22/11

excel

1. Se desea estudiar, usando el dataset *airquality* que viene por defecto en R, si existe o no una dependencia lineal entre las variables *ozono* y *viento*.

   a) Plantear el modelo teórico que se quiere estimar (es decir viento como función del ozono).
   b) Plantear la prueba que realizaría para decidir si existe o no dicha dependencia lineal
   c) Hallar en R los valores de los parámetros para este caso, usando el comando lm.
   d) Graficar 10 pares (X_i, Y_i) siendo X el ozono e Y el viento y graficar la recta estimada obtenida con los parámetros obtenidos en la parte anterior.
   e) Realizar una prueba de hipótesis para decidir a nivel 5 % (usando el comando lm) si existe la dependencia planteada en el punto b).
   f) Plantear y realizar una prueba de hipótesis (a nivel 5 %) para decidir si el término independiente es o no nulo.
   g) Hallar el valor del coeficiente de determinación.
   h) Realizar los puntos anteriores pero cambiando ozono por viento.

a) $\underbrace{Y}_{\text{Viento}} = a \overbrace{X}^{\text{Ozono}} + b + \underbrace{\varepsilon}_{\text{error}}$

ozono: viento

Tenemos una muestra $(x_1, y_1), \ldots, (x_n, y_n)$

Recordar: No conocemos a y b pero los podemos estimar:

$$\hat{a} = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}$$

$$= \frac{\overbrace{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}^{\overline{(xy)}_n} - \overbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}^{\overline{x}_n} \cdot \overbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}^{\overline{y}_n}}{\underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i^2}_{\overline{(x_n^2)}} - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}_{(\overline{x}_n)^2}}$$

$$= \frac{\overline{(XY)}_n - \overline{X}_n \, \overline{Y}_n}{\overline{X^2_n} - \overline{X}_n^{\,2}} \quad \text{estima} \quad \frac{Cov(X,Y)}{Var(X)}$$

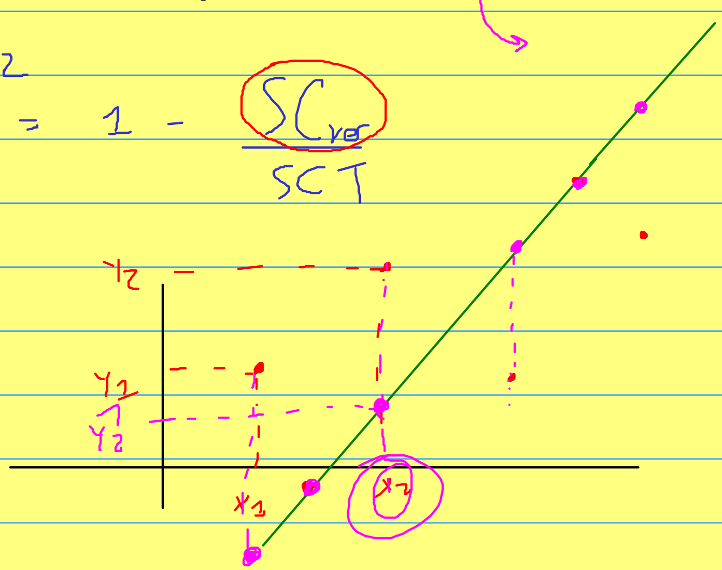
$$\hat{b} = \overline{Y}_n - \hat{a}\,\overline{X}_n$$

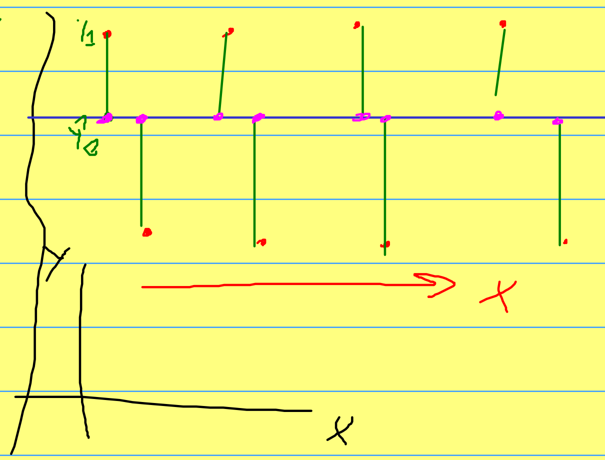
b) • $SCT = \sum_{j=1}^{n} \left(y_j - \overline{y}_n\right)^2$

• $SCR = \sum_{j=1}^{n} \left(\hat{y}_j - \overline{y}_n\right)^2$ → Regresión

• $SC_{res} = \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2$ → Residual

• $R^2 = 1 - \dfrac{SC_{res}}{SCT}$





Obs: $0 \leq R^2 \leq 1$

• Mientras más cerca de 1, mayor seguridad de que exista dependencia lineal.

$$\sum_{j=1}^{n} x_j^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

$$\left(\sum_{j=1}^{n} x_j\right)^2 = \left(x_1 + \cdots + x_n\right)^2$$

3. Sea $X \sim Unif([-2,-1] \cup [1,2])$. Definimos una variable aleatoria Y tal que $Y|X \sim N(3X-1, 1/4)$.

a) Simular una muestra iid de la variable aleatoria X de tamaño 1000.

b) Para cada dato de la muestra anterior x_i, simular una realización Y_i. Al final de este proceso, se obtiene una muestra de tamaño 1000 de la variable aleatoria Y.

c) Realizar un histograma de la muestra de Y. ¿Se puede decir que la variable aleatoria Y es normal?

d) Ajustar una regresión lineal univariada $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. ¿Es esperable que el modelo obtenido sea confiable?

e) Realizar un histograma de los errores de predicción del modelo (residuos). ¿Se puede decir que los errores son normales? Sugerencia: realizar un test de normalidad para poder decidir.