Práctica 11. El modelo de regresión lineal.

Estadística (650008). Grado en biología sanitaria. UAH.

Actualizado: 2023-12-17

Presentación del problema.

Trabajaremos con los datos de las indias Pima que han aparecido otras veces a lo largo del curso.

- `bmi`: índice de masa corporal en kg/m^2.
- `skin`: grosor del pliegue del triceps, en mm.

Se pretende establecer un modelo lineal entra ambas variables, con `Bmi` como variable explicativa y `Skin` como variable respuesta.

Carga de datos.

Teclea en un script y ejecuta las siguientes órdenes

```
library(MASS)
x = Pima.te$bmi
y = Pima.te$skin
```

Enunciado.

Ejercicio 1 Comprueba que la lectura ha sido correcta con la funcióin summary(): visualiza las primeras lineas de cada vector

```
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  19.40   28.18   32.90   33.24   37.20   67.10 
```

```
summary(y)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   7.00   22.00   29.00   29.16   36.00   63.00 
```

las variables son numéricas y no parece haber datos ausentes.

```
head(x)
```

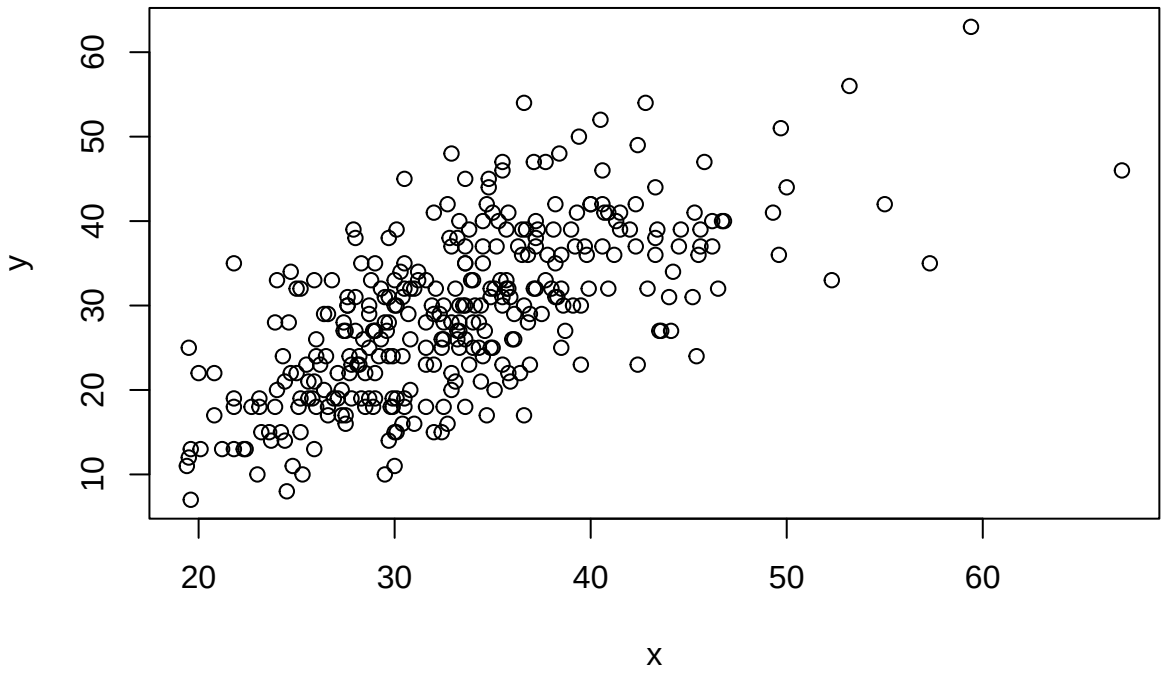
```
[1] 33.6 26.6 28.1 31.0 30.5 25.8
```

```
head(y)
```

```
[1] 35 29 23 32 45 19
```

Ejercicio 2 Visualiza la nube de puntos para determinar de forma visual si tiene sentido utilizar un modelo lineal.

```r
plot(x, y)
```



El modelo lineal parece adecuado. Recuerda poner primero la variable independiente y depués la dependiente.

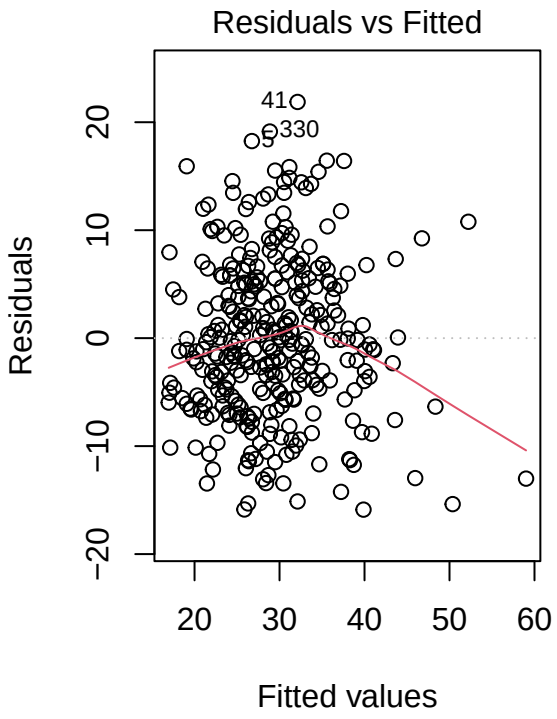
Ejercicio 3 Crea el modelo de regresión lineal y comprueba si se cumplen las hipótesis necesarias.

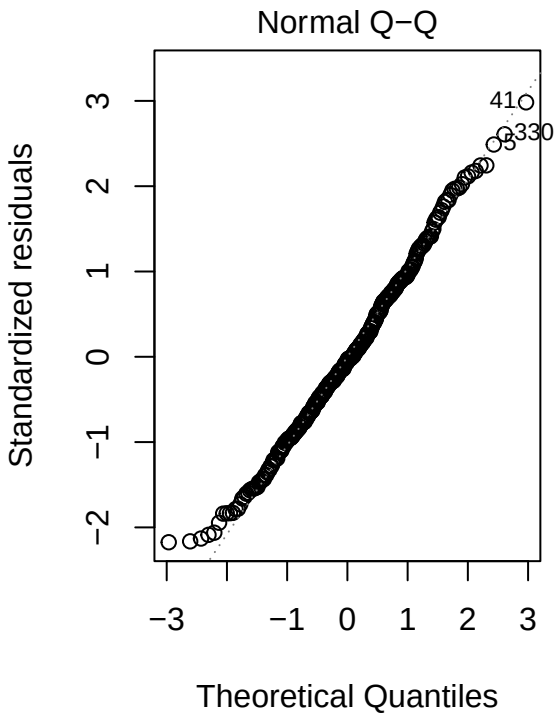
Recuerda que para crear el modelo lineal hay que poner primero la variable dependiente y después la independiente:

```r
lmXY = lm(y ~ x)
```

```r
par(mfrow = c(1, 2))

for(i in 1:2){
  plot(lmXY, which = i)
}
```






```
par(mfrow = c(1, 1))
```

Los gráficos sugieren normalidad y homocedasticidad en los residuos. Aún así, se aconseja hacer los correspondientes contrastes:

```
# contastes de normalidad
library(nortest)
lillie.test(lmXY$residuals)
```

```
    Lilliefors (Kolmogorov-Smirnov) normality test

data:  lmXY$residuals
D = 0.036715, p-value = 0.3394
```

El contraste de normalidad es satisfactorio y el de homocedasticidad

```
# ---- homocedasticityTests
# H0: homocedasticidad en poblaciones normales
library(gvlma)
library(MASS)
gvlma(lm(Pima.te$skin ~ Pima.te$bmi), alphalevel = 0.01)
```

```
Call:
lm(formula = Pima.te$skin ~ Pima.te$bmi)

Coefficients:
(Intercept)  Pima.te$bmi
```

```
    -0.1366       0.8815


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.01

Call:
 gvlma(x = lm(Pima.te$skin ~ Pima.te$bmi), alphalevel = 0.01)

                     Value  p-value                   Decision
Global Stat        12.8523 0.012021    Assumptions acceptable.
Skewness            3.1005 0.078268    Assumptions acceptable.
Kurtosis            0.9697 0.324765    Assumptions acceptable.
Link Function       8.2725 0.004025 Assumptions NOT satisfied!
Heteroscedasticity  0.5096 0.475318    Assumptions acceptable.
```

también se cumple. Recuerda que hay que fijarse en la última línea, y que en este caso el rótulo puede llevar a equívoco: H0 es "los residuos son homocedásticos" (aunque el rótulo hable de heterocedasticidad). El modelo de regresión lineal tiene sentido.

Además, las líneas `Skewness` y `Kurtosis` se refieren al grado de asimetría y apuntamiento de los residuos comparados con los de una normal, y sugieren no rechazar su normalidad.

Ejercicio 4 Calcula los coeficientes de la recta de regresión y sus intervalos de confianza al 95% de confianza Interprétalos

Los coeficientes de la recta son

```
lmXY$coefficients
```

```
(Intercept)           x
 -0.1365645   0.8814509
```

mientras que los intervalos de confianza al 95% son

```
########################################
# INFERENCIA SOBRE LA RECTA DE REGRESION

# intervalo confianza coeficientes recta
confint(lmXY, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) -3.8480022 3.5748732
x            0.7723741 0.9905277
```

- El intervalo de confianza para el término independiente, en general, no tiene interés, porque se halla fuera de la zona en la que hemos recogido valores (extrapolación).
- El intervalo de confianza para la pendiente indica que por cada unidad que aumenta el `bmi` se espera que el pliegue del tríceps aumente entre 0.77mm. y 1mm.

Ejercicio 5 Calcula, para una india con `bmi` = 18 y otra con `bmi` = 19.3

- **El valor del grosor del pliegue del tríceps predicho por el modelo.**
- **El intervalo de confianza para el valor medio del grosor del pliegue del tríceps predicho (intervalo de confianza) al nivel de confianza del 95%.**

- **El intervalo para el valor del grosor del pliegue del tríceps predicho (intervalo de predicción) al nivel de confianza del 95%.**

La predicción puntual es la que aparece bajo la etiqueta `fit`. Las etiquetas `lwr` y `upr` hacen referencia a lower y upper, los extremos inferior y superior del correspondiente intervalo:

```
x = Pima.te$bmi
y = Pima.te$skin
lmXY = lm(y ~ x)

 # intervalo de confianza valor predicho
predict(lmXY, newdata = data.frame(x = c(18, 19.3)),
        interval = "confidence", level = 0.95)
```

```
       fit      lwr     upr
1 15.72955 13.88770 17.5714
2 16.87544 15.16047 18.5904
```

Este es el intervalo de confianza para el valor medio predicho

```
# intervalo de prediccion valor predicho
predict(lmXY, newdata = data.frame(x = c(18, 19.3)),
        interval = "prediction", level = 0.95)
```

```
       fit      lwr      upr
1 15.72955 1.159900 30.29920
2 16.87544 2.321283 31.42959
```

Y este el intervalo de confianza para el valor predicho. Observa que en este caso el intervalo es mucho mayor, ya que los valores siempre están más dispersos que su promedio.