Análisis estadístico de `ventas.csv`

David Calle González
dcalleg@eafit.edu.co

Juan Sebastián Díaz Osorio
jsdiazo@eafit.edu.co

Alejandro Fernández R.
afernander@eafit.edu.co

Simón Marín Giraldo
smaring1@eafit.edu.co

Miguel Fernando Ramos García
mframosg@eafit.edu.co

Diciembre, 2020

Introducción

En el siguiente trabajo se solucionarán 9 problemas relacionados con la información contenida en `ventas.csv`, archivo seleccionado para el trabajo final del curso de estadística general en el semestre 2020-2.

Todo el aprendizaje obtenido durante el curso es aplicado aunque los procedimientos de cálculo no se ven ampliamente representados pues, cabe mencionar, se prioriza el análisis y conclusión de los problemas frente a los cálculos específicos.

Por esto, en este documento se describen las soluciones con el apoyo del software **R** para los cálculos y la generación de las gráficas. No se acompaña del código utilizado en este software para simplificar su lectura y ser más concluyente que explicativo en los métodos de cálculo.

Problema 1

Enunciado

Se desea estimar un intervalo para las ventas esperadas. **a)** ¿Qué intervalo utilizará? Explicite la fórmula. **b)** Justifique la elección anterior. **c)** ¿Cuál es el intervalo de confianza para las ventas esperadas? Comente los resultados.

Solución del problema

a) Para este caso, el intervalo más conveniente es el intervalo de confianza para la media con desviación estándar desconocida. La fórmula es:

$$\left(\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{S}{\sqrt{n}}\right)$$

b) Tenemos los datos sobre las ventas como una variable aleatoria en el archivo `ventas.csv`. Dado que buscamos las ventas esperadas, estamos hablando del valor promedio de las ventas que, por el teorema central del límite, sabemos que se distribuye normalmente.

La ausencia de información sobre la desviación estándar de esta variable hace que la debamos calcular y asumir una distribución t-student con $n-1$ grados de libertad y eso explica el uso de este intervalo.

c) Si calculamos los valores para la fórmula usando R, debemos usar los siguientes:

$$n = 200 \qquad t_{1-\frac{0.05}{2}} = 1.971957$$

$$S = 20.73091 \qquad \bar{x} = 43.13905$$

Si aplicamos la fórmula, vamos a obtener el siguiente intervalo:

$$(40.25,\ 46.03) \qquad \text{Con el } 95\,\% \text{ de confianza}$$

Y este es nuestro intervalo de confianza para las ventas esperadas, que nos dice que lo que se espera obtener por ventas en promedio está entre 40 mil y 46 mil dólares.

Problema 2

Enunciado

¿La venta promedio son mayores a 43 mil dólares? Determine si existe evidencia suficiente para afirmar esto. Justifique la elección del test y comente los resultados.

Solución del problema

Primero que nada se definen las hipotesis:

$$H_0 : \mu = 43 \text{ (miles de dólares)} \qquad H_0 : \mu > 43 \text{ (miles de dólares)}$$

En vista de que no se sabe si los datos vienen de una distribución normal, pero tenemos grandes cantidades de datos, se hace uso del estadístico de prueba:

$$z = \sqrt{n}\left(\frac{\bar{x} - \mu_0}{\sigma}\right)$$

con los siguientes valores:

$$n = 200 \qquad \bar{x} = 43.1390 \qquad \mu_0 = 43 \qquad \sigma = 20.6790$$

y, reemplazando,

$$z = \sqrt{200}\left(\frac{43.1390 - 43}{20.6790}\right) = 0.09509$$

Ahora, se debe hallar el valor-p($Z \sim N(0,1)$), para lo cual se evalua P(Z > z) en una distribución normal, de lo que se obtiene:

$$\text{valor-p} = 1 - P(Z < 0,09509) = 0.4621217$$

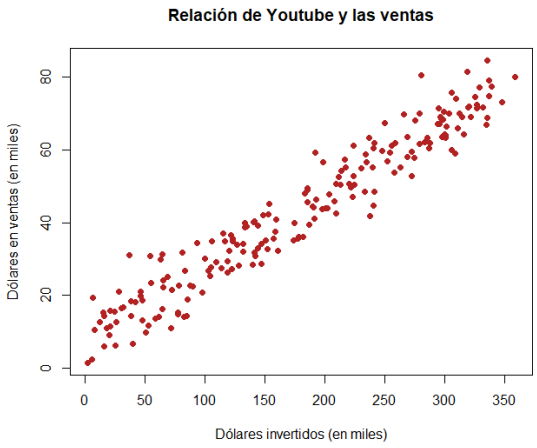
y haciendo uso de un $\alpha = 0.1$ nos damos cuenta de que el valor-p $> \alpha$, por lo que se concluye que no hay suficiente información para rechazar H_0, la venta promedio no es superior a 43 mil dolares.

Problema 3

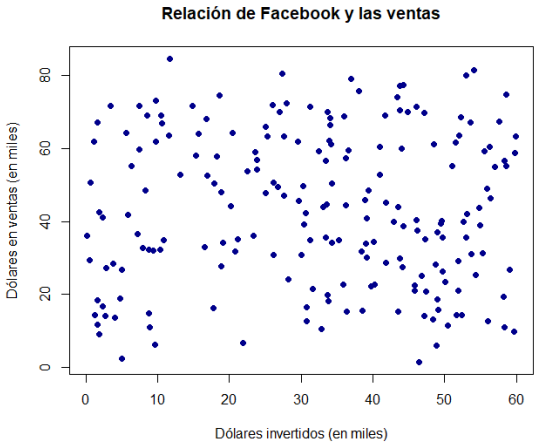
Enunciado

Realice gráficos de dispersión y calcule el coeficiente de correlación lineal para detectar posibles relaciones lineales. ¿Qué variables parecen tener una relación lineal con la variable ventas?

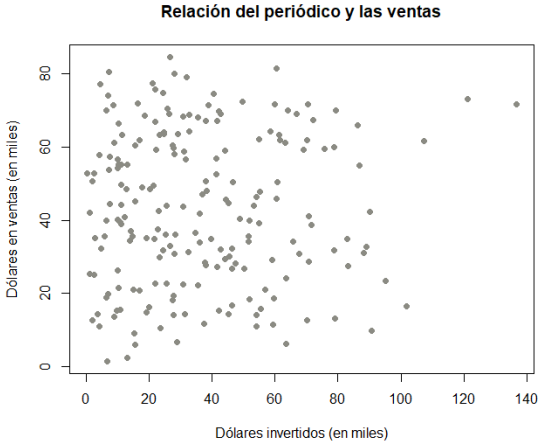
Solución del problema



(a) Youtube



(b) Facebook



(c) Newspaper

Figura 1: Gráficos de dispersión

La correlación lineal para Youtube es de **0.97**, para Facebook es de **-0.001** y para el periódico es de **0.037**.

Está claro que, ajustándose a un modelo lineal simple, la inversión en Youtube es la variable más influyente en las ventas. Facebook y el periódico ni siquiera cumplen con los criterios más básicos como linealidad, homogeneidad y homocedasticidad y por ello no se ajustan tanto a un modelo lineal simple.

Problema 4

Enunciado

Ajuste un modelo lineal entre la variable `sales` y `youtube`. Explicite el modelo, interprete los coeficientes estimados, comente el valor-p asociado a la hipótesis nula que el coefiente es igual a cero e interprete el valor del coeficiente de determinación. Concluya.

Solución del problema

Dicho modelo se puede estimar utilizando R. Siendo el modelo $\hat{y} = \beta_0 + \beta_1 x$ con $\hat{y}$ representando los valores estimados de ventas y x lo invertido en Youtube, los valores serían $\beta_0 = 7.78$ y $\beta_1 = 0.20$.

El valor de β_0 nos da información sobre el valor de las ventas cuando no se invierte nada en Youtube.

El valor de β_1 nos habla de que por cada mil dólares más que se invierte en Youtube, se puede esperar un crecimiento de 200 dólares en las ventas según el modelo lineal.

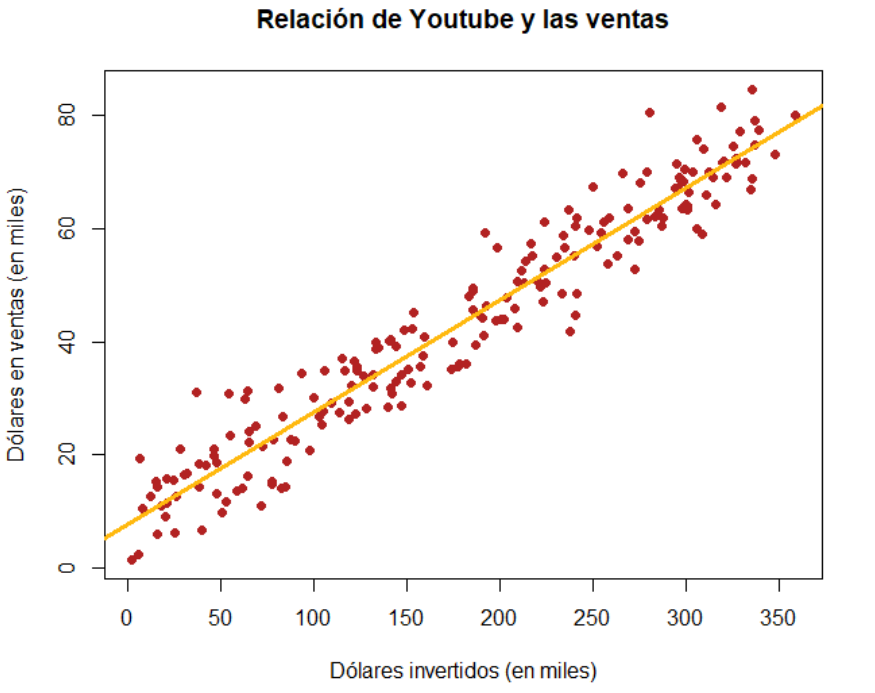


Figura 2: Gráfica de dispersión con la recta de regresión lineal

Hagamos ahora prueba de hipótesis sobre β_1 para probar si existe la relación lineal entre las variables. Calculamos el valor-p utilizando el software R, el cual nos entrega el resultado de 2.2×10^{-16}. Esto nos ayuda a **rechazar** la hipótesis nula que normalmente plantea que el modelo lineal simple no funciona.

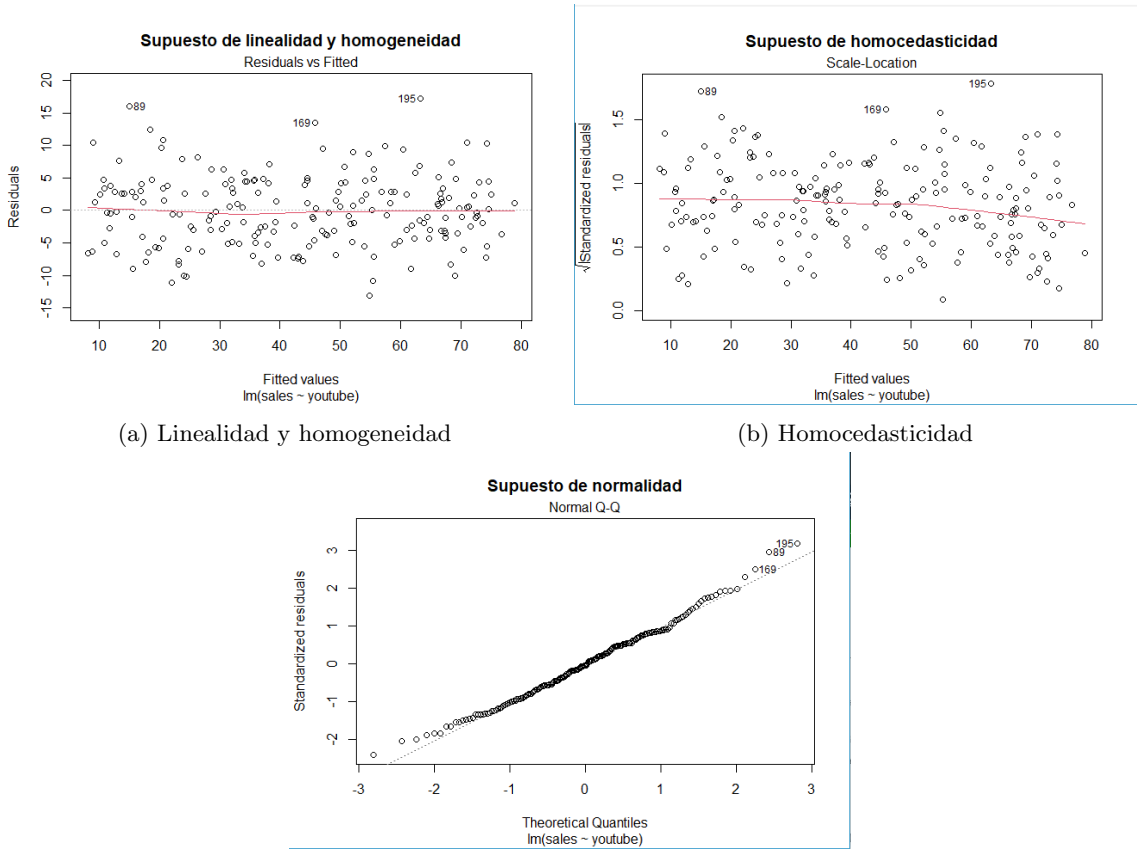
Acto seguido, podemos calcular el coeficiente de determinación usando R el cuál no da el resultado de **0.93**, un valor muy alto y que nos indica que el modelo se ajusta bastante bien a los datos entregados.

Problema 5

Enunciado

Realice un análisis gráfico sobre los supuestos del modelo ajustado.

Solución del problema



(a) Linealidad y homogeneidad

(b) Homocedasticidad

(c) Normalidad

Figura 3: Supuestos del modelo de Youtube

En la **figura 3a** vemos puntos cercanos a la linea cero, en la **figura 3b** vemos una distribución uniforme con respecto a la linea roja (lo que representa una varianza constante) y en la **figura 3c** vemos puntos siguiendo esa linea recta que justifica la normalidad de los datos. Con esto podemos ver que el modelo si cumple con las condiciones para tener linealidad, homogeneidad, homocedasticidad y normalidad.

Problema 6

Enunciado

Realice una estimación por intervalos de las ventas esperadas cuando se quiere invertir 170 mil dólares en Youtube.

Solución del problema

Recordemos la fórmula del intérvalo de confianza para valores específicos de la recta de regresión:

$$\left(\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{(n-2),\frac{\alpha}{2}} \sqrt{S_R^2 \left(\frac{1}{n} + \frac{(x^2 - \bar{x})^2}{(n-1)S_R^2}\right)}\right)$$

Vamos a utilizar los siguientes valores:

$$n = 200 \qquad t_{(n-2),\frac{\alpha}{2}} = 1.97$$

$$\beta_0 = 7.78 \qquad \beta_1 = 0.20$$

$$x^* = 170 \qquad \bar{x} = 178.23$$

$$S_R^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 29.46 \qquad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 10175.36$$

Si aplicamos entonces la ecuación, vamos a obtener el siguiente intervalo:

$$(41.64,\ 41.91) \qquad \text{Con el } 95\,\% \text{ de confianza}$$

Y este es nuestro intervalo de confianza para las ventas esperadas cuando se invierten 170 mil dólares en Youtube.

Problema 7

Enunciado

Realice una estimación por intervalos de las ventas que se tendrán cuando se quiere invertir en youtube 170 mil dólares en publicidad.

Solución del problema

Podemos responder a esta pregunta de dos formas.

La primera es que el modelo se ajusta al valor de 41.51 dentro de su intervalo de confianza, por lo que se puede esperar obtener 41 mil dólares en ventas.

La otra forma es predecir el intervalo de lo obtenido en ventas futuras que, según R, se estima en el intervalo de (30.78, 52.24), que nos indica que lo obtenido finalmente por ventas puede estar entre 30 y 52 mil dólares.

Problema 8

Enunciado

Ajuste el modelo

$$\text{sales} = \beta_0 + \beta_1 \text{facebook} + \beta_2 \text{newspaper} + \beta_3 \text{youtube} + \varepsilon$$

Interprete los coeficientes estimados, comente el valor-p asociado a la hipótesis nula que el coeficiente es igual a cero e interprete el valor del coeficiente de determinación. Concluya.

Solución del problema

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.5386  -3.3991  -0.1284   3.0606  17.6507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.241279   1.174399   3.611 0.000387 ***
youtube     0.199802   0.003655  54.660  < 2e-16 ***
facebook    0.096873   0.020905   4.634 6.53e-06 ***
newspaper   0.004712   0.014096   0.334 0.738547
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.179 on 196 degrees of freedom
Multiple R-squared:  0.9385,    Adjusted R-squared:  0.9376
F-statistic: 997.4 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figura 4: Reporte sobre el modelo relacionando ventas con Youtube, Facebook y Newspaper

Esos coeficientes nos indican que un incremento de mil dólares en la inversión en Youtube genera 200 dólares en las ventas, un incremento de mil dólares en Facebook genera 100 dólares en las ventas y un incremento de mil dólares en el periódico genera 4 dólares en las ventas. Además de esto, si no se invierte en ninguna de las opciones se tienen 4.42 mil dólares fijos en ventas.

Podemos ver que las variables Youtube y Facebook son mucho más aportantes que la variable Newspaper en las ventas.

El valor-p asociado al modelo es 2.2×10^{-16} y eso es muy cercano a cero. Esto indica que alguna de las variables analizadas en el modelo está influyendo realmente en las ventas (caso que ya observamos con Youtube).

También sabemos que el valor de R^2 es igual a 0.9376, lo cual nos dice que los resultados que obtenemos son muy parecidos al modelo ya que el valor de R^2 es muy proximo a 1.

En resumen, R nos ayuda a determinar que este modelo es realmente muy adecuado para analizar las ventas pero tal vez mejoraría si se descarta la variable Newspaper.

Problema 9

Enunciado

Ajuste el modelo

$$\text{sales} = \beta_0 + \beta_1 \text{facebook} + \beta_2 \text{youtube} + \varepsilon$$

a) Interprete los coeficientes del modelo, comente el valor-p asociado a la hipótesis nula que el coeficiente es igual a cero e interprete el valor del coeficiente de determinación. Concluya. **b)** Compruebe vía métodos gráficos los supuestos del modelo. **c)** Estime un intervalo donde se encuentre las ventas esperadas cuando se desea invertir 170 mil en youtube y 30 mil en facebook. **d)** ¿Cuánto será lo que se ganará si invierte 170 mil en youtube y 30 mil en facebook?

Solución del problema

a) Utilizando el software R, podemos obtener un reporte sobre un modelo con esa configuración (usando la variable Youtube y Facebook) que arroja lo siguiente:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.423243   1.038252   4.260 3.16e-05 ***
youtube     0.199842   0.003645  54.823  < 2e-16 ***
facebook    0.096367   0.020803   4.632 6.56e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.168 on 197 degrees of freedom
Multiple R-squared:  0.9385,    Adjusted R-squared:  0.9379
F-statistic:  1503 on 2 and 197 DF,  p-value: < 2.2e-16
```

Figura 5: Reporte sobre el modelo de Facebook y Youtube

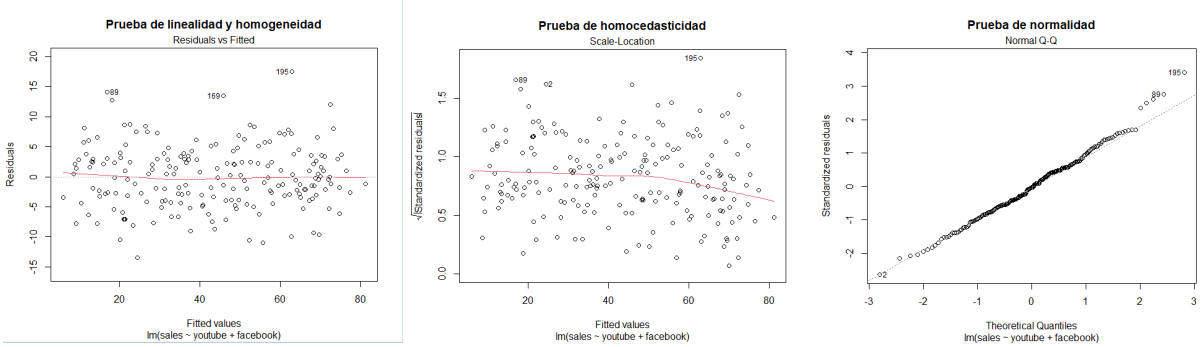
Según este reporte, podemos ver que el resultado final es sales $= 4.42 + 0.2 \cdot$ youtube $+ 0.1 \cdot$ facebook, con un coeficiente de determinación de 0.94.

Esos coeficientes nos indican que un incremento de mil dólares en la inversión en Youtube genera 200 dólares en las ventas y un incremento de mil dólares en Facebook genera 100 dólares en las ventas. Además de esto, si no se invierte en ninguna de las dos opciones se tienen 4.42 mil dólares fijos en ventas.

El valor-p asociado al modelo es 2.2×10^{-16} y eso es muy cercano a cero. Esto indica que alguna de las variables analizadas en el modelo está influyendo realmente en las ventas (caso que ya observamos con Youtube).

En conclusión, R nos ayudó a determinar que, efectivamente, este modelo es muy bueno para analizar las ventas.

b) Veamos ahora los supuestos del modelo graficados en R:



(a) Linealidad y homogeneidad (b) Homocedasticidad (c) Normalidad

Figura 6: Supuestos del modelo

En la **figura 6a** vemos puntos cercanos a la linea cero, en la **figura 6b** vemos una distribución uniforme con respecto a la linea roja (lo que representa una varianza constante) y en la **figura 6c** vemos puntos siguiendo esa linea recta que justifica la normalidad de los datos.

Con esto podemos ver que el modelo si cumple con las condiciones para tener linealidad, homogeneidad, homocedasticidad y normalidad.

c) Dado que el modelo es de varias variables y se complica usar la fórmula del problema 6, podemos calcular este intervalo usando R el cual determina intervalos para los valores que se entreguen.

Tenemos entonces que el intervalo de ganancias en ventas está entre $(40.56,\ 42.01)$ para esos valores invertidos en Youtube y Facebook

d) Podemos responder de dos maneras a esta pregunta.

La primera es que el modelo se ajusta a un valor de 41.29 dentro de su intervalo de confianza y puede esperarse eso.

La segunda es que se puede predecir un intervalo futuro de ganancia en ventas que, según R, se estima en el intervalo de $(31.07,\ 51.50)$ que, en miles, sería el intervalo de lo que se espera.