el estadístico de prueba y región critica al nivel de 0.05, realice la prueba y concluya.

La igualdad de las pendientes de las rectas para cada nivel de la variable `drivewheel` implica que:

$$\beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 \iff \beta_{1,1} = \beta_{1,2} = 0$$

Luego se debe probar que:

$$H_0 : \beta_{1,1} = \beta_{1,2} = 0 \quad vs \quad H_1 : \text{Al menos uno no es cero } \beta_{1,1}, \beta_{1,2}$$

Estadístico de prueba:

$$F_0 = \frac{[SSE_{(MR)-SSE_{(MF)}}]/r}{MSE_{(MF)}}$$

Con

- $r = gl(SSE_{MR}) - gl(SSE_{MF}) = 201 - 199 = 2$
- SSE_{MR} es el SSE de modelo (MR) bajo H_0
- $SSE_{(MF)}$ es el SSE del modelo (MF) con todas las k variables

Se rechaza H_0 si:

- Si $F_0 > f_{0.05,2,199}$

```
linearHypothesis(modelo,c("carwidth:drivewheelfwd=0","carwidth:drivewheelrwd=0"))
```

```
Linear hypothesis test:
carwidth:drivewheelfwd = 0
carwidth:drivewheelrwd = 0

Model 1: restricted model
Model 2: price ~ carwidth * drivewheel

  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    201 4384832323
2    199 4271513634  2 113318689 2.6396 0.07389 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

No encontramos evidencia estadística suficiente para afirmar que las pendientes de `carwidth` son significativamente diferentes entre los niveles de `drivewheel`. Es decir, la interacción entre `carwidth` y `drivewheel` no parece ser significativa al nivel del 0.05.

Punto 7: Test lineal general

Teniendo en cuenta los resultados anteriores realice una prueba de suma de cuadrados extra con test lineal general. Plantee y justifique la hipótesis a probar, el estadístico de prueba región critica al nivel de 0.05, realice la prueba y concluya.

Dependiendo de los resultados obtenidos anteriormente se quiere probar la siguiente prueba de hipótesis.

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

De forma que $L\beta$ corresponde a

$$L = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_{1,1} \\ \beta_{1,2} \end{bmatrix} = \begin{bmatrix} \beta 1, 1 - \beta_{1,2} \\ \beta_2 - \beta_3 \end{bmatrix}$$

El estadístico de prueba es:

$$F_0 = \frac{[SSE_{(MR)-SSE_{(MF)}}]/r}{MSE_{(MF)}}$$

Se rechazaría a un nivel de significancia α si $F_0 > f_{\alpha,r,n-k-1}$

```
linearHypothesis(modelo,c("carwidth:drivewheelfwd=carwidth:drivewheelrwd",
                          "drivewheelfwd=drivewheelrwd"))
```

```
Linear hypothesis test:
carwidth:drivewheelfwd - carwidth:drivewheelrwd = 0
drivewheelfwd - drivewheelrwd = 0

Model 1: restricted model
Model 2: price ~ carwidth * drivewheel

  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    201 5509805084
2    199 4271513634  2 1238291450 28.845 9.996e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La prueba indica que existen diferencias significativas entre las pendientes y los efectos base de `drivewheel`. Esto sugiere que la relación entre `carwidth` y `price` depende significativamente del tipo de tracción del vehículo.

Punto 8: Prueba de hipótesis

Si se quiere probar que la recta de Y vs. X es diferente para cada niveles de Z, plantee la hipótesis a probar, el estadístico de prueba y región critica al nivel de 0.05, realice la prueba y concluya.

Las rectas serán iguales si coinciden sus interceptos y sus pendientes, entonces.

$$\text{Se require que } \beta_0 + \beta_2 = \beta_0 + \beta_3 = \beta_0 \iff \beta_2 = \beta_3 = 0$$

$$\text{tambien que } \beta_1 + \beta_{1,1} = \beta_1 + \beta_{1,2} = \beta_1 \iff \beta_{1,1} = \beta_{1,2} = 0$$

Luego se debe probar:

$$H_0 : \beta_2 = \beta_3 = \beta_{1,1} = \beta_{1,2} = 0 \quad vs \quad H_1 : \text{ Al menos uno es diferente de cero}$$

Estadístico de prueba:

$$F_0 = \frac{[SSE_{(MR)-SSE_{(MF)}}]/r}{MSE_{(MF)}}$$

Con

- $r = gl(SSE_{MR}) - gl(SSE_{MF})$
- SSE_{MR} es el SSE de modelo (MR) bajo H_0
- $SSE_{(MF)}$ es el SSE del modelo (MF) con todas las k variables

Se rechaza H_0 si:

- Si $F_0 > f_{0.05,4,199}$

```
linearHypothesis(modelo,c("drivewheelfwd=0","drivewheelrwd=0",
                          "carwidth:drivewheelfwd=0","carwidth:drivewheelrwd=0"))
```

```
Linear hypothesis test:
drivewheelfwd = 0
drivewheelrwd = 0
carwidth:drivewheelfwd = 0
carwidth:drivewheelrwd = 0

Model 1: restricted model
Model 2: price ~ carwidth * drivewheel

  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1    203 5512841958
2    199 4271513634  4 1241328324 14.458 2.215e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dado que rechazamos la hipótesis nula, no hay suficiente significancia estadística para aceptarla decimos que las rectas son diferentes para cada nivel de la variable categórica.

Importante: Desde que se ajusto el modelo con las variables involucradas en el modelo notamos que las variables indicadoras junto con las interacciones resultaron no ser significativas.