**Fundación Universitaria Los Libertadores**

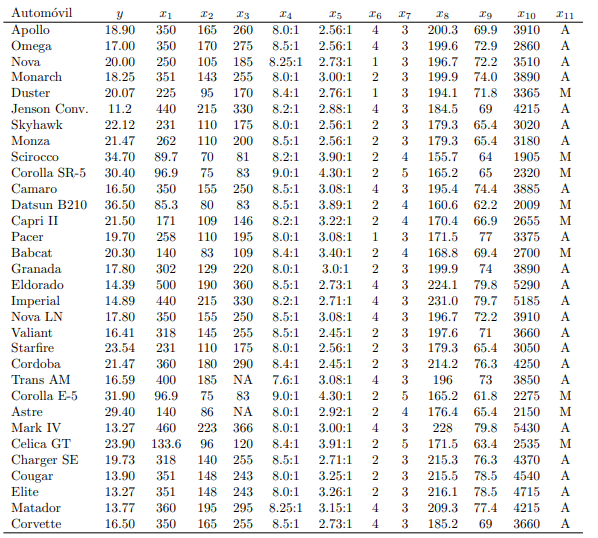
**Diego Alexander Torres**

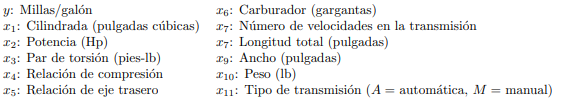
**Taller 4: Regresión Aplicada**

**Especialización en Estadística Aplicada**

**PARTE l**

El siguiente taller incluye un problema aplicado sobre análisis residual en el modelo de regresión lineal múltiple con los siguientes datos.





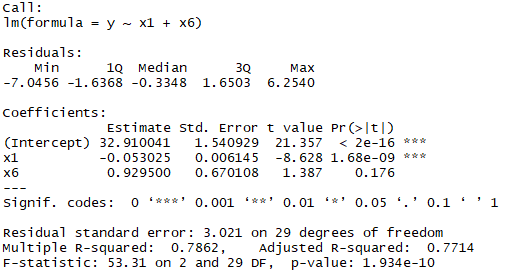
La base de datos va a estar ligada a la variable T4.



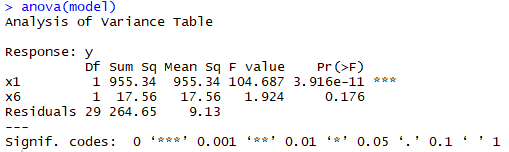
**Problemas**

1. **Ajustar un modelo de regresión lineal múltiple que relacione el rendimiento de la gasolina y, en millas por galón, la cilindrada del motor x1 y la cantidad de gargantas del carburador, x6.**





1. **Formar la tabla de análisis de varianza, y probar la significancia de la regresión.**

****

Se rechaza y se comprueba que efectivamente existe diferencia entre los grupos, aunque no se sepa en cuales. Entonces podremos afirmar que el efecto observado es demasiado grande para poder ser explicado por el azar (error aleatorio).

1. **Calcular y para este modelo. Interprete los resultados.**

De la regresión obtenemos que:

Multiple R-squared: 0.7862, Adjusted R-squared: 0.7714

Los valores obtenidos indican que el modelo, con todas las variables introducidas como predictores, tiene un  alta (0.7862), es capaz de explicar el 78,62% de la variabilidad observada en el desempeño del combustible.

1. **Determinar un intervalo de confianza de 95% para β1.**

****

1. **Calcular el estadístico t para probar :** **β1 = 0 y : β6 = 0. ¿Qué conclusiones se pueden sacar?**

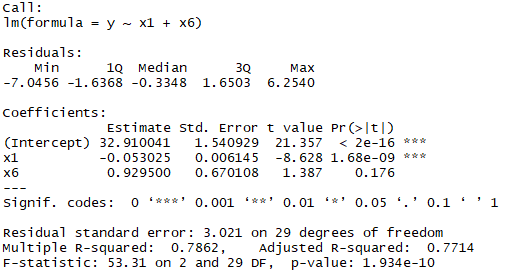
β1 = 0

β1 ≠ 0

β6 = 0

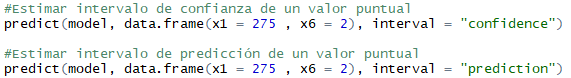
β6 ≠ 0

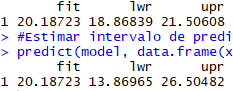
De la regresión obtenemos:



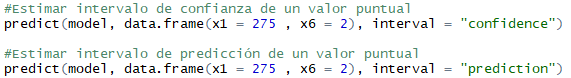
Del valor del estadístico t para los parámetros β1 y β6 es posible concluir que β1 es significativo y aporta al modelo, mientras que β6 no lo es, no aporta mucha información al modelo.

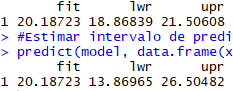
1. **Determinar un intervalo de confianza de 95% para el rendimiento promedio de la gasolina, cuando x1 = 275 pulgadas cúbicas y x6 = 2 gargantas.**

****

****

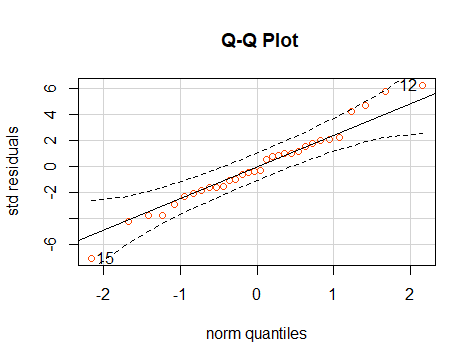
1. **Determinar un intervalo de predicción de 95% para una nueva observación de rendimiento de gasolina cuando x1 = 275 pulgadas cubicas y x6 = 2 gargantas.**

****

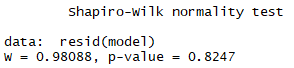
****

1. **Trazar una gráfica de probabilidad normal de los residuales. ¿Parece haber**

**algún problema con la suposición de normalidad?**

****

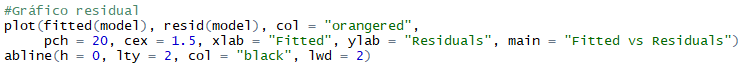


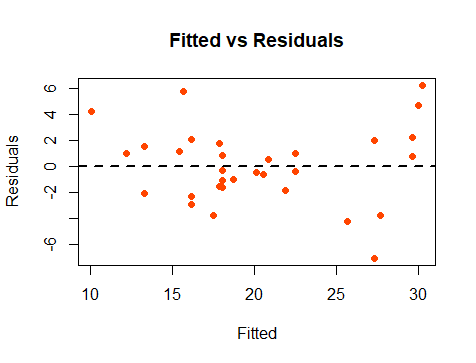


Al hacer la prueba Shapiro-Wilk, se puede observar que efectivamente se está cumpliendo el requisito de normalidad de los residuos. Sin embargo, en la gráfica se evidencia que existen ciertos valores residuales como el 12 o el 15 que pueden generar conflicto en el modelo y valdría la pena revisarlos.

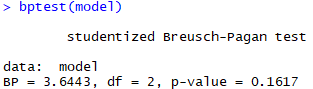
1. **Trazar e interpretar una gráfica de los residuales en función de la respuesta**

**predicha.**

****

****

El gráfico demuestra cierta heteroscedasticidad en los residuos del modelo, ya que parece que los residuos no se encuentran dentro de un intervalo constante. Para esto, se realizará una prueba Breusch-Pagan.



Según la prueba, hay homocedasticidad en los datos residuales.

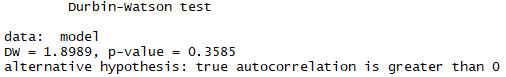
1. **Realice una verificación de la adecuación modelo, que incluya calcular los**

**residuos estudentizados y la medida de distancia de Cook para cada una**

**de las observaciones. Comente sus resultados.**

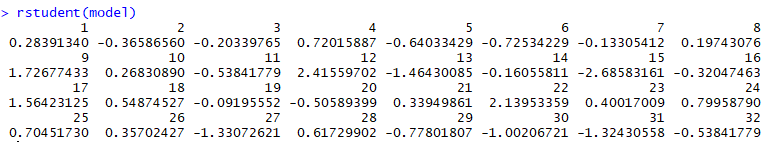
1. Siguiendo con el procedimiento de verificación del modelo, ya tenemos que los supuestos de normalidad y homocedasticidad de los residuales están bien, se han cumplido. Ahora, buscaremos demostrar la independencia en el modelo.





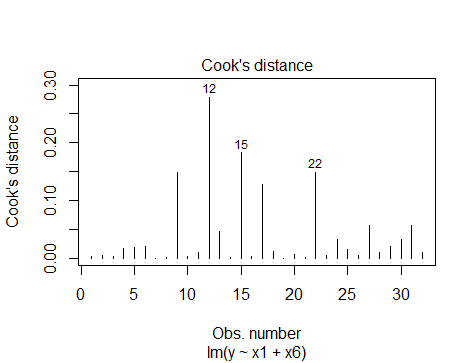
Se confirma que la covarianza entre los residuos del modelo son cero.

1. A continuación, los residuos ponderados (estudentizados)



1. Ahora, bien, con respecto a la distancia de Cook de los residuales, tenemos:

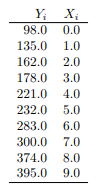




El gráfico de las distancias de Cook para los datos residuales del modelo indica que, tal como se esperaba, hay una alta influencia de los puntos 12, 15 y también del 22 en el análisis de regresión.

**PARTE ll**

**Crecimiento de las ventas:** Un investigador de mercado estudió las ventas anuales de un producto que se introdujo hace 10 años. Los datos son los siguientes, donde X es el año (codificado) e Y es ventas en miles de unidades:

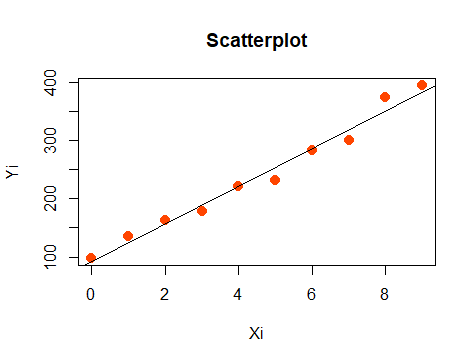


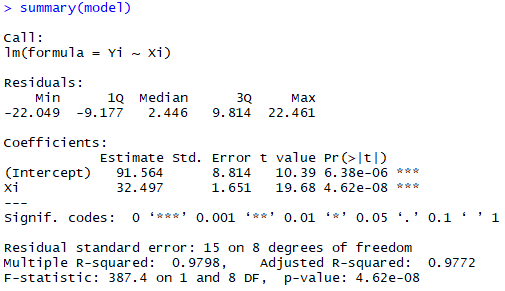
****

**Problemas**

1. **Haga un diagrama de dispersión de los datos. ¿Considera que es adecuada una relación lineal?**

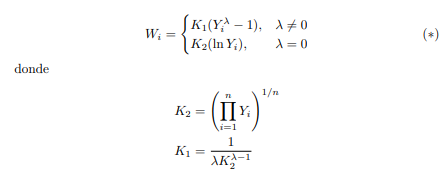


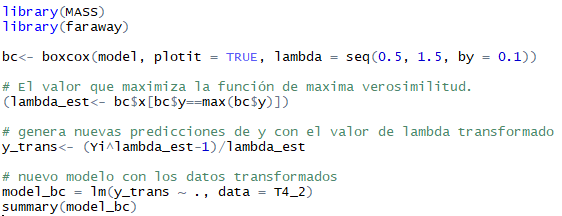




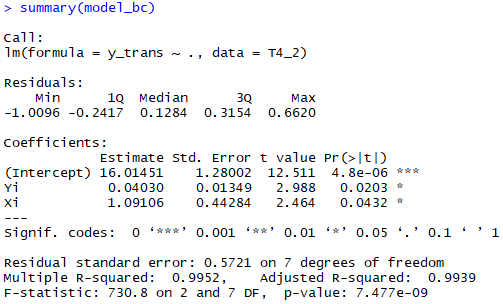
Revisando el modelo, considero que se ajusta bastante bien a una relación lineal entre las variables. Diría que es hasta inmejorable.

1. **Use el procedimiento de Box-Cox y la estandarización (∗) para encontrar una transformación de potencia apropiada de Y. evalúe la SCE para λ = 0.3, 0.4, 0.5, 0.6, 0.7. ¿Qué transformación de Y se sugiere?**

****

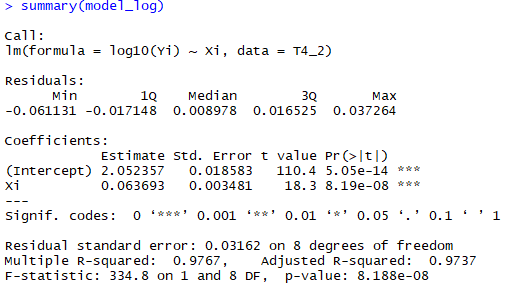
****

El valor de lambda\_ est es de 05202.



1. **Use la transformación Y 0 = log10 Y y obtenga la función de regresión lineal estimada para los datos transformados.**

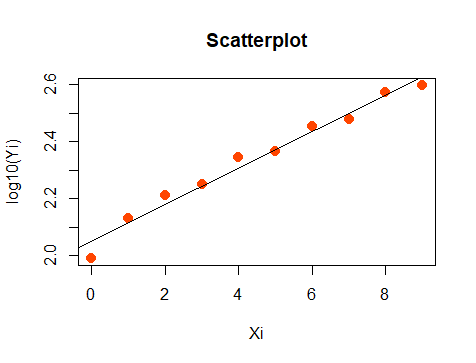
****

****

La ecuación para la regresión lineal de este modelo es de Yi = 10^(2.05 + 0.0636Xi)

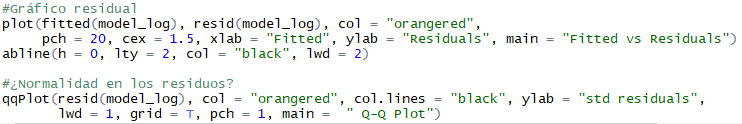
1. **Trace la recta de regresión estimada y los datos transformados. ¿La recta de regresión parece ajustarse bien a los datos transformados?**

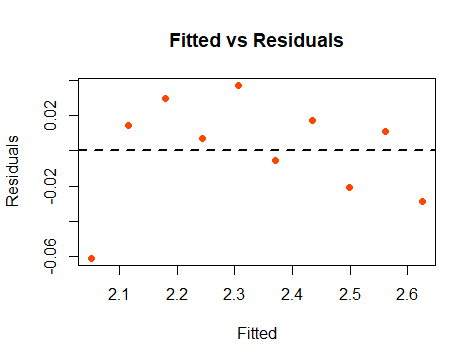
****

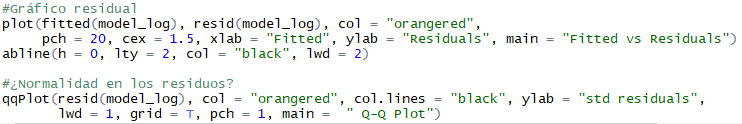


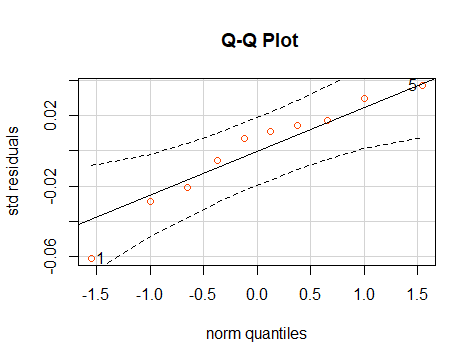
La recta de la regresión lineal entre el log10(Yi) y Xi parece ajustarse bien a los datos transformados.

1. **Obtenga los residuales y grafíquelos contra los valores estimados. También haga un diagrama de probabilidad normal. ¿Qué muestran estas gráficas?**

****

****





Las gráficas demuestran que el modelo es coherente, ya que es normal y homoscedástico.

1. **Exprese la función de regresión estimada en las unidades originales.**

Yi = 10^(2.05 + 0.0636Xi)