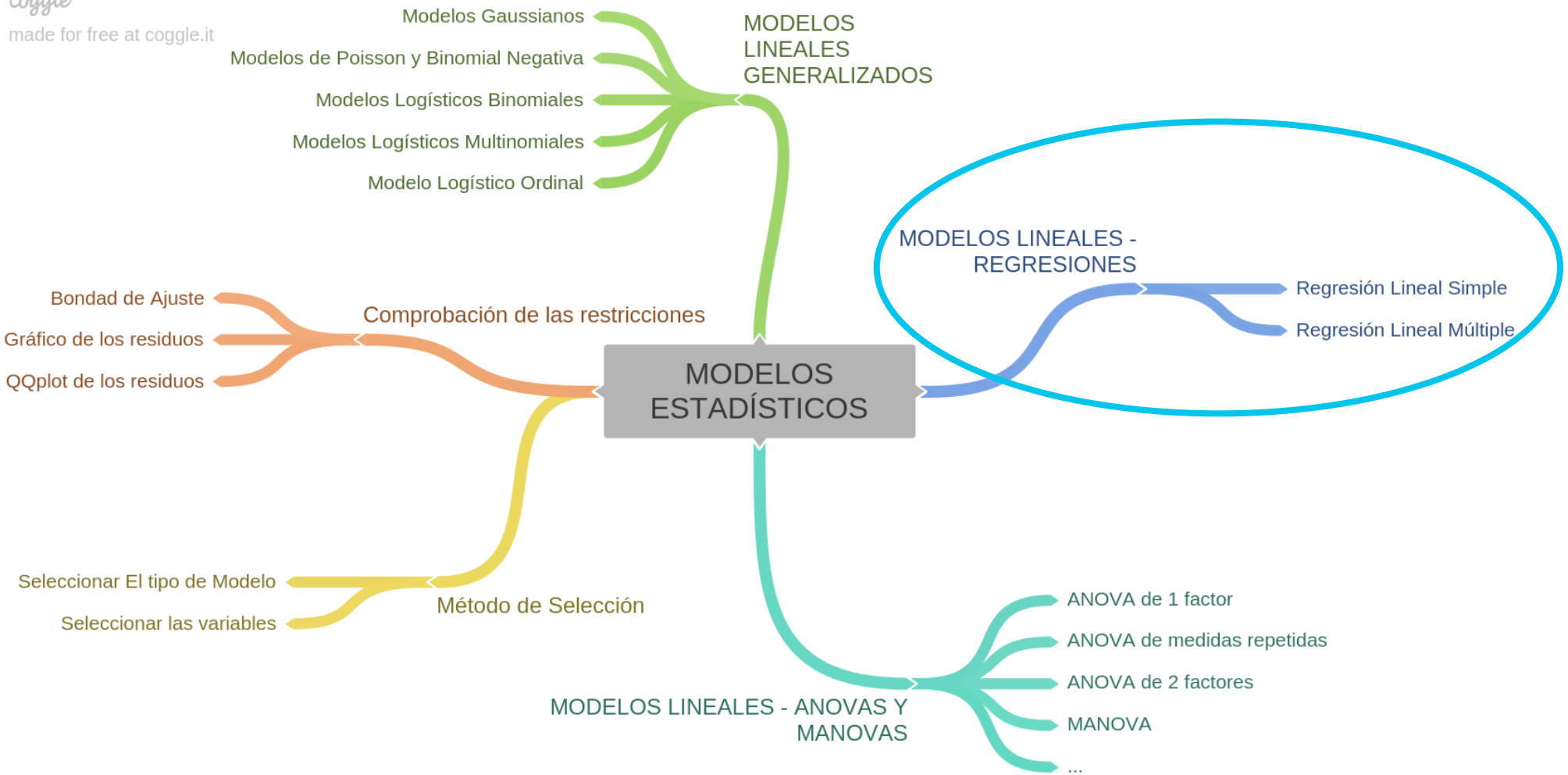
Regresión Lineal

El comienzo de los modelos lineales: las regresiones

¿Que vas a ver en este bloque?

- La regresión Lineal Simple
- Los errores o residuos
- La regresión Lineal Múltiple
- La comparación de modelos con el BIC



La regresión lineal simple

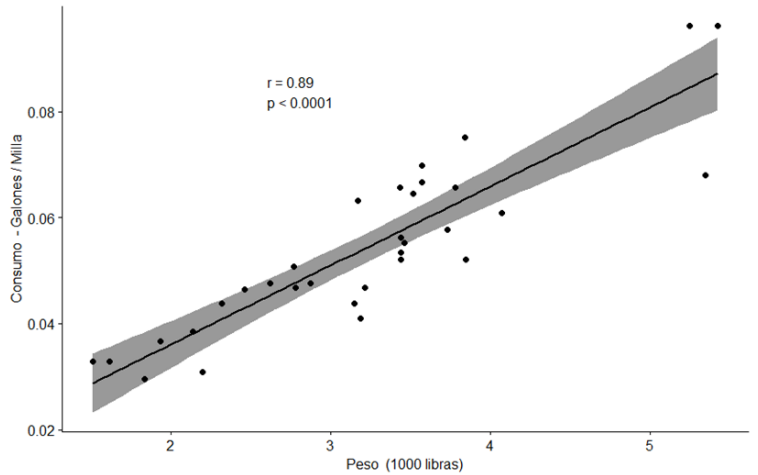
El principio de los modelos

La regresión lineal Simple

OBJETIVO:
Explicar los datos como una relación lineal entre 1 entrada y 1 salida de medidas continuas

Variable de Entrada Variable Independientes	VARIABLE CONTÍNUA Ej: peso
Variable de Salida Variable Dependiente	VARIABLE CONTÍNUA Ej: consumo
Modelo Función matemática	Y = aX+b Ej: Consumo = a·peso + b + error
Restricciones	Variables continuas La relación sea lineal Los residuos normales con media 0 Varianzas de los residuos constantes

Consumo = 0.014938 * Peso + 0.006169 + Error



Conceptosclaros

La regresión lineal Simple

OBJETIVO:
Explicar los datos como una relación lineal entre 1 entrada y 1 salida de medidas continuas

Variable de Entrada Variable Independientes	VARIABLE CONTÍNUA Ej: peso
Variable de Salida Variable Dependiente	VARIABLE CONTÍNUA Ej: consumo
Modelo Función matemática	Y = aX+b Ej: Consumo = a·peso + b + error
Restricciones	*(Variable normal de entrada = se traduce en la normalidad de los residuos)* Variables continuas La relación sea lineal Los residuos normales con media 0 Varianzas de los residuos constantes

Consumo = 0.014938 * Peso + 0.006169 + Error

```
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
```

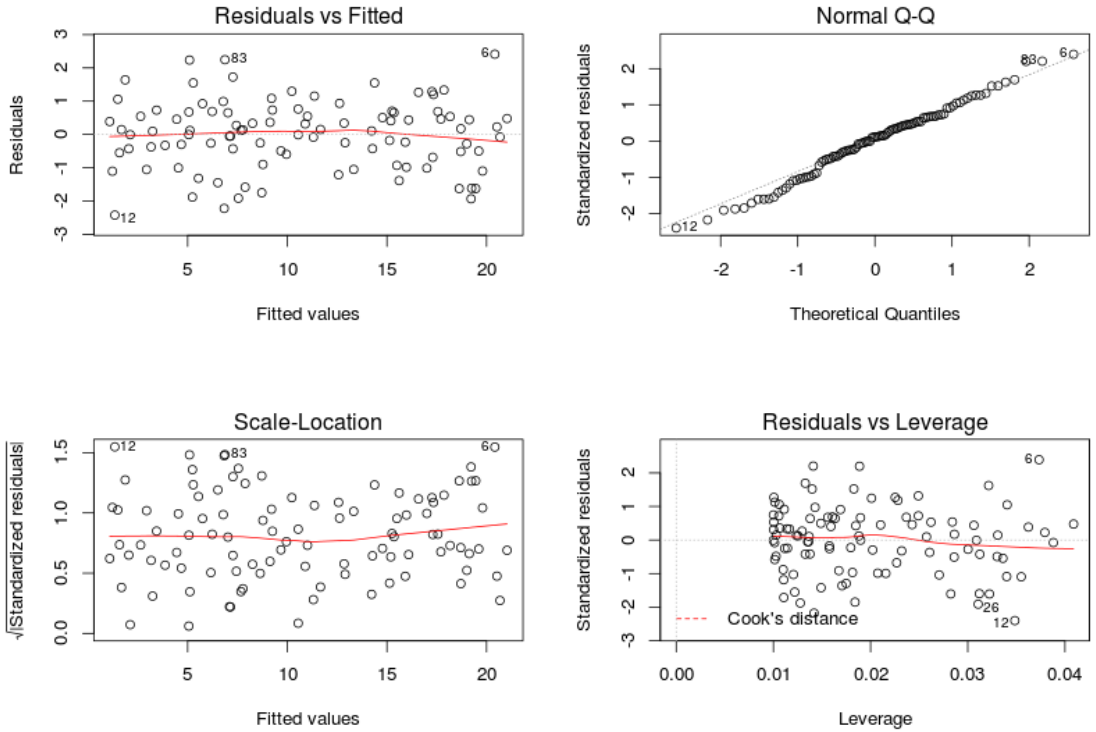
Los errores o residuos

¿Cómo podemos revisar si la regresión lineal es buena?

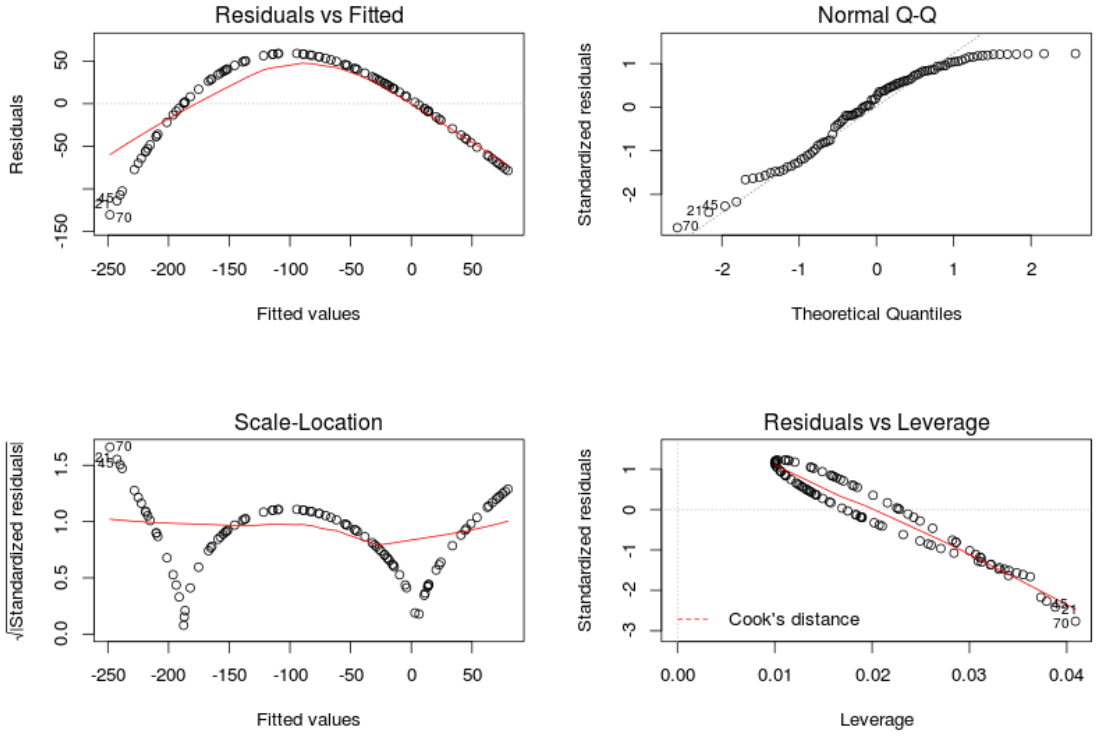
Los errores o residuos

Las tres restricciones de los residuos:
- Igualdad de dispersión en todo el rango
- Normales
- Media 0

- Residuos **ok**



- Residuos que **NO ok**



La regresión lineal múltiple

Aportando más información de variables de entrada

La regresión lineal Múltiple

OBJETIVO:
Explicar los datos como una relación lineal entre 1 entrada y más de una variable de salida de medidas continuas

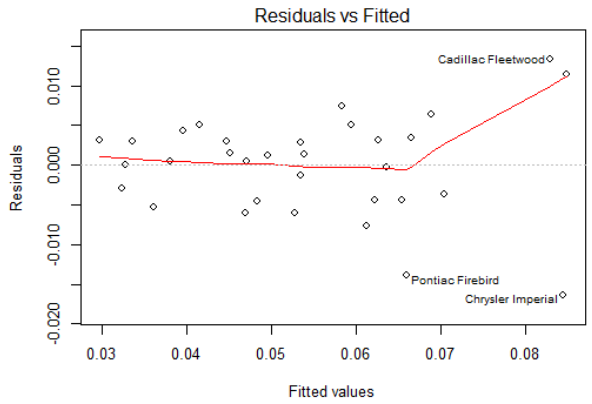
Variables de Entrada Variables Independientes	VARIABLE CONTÍNUA Ej: peso, caballos y cilindrada
Variable de Salida Variable Dependiente	VARIABLE CONTÍNUA Ej: consumo
Modelo Función matemática	Y = a1·X1 + a2·X2 + a3·X3 +b Ej: Consumo = a1·peso + a2·caballos + a3·Cilindrada + b + error
Restricciones	Variables continuas La relación sea lineal Los residuos normales con media 0 Varianzas de los residuos constantes No presentan colinealidad . VIF < 4

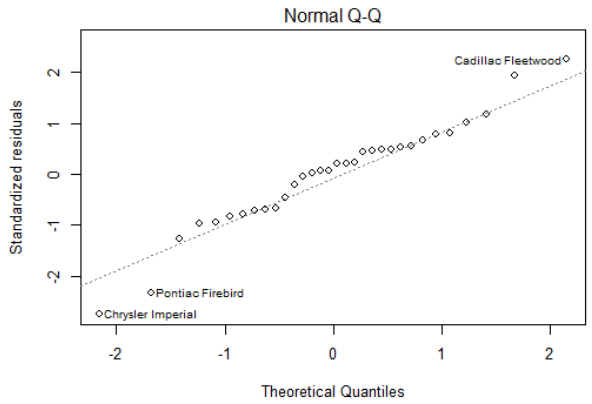
Consumo = 9.9496e-03 * Peso + 5.864e-05·Caballos +2.456e-05·Cilindrada + Error

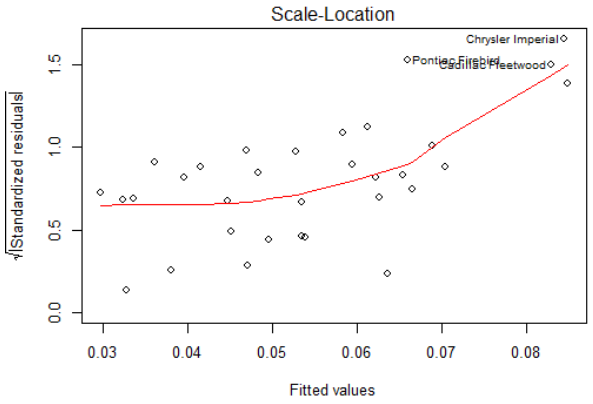
```
Call:
lm(formula = consumo ~ peso + Caballos + cilindrada, data = df)

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 .
cilindrada  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
```

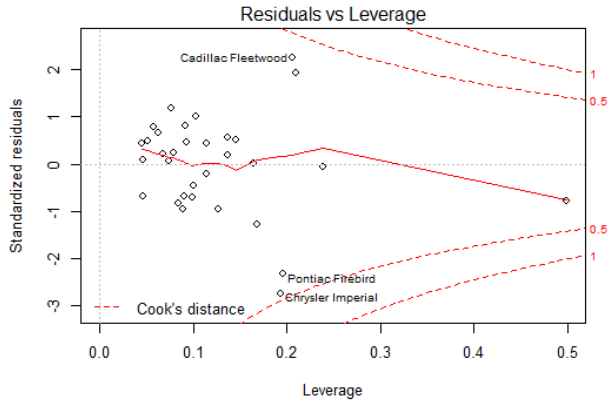
La regresión lineal Múltiple

OBJETIVO:
Explicar los datos como una relación lineal entre 1 entrada y más de una variable de salida de medidas continuas










```
vif(modelo) # parece que hay colinealidad --> quitaremos la cilindrada
    peso   Caballos cilindrada
4.844618   2.736633   7.324517
```

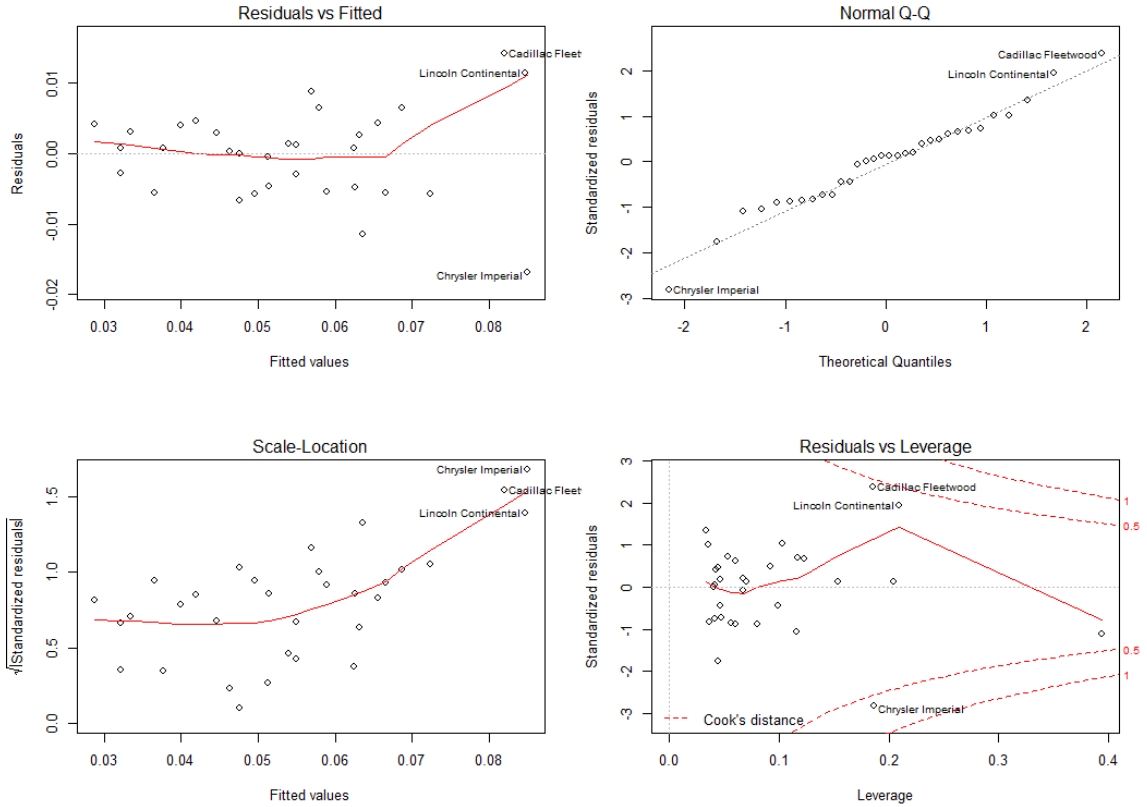
Peso → MODELO Y = F(X) → Consumo

Consumo = 9.9496e-03 * Peso + 5.864e-05·Caballos +2.456e-05·Cilindrada + Error

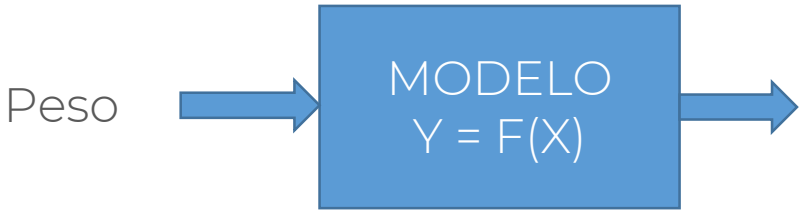
```
Call:
lm(formula = consumo ~ peso + Caballos + cilindrada, data = df)

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 .
cilindrada  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
```

La regresión lineal Múltiple

OBJETIVO:
Explicar los datos como una relación lineal entre 1 entrada y más de una variable de salida de medidas continuas




```
> vif(modelo2) # <4 vamos bien!!!
    peso Caballos
1.766625 1.766625
```



Consumo = 1.149e-02 * Peso + 7.479e-05·Caballos + 6.305e-03 + Error

```
Call:
lm(formula = consumo ~ peso + Caballos, data = df)

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
```

Comparando modelos con el BIC o AIC

Dos índices muy buenos para comparar modelos lineales

Cómo comparar modelos y decidor cuál de ellos es el mejor

- R^2 Ajustado
- BIC: bayesian index criteria
- AIC: Akaike index criteria

- Cuanto menor es el BIC mejor modelo es

Modelo 1

```
Call:
lm(formula = consumo ~ peso + caballos + cilindrada, data = df)
```

Modelo 2

```
Call:
lm(formula = consumo ~ peso + caballos, data = df)
```

```
> BIC(modelo)
[1] -216.9398
> BIC(modelo2)
[1] -219.4091
```

- Los modelos se pueden calcular de dos formas:
  - Regresiones = minimizando el error → mínimos cuadrados
    - El mayor R cuadrado Ajustado
  - GLM = máxima verosimilitud (en inglés: likelihood) → minimizando la log likelihood
    - El menor BIC

Take away

El resumen de la ½ lección

Lo más importante de la lección

- La regresión lineal es un modelo que minimiza el error dibujando una línea recta en los datos
- Las restricciones son:
  - Relación lineal entre la entrada y salida
  - Residuos: normales, varianza igual para todos los valores y media 0
  - No colinealidad en las entradas VIF < 4
- Una manera de utilizar la regresión lineal múltiple es quitar las variables con mayor p-valor hasta encontrar un modelo con todos los coeficientes significativos
- Para comparar dos o más modelos lo podemos hacer con el BIC (Bayesian Index Criteria) -> El modelo con menor BIC es el ganador

Tú turno

A por tus primeros test estadísticos

A poner en práctica lo que has visto

- Descarga la hoja de trabajo
- Trabaja con los datos que te propongo y entiende el proceso de la regresión lineal

- ¡Te ayudará a ponerte en marcha!