Labbrapport i Statistik

Laboration 4

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-09-23

Innehåll

1 Introduktion **1**

2 Databehandling **2**

3 Uppgifter **3**

3.1 8.8 Refer to Commercial properties Problems 6.18 and 7.7. 3

3.1.1 a) The age of the property (X1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (x1c), the square of centered property age (x1^2), operating expenses and taxes (X2), and total square footage (X4). Plot the Y observations against the fitted values. Does the response function provide a good fit? 3

3.1.2 b) Calculate R2-adj 8

3.1.3 c) Test wheter or not The square of centered property age (x^21) can be dropped from the model. 9

3.1.4 d) Estimate the mean rental rate when X1=8, X2 = 16 and X4 = 250,000; Use a 95 percent confidence interval. Interpret your interval. 9

3.2 8.40 Refer to the SENIC data set in Appendix C.1. Infection risk (Y) is to be regressed against length of stay (X1), age (X2), routine chest X-ray ratio (X3), and medical school affiliation (X4). 10

3.2.1 b) Estimate the effect of medical school affilation on infection risk using a 98 percent confidence interval. interpret your interval estimate 11

3.2.2 c) It has been suggested that the effect of medical school affiliation on infection risk may interactwith the effects of age and routine chest X-ray ratio. Add appropriale interaction terms to the regression model, fit the revised regression model, and test whether the interaction terms are helpful; use a = .10. State the alternatives, decision rule, and conclusion. 12

3.2.3 8.41 Refer to the SENIC data set in Appendix C.). Length of stay (Y) is to be regressed on age(X1), routine culturing raio (X2), average daily census (X3), available facilities and services(X.), and region(X5,X6,X7). 13

3.2.4 b) Test whether the routine culturing ratio can be dropped from the model; use a level of significance of .05. State the alternatives, decision rule, and conclusion. 14

3.2.5 c) Examine whether the effect on length of stay for hospitals located in the western region differs from that for hospitals located in the other three regions by constructing an appropriate confidence interval for each pairwise comparison. Use the Bonferroni procedure with a 95 percent family confidence coefficient, Summarize your findings. 15

4 Lärdomar **16**

1 Introduktion

I denna laboration kommer två dataset användas I det första som kommer 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).

I det andra datasetet som består av ett OSU av 113 sjukhus från 338 sjukhus som har undersökts, kommer variabler såsom; medellängden på besöken, åldern, sannolikheten att få en infektion och medelsumman av antal sjukhussängar m.fl. att analyseras och användas i modeller.

Målen med laborationen är att bemästra polynomial regression med vanliga mjukvaror. Använda kvalitativa förutspåelser i multipel regression genom att definera lämpliga dummy-variabler. Se samspel med olika variabler, specifikt interaktionen mellan kvalitativa och kvantitativa variabler.

2 Databehandling

```
Commercial_properties <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH06PR18.t
Commercial_properties = Commercial_properties[,-4]
cols1 <-c("y","x1","x2","x4")
colnames(Commercial_properties) <-cols1
options(digits = 4)


DS_C1 <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/Data%20set%20C.1")

cols2 <-c("NMR","length","Age","infectionRisk","RoutineCR","RoutineChXrR","Nmbrofbeds",
          "Medschal","Reg","ADC","NmbrofNurs","AvaFAS")
colnames(DS_C1) <-cols2
```

3 Uppgifter

3.1 8.8 Refer to Commercial properties Problems 6.18 and 7.7.

3.1.1 a) The age of the property (X1) appears to exhibit some curvature when plotted against the rental rates (Y). Fit a polynomial regression model with centered property age (x1c), the square of centered property age (x1^2), operating expenses and taxes (X2), and total square footage (X4). Plot the Y observations against the fitted values. Does the response function provide a good fit?

```
x1_c <-Commercial_properties$x1-mean(Commercial_properties$x1) # Subraherar medelvärdet från alla värden
x1_c2 <-x1_c^2 #Tar de subtraherade medelvärdet för att centrera värdena.
Commercial_properties <-cbind(Commercial_properties,x1_c) #Binder ihop datasetet

#Skapar en model med centrerad x1
model1 <-lm(y~x1_c+x1_c2+x2+x4,data = Commercial_properties)
summary(model1)#Returnerar en sammanfattning
```

```
##
## Call:
## lm(formula = y ~ x1_c + x1_c2 + x2 + x4, data = Commercial_properties)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8960 -0.6255 -0.0891  0.6279  2.6831
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.02e+01   6.71e-01   15.19  < 2e-16 ***
## x1_c        -1.82e-01   2.55e-02   -7.13  5.1e-10 ***
## x1_c2        1.41e-02   5.82e-03    2.43    0.017 *
## x2           3.14e-01   5.88e-02    5.34  9.3e-07 ***
## x4           8.05e-06   1.27e-06    6.35  1.4e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.1 on 76 degrees of freedom
## Multiple R-squared:  0.613,  Adjusted R-squared:  0.593
## F-statistic: 30.1 on 4 and 76 DF,  p-value: 5.2e-15
```

```r
#Skapar en modell utan centrerad x1
model2 <- lm(y~x1+x2+x4, data = Commercial_properties)
summary(model2)
```

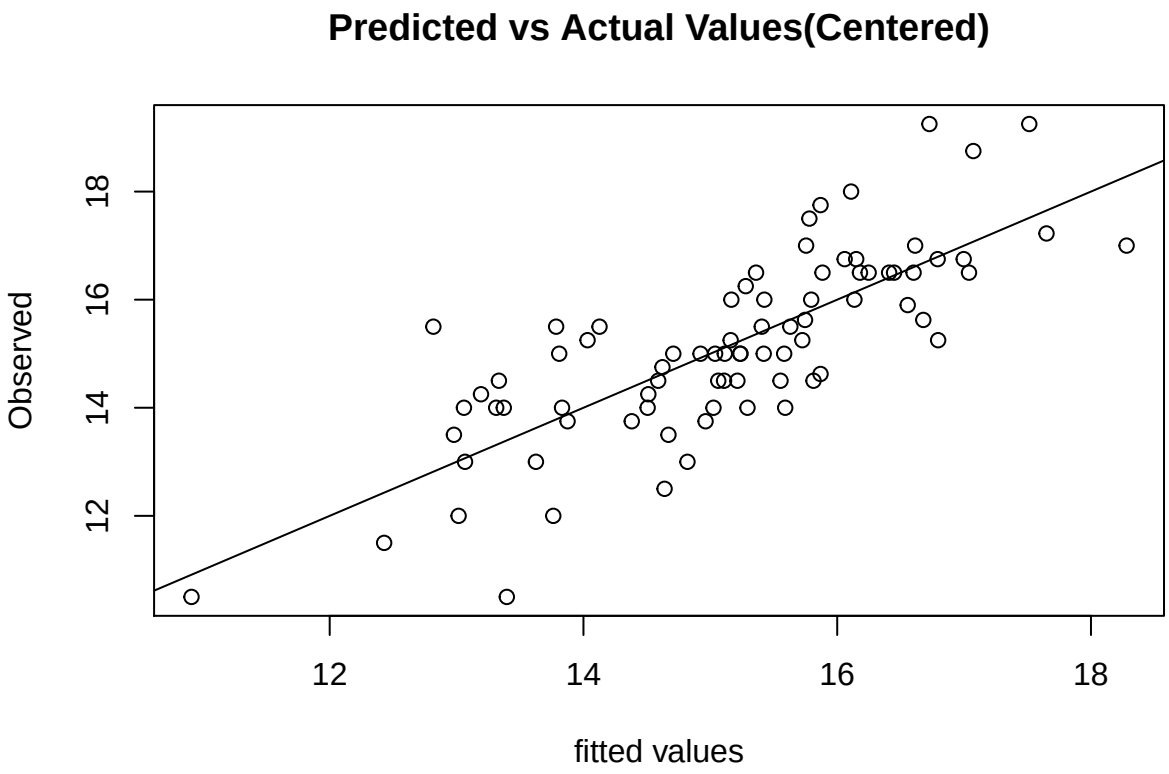
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x4, data = Commercial_properties)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.062 -0.644 -0.101  0.567  2.958
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.24e+01   4.93e-01   25.10  < 2e-16 ***
## x1          -1.44e-01   2.09e-02   -6.89  1.3e-09 ***
## x2           2.67e-01   5.73e-02    4.66  1.3e-05 ***
## x4           8.18e-06   1.31e-06    6.27  2.0e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.13 on 77 degrees of freedom
## Multiple R-squared:  0.583,  Adjusted R-squared:  0.567
## F-statistic: 35.9 on 3 and 77 DF,  p-value: 1.3e-14
```

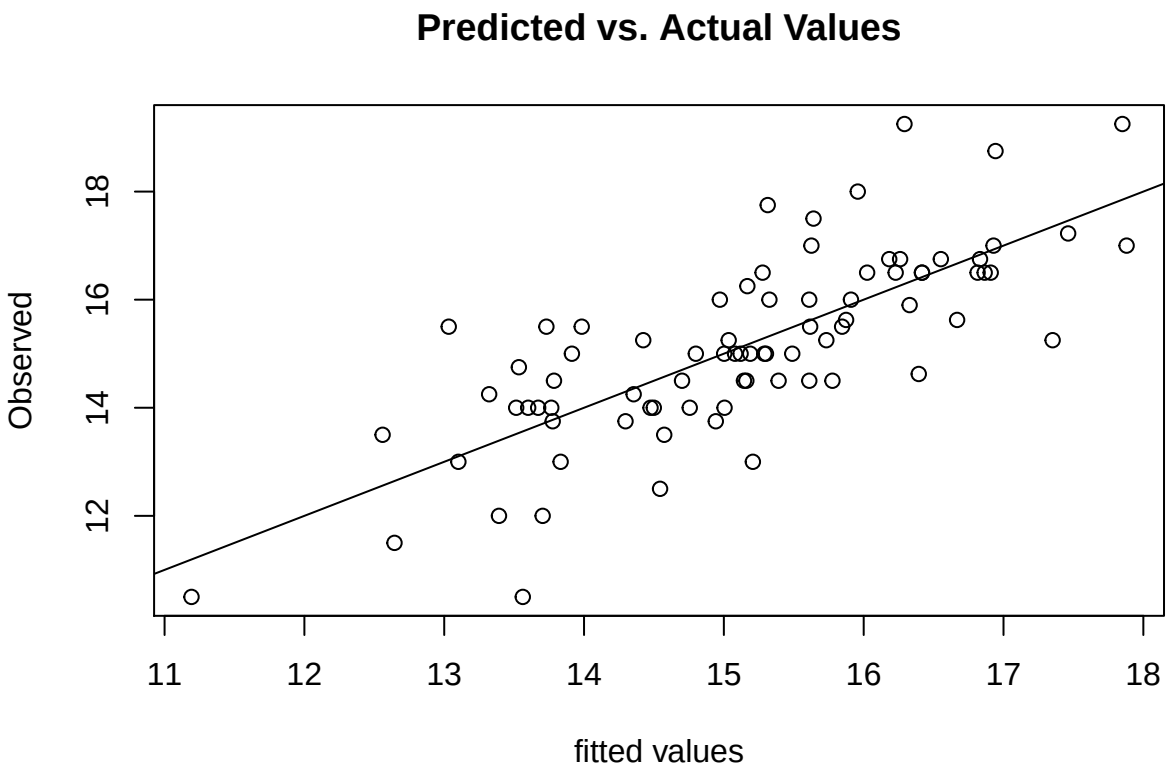
```r
model2$coefficients
```

```
## (Intercept)          x1          x2          x4
##   1.237e+01  -1.442e-01   2.672e-01   8.178e-06
```

Modellernas som returneras har samtliga variabler signifikanta, en moderat stark förklaringsgrad, en stark f-statistika, ett väldigt lågt p-värde. Rent generellt är modellen helt okej och kan förbättras.

Det man kan säga om skillnaden mellan de olika modellerna är att det centrerade har både lägre f statistika och p-värde än den ocentrerade modelllen.


Predicted vs Actual Values(Centered)
Observed
12
14
16
18
12
14
16
18
fitted values



När man jämför de två olika plottarna så ser man att plottarna är väldigt lika, däremot har modellen med centrerade x1 en mindre spridning mellan fitted values och observationerna vilket innebär att det sambanden är starkare än det som inte är observerat.

```r
#VIF värden
Centrerad <- vif(model1)
Ej_Centrerad <- vif(model2)
Centrerad
```

```
##  x1_c x1_c2    x2    x4
## 1.902 1.609 1.533 1.269
```

```r
Ej_Centrerad
```

```
##    x1    x2    x4
## 1.202 1.368 1.266
```

Man ser tydligen att det centrerade vif värdena har en högre korrelation mellan prediktionsvariablerna men att de båda VIF värdena visar på en väldigt låg korrelation.

```r
cor(x=Commercial_properties$y,y=model1$fitted.values)
```

```
## [1] 0.783
```

```r
cor(x=Commercial_properties$y,y=model2$fitted.values)
```

```
## [1] 0.7635
```

Här kan man se att det centrerade modellen har en högre korrelation än den modellen som ej har blivit centrerad.

Slutsatsen med jämförelse mellan dessa modeller är att ett centrererad värde ger en bättre model än om man ej hade centrerat. På alla de olika testerna och plottar ser man att den centrerade modellen har bättre värden och spridning.

3.1.2 b) Calculate R2-adj

$$\bar{R}^2 = 1 - \left(1 - R^2\right)\left[\frac{n-1}{n-(k+1)}\right] = 0.5927$$

Förklaringsgraden för modellen är varken hög eller låg och befinner sig i ett mellanstadium, den justerade förklaringsgraden skiljer sig med .02 vilket är väldigt lite. Sammanfattningsvis går det att säga att modellen endast kan förklara 59.3 % av variationerna i den beroende variabeln.

3.1.3 c) Test wheter or not The square of centered property age (xˆ21) can be dropped from the model.

H0: x21=0

H1: x21 $\neq$ 0

```
x1_c  <-1.41e-02
sbx1_c <-5.82e-03
t_test <-x1_c/sbx1_c
t_krit <-qt(0.975,76)

# Ifall teststatistiskan skulle överksrida det kritiska t-värdet kan vi förkasta nollhypotesen om att Xˆ
```

Eftersom att teststatistikan (2.4227) är högre än det kritiska värdet (1.9917) kan vi förkasta nollhypotesen om att kvadrade fastighetsåldern summerar till noll, vilket innebär betyder vi ej kan droppa variabeln från modellen då den är statistiskt signfikant.

3.1.4 d) Estimate the mean rental rate when X1=8, X2 = 16 and X4 = 250,000; Use a 95 percent confidence interval. Interpret your interval.

```
new.dat2 <- data.frame(x1=8,x11=0,x2=16,x4=250000) #Matar in värdena som ska användas
punkt.skatt2 <-predict(mm, newdata = new.dat2, interval = 'confidence',level = 0.95)
punkt.skatt2
```

```
##    fit   lwr   upr
## 1 17.2 16.46 17.94
```

Med 95 % konfidens kan vi säga att medelhyran för en fastighet som är 8 år gammal, har 16 i driftkostnader och skatter samt en yta på 250 000 kvadratmeter, kommer ligga i ett interval mellan 16.46 och 17.94 prisenheter.

3.2 8.40 Refer to the SENIC data set in Appendix C.1. Infection risk (Y) is to be regressed against length of stay (X1), age (X2), routine chest X-ray ratio (X3), and medical school affiliation (X4).

3.2.0.1 a) Fit a first—order regression model. Let X4 = 1 if hospital has medical school affiliation and 0 if not.

```
DS_C1$Medschal <- ifelse(DS_C1$Medschal == "1", 1, 0)
model_1 <-lm(infectionRisk~length+Age+RoutineChXrR+Medschal,data = DS_C1)
summary(model_1)
```

```
##
## Call:
## lm(formula = infectionRisk ~ length + Age + RoutineChXrR + Medschal,
##     data = DS_C1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7467 -0.7665 -0.0028  0.7727  2.5970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.85738    1.32434    0.65  0.51874
## length        0.28882    0.06291    4.59  1.2e-05 ***
## Age          -0.01805    0.02411   -0.75  0.45569
## RoutineChXrR  0.01995    0.00577    3.46  0.00078 ***
## Medschal      0.28782    0.30668    0.94  0.35009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 108 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.345
## F-statistic: 15.7 on 4 and 108 DF,  p-value: 3.57e-10
```

Modellen som returneras har främst en väldigt låg förklaringsgrad och tillsammans med en låg f-statistika. Enbart två värden är statistiskt signifikanta och har en signifikant påverkan på responsvariabeln. Det som sticker ut i modellen är det extremt låga p-värdet.

3.2.1 b) Estimate the effect of medical school affilation on infection risk using a 98 percent confidence interval. interpret your interval estimate

H0: B4=0

H1: B4 $\neq$ 0

```
##
## Call:
## lm(formula = infectionRisk ~ length + Age + RoutineChXrR + Medschal,
##     data = DS_C1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7467 -0.7665 -0.0028  0.7727  2.5970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.85738    1.32434    0.65  0.51874
## length        0.28882    0.06291    4.59  1.2e-05 ***
## Age          -0.01805    0.02411   -0.75  0.45569
## RoutineChXrR  0.01995    0.00577    3.46  0.00078 ***
## Medschal      0.28782    0.30668    0.94  0.35009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 108 degrees of freedom
## Multiple R-squared:  0.368,  Adjusted R-squared:  0.345
## F-statistic: 15.7 on 4 and 108 DF,  p-value: 3.57e-10
```

Formeln för konfidensintervall för lutningsparameter. $b_i \pm t(a/2) \cdot s_{b_1}$

$$T = (0.99, 2, 16) = 2.361$$

$$0.2878_4 \pm 2.361 * 0.3066 = -0.4364 : 1.012$$

Med 98 % konfidens kommer Medical school affilation ha en påverkan på infection risk som kommer att ligga mellan -0.4364 och 1.012 enheter. Alltså kan Medical school affilation både öka och sänka infektionsrisken, vilket kan innebära att det kan vara en osäker parameter att ha med samtidigt som att p-värdet för parametern är väldigt högt.

3.2.2 c) It has been suggested that the effect of medical school affiliation on infection risk may interactwith the effects of age and routine chest X-ray ratio. Add appropriale interaction terms to the regression model, fit the revised regression model, and test whether the interaction terms are helpful; use a = .10. State the alternatives, decision rule, and conclusion.

$$\text{H0: B5=B6=0}$$

$$\text{H1: B5=B6} \neq 0$$

$$F = (192.305556/1.1380 = 2.257)$$

I fall av att teststatistikan överstiger det kritiska värdet ($F > 4.757$) förkastas $ H_{0} $ om att B5 och B6 = 0

$$F(.90, 2, 106) = 2.353$$

I detta fall då ($2.257 < 2.353$) samtidigt som att vi får ett p-värde som summeras till ($3.4934e-065$) kan vi med 95 % konfidens förkasta H_0 om om att B5 och B6 = 0.

```
model_2 <-lm(infectionRisk~+length+Age+RoutineChXrR+Medschal+Age*Medschal+RoutineChXrR*Medschal,data = D
summary(model_2)
```

```
##
## Call:
## lm(formula = infectionRisk ~ +length + Age + RoutineChXrR + Medschal +
##     Age * Medschal + RoutineChXrR * Medschal, data = DS_C1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7507 -0.7032 -0.0747  0.7647  2.6090
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.99413    1.39446    0.71  0.47746
## length                 0.26414    0.06337    4.17  6.3e-05 ***
## Age                   -0.02283    0.02470   -0.92  0.35743
## RoutineChXrR           0.02429    0.00648    3.75  0.00029 ***
## Medschal              -5.69520    4.60096   -1.24  0.21851
## Age:Medschal           0.15576    0.09268    1.68  0.09578 .
## RoutineChXrR:Medschal -0.02406    0.01389   -1.73  0.08623 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 106 degrees of freedom
## Multiple R-squared:  0.394,  Adjusted R-squared:  0.36
## F-statistic: 11.5 on 6 and 106 DF,  p-value: 7.25e-10
```

Rent allmänt har modellen väldigt låg förklaringsgrad, samtidigt som att väldigt få variabler är signifikanta, detta i syns i de höga p-värdena och de låga t-värdena.

3.2.3 8.41 Refer to the SENIC data set in Appendix C.). Length of stay (Y) is to be regressed on age(X1), routine culturing raio (X2), average daily census (X3), available facilities and services(X.), and region(X5,X6,X7).

```
##
## Call:
## lm(formula = length ~ +Age + RoutineCR + ADC + AvaFAS + Reg1 +
##     Reg2 + Reg3, data = DS_C2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.794 -0.730  0.004  0.539  7.723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.0478     1.8130    1.13   0.2612
## Age           0.1037     0.0315    3.30   0.0013 **
## RoutineCR     0.0403     0.0143    2.82   0.0058 **
## ADC           0.0066     0.0014    4.70  7.9e-06 ***
## AvaFAS       -0.0208     0.0144   -1.44   0.1515
## Reg1          2.1500     0.4615    4.66  9.4e-06 ***
## Reg2          1.1903     0.4371    2.72   0.0076 **
## Reg3          0.6335     0.4275    1.48   0.1414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.4 on 105 degrees of freedom
## Multiple R-squared:  0.498,  Adjusted R-squared:  0.465
## F-statistic: 14.9 on 7 and 105 DF,  p-value: 2.28e-13
```

$$\hat{Y} = 2.0478 + .1037X_1 + .0403X_2 + .0066X_3 - .0208X_4 + 2.15X_5 + 1.1903X_6 + 0.6335X_7$$

Ovan visas ekvationen för regressionsmodellen som anpassats, vad gäller de 3 sista parametrarna är det enbart en som används beroende på vilken region som syftas på, för region 1 multipliceras den första av de tre sista med 1 och resterande med 0 detta gäller för samtliga av de sista parametrarna.
I modellen som främst syftar in sig på region kan vi se att två regioner är signifkanta medan den tredje ej är signifkant. P-värdet är väldigt lågt och f-statistikan är också svag. Modellen behöver definitivt förbättras.

3.2.4 b) Test whether the routine culturing ratio can be dropped from the model; use a level of significance of .05. State the alternatives, decision rule, and conclusion.

H0: x2=0

H1: x2 $\neq 0$

$$t = (0.975, 105) = 1.983$$

$$T_1 = \frac{\beta 3}{s\beta 3} = \frac{0.0403}{0.0143} = 2.818$$

Då teststatistikan 2.818 överstiger det kritiska värdet 1.983 kan vi förkasta nollhypotesen om att Routine Culturing ratio kan droppas från modellen och bekräftar att det är en statistiskt signifikant variabel som bör vara kvar.

3.2.5 c) Examine whether the effect on length of stay for hospitals located in the western region differs from that for hospitals located in the other three regions by constructing an appropriate confidence interval for each pairwise comparison. Use the Bonferroni procedure with a 95 percent family confidence coefficient, Summarize your findings.

```
fit <-(model_3$coefficients[6:8]) # Tar ut koefficenterna som en vektor
sd_1 <-smry_model_3$coefficients[6:8,2] #Tar ut standardfelen som en vektor
B <-qt(1-(.05/6),105) #Det kritiska värdet
bon_upper <- fit + B * sd_1
bon_lower <- fit - B * sd_1
bon_f <-cbind(bon_lower,bon_upper,fit)
cols_3 <-c("2.5%","97.5%","fit")
colnames(bon_f) <-cols_3
rows_1 <-c("Region1","Region2","Region3")
row.names(bon_f) <-rows_1
bon_f
```

```
##            2.5% 97.5%    fit
## Region1  1.0271 3.273 2.1500
## Region2  0.1270 2.254 1.1903
## Region3 -0.4067 1.674 0.6335
```

Ovan i tabellen så syns ett 95% konfidensintervall som har tagits fram med bonferroni procedur. I den vänstra kolumnen syns LCL, den mittersta UCL och den högersta kolumnen fit.

4 Lärdomar

I denna laboration har vi bekantat oss med polymial regression, använda kvalitativa förutspåelser i multipel regression genom att definera lämpliga dummy-variabler. Se samspel med olika variabler, specifikt interaktionen mellan kvalitativa och kvantitativa variabler.