

PRACTICANDO LAS REGRESIONES LINEALES

Empezamos con los modelos y vamos a trabajar con las regresiones lineales.

Para ir familiarizándose propongo dos ejercicios.

Una regresión lineal simple

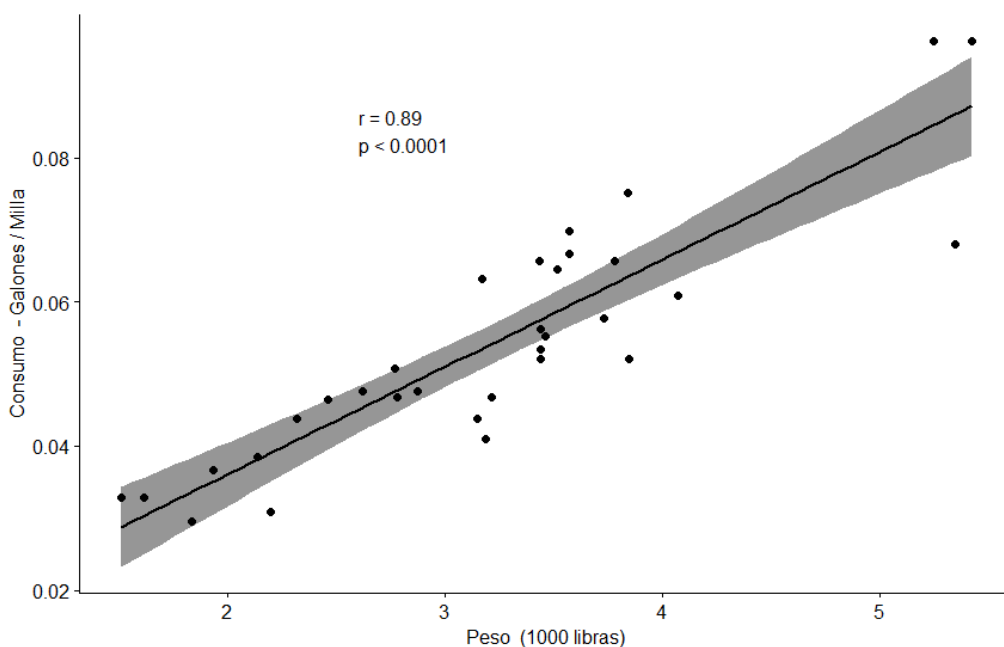
Una regresión lineal múltiple

Sigue los vídeos tech Zen porque será simplemente cambiar algunas variables.

Te escribo el ejemplo de una regresión lineal simple para que te facilite el trabajo:

Queremos encontrar un modelo que permita obtener el consumo del coche solo sabiendo el peso. Este modelo lo podemos encontrar utilizando datos de consumo y peso de distintos coches.

Crear un gráfico de dispersión



consumo	peso
0.04761905	2.620
0.04761905	2.875
0.04385965	2.320
0.04672897	3.215
0.05347594	3.440
0.05524862	3.460
0.06993007	3.570
0.04098361	3.190
0.04385965	3.150
0.05208333	3.440
0.05617978	3.440
0.06097561	4.070
0.05780347	3.730
0.06578947	3.780
0.09615385	5.250
0.09615385	5.424
0.06802721	5.345
0.03086420	2.200
0.03280474	1.615

El significado del p-valor

- H1: los coeficientes son significativamente diferentes de cero
- H0: los coeficientes son igual a 0

Establece el nivel de significación o el margen de error

- 5%

Elige una prueba o test estadístico

PRACTICANDO LAS REGRESIONES LINEALES

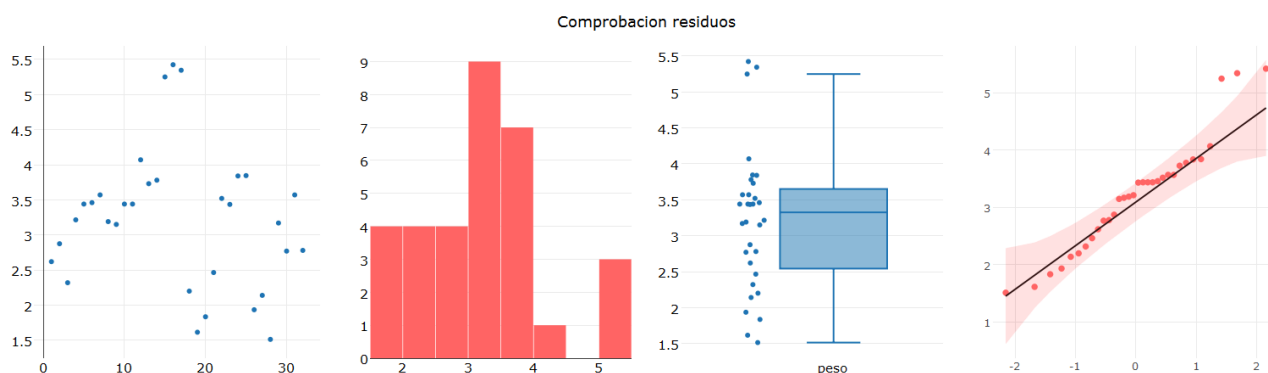
d. Regresión Lineal Simple

Calcula el p-valor y el coeficiente de correlación:

- e. Intercepto "a" = 0.006169 con un p-valor = 0.199 >> Debería ser 0 pero siempre se deja
- f. Pendiente o "b" = 0.014938 con un p-valor = 9.57e-12 >> significativo
- g. El modelo es:

$$\text{Consumo} = 0.014938 * \text{Peso} + 0.006169$$

- h. Los residuos son normales y NO presentan autocorrelación (gráficos).



Los gráficos reflejan la normalidad de los residuos y no presentan formas raras en el diagrama de dispersión.

Conclusión:

El modelo parece bueno aunque el p-valor de la b sea menos que 0.05. Normalmente el "intercept" se deja a no ser que estés muy seguro que ninguna otra variable tiene influencia con el consumo.

Este es el resultado de esta recta de regresión. Lo más importante está en la parte de coeficientes. Puedes ver el valor (Estimate) y el p-valor.

PRACTICANDO LAS REGRESIONES LINEALES

```

call:
lm(formula = consumo ~ peso, data = varEstudio)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0179837 -0.0044301  0.0009413  0.0045354  0.0116583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006169   0.004695   1.314   0.199
peso         0.014938   0.001398  10.685 9.57e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007616 on 30 degrees of freedom
Multiple R-squared:  0.7919,    Adjusted R-squared:  0.785
F-statistic: 114.2 on 1 and 30 DF, p-value: 9.566e-12

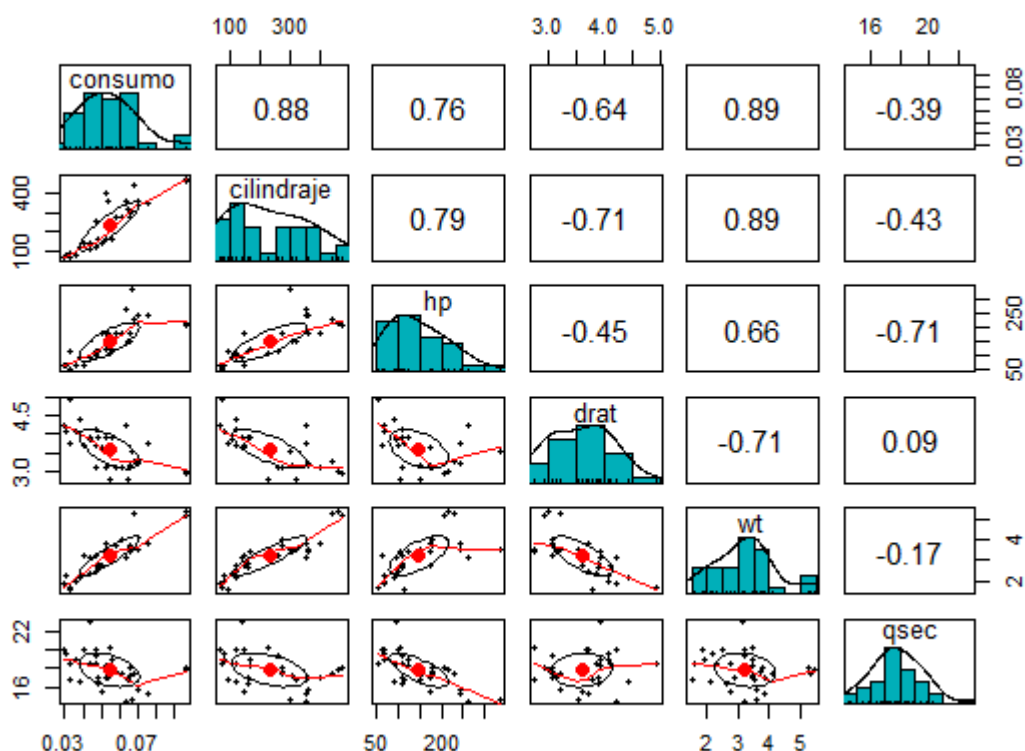
```

¡Bien! Ahora es tu turno.

REGRESIÓN LINEAL SIMPLE

Vamos a cambiar la variable de entrada a ver si conseguimos un buen modelo de regresión lineal.

Calcula el matrixplot de todas las variables:



PRACTICANDO LAS REGRESIONES LINEALES

Escoge una de las variables que, no sea el peso, que según el matrixplot y el coeficiente de correlación está más relacionada con el consumo (que no sea el peso)

Calcula el modelo y copia los resultados aquí:

Variable seleccionada = cilindraje (disp)

Los resultados son:

- La ecuación del modelo

$$\text{Consumo} = 0.0001166 (\text{cilindraje}) + 0.0273265$$

- Los resultados de la tabla de coeficientes y del modelo

```
> regresionLinealSimple <- lm(formula = consumo ~ cilindraje, varEstudio)
> summary(regresionLinealSimple)

call:
lm(formula = consumo ~ cilindraje, data = varEstudio)

Residuals:
    Min       1Q   Median       3Q      Max
-0.021881 -0.003825  0.001061  0.005191  0.015194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0273265   0.0030013    9.105 3.88e-10 ***
cilindraje   0.0001166   0.0000115   10.139 3.32e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007936 on 30 degrees of freedom
Multiple R-squared:  0.7741,    Adjusted R-squared:  0.7666
F-statistic: 102.8 on 1 and 30 DF,  p-value: 3.316e-11
```

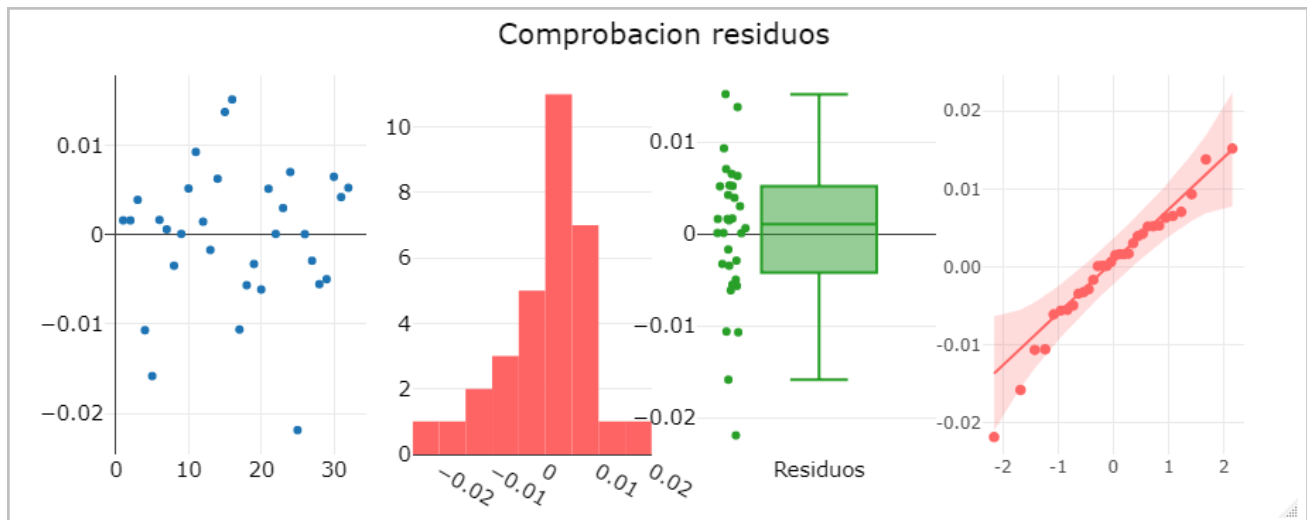
- La comprobación de los residuos

```
> # Prueba Distribución Normal
> shapiro.test(residuos)

      shapiro-wilk normality test

data:  residuos
W = 0.96228, p-value = 0.3167
```

PRACTICANDO LAS REGRESIONES LINEALES



¿Qué puedes interpretar según los resultados del modelo?

Describe lo que ves en los resultados.

Los residuos del modelo están en torno a la media cero se nota que hay balance. y hay normalidad la estimación de los parámetros son números pequeños pero son significativos tanto en el intercepto como en la pendiente.

En conclusión conociendo el volumen total de los cilindros del motor se puede predecir el consumo de combustible porque hay una relación lineal.

REGRESIÓN LINEAL MÚLTIPLE

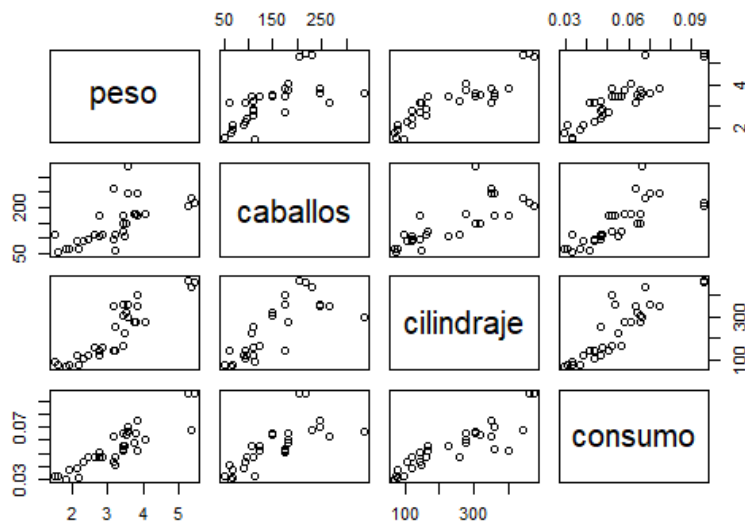
Calcula dos modelos lineales múltiples:

Modelo 1: Consumo = $a_1 \cdot \text{Peso} + a_2 \cdot \text{Caballos} + a_3 \cdot \text{Cilindrada} + b$

Modelo 2: Consumo = $a_1 \cdot \text{Peso} + a_2 \cdot \text{Caballos} + b$

Copia los resultados del modelo 1

PRACTICANDO LAS REGRESIONES LINEALES



Ecuación del Modelo 1

$$\text{Consumo} = 9.469\text{e-}03 \cdot \text{Peso} + 5.864\text{e-}05 \cdot \text{Caballos} + 2.456\text{e-}05 \cdot \text{Cilindrada} + 9.496\text{e-}03$$

```
> regresionLinealMultiple <- lm(formula = consumo ~ peso + caballos + cilindraje, varEstudio)
> summary(regresionLinealMultiple)
```

Call:

```
lm(formula = consumo ~ peso + caballos + cilindraje, data = varEstudio)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0163719	-0.0043511	0.0008672	0.0032544	0.0133345

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.496e-03	5.322e-03	1.784	0.08521 .
peso	9.469e-03	2.688e-03	3.522	0.00149 **
caballos	5.864e-05	2.883e-05	2.034	0.05155 .
cilindraje	2.456e-05	2.609e-05	0.941	0.35472

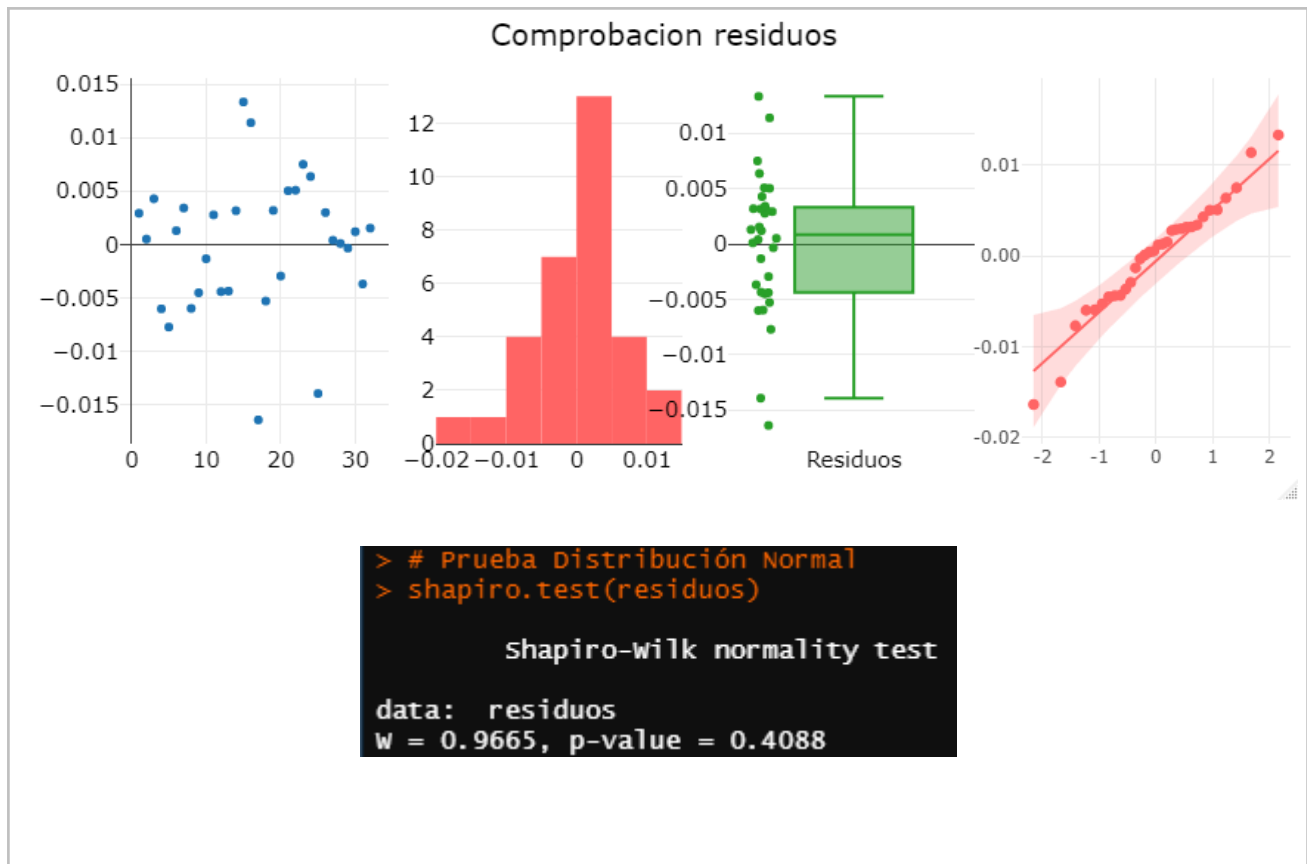
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.006653 on 28 degrees of freedom

Multiple R-squared: 0.8518, Adjusted R-squared: 0.8359

F-statistic: 53.63 on 3 and 28 DF, p-value: 9.94e-12

PRACTICANDO LAS REGRESIONES LINEALES



Copia los resultados del modelo 2

Ecuación del Modelo 2

$$\text{Consumo} = 1.149\text{e-}02 \cdot \text{Peso} + 7.479\text{e-}05 \cdot \text{Caballos} + 6.305\text{e-}03$$

```

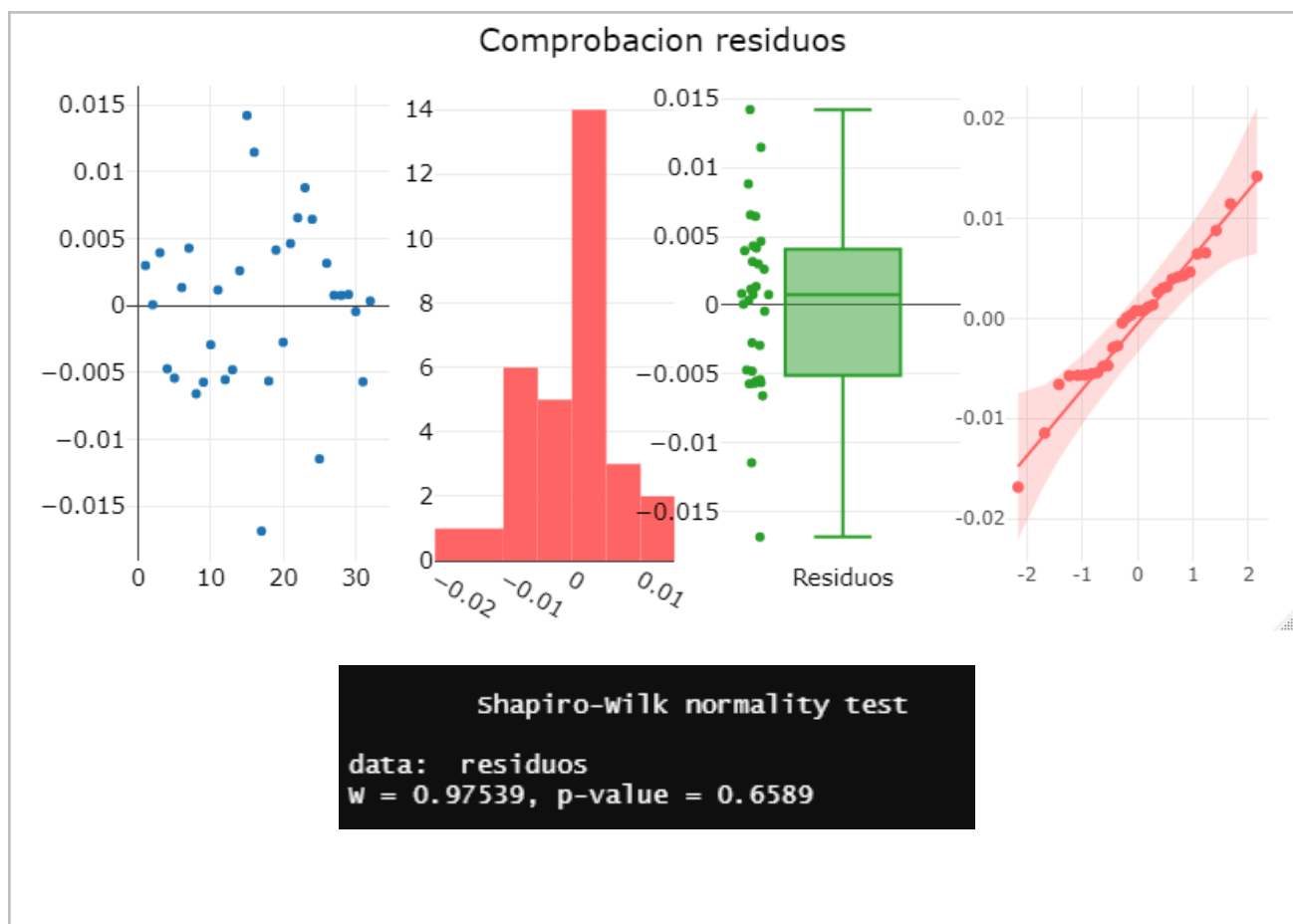
call:
lm(formula = consumo ~ peso + caballos, data = varEstudio)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0168690 -0.0049601  0.0007663  0.0040027  0.0142186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.305e-03  4.094e-03   1.540  0.13435
peso         1.149e-02  1.620e-03   7.089 8.45e-08 ***
caballos      7.479e-05  2.312e-05   3.235  0.00303 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.00664 on 29 degrees of freedom
Multiple R-squared:  0.8471,    Adjusted R-squared:  0.8365
F-statistic: 80.33 on 2 and 29 DF, p-value: 1.494e-12
  
```

PRACTICANDO LAS REGRESIONES LINEALES



Compara los dos modelos con el BIC o AIC y elige uno de ellos.

¿Qué puedes interpretar de este análisis?

```
> BIC(regresionLinealMultiple1)
[1] -216.9398
> BIC(regresionLinealMultiple2)
[1] -219.4091
```

Con el criterio BIC se elige el modelo 2

El p-valor para el modelo 2 es más significativo que en el modelo 1.

Eliminar la variable cilindraje del modelo hace que los residuos sean más equilibrados esto hace que el modelo sea más preciso que el modelo 1.

En conclusión más variables no implican una mejor precisión, más bien pueden distorsionar el modelo.

¡Tus primeras experiencias con modelos!

La siguiente es flexibilizar un poquito el modelo lineal ☺ y relajar las condiciones de uso.

Estás cerca de los modelos lineales generalizados