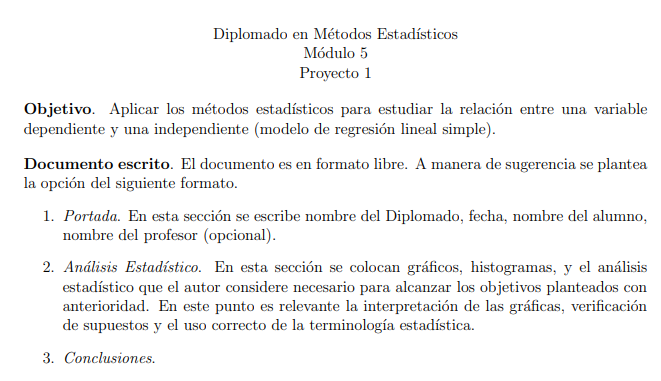
Diplomado en métodos estadísticos aplicados

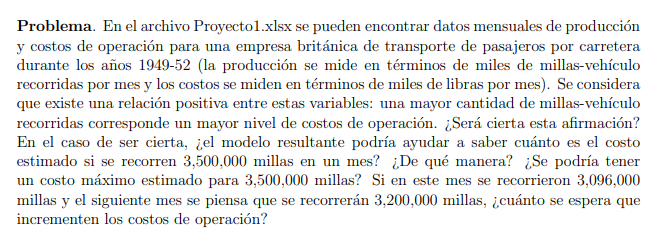
Módulo 5 – Relación entre variables

25 de septiembre de 2021

Alumno: Guillermo Gómez Sánchez

Profesor: Henry Gaspar Panti Trejo



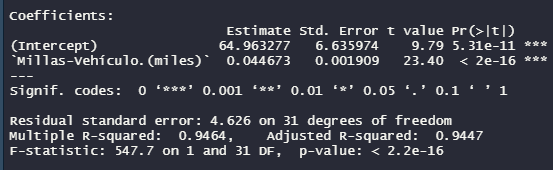


Análisis estadístico

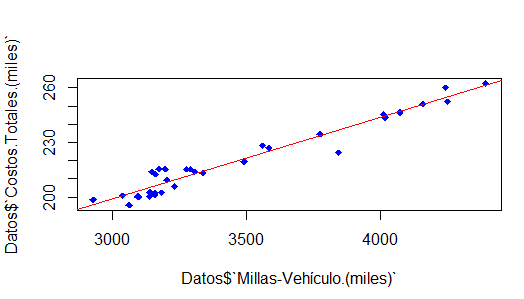
Debido a que usaremos los datos obtenidos de la regresión lineal simple para hacer estimaciones, procederemos a checar que se cumplan las siguientes condiciones:

1. La relación entre las variable X e Y es lineal
2. Los errores tienen distribución normal
3. El error tiene media cero
4. El error tiene varianza constante

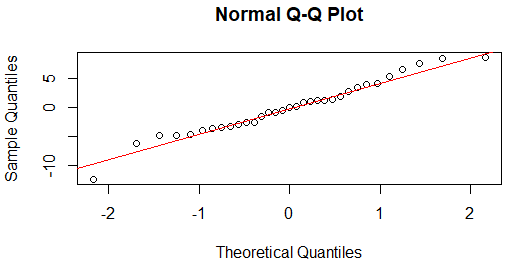
1.- Para probar que la relación entre la variable X (Millas de vehículo por mes) y la variable Y (Costo por mes) es lineal se realiza una regresión lineal.



Al ser la R2 .9464 es una buena aproximación lineal ya que es muy cercano a 1, esto se puede ver en el siguiente gráfico, donde los puntos azules son los datos a evaluar y la línea roja es la línea obtenida por medio de la regresión lineal simple. Tanto β1 como β0 tiene como hipótesis nula que son iguales a cero y como hipótesis alternativa que son diferentes de cero. Como tanto el P-valor de β1 (< 2\*10-16) como β0 (5.31\*10-11) son menores a un nivel de significancia de 0.001 por lo que son diferentes a cero. Al ser β1 diferente de cero hay relación significativa entre las variables y β0 tiene sentido, ya que es el costo de operación de la empresa de transporte sin recorrer ninguna milla con algún vehículo.



2.- La gráfica de distribución normal es la siguiente.

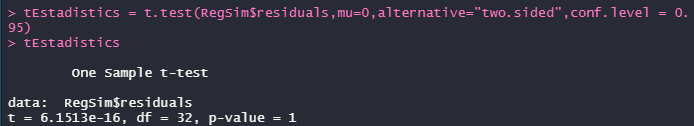


Esta gráfica parece indicarnos una buena aproximación a la normal, pero para estar completamente seguros utilizamos los siguientes tests de normalidad:

* Shapiro-Wilk. P-valor: 0.5768
* Anderson-Darling. P-valor: 0.6766
* Cramer-von Mises. P-valor: 0.8221
* Lilliefors (Kolmogorov-Smirnov) . P-valor: 0.7788
* Pearson chi cuadrada. P-valor: 0.6276
* Shapiro-Francia. P-valor: 0.4149

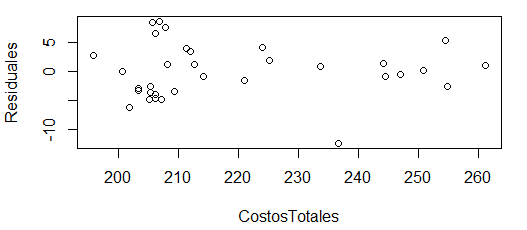
En estas pruebas h0 = tiene distribución normal y ha = no tiene distribución normal, por lo que podemos concluir que para todas las pruebas los errores tienen distribución normal para un nivel de significancia de 0.05.

3.- Dado que checamos que los errores tienen distribución normal utilizaremos una prueba de t de student con un nivel de confianza del 95% para checar si la media es igual o diferente de cero con un h0: media=0 y ha: media ≠ 0.



Al ser el P-valor = 1 no se rechaza h0 con un nivel de significancia de 0.05 por lo que la media del error es igual a 0.

4.- Usando el test de varianza no constante (P-valor = 0.57699) y el test de Breusch-Pagan estudiantizado (P-valor = 0.599) con h0 = varianza constante y ha=varianza no constante obtenemos que la varianza es constante para ambas pruebas. Esta es la gráfica.



Al cumplirse todas las condiciones de mencionadas al principio, concluimos que el modelo es suficientemente bueno para dar buenas estimaciones a las preguntas que se nos pidan.

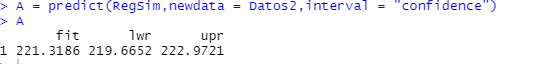
Para resolver estas preguntas las analizaremos por separado mediante incisos, por lo que empezaremos con:

1. ¿Será cierto que existe una relación positiva entre las variables?

Para resolver esta pregunta tenemos que checar que sea mayor a cero. Al checar rstudio obtenemos que R2 = 0.9464 por lo que la regresión lineal es una buena aproximación y podemos fiarnos de ella. Al checar el p-valor de obtenemos que es menor a 2\*10-16 y siendo h0: = 0 y ha: ≠ 0; podemos concluir que es diferente de cero con un nivel de significancia de .05 y al checar su intervalo de confianza con el mismo nivel de significancia obtenemos que es [0.04078, 0.04856592] por lo que podemos asegurar que al no incluir el cero el intervalo de confianza y su valor mínimo ser mayor a 0, la relación entre las variables es positiva ya que al incrementar una variable la otra también lo hace.

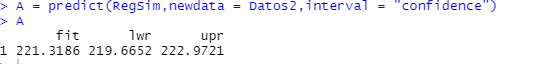
1. ¿El modelo resultante podría ayudar a saber cuánto es el costo estimado si se recorren 3,500,000 millas en un mes? ¿de qué manera?

Si serviría ya que al checar los datos de los costos (que vienen siendo estos en miles [3147,3160,3197,3173,3292,3561,4013,4244,4159,3776,3232,3141,2928,3063,3096,3096,3158,3338,3492,4019,4394,4251,3844,3276,3184,3037,3142,3159,3139,3203,3307,3585,4073]) podemos ver que el mínimo valor proporcionado es 2,928,000 millas por mes y el máximo valor proporcionado es 4,394,000 millas por mes y al encontrarse el 3,500,000 entre estos dos valores se puede estimar el valor correctamente ya que se conoce la tendencia que los datos tienen entre esos valores. Al estimar el costo con el software mediante la función predict obtenemos un costo estimado de 221.3186 con un intervalo de estimación de [219.6652, 222.9721].



1. ¿Se podría tener un costo máximo estimado para 3,500,000 millas?

Sí, al estimar el costo con el software mediante la función predict obtenemos un costo máximo estimado de 222.9721 miles de libras por mes.



1. Si en este mes se recorrieron 3,096,000 millas y el siguiente mes se piensa que se recorrerán 3,200,000 millas, ¿cuánto se espera que incrementen los costos de operación?

Para saber el intervalo de incremento debemos saber desde el menor incremento posible al mayor incremente posible. El menor incremento posible sería cuando en el primer mes se obtiene el costo máximo estimado y en el segundo mes el mínimo costo estimado. El mayor incremento posible sería cuando en el primer mes se obtiene el costo mínimo estimado y en el segundo mes el máximo costo estimado.

Para 3,096,000 millas el intervalo de estimación se encuentra entre [201.1246,205.4169] mientras que para 3,200,000 millas el intervalo de estimación se encuentra entre [206.006,209.8275]. Tomando en cuenta lo escrito en el párrafo anterior el intervalo de incremento sería de [0.5891, 8.7029] y el incremento esperado es de 4.6459 (Δy = Δx para una recta. Entonces Δy = 0.044673(3,200,000 – 3,096,000)) miles de libras por mes.