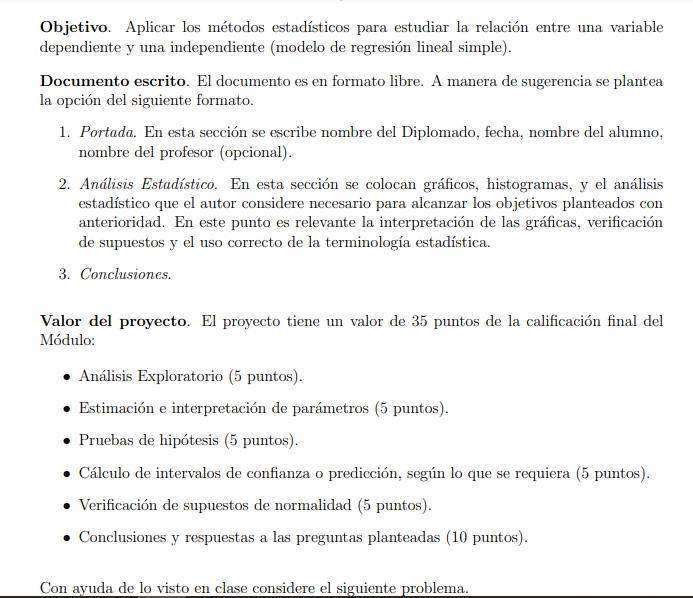
Diplomado en métodos estadísticos aplicados

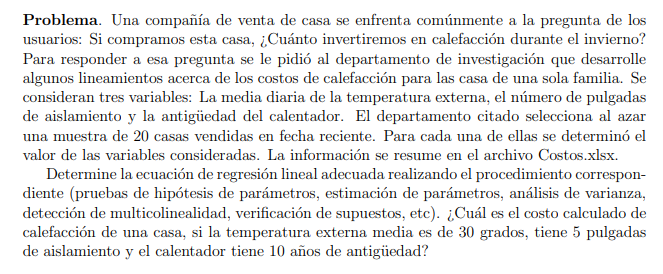
Módulo 5 – Relación entre variables

25 de septiembre de 2021

Alumno: Guillermo Gómez Sánchez

Profesor: Henry Gaspar Panti Trejo



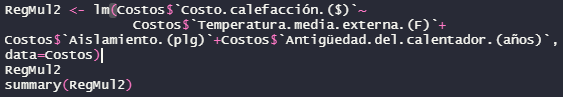


Análisis estadístico

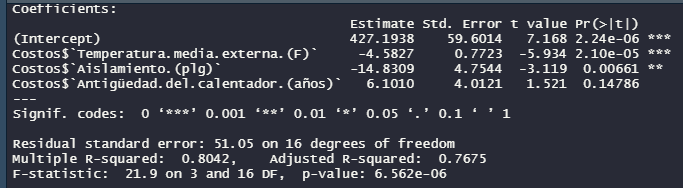
Debido a que usaremos los datos obtenidos de la regresión lineal múltiple para hacer estimaciones, procederemos a checar que se cumplan las siguientes condiciones:

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
5. Las variables independientes no poseen multicolinealidad

Para probar que la relación entre las variables independientes y la variable dependiente se puede aproximar mediante una regresión lineal múltiple utilizamos el siguiente comando.

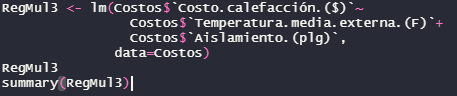


El resultado regresado por este comando es el siguiente:

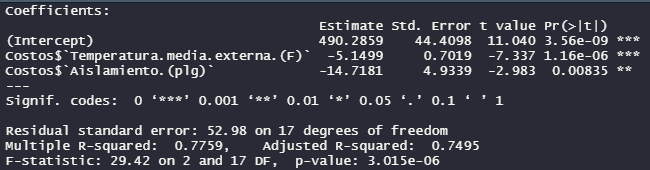


Antes de continuar analizando este modelo, checaremos si todas las variables resultan ser significativas para ver si continuamos con el análisis del modelo. En este caso podemos ver que la antigüedad del calentador no es significativa, debido a que su p-valor de 0.14786, así que realizaré la regresión lineal múltiple sin esta variable en esta ocasión.

Al eliminar esta variable del modelo de regresión linear múltiple como se ve en la siguiente imagen



El resultado regresado por este comando es el siguiente:



β0 = 490.2859 de costo y tiene sentido ya que es el costo de calefacción de una casa cuanto la temperatura externa media es 0 grados y no tiene aislamiento.

β1 = -5.1499 e indica un decremento de 5.1499 de costo de calefacción de una casa cuando aumenta en 1 grado Fahrenheit la temperatura media externa manteniendo constante el grosor del aislamiento.

β2 = -14.7181 e indica un decremento de 14.7181 de costo de calefacción de una casa cuando aumenta en una pulgada el aislamiento manteniendo constante la temperatura media externa.

El aislamiento es la variable que tiene mayor efecto sobre el modelo ya que tiene el parámetro con mayor magnitud (14.7181).

El p-valor de β0 indica que para un nivel de significancia de 0.05 β0 no es significativamente diferente de cero, por lo que la línea obtenida por la regresión lineal pasa por el origen.

El p-valor de temperatura media externa indica que para un nivel de significancia de 0.05 β1 es significativamente diferente de cero, por lo que existe una relación significativa entre las variables temperatura media externa y los costos de calefacción.

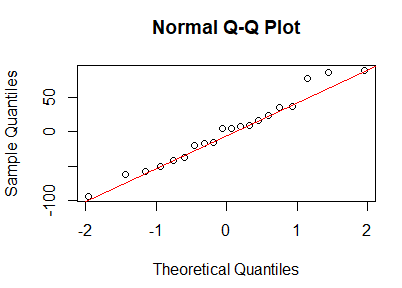
El p-valor del aislamiento indica que para un nivel de significancia de 0.05 β2 es significativamente diferente de cero, por lo que existe una relación significativa entre las variables aislamiento y los costos de calefacción.

El R2ajustada es igual a 0.7495 por lo que es una buena aproximación lineal múltiple.

Los intervalos de confianza de los parámetros con un nivel de confianza del 95% son:

β0 = [396.589346, 583.982507], β1 = [-6.630735, -3.669032], β2 = [-25.127806, -4.308491]

2.- 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: 0.7359

• Anderson-Darling. P-valor: 0.7708

• Cramer-von Mises. P-valor: 0.8394

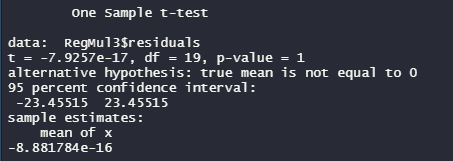
• Lilliefors (Kolmogorov-Smirnov). P-valor: 0.9407

• Pearson chi cuadrada. P-valor: 0.9898

• Shapiro-Francia. P-valor: 0.7881

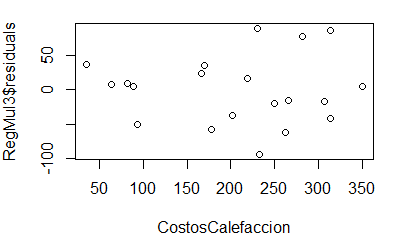
En estas pruebas h0 = tiene distribución normal y ha = no tiene distribución normal, por lo que podemos concluir que para todas las pruebas los errores tienen distribución normal para un nivel de significancia de 0.05.

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.

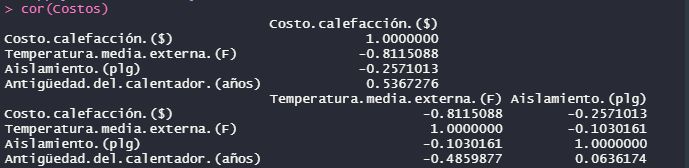


Al ser el P-valor = 1 no se rechaza h0 con un nivel de significancia de 0.05 por lo que la media del error es igual a 0. Esto igual puede confirmarse viendo como en el intervalo de confianza el cero se encuentra incluido.

4.- Usando el test de varianza no constante (P-valor = 0.27752) y el test de Breusch-Pagan estudiantizado (P-valor = 0.1808) con h0 = varianza constante y ha=varianza no constante obtenemos que la varianza es constante para ambas pruebas. Esta es la gráfica.



5.- Se realizó el comando de correlación para verificar que efectivamente



No hay correlaciones altas para poder afirmar que la temperatura media externa es múltiplo del aislamiento (en este caso no tomamos en cuenta la antigüedad del calentador debido a que fue eliminado para el modelo final).

Habiendo cumplido estos 5 supuestos podemos decir con seguridad que es posible realizar una aproximación fiable con nuestro modelo.

La ecuación de la recta del modelo es ‘Costos Calefacción = 490.2859 -5.1499 Temperatura Media Externa -14.7181 Aislamiento‘.

¿Cuál es el costo calculado de calefacción de una casa, si la temperatura externa media es de 30 grados, tiene 5 pulgadas de aislamiento y el calentador tiene 10 años de antigüedad?

En este caso utilizaremos la ecuación lineal que acabamos de definir para hacer un cálculo de cuanto sería el costo de calefacción (sin tomar en cuenta los errores de predicción y estimación). Debido a que en el modelo final resultó que la antigüedad del calentador no era significativa, el dato que nos provee la pregunta no será tomado en cuenta.

Al sustituir obtenemos que el costo de calefacción es 490.2859 -5.1499\*30 -14.7181\*5 = 262.1984 de costo.

Ahora tomando en cuenta errores de estimación y predicción procederé a dar un intervalo de confianza para los valores. Antes de poder hacer esto debo mostrar que los datos utilizados para la estimación se encuentran en el intervalo de los datos muestreados para poder hacer una buena predicción.

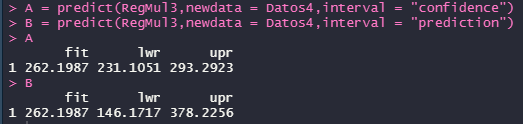
Los datos de temperatura media externa son los siguientes [35,29,36,60,65,30,10,7,21,55,54,48,20,39,60,20,58,40,27,30] siendo la temperatura mínima de 7 y la máxima de 65, el valor de 30 se encuentra contenido en el intervalo.

Los datos de aislamiento son los siguientes [3,4,7,6,5,5,6,10,9,2,12,5,5,4,8,5,7,8,9,7] siendo el aislamiento mínimo de 2 y el máximo de 12, el valor de 5 se encuentra contenido en el intervalo.

Debido a que ambos valores se encuentran dentro de sus respectivos intervalos, procedemos a hacer las predicciones y estimaciones.

Al utilizar la función predict (‘a’ siendo la temperatura media externa y ‘b’ el aislamiento) para estimar y predecir el costo de calefacción obtenemos la siguiente respuesta del software





Esto nos dice que tenemos un costo de calefacción medio de 262.1987, mientras que el intervalo de confianza de la estimación es de [231.1051,293.2923] y el intervalo de confianza de la predicción es de [146.1717,378.2256].