**Ejercicio Propuesto1**

**Dada la cantidad de gasolina en porcentaje con respecto a la cantidad del petróleo en crudo, denotada por y. Se quiere saber si se puede expresar como combinación lineal de cuatro variables: la gravedad del crudo, la presión del vapor del crudo, la temperatura para la cual se ha evaporado un 10% y la temperatura para la cual se ha evaporado el 100%, a partir de los siguientes datos:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Y****Cant\_gasolina** | **Gravedad****crudo** | **Presión****vapor** | **Temperatura****Vapor 10%** | **Temperatura****Vapor 100%** |
| **6,90****14,40****7,40****8,50****8,00****2,80****5,00****12,20****10,00****15,20****26,80****14,00****14,70****6,40****17,60****22,30****24,80****26,00****34,90****18,20****23,20****18,00****13,10****16,10****32,10****34,70****31,70****33,60****30,40****26,60****27,80****45,70** | **38,40****40,30****40,00****31,80****40,80****41,30****38,10****50,80****32,20****38,40****40,30****32,20****31,80****41,30****38,10****50,80****32,20****38,40****40,30****40,00****32,20****31,80****40,80****41,30****38,10****50,80****32,20****38,40****40,00****40,80****41,30****50,80** | **6,10****4,80****6,10****,20****3,50****1,80****1,20****8,60****5,20****6,10****4,80****2,40****,20****1,80****1,20****8,60****5,20****6,10****4,80****6,10****2,40****,20****3,50****1,80****1,20****8,60****5,20****6,10****6,10****3,50****1,80****8,60** | **220****231****217****316****210****267****274****190****236****220****213****284****316****267****274****190****236****220****231****217****284****316****210****267****274****190****236****220****217****210****267****190** | **235****307****212****365****218****235****285****205****267****300****367****351****379****275****365****275****360****365****395****272****424****428****273****358****444****345****402****410****340****347****416****407** |

[Tabla de datos del ejercicio propuesto 3](https://wpd.ugr.es/~bioestad/wp-content/uploads/mul2.txt)

**Se pide:**

1. **Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación**
2. **Contrastar la significación del modelo propuesto**
3. **¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales**
4. **Coeficiente de determinación y de determinación corregido.**

### **Solución****1. Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación**

Comenzamos introduciendo los datos en R. Para ello, creamos un vector numérico que recoja los datos de cada variable y, a continuación, un data.frame que agrupe a los cuatro vectores creados.

> y <- c(6.9, 14.4, 7.4, 8.5, 8.0, 2.8, 5, 12.2, 10, 15.2, 26.8, 14, 14.7, 6.4, 17.6, 22.3, 24.8, 26, 34.9, 18.2, 23.2, 18, 13.1, 16.1, 32.1, 34.7, 31.7, 33.6, 30.4, 26.6, 27.8, 45.7)

> Gravedad <- c(38.4, 40.3, 40, 31.8, 40.8, 41.3, 38.1, 50.8, 32.2, 38.4, 40.3, 32.2, 31.8, 41.3, 38.1, 50.8, 32.2, 38.4, 40.3, 40, 32.2, 31.8, 40.8, 41.3, 38.1, 50.8, 32.2, 38.4, 40, 40.8, 41.3, 50.8)

> Presion <- c(6.1, 4.8, 6.1, 0.2, 3.5, 1.8, 1.2, 8.6, 5.2, 6.1, 4.8, 2.4, 0.2, 1.8, 1.2, 8.6, 5.2, 6.1, 4.8, 6.1, 2.4, 0.2, 3.5, 1.8, 1.2, 8.6, 5.2, 6.1, 6.1, 3.5, 1.8, 8.6)

> Temp\_Evap\_10 <- c(220, 231, 217, 316, 210, 267, 274, 190, 236, 220, 213, 284, 316, 267, 274, 190, 236, 220, 231, 217, 284, 316, 210, 267, 274, 190, 236, 220, 217, 210, 267, 190)

> Temp\_Evap\_100 <- c(235, 307, 212, 365, 218, 235, 285, 205, 267, 300, 367, 351, 379, 275, 365, 275, 360, 365, 395, 272, 424, 428, 273, 358, 444, 345, 402, 410, 340, 347, 416, 407)

> datos <- data.frame(y, Gravedad, Presion, Temp\_Evap\_10, Temp\_Evap\_100)

Una vez creado el data frame con los datos consultamos sus primeros registros mediante la orden head.

> head(datos)

 **Gravedad Presion Temp\_Evap\_10 Temp\_Evap\_100**

**1 38.4 6.1 220 235**

**2 40.3 4.8 231 307**

**3 40.0 6.1 217 212**

**4 31.8 0.2 316 365**

**5 40.8 3.5 210 218**

**6 41.3 1.8 267 235**

A continuación, vamos a crear el modelo de regresión múltiple utilizando el comando **lm**.

> reg\_lin\_mul <- lm(y ~ Gravedad + Presion + Temp\_Evap\_10 + Temp\_Evap\_100, data = datos)

> summary(reg\_lin\_mul)

**Call:**

**lm(formula = y ~ Gravedad + Presion + Temp\_Evap\_10 + Temp\_Evap\_100,**

 **data = datos)**

**Residuals:**

 **Min 1Q Median 3Q Max**

**-3.513 -1.709 0.087 1.610 4.763**

**Coefficients:**

 **Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -9.809720 10.160831 -0.965 0.3429**

**Gravedad 0.237989 0.102665 2.318 0.0283 \***

**Presion 0.657347 0.372128 1.766 0.0886 .**

**Temp\_Evap\_10 -0.138987 0.028950 -4.801 5.2e-05 \*\*\***

**Temp\_Evap\_100 0.153170 0.006574 23.299 < 2e-16 \*\*\***

**---**

**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**Residual standard error: 2.303 on 27 degrees of freedom**

**Multiple R-squared: 0.9598, Adjusted R-squared: 0.9539**

**F-statistic: 161.2 on 4 and 27 DF, p-value: < 2.2e-16**

El modelo de regresión lineal múltiple es, por tanto:

La interpretación de los parámetros del modelo es la que sigue:

* La cantidad del petróleo en crudo cuando su gravedad, su presión, su temperatura de evaporación del 10% y su temperatura de evaporación del 100% son iguales a 0 sería igual a -9.809 unidades. En este caso vemos que la interpretación del parámetro constante no tiene sentido.
* Cuando la gravedad del crudo aumenta en una unidad y el resto de variables permanecen invariantes, la cantidad esperada de petróleo en crudo aumenta también en 0.237 unidades.
* Cuando la presión del crudo aumenta en una unidad y el resto de variables permanecen invariantes, la cantidad esperada de petróleo aumenta en 0.657 unidades.
* Cuando la temperatura de evaporación del 10% de crudo aumenta en una unidad y el resto de variables permanecen invariantes, la cantidad esperada de petróleo disminuye en 0.138 unidades.
* Y, por último, cuando la temperatura de evaporación del 100% de crudo aumenta en una unidad y el resto de variables permanecen invariantes, la cantidad esperada de petróleo aumenta en 0.153 unidades.

**2. Contrastar la significación del modelo propuesto**

Vamos a comprobar ahora si, al menos, una de las variables independientes del modelo guarda relación lineal con la variable dependiente. En tal caso, concluiremos que la regresión lineal tiene sentido en este caso.

El contraste que debemos resolver es el siguiente:

El estadístico del contraste y el p-valor asociado a este contraste aparecen en la parte final de la salida de la regresión lineal múltiple.

**F-statistic: 161.2 on 4 and 27 DF, p-value: < 2.2e-16**

En este ejemplo, el estadístico de contraste vale 161.2. El p-valor asociado es un valor muy pequeño (muy cercano a 0) y, consecuentemente, menor que 0.05. Por ello, considerando un nivel de significación del 5%, podemos concluir la significación de la regresión. Es decir, al menos una de las variables independientes estudiadas tiene una relación lineal significativa con la variable dependiente.

**3. ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales**

Vamos a contrastar ahora si todas las variables independientes son relevantes a la hora de explicar la variable independiente o si, por el contrario, podemos prescindir de alguna. Para ello resolveremos los contrastes:

Los estadísticos de contraste y los p-valores para resolver estos contrastes se encuentran en la salida de R al lado de las estimaciones de los coeficientes beta.

**Coefficients:**

 **Estimate Std. Error t value Pr(>|t|)**

**(Intercept) -9.809720 10.160831 -0.965 0.3429**

**Gravedad 0.237989 0.102665 2.318 0.0283 \***

**Presion 0.657347 0.372128 1.766 0.0886 .**

**Temp\_Evap\_10 -0.138987 0.028950 -4.801 5.2e-05 \*\*\***

**Temp\_Evap\_100 0.153170 0.006574 23.299 < 2e-16 \*\*\***

Los p-valores asociados a las variables *Gravedad*, *Temp\_Evap\_10* y *Temp\_Evap\_100* son inferiores a 0.05. Consecuentemente, considerando un nivel de significación del 5% rechazamos la hipótesis nula correspondiente y concluimos que el coeficiente asociado a dichas variables es significativamente distinto de 0. Es decir, las tres variables son relevantes a la hora de explicar la variable dependiente. Por el contrario, los p-valores asociados al parámetro constante y a la variable *Presion* superan el valor 0.05, lo que nos lleva a concluir que ambos parámetros pueden asumirse iguales a 0. Esto equivale a decir que no existe ningún tipo de asociación lineal entre la variable *Presion* y la variable dependiente, de manera que no es aconsejable utilizar la variable *Presion* para explicar la variable dependiente. Convendría, por tanto, ajustar un nuevo modelo eliminando el parámetro constante y la variable *Presion*.

**4. Coeficiente de determinación y de determinación corregido**

Por último, comprobemos mediante el coeficiente de determinación y el coeficiente de determinación corregido la bondad de ajuste de los datos al modelo planteado.

**Multiple R-squared: 0.9598, Adjusted R-squared: 0.9539**

El valor del coeficiente de determinación es 0.959 y el de determinación corregido es 0.953. Esto significa que el 95.9% de la variabilidad presente en la variable dependiente puede explicarse por el modelo propuesto. El ajuste de los datos al modelo es, por tanto, excelente.