Estadística II - Solución Taller 01 Semestre: 2024-01

Profesores: Johnatan Cardona Jimenez, Freddy Hernández Barajas, Raul Alberto Perez

Monitor: Ronald Palencia

solución primer punto

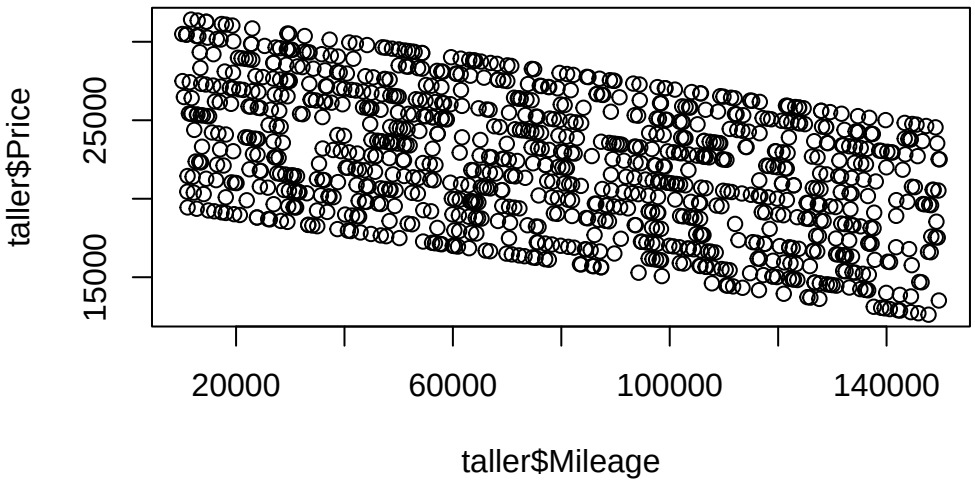
El siguiente còdigo permite leer la base datos

```
library(tidyverse)
data = read.csv("CarPricesPrediction.csv")


taller = data %>%
  select(Price, Mileage)
```

solución punto 2

```
plot(taller$Mileage, taller$Price)
```



En un diagrama de dispersión como este, cada punto representa un vehículo específico, con su kilometraje correspondiente en el eje horizontal y su precio en el eje vertical. De un vistazo rápido, parece que no hay una relación clara o lineal entre el kilometraje y el precio de los vehículos. Los puntos están bastante dispersos y no forman un patrón discernible que indique que a mayor kilometraje, menor es el precio, o viceversa, lo cual es común en datos de vehículos usados, ya que el precio no depende únicamente del kilometraje sino también de factores como la marca, el modelo, el año, el estado del vehículo, entre otros.

Además, se observa que hay vehículos con un rango de kilometraje muy amplio, desde casi nuevos hasta vehículos con más de 140,000 millas. Del mismo modo, el rango de precios también es amplio, con algunos vehículos costando menos de $15,000 y otros más de 30,000.

Solución punto 3

Ecuaciòn del modelo ajustado es:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Primero, la función lm() en R es usada para ajustar modelos lineales. Esta es la sintaxis básica de cómo se utiliza la función:

modelo1 = lm(dependent_variable ~ independent_variable, data = my_data)

```
modelo1 = lm(taller$Price~ taller$Mileage)

summary(modelo1)
```

```
Call:
lm(formula = taller$Price ~ taller$Mileage)

Residuals:
    Min      1Q  Median      3Q     Max
-6245.2 -3169.0  -102.9  2951.7  5972.4

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.601e+04  2.653e+02   98.04   <2e-16 ***
taller$Mileage -4.840e-02  3.005e-03  -16.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3784 on 998 degrees of freedom
Multiple R-squared:  0.2063,    Adjusted R-squared:  0.2055
F-statistic: 259.5 on 1 and 998 DF,  p-value: < 2.2e-16
```

Aquí, taller_Price es la variable dependiente (es decir, la que queremos predecir), y taller_Mileage es la variable independiente (la que usamos para hacer la predicción). No se especificó un conjunto de datos (data), por lo que R asume que taller_Price y taller_Mileage están en el espacio de trabajo actual.

Ecuación del modelo ajustado

$$\text{Precio} = 26010 - 0.0484 \times \text{Kilometraje}$$

- Precio: es la variable dependiente, que representa el precio de los vehículos en el conjunto de datos `taller`.
- 26010: es el valor estimado del intercepto (β_0), que indica el precio estimado cuando el kilometraje es cero.
- 0.0484: es el valor estimado de la pendiente (β_1), que representa el cambio en el precio estimado por cada unidad de cambio en el kilometraje.
- Kilometraje: es la variable independiente del modelo.

Solción punto 4 y 5

Interpretación

Para realizar la interpretaciòn debemos tener en cuenta lo siguiente:

En general, se puede hacer la interpretación de β_0 y β_1, de acuerdo a lo siguiente:

- Interpretación de β_0. Es el valor promedio de la respuesta cuando la variable predictora toma el valor de cero. Esto sólo si $X = 0 \in [X_{\min}, X_{\max}]$.
- Interpretación de β_1. Es el efecto de la variable predictora sobre la respuesta, en otras palabras, se dice que por cada unidad de aumento en la predictora, el promedio de la respuesta cambia en β_1 unidades.

```
min(taller$Mileage)
```

```
[1] 10079
```

```
max(taller$Mileage)
```

```
[1] 149794
```

- **Interpretación de** β_0: El valor estimado de β_0 es 26,010. Este valor representa el precio promedio estimado de un vehículo cuando el kilometraje es cero. Sin embargo, dado que el rango mínimo de kilometraje en tus datos es 10,079, la interpretación directa de β_0 como el precio de un vehículo nuevo o sin uso no es del todo aplicable en este contexto específico. En su lugar, β_0 actúa más como un ajuste en la línea de regresión para el rango de datos observados, y su interpretación práctica puede ser limitada.
- **Interpretación de** β_1: El coeficiente β_1 tiene un valor estimado de -0.0484. Esto indica que, en promedio, por cada unidad adicional en el kilometraje, el precio del vehículo disminuye en 0.0484 unidades monetarias. Esta interpretación es válida dentro del rango observado de la variable predictora, que en tus datos varía entre 10,079 y 149,794 millas.

Prueba de significancia

Las pruebas de hipótesis para los parámetros del modelo de regresión lineal son las siguientes:

Para el intercepto β_0:

- Hipótesis nula (H_0): $\beta_0 = 0$. No hay efecto del intercepto en el precio.
- Hipótesis alternativa (H_1): $\beta_0 \neq 0$. El intercepto tiene un efecto en el precio.

Para la pendiente β_1:

- Hipótesis nula (H_0): $\beta_1 = 0$. No hay relación entre el kilometraje y el precio.
- Hipótesis alternativa (H_1): $\beta_1 \neq 0$. Existe una relación entre el kilometraje y el precio.

Dado que los valores p asociados con ambos β_0 y β_1 son menores que 2×10^{-16}, lo cual es mucho menor que el nivel de significancia de $\alpha = 0.05$, rechazamos ambas hipótesis nulas. Esto indica que tanto el intercepto como la pendiente son estadísticamente significativos en el modelo.

Solución punto 6

Estimación de los Parámetros MCO

6.1 Estimación de $\hat{\beta}_1$

La fórmula para la estimación de $\hat{\beta}_1$ en un modelo de regresión lineal mediante MCO es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Con los datos proporcionados:

$$\hat{\beta}_1 = \frac{-0.01}{0.050} = -0.2$$

6.2 Estimación de $\hat{\beta}_0$

Una vez que tenemos $\hat{\beta}_1$, podemos calcular $\hat{\beta}_0$ usando la fórmula:

$$\hat{\beta}_0 = \bar{z} - \hat{\beta}_1 \bar{x}$$

Asumiendo que los promedios de z_i y x_i son cero, la fórmula se simplifica a:

$$\hat{\beta}_0 = 0 - (-0.2 \times 0) = 0$$

Por lo tanto, la estimación para $\hat{\beta}_0$ es 0.

La ecuación del modelo ajustado, utilizando las estimaciones de los parámetros, es:

$$Z = \hat{\beta}_0 + \hat{\beta}_1 x$$

Sustituyendo las estimaciones $\hat{\beta}_0 = 0$ y $\hat{\beta}_1 = -0.2$, obtenemos:

$$Z = 0 - 0.2x$$

Esta ecuación sugiere que el rendimiento diario de la compañía ABC (Z) disminuye en 0.2 unidades por cada unidad de aumento en el rendimiento diario del índice del mercado (x).

Solución punto 7

7.1 Falso o Verdadero: Mínimos Cuadrados y Estimadores Lineales Insesgados

Respuesta: Verdadero.

Argumento: El método de Mínimos Cuadrados Ordinarios (MCO) es conocido por proporcionar los Mejores Estimadores Lineales Insesgados (BLUE, por sus siglas en inglés) para los parámetros en un modelo de regresión lineal, bajo ciertas condiciones (Teorema de Gauss-Markov). Estos estimadores son "mejores" en el sentido de tener la menor varianza entre todos los estimadores lineales insesgados. Este método no requiere suposiciones distribucionales específicas sobre los términos de error para la estimación de los coeficientes; los supuestos se relacionan más con la estructura del modelo (como la linealidad) y las propiedades de los errores (como la homocedasticidad y la no correlación).

7.2 Falso o Verdadero: Necesidad de Supuestos para Pruebas de Hipótesis e Intervalos de Confianza

Respuesta: Verdadero.

Argumento: Para realizar pruebas de hipótesis y construir intervalos de confianza en el contexto de la regresión lineal, es necesario validar ciertos supuestos. Estos incluyen la normalidad de los errores, la homocedasticidad (varianza constante de los errores), y la independencia de los errores. Estos supuestos son cruciales para la validez de las pruebas de hipótesis y para la construcción de intervalos de confianza fiables. Si bien el MCO puede estimar coeficientes sin estas consideraciones distribucionales, las pruebas de hipótesis y la construcción de intervalos de confianza requieren que estos supuestos se cumplan para asegurar la precisión y fiabilidad de los resultados inferenciales.

Estas respuestas toman en cuenta la naturaleza tanto no paramétrica del MCO en términos de estimación de coeficientes, como la necesidad de cumplir con supuestos distribucionales para la inferencia estadística en el marco del modelo lineal (lm).