PAC5 (Pràctica)

Josep Andreu Miralles

12/12/2020

EXERCICI 1

a) Feu amb R el diagrama de dispersió del núvol de punts de la variable (X) valor del lloguer mitjà mensual dels contractes signats l'any 2015 als municipis/districtes i la variable (Y) valor del lloguer mitjà mensual dels contractes signats l'any 2018 als municipis/districtes. Què trobeu d'estrany en aquesta gràfica?

b) Per a no tenir en compte aquestes observacions anòmales trobades anteriorment, crearem dues noves variables M_2015_bis i M_2018_bis en les quals només conservarem els registres que tenen mitjana de lloguers superiors a zero tots dos anys. Feu de nou el diagrama de dispersió i a més, ara afegiu la recta de regressió. Calculeu amb R la recta de regressió de la variable M_2018_bis en funció de la variable M_2015_bis. Heu de donar el pendent i l'ordenada a l'origen. Interpreteu el valor del pendent obtingut.

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

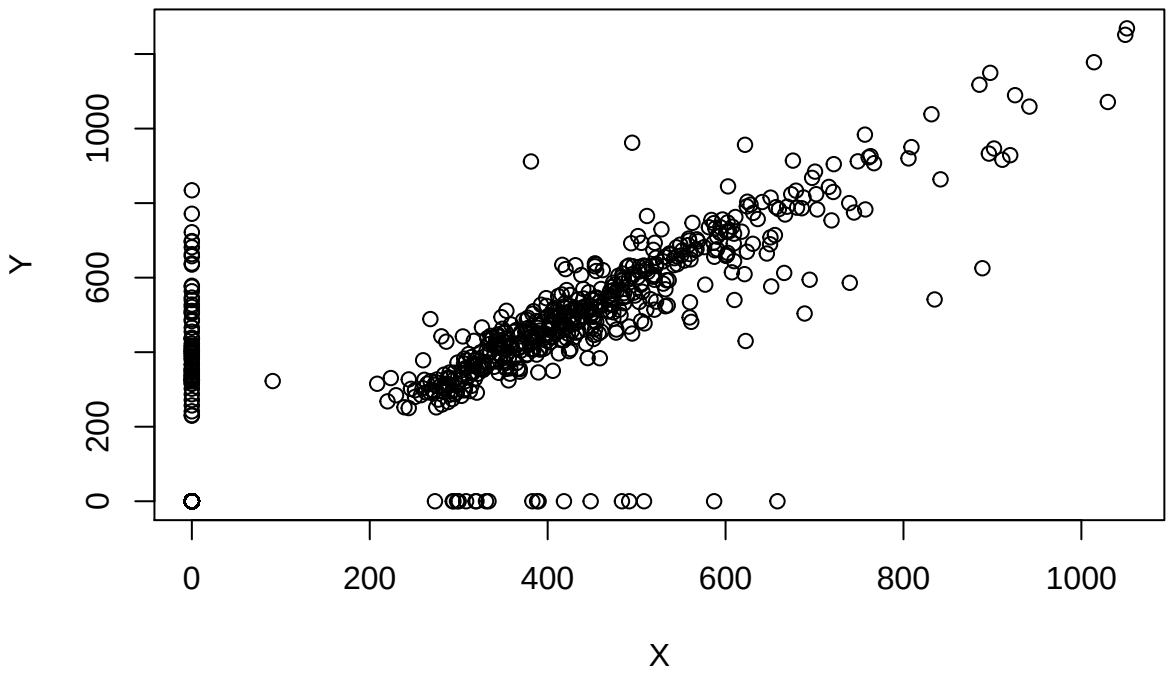
d) Estimeu el lloguer mitjà en el 2018 en un municipi que en el 2015 va tenir un lloguer mitjà de 500 euros/mes.

e) Volem fer un contrast sobre el pendent de la recta de regressió per saber si la variable M_2015_bis és explicativa. Indiqueu les hipòtesis nul · la i alternativa, el p-valor i la conclusió a que arribeu.

Solució

a) Feu amb R el diagrama de dispersió del núvol de punts de la variable (X) valor del lloguer mitjà mensual dels contractes signats l'any 2015 als municipis/districtes i la variable (Y) valor del lloguer mitjà mensual dels contractes signats l'any 2018 als municipis/districtes. Què trobeu d'estrany en aquesta gràfica?

```
library(readr)
dades<-read.csv("PIS_MUN2.csv", header=TRUE, sep=";",dec=",")
X <- dades$M2015
Y <- dades$M2018
plot (X,Y)
```

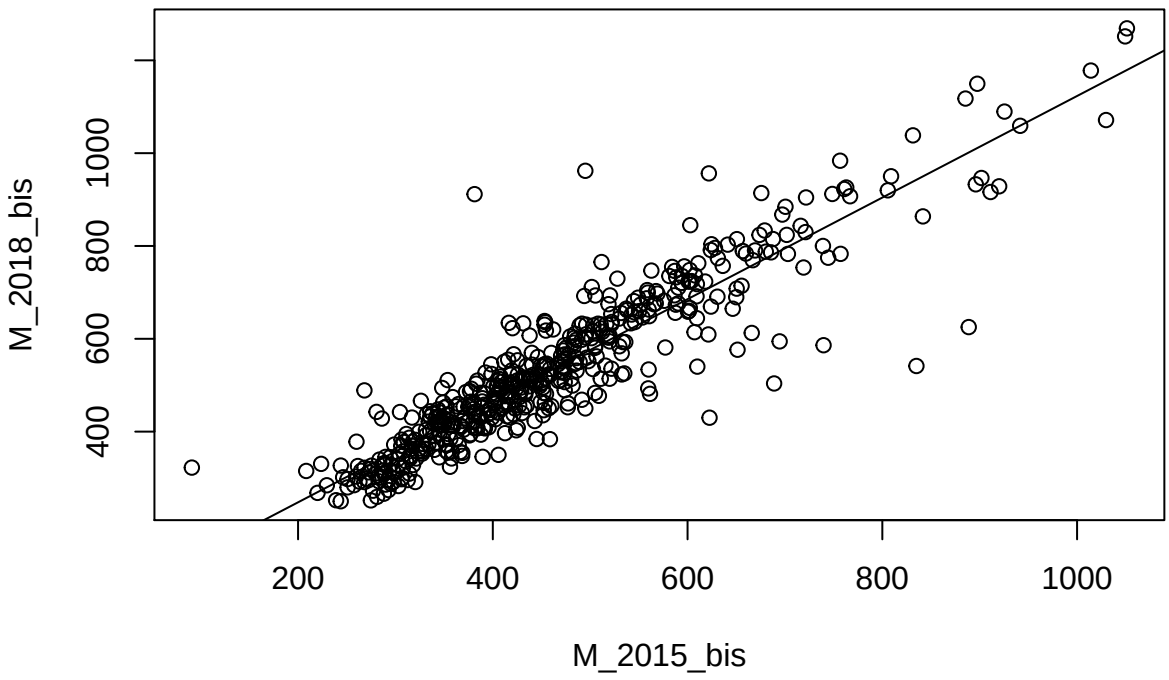


Hi ha una gran quantitat de valors atípics. Ja sigui en la columna M2015 o M2018 hi ha moltes dades que tenen un valor igual a 0.

b) Per a no tenir en compte aquestes observacions anòmales trobades anteriorment, crearem dues noves variables M_2015_bis i M_2018_bis en les quals només conservarem els registres que tenen mitjana de lloguers superiors a zero tots dos anys. Feu de nou el diagrama de dispersió i a més, ara afegiu la recta de regressió. Calculeu amb R la recta de regressió de la variable M_2018_bis en funció de la variable M_2015_bis. Heu de donar el pendent i l'ordenada a l'origen. Interpreteu el valor del pendent obtingut.

```
M_2015_bis<- dades$M2015[dades$M2015>0 & dades$M2018>0]
M_2018_bis<- dades$M2018[dades$M2018>0 & dades$M2015>0]
plot (M_2015_bis,M_2018_bis)

lm_M15_18<- lm(M_2018_bis~M_2015_bis)
abline(lm_M15_18)
```




```
summary(lm_M15_18)
```

```
##
## Call:
## lm(formula = M_2018_bis ~ M_2015_bis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -401.19  -33.44    1.04   36.19  465.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.84361   10.28818   2.804  0.00525 **
## M_2015_bis   1.09430    0.02132  51.333  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.93 on 502 degrees of freedom
## Multiple R-squared:   0.84,  Adjusted R-squared:  0.8397
## F-statistic:  2635 on 1 and 502 DF,  p-value: < 2.2e-16
```

La recta de regressió de la variable M_2018_bis en funció de la variable M_2015_bis és: Y = 28.84361 + 1.09430X.

Així doncs el seu valor a l'origen és 28.84361, i la pendent és igual a 1.09430. La pendent té un valor positu, de manera que com més gran és el valor mitjà del lloguer en l'any 2015 més gran és el valor miyjà del lloguer en l'any 2018.

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

```
r<- cov(M_2015_bis,M_2018_bis)/sqrt(var(M_2018_bis)*var(M_2015_bis))
R2<- r^2
R2
```

```
## [1] 0.8399774
```

```
r
```

```
## [1] 0.9165028
```

El valor del coeficient de determinació és 0.8399774, i el valor del coeicient de correlació és 0.9165028.

El que ens indica el coeficient de correlació és que l'ajust és molt bo. Mentre que el coeficient de determinació ens indica que el nostre model explica el 84% de la variabilitat de les observacions. Si consultem les dades obtingudes amb la funció summary veiem que el Multiple R-squared és 0.84, i com veiem coincideix amb el valor obtingut amb les covariancies.

d) Estimeu el lloguer mitjà en el 2018 en un municipi que en el 2015 va tenir un lloguer mitjà de 500 euros/mes.

```
Y <- 28.84361 + 1.09430*500
Y
```

```
## [1] 575.9936
```

Per un lloguer mitjà de 500 euros/mes en l'any 2015, podem esperar un lloguer mitjà en el 2018 de 575.99 euros/mes.

e) Volem fer un contrast sobre el pendent de la recta de regressió per saber si la variable M_2015_bis és explicativa. Indiqueu les hipòtesis nul · la i alternativa, el p-valor i la conclusió a que arribeu.

Hipòtesi nula: $H_0 : \beta_1 = 0$, és a dir la variable M_2015_bis no és explicativa. Hipòtesi alternativa: $H_1 : \beta_1$ no $= 0$, és a dir la variable M_2015_bis és explicativa. Nivell significatiu: $\alpha = 0.05$.

El p-valor obtingut amb la funció summary és 2.2e-16, i 2.2e-16 <= 0.05, pel que es rebutja la hipótesi nula i es conclou que la variable M_2015_bis és explicativa i per tant, també ho és el model.

EXERCICI 2

Considerem ara les variables M_2015_bis2, M_2016_bis2, M_2017_bis2 i M_2018_bis2 en els quals només conservem els registres que tenen totes les mitjanes de lloguers superiors a zero. a) Calculeu amb R el model de regressió múltiple per explicar la variable M_2018_bis2 en funció de totes les altres variables. Heu de donar explícitament el model. b) Quina variabilitat de la variable M_2018_bis2 queda explicada per aquest model. c) Indiqueu si algun dels coeficients del model de regressió múltiple no és significatiu.

Solució

a) Calculeu amb R el model de regressió múltiple per explicar la variable M_2018_bis2 en funció de totes les altres variables. Heu de donar explícitament el model.

```
M_2015_bis2<- dades$M2015[dades$M2015>0 & dades$M2016>0 & dades$M2017>0 & dades$M2018>0]
M_2016_bis2<- dades$M2016[dades$M2015>0 & dades$M2016>0 & dades$M2017>0 & dades$M2018>0]
M_2017_bis2<- dades$M2017[dades$M2015>0 & dades$M2016>0 & dades$M2017>0 & dades$M2018>0]
M_2018_bis2<- dades$M2018[dades$M2015>0 & dades$M2016>0 & dades$M2017>0 & dades$M2018>0]

model <- lm(M_2018_bis2~M_2015_bis2 + M_2016_bis2 + M_2017_bis2)
summary(model)
```

```
##
## Call:
## lm(formula = M_2018_bis2 ~ M_2015_bis2 + M_2016_bis2 + M_2017_bis2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -225.38  -23.27    0.56   22.96  334.58
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.18327    7.37148   0.432    0.666
## M_2015_bis2  0.30269    0.04379   6.912 1.55e-11 ***
## M_2016_bis2  0.19050    0.03681   5.175 3.37e-07 ***
## M_2017_bis2  0.59022    0.04465  13.219  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.2 on 472 degrees of freedom
## Multiple R-squared:  0.9262, Adjusted R-squared:  0.9257
## F-statistic:  1974 on 3 and 472 DF,  p-value: < 2.2e-16
```

La recta de regressió de la variable M_2018_bis2 en funció de les variables M_2015_bis2, M_2016_bis2 i M_2017_bis2 és: M_2018_bis2 = 3.18327 + 0.30269M_2015_bis2 + 0.19050M_2016_bis2 + 0.59022M_2017_bis2.

b) Quina variabilitat de la variable M_2018_bis2 queda explicada per aquest model.

El model amb les variables introduïdes como a predictors té un R2 alt (0.9262), és capaç d'explicar doncs el 92,62% de la variabilitat observada en la mitjana de lloguers de l'any 2018.

c) Indiqueu si algun dels coeficients del model de regressió múltiple no és significatiu.

Segons el model calculat tots els coeficients del model de regressió són significatius.

EXERCICI 3

Feu un ANOVA amb R per estudiar si hi ha diferències entre les mitjanes dels lloguers de l'any 2018, M_2018_bis2, de les 4 províncies (Barcelona, Lleida, Girona i Tarragona). A quina conclusió podem arribar?

Hipòtesi nula: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, és a dir les mitjanes són iguals. Hipòtesi alternativa: H_1 : les mitjanes no són iguals.

```
prov<- dades$PROV[dades$M2015>0 & dades$M2016>0 & dades$M2017>0 & dades$M2018>0]
ANOVA = aov(M_2018_bis2 ~ prov)
summary (ANOVA)
```

```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## prov          3  4233515 1411172   62.72 <2e-16 ***
## Residuals   472 10620471   22501
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Segons podem veure el valor de l'estadístic F és de 62.72, i el p-valor és aproximadament 0 (<2e-16). Així doncs, com que el p-valor és inferior a 0.05 podem assegurar que la mitjana dels preus de lloguer de l'any 2018 depèn de la província.