Labbrapport i Statistik

Laboration 2

732G46

Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-09-09

Innehåll

1 Introduktion

I denna laboration kommer det första datasetet bestående av 10 transporter där antalet gånger en kartong blivit överförd från ett flygplan till ett annat över rutten (X) och summan av antalet ampuler som gått sönder under transporten (Y).

I det andra kommer data över kommersiella fastigheteters hyreskostnad (Y) och olika bakgrundsvariabler att behandlas.

Målen med laborationen är att skriva en Multipel linjär regressionsmodell i matrisform och genomföra matrisoperationer. Genomföra statistisk inferens , räkna och tolka olika konfidensintervall for förväntat och förklarande variabler.

2 Databehandling

```
upp_121 <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH01PR21.txt") #Laddar i
cols <-c("Ampuler","Transfers") #Namnger kolumnerna
colnames(upp_121) <-cols

Commercial_properties <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH06PR18.t
cols1 <-c("Y","x1","x2","x3","x4")
colnames(Commercial_properties) <-cols1
options(digits = 3)
```

3 Uppgifter

3.1 Uppgift 1

3.1.1 c) Compute and print the matrix $(X^T X)^{-}1$

```
c <-solve(TMatrisX%*%MatrisX) #Tar fram inversen
```

$$(X^T X)^{-}1 = \begin{bmatrix} 0.2 & -0.1 \\ -0.1 & 0.1 \end{bmatrix}$$

Ovan visas inversen till transponatet till matris X multiplicerat med matrisen X. Detta är en del av kovariansmatrisen.

3.1.2 d) Estimate the intercept and slope parameters of the regression model by computing $b = \left(X^T X\right)^{-1} X^T Y$

```
B <-solve(TMatrisX%*%MatrisX)%*%TMatrisX%*%Y
rownames(B) <-c("Intercept","X")
B
```

```
##           [,1]
## Intercept 10.2
## X          4.0
```

I detta fall där minsta-kvadratmetoden använts returneras en modell där interceptet summerar till 10, alltså kommer i snitt 10 ampuler gå sönder utan en enda transfer. Att X summerar till 4 innebär att för varje transfer som sker kommer antalet ampuler som går sönder att öka med 4.

3.1.3 e) Estimate the variance of the error terms by computing $MSE = \frac{e^T e}{(n-2)}$, where e is the 10 x 1 matrix (vector) of residuals and n = 10.

```
e <-matrix(ncol = 1,nrow = 10)
E <- Y-MatrisX%*%B
n <-10
MSE <-(t(E) %*% E)/(n-2)
MSE
```

```
     [,1]
```

[1,] 2.2

Ovan visas MSE utträknat med matrisekvationer, MSE symboliserar medelfelet kvadrerat.

3.1.4 f) Finally, estimate the covariance matrix of b by computing $MSE \cdot \left(X^T X\right)^{-1}$ Try to explain why the estimates of the intercept and the slope are correlated.

```
cov_b <-MSE[1,1]*(solve(TMatrisX%*%MatrisX))
cov_b
```

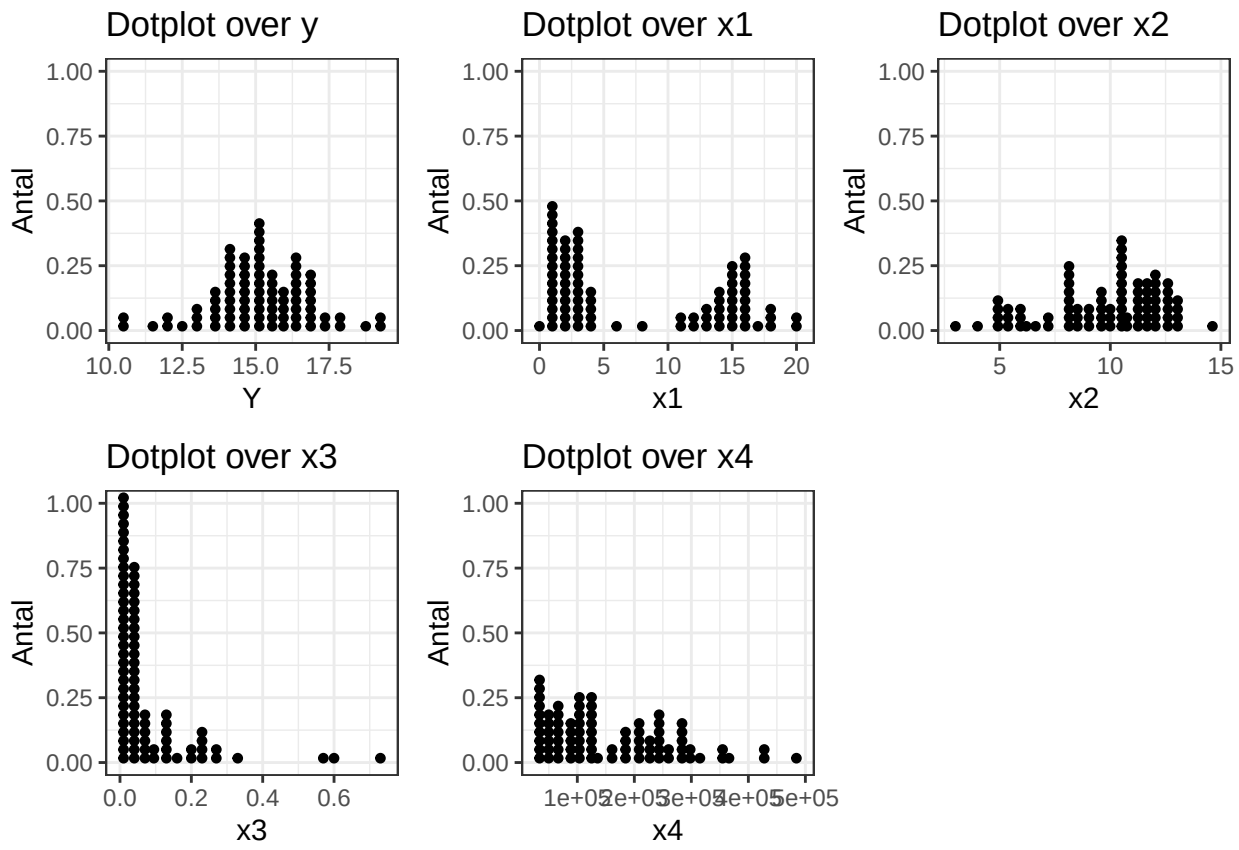
```
##                 Transfers
##            0.44     -0.22
## Transfers -0.22      0.22
```

Det som estimatorerna förklarar är att i samband med fler transfers går fler ampuler sönder.

3.2 Uppgift 2

3.2.1 6.18

3.2.2 a) Prepare a dot plot for each predictor variable, what information do these plots provide?

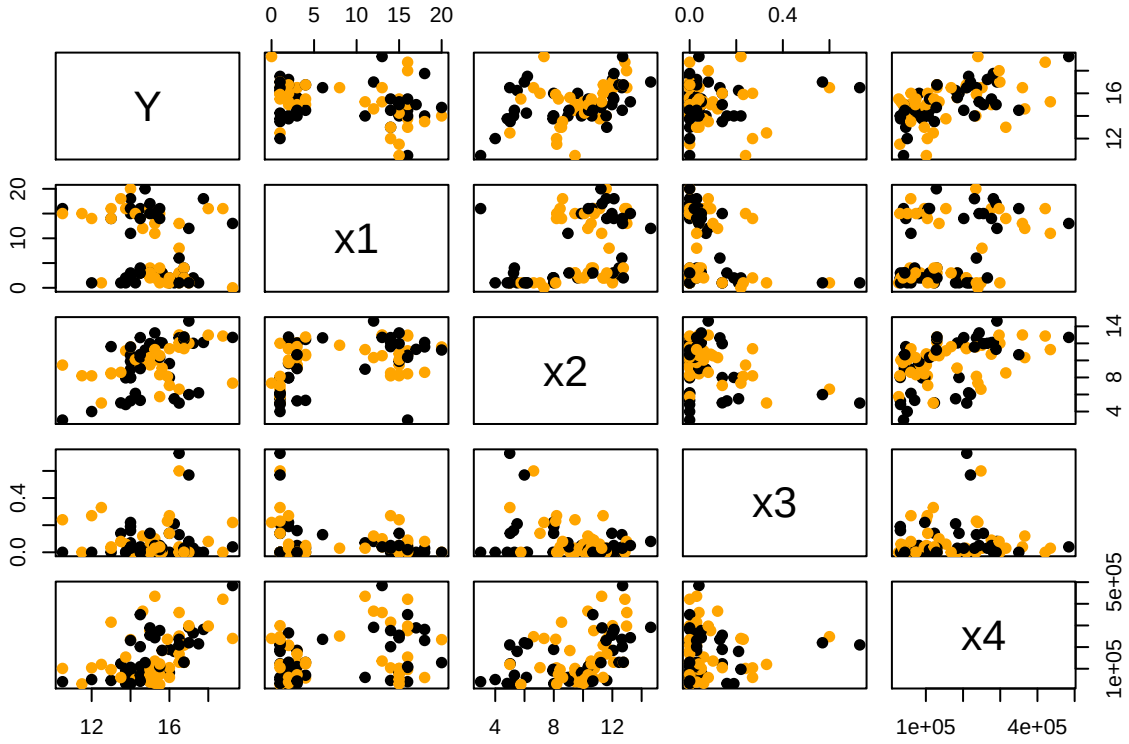


Den information som går att utläsa från graferna är spridning/fördelning hos respons och bakgrundsvariablerna. Responsvariabeln i vårat fall (Y) är hyrekostnaden för de kommersiella fastigheterna. X1 är ålder på fastigheten, X2 driftkostnader och skatter, X3 vakansgraden, och X4 storleken mätt i kvadrafot.

Responsvariabeln Y som är fördelning kan bedömas vara normalfördelad sett till den visuella spridningen och den stora urvalsstorlek som är större än 30. För bakgrundsvariablerna (x1:4) sker spridningen väldigt olika ut, där den variabeln som har bäst spridning är x2 och den som har sämst x3.

3.2.3 b) Obtain a scatter plot matrix and the correlation matrix. Interpret these and state your principal findings

```
paren <-pairs(Commercial_properties[,1:5], pch = 19,col=c("black","orange"))
```



Ovan syns ett punktdiagram över de parvisa jämförelserna mellan de förklarande variablerna och responsvariabeln, det som syns är 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, 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, samt x4 och Y.

```
cor(Commercial_properties)
```

```
##          Y     x1     x2      x3     x4
## Y   1.0000 -0.250  0.414  0.0665 0.5353
## x1 -0.2503  1.000  0.389 -0.2527 0.2886
## x2  0.4138  0.389  1.000 -0.3798 0.4407
## x3  0.0665 -0.253 -0.380  1.0000 0.0806
## x4  0.5353  0.289  0.441  0.0806 1.0000
```

I denna korrelationsmatris uppvisas sambandet mellan samtliga variabler varandra emellan, både bakgrundsvariabler och responsvariabler, anmärkningsvärt är även att en linje som visar formen på spridningen för variablerna finns.

Högst korrelation är det mellan x4 som är square footage på fastigheten, där korrelationen uppgår till 0.54 som är moderat stark men inte övertygande.

Den variabelkombination som har svagast korrelation är x3 och y där det knappt exsterar någon form av korrelation vilket är en bra förklaring till att eventuellt testa en modell utan den variabeln.

3.2.4 c) Fit regresion model (6.5) for four predictor~variabler to the data. State the estimated regression function.

```
linjar_modell <-lm(Y~x1+x2+x3+x4,data = Commercial_properties)
summary(linjar_modell)
```

```
##
## Call:
## lm(formula = Y ~ x1 + x2 + x3 + x4, data = Commercial_properties)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.187 -0.591 -0.091  0.558  2.944
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.22e+01   5.78e-01   21.11  < 2e-16 ***
## x1          -1.42e-01   2.13e-02   -6.65  3.9e-09 ***
## x2           2.82e-01   6.32e-02    4.46  2.7e-05 ***
## x3           6.19e-01   1.09e+00    0.57     0.57
## x4           7.92e-06   1.38e-06    5.72  2.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.14 on 76 degrees of freedom
## Multiple R-squared:  0.585,  Adjusted R-squared:  0.563
## F-statistic: 26.8 on 4 and 76 DF,  p-value: 7.27e-14
```

Den generella modellen för linjär regression, där β_0 symboliserar interceptet (startvärdet) och variablerna x kännetecknar förklarande variabler.

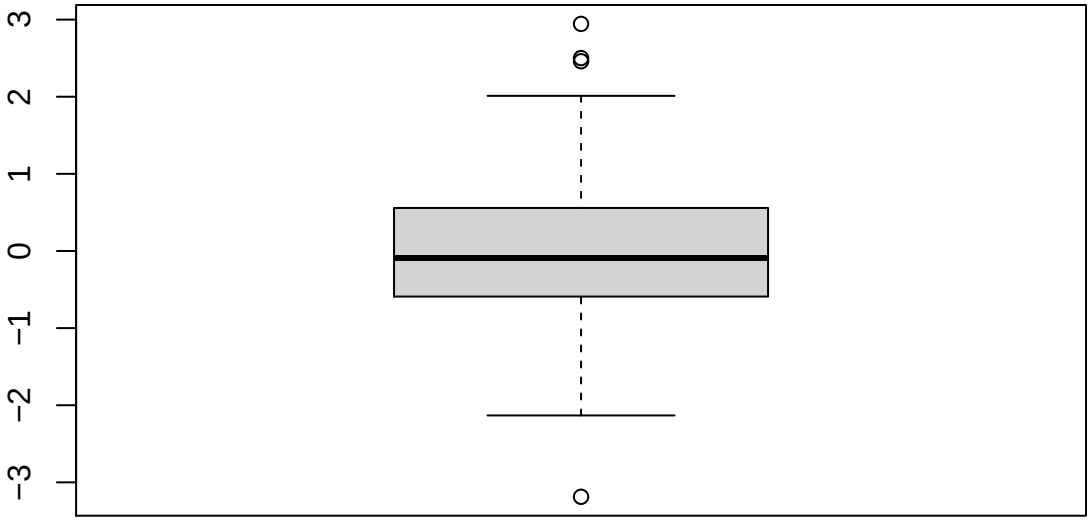
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

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

I regressionsfuktionen som anpassas syns flera faktum, för x1 som kännetecknar ålder på fastigheten, påverkas hyrespriset enligt modellen, negativt i samband med att åldern på fastigheten ökar. Bortsett från x1 ökar resterande variabler hyrespriset, x4 som kännetecknar storlek på fastigheten är väldigt låg, vilket kan förklaras med att koefficienten multipliceras med stora värden, då det är komersiella fastigheten som analyseras. x2 och x3 summerar till 0.282 och 0.613.

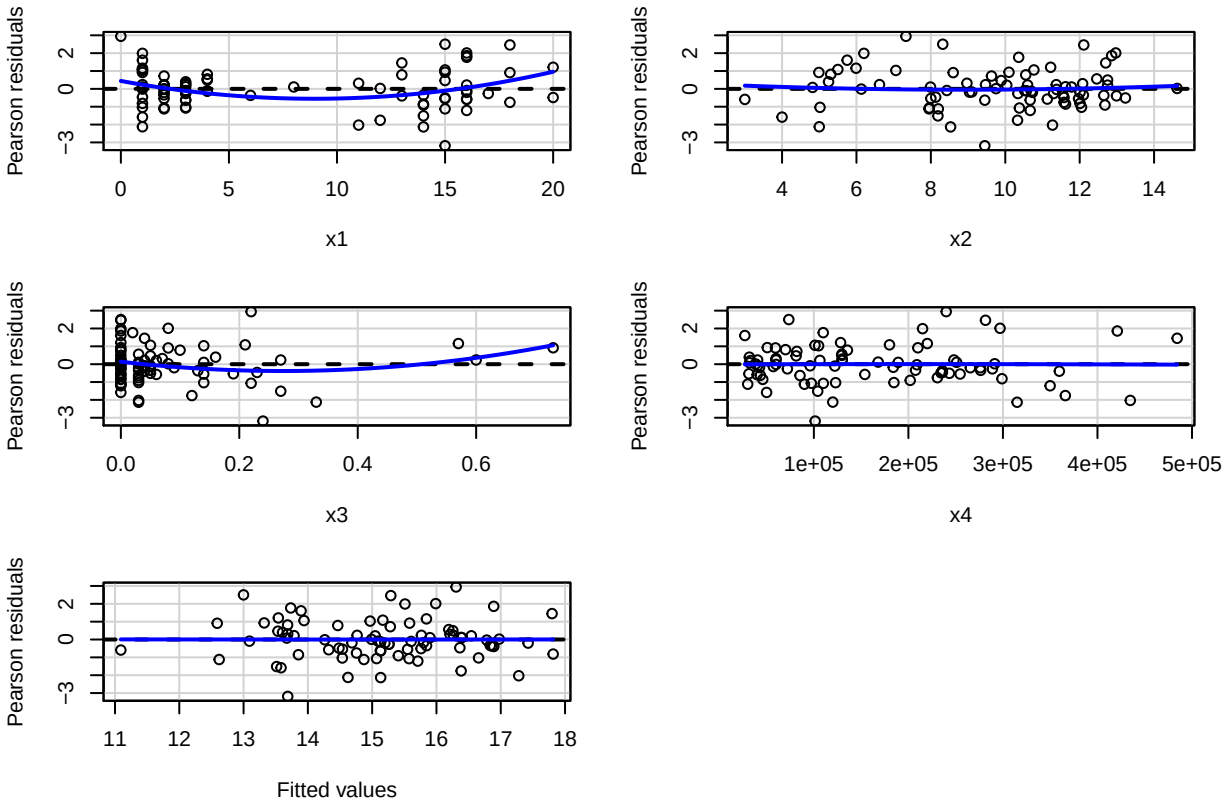
3.2.5 d) Obtain the resiudals and prepare a box plot of the residuals? Does the distrubition seem to be fairly symmetrical?

```
boxplot(resid(linjar_modell))
```



I boxploten ovan visas hur spridningen bland residualerna för modellen ser ut. En tydlig median vid 0 syns samtidigt som tre stycken avvikande observationer visas, en under medianen och två över medianen.

3.2.6 e) Plot the resiudals against Y_hat, each predictor variable and each two-factor interaction term on separete graphs.



I diagrammen ovan syns det hur residualerna är utspridda och skiljer sig åt från varandra vilket innebär att de är oberoende. Bland förklaringsgraderna är det X4 följt av x2,x1 och sist x3 sett till bäst spridning av residualer. Vad gäller residualerna gentemot responsvariabeln syns en fin fördelning som tyvärr dock inte är normalfördelade vilket innebära att hänsyn måste tas till de i modelleringen då vårat data inte är perfekt.

3.3 6.19

3.3.1 a) Test whether there is a regression relation.

$$\text{H0: B1} = \text{B2} = \text{B3} = \text{B4} = 0.$$

$$\text{H1: B1} = \text{B2} = \text{B3} = \text{B4} \neq 0.$$

```
Anova(linjar_modell)
```

```
## Anova Table (Type II tests)
##
## Response: Y
##           Sum Sq Df F value  Pr(>F)
## x1          57.2  1   44.29 3.9e-09 ***
## x2          25.8  1   19.93 2.7e-05 ***
## x3           0.4  1    0.32    0.57
## x4          42.3  1   32.75 2.0e-07 ***
## Residuals   98.2 76
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSR = (sum((fitted(linjar_modell) - mean(Commercial_properties$Y))^2))
SSE= sum((fitted(linjar_modell) - Commercial_properties$Y)^2)
```

$F = \frac{MSR}{MSE} = \frac{138.32/4}{98.23/76} = \frac{34.58}{1.293} = 26.744$

$$F(.95, 4, 76) = 2.492$$

I fall av att F > 2.492 förkastas H_0, om inte, antas H_0.

Då F överstiger det kritiska värdet ($26.744 > 2.492$), kan vi på 5 % signifikansnivå förkasta H_0 om att det inte finns en regressions relation.

3.3.2 b) Esimate the coefficients

```
## $coefficients
##                   Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  1.220059e+01 5.779562e-01 21.1098807 1.601720e-33
## x1          -1.420336e-01 2.134261e-02 -6.6549332 3.894322e-09
## x2           2.820165e-01 6.317235e-02  4.4642400 2.747396e-05
## x3           6.193435e-01 1.086813e+00  0.5698714 5.704457e-01
## x4           7.924302e-06 1.384775e-06  5.7224457 1.975990e-07
```

$$BT(1 - (0.05/(2 * 4)), 76) = 2.5585$$

$b_i \pm bf \cdot s_{b_1}$ Formel för konfidensintervall för koefficenter med bonferroni metoden.

Variabeln ålder på fastigheten

$$B1 = -.142 \pm 2.5585(.02134) = -.1966 : . - 0874$$

Då både översta och nedersta konfidensgränserna är negativa, kan slutsatsen kring att ålder på fastigheten alltid kommer att påverka hyrespriset negativt.

Variabeln Driftkostnad och skatter

$$B2 = 2.282 \pm 2.5585(.06317) = .1204 : .4436$$

Koefficenten är positiv i övre och nedre konfidensgränsen vilket innebär att den alltid kommer att bidra till ett ökat hyrespris.

Variabeln Vakansgrad

$$B3 = .6193 \pm 2.5585(1.08681) = -2.1613 : 3.4$$

Konfidensintervallet för variabeln är väldigt stort, vilket medför en stor osäkerhet i modellen, samtidigt som att detta påverkar prediktioner och kan ge felaktiga intervall.

Variabeln kvadratfot

$$B4 = .0000079 \pm 2.5585(.00000138) = .00000138 : .0000114$$

Ett väldigt litet värde för koefficenten returneras samtidigt som att väldigt stora intervall har bildats, det lilla värdet kan förklaras med att koefficienten multipliceras med stora värden, medelvärde för fastigheternas storlek summerar till 160633.3 vilket kan förklara det lilla värdet.

3.3.3 c) Calculate R2 and interpret your measure

```
SSR = (sum((fitted(linjar_modell) - mean(Commercial_properties$Y))^2))
SSE= sum((fitted(linjar_modell) - Commercial_properties$Y)^2)
SST = SSR+SSE
anova(linjar_modell)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14.819  14.819 11.4649  0.001125 **
## x2         1 72.802  72.802 56.3262 9.699e-11 ***
## x3         1  8.381   8.381  6.4846  0.012904 *
## x4         1 42.325  42.325 32.7464 1.976e-07 ***
## Residuals 76 98.231   1.293
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2 = 1 - \frac{SSR}{(SST)} = \frac{138.32}{236.56} = 0.5847$$

Alltså kan variablerna förklara 58.5 % av variationen i Y. Detta är ett moderat värde, men långt ifrån starkt, därav kan de vara tillfälle att se över andra modeller, där en eller flera variabler i nuvarande modell utesluts.

3.3.4 Uppgift 6.20 Obtain the family of estimates using a 95 percent CI coefficent.

```
options(scipen=0, digits=5)
property_one <- data.frame(Ålder=5,Driftkostnader_skatt=8.25,vakansgrad=0,yta=250000)
property_one
```

```
##   Ålder Driftkostnader_skatt vakansgrad    yta
## 1     5                 8.25          0 250000
```

Formel för konfidensintervall

$$\hat{y}_h \pm t_{(1- \ a/2,n-2)} \times \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

$$KI = 15.80 \pm 1.99(0.2713) = 15.26 : 16.34$$

I ett 95 % konfidensintervall kommer priset för fastighet 1 att ligga mellan 15.26 och 16.34, detta är ett relativt litet intervall trots en konfidensgrad på 95 %.

Formel för prediktionsintervall

$$\hat{y}_h \pm t_{(1- \ a/2,n-2)} \times \sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}$$

$$PI = 15.80 \pm 1.99(1.08681) = 13.51 : 18.08$$

I ett 95 % prediktionsintervall kommer hyrespriset för fastighet 1 att ligga mellan 13.51 och 18.08 prisenheter. Detta är ett relativt brett intervall, och modellen som används för anpassas för bättre prediktion.

3.4 Uppgift 3

Hyrespris	X0	Ålder	Driftkost_skatter	vakansgrad	storlek
13.5	1	1	5.02	0.14	123000
12.0	1	14	8.19	0.27	104079
10.5	1	16	3.00	0.00	39998
15.0	1	4	10.70	0.05	57112
14.0	1	11	8.97	0.07	60000

Ovan visas de fem första värden för variablerna i datat som kommer från uppgift 6.18.

```
##                   Intercept       x1      x2      x3         x4
## Matrisekvation       12.201 -0.14203 0.28202 0.61934 7.9243e-06
## Linjär regression    12.201 -0.14203 0.28202 0.61934 7.9243e-06
```

Samstämmiga koefficenter returneras för de båda modellerna, vilket tyder på att matrisekvationen fungerar. Matrisekvation skulle jag personligen inte klassa som svårt, däremot är det ett smidigare alternativ att använda sig av datorkraft. Med de sagt anser jag ändå att matrisekvationen är bra att ha med sig för att kunna kontrollera huruvida en linjär regressionsmodel gjord med datorkraft samstämmer eller inte.

3.5 Uppgift 4

3.5.1 Fit the regression model with x1, x2 and x4 as explanatory variables.Compare the obtained values of the F-test and R2 from this model with the model obtained from assignment 1 and comment.

```
##               Model_med_x3 Model_utan_x3
## r.squared     0.58475      0.58298
## adj.r.squared 0.56289      0.56673
## value         26.756       35.88
```

Från den modellen utan x3 fås en högre F-statistiska ut, vilket innebär att variablerna såsom kvadratfot, hyrespriser, och driftkostnad har en högre inverkan på Y (Hyrespriset) , till skillnad från modellen där variabeln vakansgrad inkluderas.

I den modellen utan x3 är det även mindre spridning bland residualer, två tusendelars sämre R2, det som däremot är viktigt att ta i beaktan är att det är en variabel mindre. Desto fler variabler en modell innehåller resulterar i att fler variabler har möjlighet att förklara variationen i responsvariabeln, med de sagt kan fler variabler innebära ett högre förklaringsgrad, trots att variablerna inte är signifikanta.

Rent generellt påstår jag att den exkluderande modellen är bättre, en högre f-statistiska, med en i princip oförändrad förklaringsgrad samtidigt som att modellen har en variabel mindre.

4 Lärdomar, problem, övriga kommentarer

Främst av den frekventa användning av LATEX bidragit till förbättrade kunskaper. Denna inlämningsuppgift har skapat tillfällen till att använda R's inbyggda funktioner för smidigt lösande av uppgiften.

5 Referenser