PAC6 Primavera 2021. Solució

UOC

Us pot ser útil consultar el següent material:

1. Mòdul Regressió lineal simple de les notes d'estudi
2. Activitats Resoltes del Repte 5 (Regressió lineal)

Utilitzarem dues bases de dades. Per a comparar-les, les escriurem en minúscules i les unirem. Adjuntem les dues taules de dades a fer servir en aquesta PAC.

Taula de dades -1

Les PACs es basaran en una taula de dades obtinguda a partir del repositori de micro-dades del "Banc Mundial" a https://microdata.worldbank.org/index.php/catalog/424/get-microdata

Conté indicacions, entre d'altres de

1. *City* = Nom de la ciutat
2. *Country* = País
3. *Population2000* = Població de la ciutat a l'any 2000.
4. *PM10Concentration1999 = PM10 concentrations (micro grams per cubic meter) in residential areas of cities larger than 100,000*, l'any 1999
5. *Region* = Classificació en regió geogràfica
6. *IncomeGroup* = Classificació segons nivell d'ingressos del país.

Per importar les dades podem utilitzar la següent instrucció. A part d'importar les taules de dades necessitem convertir el contingut dels camps a minúscules amb la funció tolower tal com es veu a continuació.

```
dadesPM10_0<-read.table("AirPollution2000WB_UOC2.csv", header=TRUE,
          sep=";",na.strings="NA",
          fileEncoding = "UTF-8", quote = "\"",
          colClasses=c(rep("character",4),rep("numeric",2),
                       rep("character",2)))
```

```
dadesPM10<-data.frame(apply(dadesPM10_0,2,tolower))#Dataset
#en minúscules
dadesPM10$PM10<-as.numeric(as.character(dadesPM10$PM10))
```

Taula de dades -2

En aquesta PAC continuarem usant la taula de dades habitual però estudiarem si existeix relació lineal entre la concentració de PM10 de 1999 i la concentració de PM10 de 2013 que tindrem en una nova base de dades obtinguda a partir de la web de la "Organització Mundial de la Salut" (*Ambient Air Quality Database*) https://whoairquality.shinyapps.io/AmbientAirQualityDatabase/ i que conté indicacions, entre d'altres de

1. City = Nom de la ciutat
2. Year = Any de la mesura
3. Country = País
4. Region = Àrea geogràfica del país.
5. pm10 = Concentració de partícules pm10.

A part d'importar les taules de dades necessitem convertir el contingut dels camps a minúscules amb la funció tolower tal com es veu a continuació.

```
dadesAir_0<-read.table("WHO_AirQuality_Database_2013_UOC.CSV", header=TRUE,
          sep=";",dec=".",na.strings="NA",skip=9,
          fileEncoding = "UTF-8", quote = "\"",fill=TRUE,
          )

dadesAir_0<-dadesAir_0[-6047,]#Ciutat repetida
dadesAir<-data.frame(apply(dadesAir_0,2,tolower)) #Dataset
#en minúscules
dadesAir$pm10<-as.numeric(as.character(dadesAir$pm10))
names(dadesAir)<-names(dadesAir_0)
```

Unim les taules de dades

Per poder comparar les dues taules de dades utilitzem la funció *merge* (amb by=c("City", "Country")) i considerem només les ciutats per a les que tenim valors de PM10 a les dues bases de dades. Seleccionem les dades tals que per a la mateixa ciutat les concentracions siguin les dues superiors a zero.

```
nova2<-merge(dadesPM10,subset(dadesAir,Year=="2013"),by=c("City", "Country"))
head(nova2,3)
```

```
##        City              Country Cod Citycode Population2000
## 1    aachen              germany deu  2760032         253040
## 2  aberdeen       united kingdom gbr  8260035         193694
## 3 abu dhabi united arab emirates are  7840001         904000
##   PM10Concentration1999                  Region.x IncomeGroup PM10 iso3
## 1                    18      europe & central asia high income   18  deu
## 2                    14      europe & central asia high income   14  gbr
## 3                    63 middle east & north africa high income   63  are
##        pm10 Year  Population date_compiled                   Region.y
## 1  18.04650 2013   259160.00          2018                europe (hic)
## 2  13.48377 2013   208361.00          2018                europe (hic)
## 3 143.75000 2013  1144933.00          2018 eastern mediterranean (hic)
##   region_abbr
## 1   eur (hic)
## 2   eur (hic)
## 3   emr (hic)
```

```
PM10_1999<-nova2$PM10[nova2$PM10>0 & nova2$pm10>0]
PM10_2013<-nova2$pm10[nova2$PM10>0 & nova2$pm10>0]
```

Per tant:

1. *nova2* conté la informació de les dues taules de dades per a les ciutats que estan als dos datasets.
2. *PM10_1999* conté la concentració de partícules a l'any 1999 per a les ciutats que estan als dos datasets.
3. *PM10_2013* conté la concentració de partícules a l'any 2013 per a les ciutats que estan als dos datasets.

PM10_1999 i *PM10_2013* són les variables que cal utilitzar per a resoldre aquesta PAC.

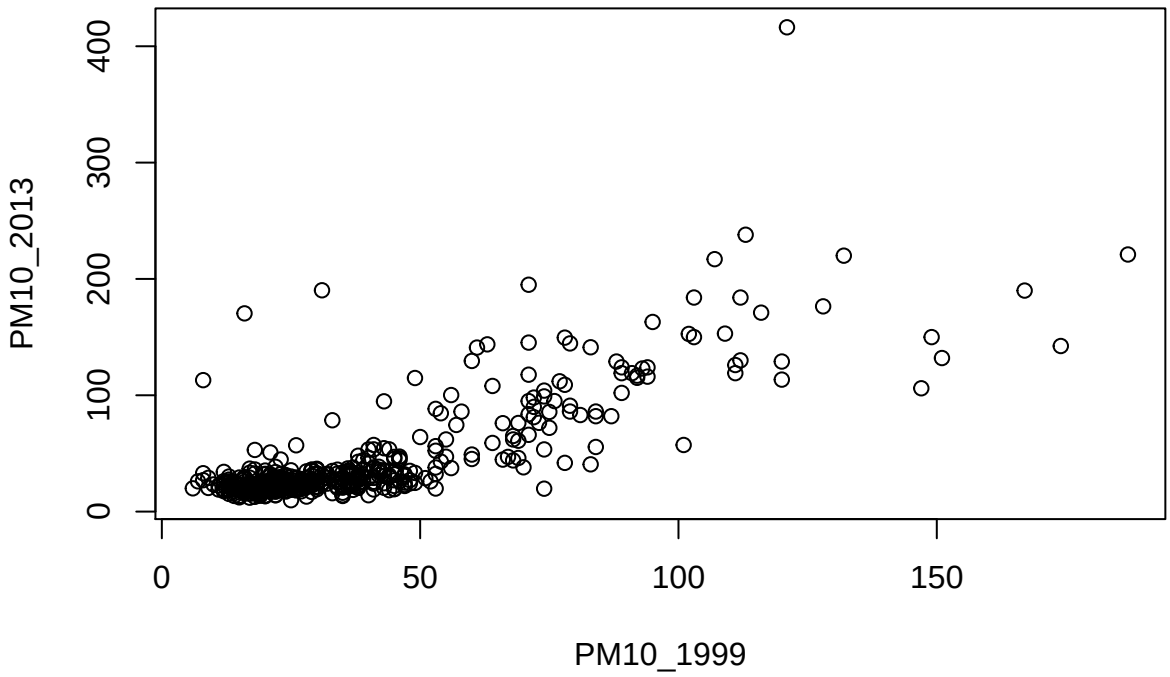
NOMBRE: Solució

PAC6

Pregunta-1 (15%)

Feu amb R el diagrama de dispersió del núvol de punts de la variable concentració de PM10 de l'any 2013 (en l'eix d'ordenades) en funció de la variable concentració de PM10 de l'any 1999 (en l'eix d'abscisses). Comenteu la gràfica obtinguda.

```
plot(PM10_1999,PM10_2013)
```



Observem que el núvol te forma allargada i per tant, cert comportament lineal. A mesura que augmenta la variable la concentració de PM10 de 1999 també augmenta la del 2013. Aquest comportament és més present a partir de concentracions de PM10 de 1999 superiors a $50\mu gr/m^3$.

Pregunta-2 (25%)

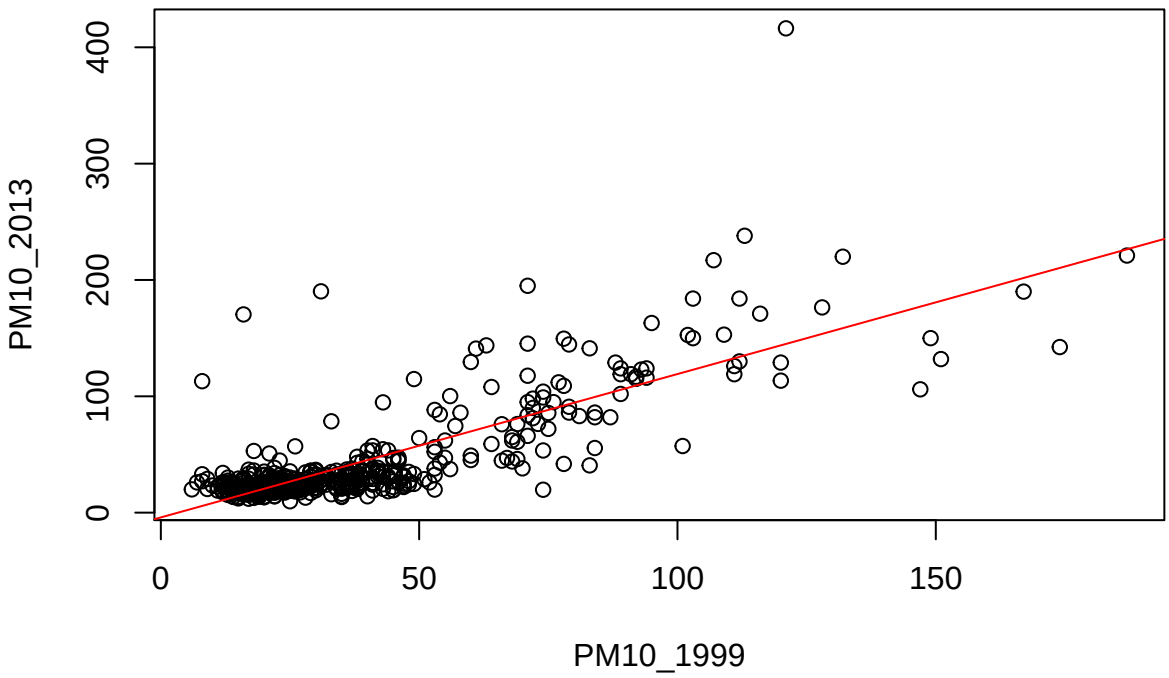
Calculeu amb R la recta de regressió de la variable concentració de PM10 de l'any 2013 en funció de la variable concentració de PM10 de l'any 1999. Doneu el pendent i l'ordenada a l'origen. Interpreteu el valor del pendent obtingut.

Podem fer de nou el diagrama de dispersió, però afegint ara la recta de regressió:

```
summary(lm(PM10_2013~PM10_1999))
```

```
##
## Call:
## lm(formula = PM10_2013 ~ PM10_1999)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.055 -11.270  -1.749   5.975 271.287
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.10764    2.12504  -1.933   0.0539 .
## PM10_1999    1.23240    0.04389  28.079   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.71 on 451 degrees of freedom
## Multiple R-squared:  0.6361, Adjusted R-squared:  0.6353
## F-statistic: 788.4 on 1 and 451 DF,  p-value: < 2.2e-16
```

```
plot(PM10_1999,PM10_2013)
abline(lm(PM10_2013~PM10_1999),col="red")
```



Així doncs, la recta de regressió és:

$$PM10_{2013} = -4.10764 + 1.23240 \cdot PM10_{1999}$$

El pendent ens diu que per cada $1\mu gr/m^3$ de PM10_1999 tindrem un augment de $1.23240\mu gr/m^3$ de PM10_2013. Aquest resultat no és gaire optimista ja que pel que sembla, la concentració de PM10 entre l'any 1999 i el 2013 ha anat, en mitjana, en augment tot i les diverses polítiques de protecció al medi ambient que han hagut durant aquest periode de temps en alguns països.

A partir de la sortida de R obtinguda en la pregunta (2), contesteu les següents preguntes:

Pregunta-3 (25%)

Quin és el valor del coeficient de determinació? I el valor del coeficient de correlació? Que podeu dir sobre la bondat de l'ajust?

El coeficient de determinació és $R^2 = 0.6361$ que ens diu que el model de regressió lineal explica un 63.6% de la variabilitat de la concentració de PM10_2013. El coeficient de

correlació lineal val $r = +\sqrt{0.6361} = 0.7975588$, per tant, tenim una bondat de l'ajust moderada. El signe del coeficient de correlació és positiu ja que el pendent de la recta és positiu.

Pregunta-4 (10%)

Estimeu el valor esperat de concentració de PM10 de l'any 2013 quan la concentració de PM10 de l'any 1999 és de $60\mu gr/m^3$.

$$PM10_{2013}(60) = -4.10764 + 1.23240 \cdot 60 = 69.83636\,\mu gr/m^3$$

Evidentment, s'ha d'interpretar con un valor promig.

Pregunta-5 (25%)

Volem fer un contrast d'hipòtesis amb un nivell de significació del 0.05 sobre el pendent de la recta de regressió obtinguda per saber si la variable X és explicativa. Indiqueu les hipòtesis nul · la i alternativa, el p-valor i la conclusió a la que arribeu.

Hipòtesi nul · la: $H_0 : \beta_1 = 0$, hipòtesi alternativa: $H_1 : \beta_1 \neq 0$

El p-valor és més petit que 2e-16 un valor molt petit (molt més petit que 0.05), per tant, rebutgem la hipòtesi nul · la i arribem a la conclusió que la variable PM10_1999 és una variable explicativa de la variable PM10_2013.