Labbrapport i Statistik

Laboration 3

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-09-19

Innehåll

1 Introduktion

I denna laboration kommer data över kommersiella fastigheteters pris (y) att behandlas och jämföras mot sina bakgrundsvariabler; ålder på fastigheten (x1), driftkostnader och skatter (x2), ledighetsgrader (x3), storlek på fastigheten(x4).

Målen med laborationen är att lära sig jämföra olika modeller med formella tester, använda sig av koncept där partiell beslut och korrelation förekommer. Ta till sig de viktiga koncepten som kommer med multikollinaritet and använda sig av metoder för att hitta dessa problem.

2 Databehandling

```
Commercial_properties <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH06PR18.t
cols1 <-c("y","x1","x2","x3","x4")
colnames(Commercial_properties) <-cols1
```

3 Uppgifter

3.1 7.7 Refer to Commercial properties Problem 6.18

3.1.1 a) Obtain the anova table that decompose the regression sum om squares into extra sums of squares.

```
##      SSR(X4) SSR(X1|X4) SSR(X2|X1, X4) SSR(X3|X1, X2, X4) SSE(X1,X2,X3,X4)
## [1,] 67.7751   42.27457       27.85749          0.4197463         98.23059
```

Hur fler eller mindre variabler påverkar SSR syns tydligt i tabellen ovan, i det första exemplet där det enbart är en variabel så är SSR relativt högt, det som syns är att ju fler variabler modellen innehåller, desto lägre blir SSR-värdet. Det som påverkar hur mycket lägre det blir, är hur pass mycket variabeln kan förklaravariationen i modellen. SSR är suman av alla residualer kvadrade och SSE är summan av skillnaden mellan de observerade värdena och de förutspådda.

3.1.2 b) Test weheter X3 can be dropped from the regression model. Use the F-test statistics and level of significance .01.

H0: X3=0

H1: X3 $\neq$ 0

$$F = \frac{MSR}{MSE} = \frac{RSS/k}{SSE/n-k-1} = \frac{0.42/1}{98.2306/76} = 0.3249$$

$$F(.99, 1, 76) = 6.98$$

I fall av att F $>$ 6.98 förkastas H_0, om inte antas H_0.

Då F inte överstiger det kritiska värdet ($0.3249 < 6.98$), samt att p-värdet summerar till 0.57045 som inte är signifikant på 1 % signifikansnivå kan vi inte förkasta H_0 om att x3 = 0. Alltså kan x3 droppas från modellen.

3.2 7.8 Test whether both X2 and X3 Can be dropped from the regression model given that X1 and X 4 are retained; use a = .01. State alternatives, decision rule, and conclusion. What is the P-value of the test?

H0: B2 =B3=0

H1: B2=B3≠ 0

```
Full_model_ <-lm(y~x1+x2+x3+x4,data = Commercial_properties)
reduced_model <-lm(y~x1+ x4, data = Commercial_properties)
anova(reduced_model,Full_model_)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x4
## Model 2: y ~ x1 + x2 + x3 + x4
##   Res.Df     RSS Df Sum of Sq      F     Pr(>F)
## 1     78 126.508
## 2     76  98.231  2    28.277 10.939 0.00006682 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{SSR}{SSE} = \frac{28.277/2}{98.2306/76} = 10.9388$$

$$F(.99, 2, 76) = 4.9$$

I fall av att F > 4.9 förkastas H_0, om inte antas H_0.

Då F överstiger det kritiska värdet ($10.9388 > 6.98$), samt att p-värdet summerar till 0.000067 som är signifikant på 1 % signifikansnivå kan vi på 1 % signifikansnivå förkasta H_0 om att x3 $= 0$.

3.3 7.10 Test whether B1 = -. 1 and B2 = .4, a=.01. State the alternatives, full and reduced models, decision rule, and conclusion.

H0: B1=-.1, B2 =.4

H1: B1≠-.1, B2≠.4

```
Full_model_0 <-(lm(y~x2+x3+x1+x4,data = Commercial_properties))
reduced_model_0 <-(lm(y +.1*x1 - .4*x2 ~ x3 + x4, data = Commercial_properties))
Anova(Full_model_0)
```

```
## Anova Table (Type II tests)
##
## Response: y
##           Sum Sq Df F value         Pr(>F)
## x2        25.759  1 19.9294 0.000027473960 ***
## x3         0.420  1  0.3248         0.5704
## x1        57.243  1 44.2881 0.000000003894 ***
## x4        42.325  1 32.7464 0.000000197599 ***
## Residuals 98.231 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(reduced_model_0)
```

```
## Anova Table (Type II tests)
##
## Response: y + 0.1 * x1 - 0.4 * x2
##            Sum Sq Df F value      Pr(>F)
## x3          6.600  1  4.6738     0.03369 *
## x4         31.872  1 22.5713 0.000009058 ***
## Residuals 110.141 78
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \left(\frac{\text{SSE}(R) - \text{SSE}\,(F')}{df_R - df_F}\right) \div \left(\frac{\text{SSE}(F)}{df_F}\right) = \frac{110.141 - 98.2306}{78 - 76} \div \frac{98.2306}{76} = 4.607$$

$$F(.99, 2, 76) = 4.9$$

I fall av att F $>$ 4.9 förkastas H_0, om inte, antas H_0.

Då F understiger det kritiska värdet (4.607 $<$ 4.9), kan vi på 1 % signifikansnivå inte förkasta H_0 om att B1=-.1, B2 =.4.

3.4 7.27 Refer to Commercial properties Problem 6.18

3.4.1 a) Fit first-order linear regression model (6.1) for relating rental rates(Y) to property age(X1) and size (X4). State the fitted regression function.

```
options(scipen = 3)
model <- lm(y~x1+x4,data=Commercial_properties)#Räknar hyrespriser med fastighetsålder och storlek
```

$$\hat{y} = 14.361 - 0.114x1 + 0.00001x4$$

I regressionsekvation symboliserar 14.361 startvärdet för hyrespriset för fastigheten, x1 symboliserar ålder på fastigheten och åldern på fastigheten påverkar priset negativt. X4 kännetecknar kvadratmeter på fastigheten, vilket förklarar den lilla talet som multipliceras med i snitt 160 000 enheter.

3.4.2 b) Compare the estimated regression coefficients for property age and size with the corresponding obtained in problem 6.18c. What do you find?

$$\hat{y} = 14.361 - 0.114x1 + 0.00001x4$$

$$\hat{y} = 12.20 - 0.142x1 + 0.282x2 + 0.6193x3 + 0.000007x4$$

I en jämförelse mellan de två modellerna, där den översta enbart har storlek och ålder på fastigheten medan den nedre också tar variabler som driftkostnader,skatt och vakansgrad i beaktan, syns ett betydligt lägre startvärde i den nedre modellen, samtidigt som att ålder har en större negativ påverkan på priset och storleken har en mindre betydelse.

3.4.3 c) Does SSR(X4) equal SSR(X4|X3) here? Does SSR(X1) equal SSR(X1|X3)

```
##       SSR(X4) SSR(X4|X3)
## [1,] 67.7751   66.85829
```

```
##       SSR(X1) SSR(X1|X3)
## [1,] 14.81852    13.7743
```

Värdena är ej detsamma och anledningen till det är, att i samband med att ännu en variabel adderas till modellen förklaras variationen inom datat utav två istället för en variabel. I dessa fall hade SSR sjunkit ännu mer om X3 innehållit data som betyder mer för modellen.

3.4.4 d) Refer to the correlation matrix obtained in problem 6.18b. What bearing does this have on your findings in parts (b) and (C)?

```
options(digits = 2)
cor(Commercial_properties)#Korrelationsmatris för samtliga variabler
```

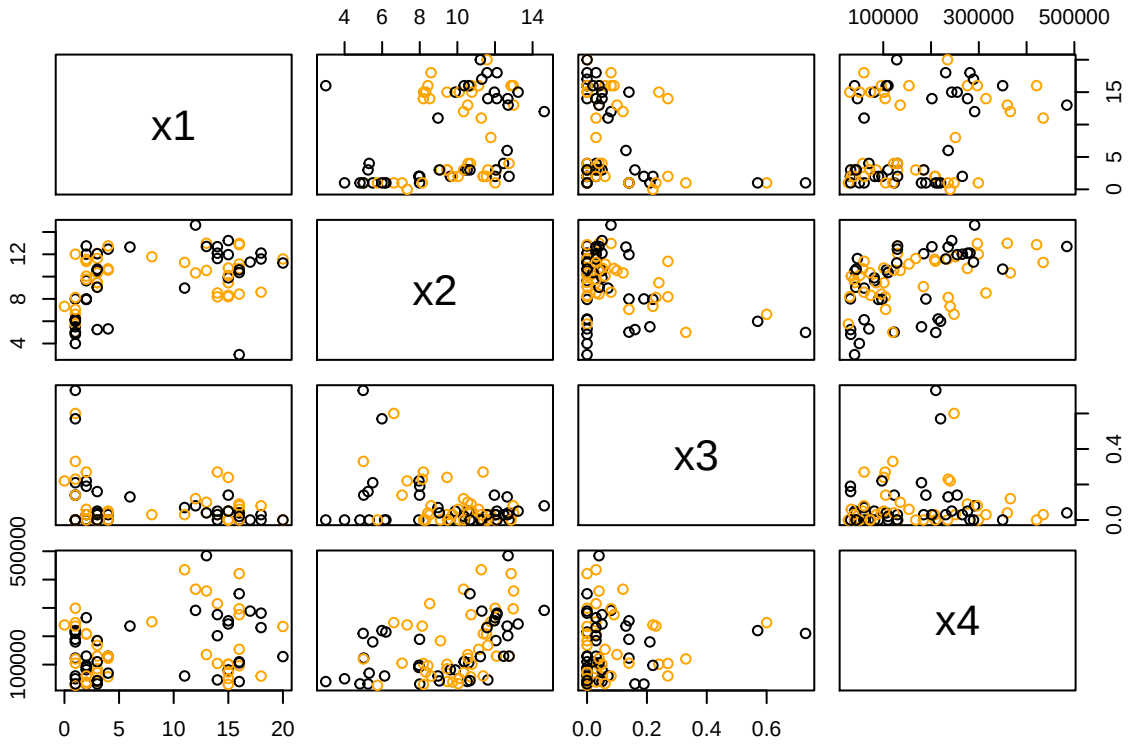
```
##         y    x1    x2     x3    x4
## y   1.000 -0.25  0.41  0.067 0.535
## x1 -0.250  1.00  0.39 -0.253 0.289
## x2  0.414  0.39  1.00 -0.380 0.441
## x3  0.067 -0.25 -0.38  1.000 0.081
## x4  0.535  0.29  0.44  0.081 1.000
```

I matrisen ovan kan syns korrelationerna mellan variablerna i modellerna. Det mest utmärkande i matrisen är variabel x3 låga korrelation med responsvariabeln som nästan summerar till 0. Bara utifrån detta kan man nästan fråga sig själv om x3 ska vara med i modellen.

3.5 10.18 Refer to Commercial properties Problem 6.18b

3.5.1 a) What do the scatter plot matrix and the correlation matrix show about pairwise linear associations among the predictor variables?

```
pairs(Commercial_properties[,2:5],col=c("black","orange"))
```



Ovan syns ett punktdiagram över de parvisa jämförelserna mellan de förklarande variablerna, det som syns är återigen att oberoende vilken variabeln den jämförs med ser inte spridningen för x3 bra ut alls, en stark koncentration i form av en sidoklump.

För resterande variabler går korrelationen aldrig under 0.28 vilket tydligt syns i plottarna, någorlunda linjära samband syns, däremot är de inte helt linjära och den enda korrelationen som faktiskt ser riktigt bra ut sett till datat är mellan x2 och x4.

3.5.2 b) Obtain the four variance inflation factors. Do they indicate that a serious multicollinearity problem exsts here?

```
model1 <- lm(y~x1+x2+x3+x4,data=Commercial_properties)
vif(model1)#Får ut VIF värdena
```

```
##  x1  x2  x3  x4
## 1.2 1.6 1.3 1.4
```

VIF värdena på de 4 förklarande variablerna är alla under 5 och är nära 1 vilket man kan anta att det ej är problem med multikollinearitet.

4 Lärdomar

I denna laboration har vi bekantat oss med hur man kan använda F-test i större utsträckning, vilken funktion SSR kan ha och hur mycket värdet kan variera beroende på hur många av de ursprungliga förklaringsvariablerna man har med i modellen.