Labbrapport i Statistik

Laboration 1

732G46

Mattias Hällgren
Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-09-01

Innehåll

1 Introduktion

I denna laboration kommer två datamaterial att användas, det första datamaterialet behandlar Kopieringsmaskiner och hur lång tid det tar för servicepersonal att utföra service beroende på hur många maskiner som servas. I det andra datamaterialet som används har antalet skadade ampuler som påträffats i samband med transport jämförts beroende på hur många transfers som skett. Målet med uppgifterna är att bli bekväm och kunna beräkna olika typer av spridningar i form av prediktions och konfidensintervall för variabler och summor.

För smidig datahantering läggs samtlig data som behövs upp i min (Michael Debebe's) github, vilket gör det möjligt att enkelt kunna ladda ner data till vilken dator som helst.

1.1 Databehandling

```
Uppgiftetttjugo <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH01PR20.txt") #
cols <-c("Minuter","Maskiner") #Namnger kolumnerna
colnames(Uppgiftetttjugo) <-cols
```

2 Uppgifter

2.1 Uppgift 1

2.1.1 d) How large is the estimated standard deviation and variance of the error terms in the regression model?

```
##
## Call:
## lm(formula = Minuter ~ Maskiner, data = Uppgiftetttjugo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Maskiner     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Från utskriften går det att läsa ut de uppskattade standardavvikelserna för regressionsmodellen. För β_1 (Interceptet) är standardavvikelsen 2.8, för β_0 är standardavvikelsen 0.4831

$Varians\ av\ residualerna\ = (\frac{\text{SumSq}}{\text{df}})^2 = (\frac{3416}{43})^2 = (8.91)^2 = 79.45$

2.1.2 g) A customer with six copying machines calls for service. Compute a 95% prediction interval for the time it will take for that service.

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

$$89.6 \pm 18.2$$

$$[71.4 : 107.8]$$

Alltså kommer det att ta mellan 71.4 och 107.8 minuter för att serva 6 kopieringsmaskiner.

2.1.3 i) Consider the ANOVA table in for the regression model. Which parameter is estimated by MSE? How is MSE related to s? Hint: anova()

```
anova(linjar_modell)
```

```
## Analysis of Variance Table
##
## Response: Minuter
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Maskiner   1  76960   76960  968.66 < 2.2e-16 ***
## Residuals 43   3416      79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Den parametern som är uppskattad genom MSE (Mean Squared Error), alltså hur mycket varje enskild observation skiljer sig från medelvärdet i kvadrat.

MSE är relaterat till s på så sätt att MSE är S upphöjt i 2, S är standardavvikelsen för hela modellen.

2.1.4 j) Which hypothesis is tested in the F-test? (use summary()) How is the outcome of this test related to the confidence interval computed in exercise

```
summary(linjar_modell)
```

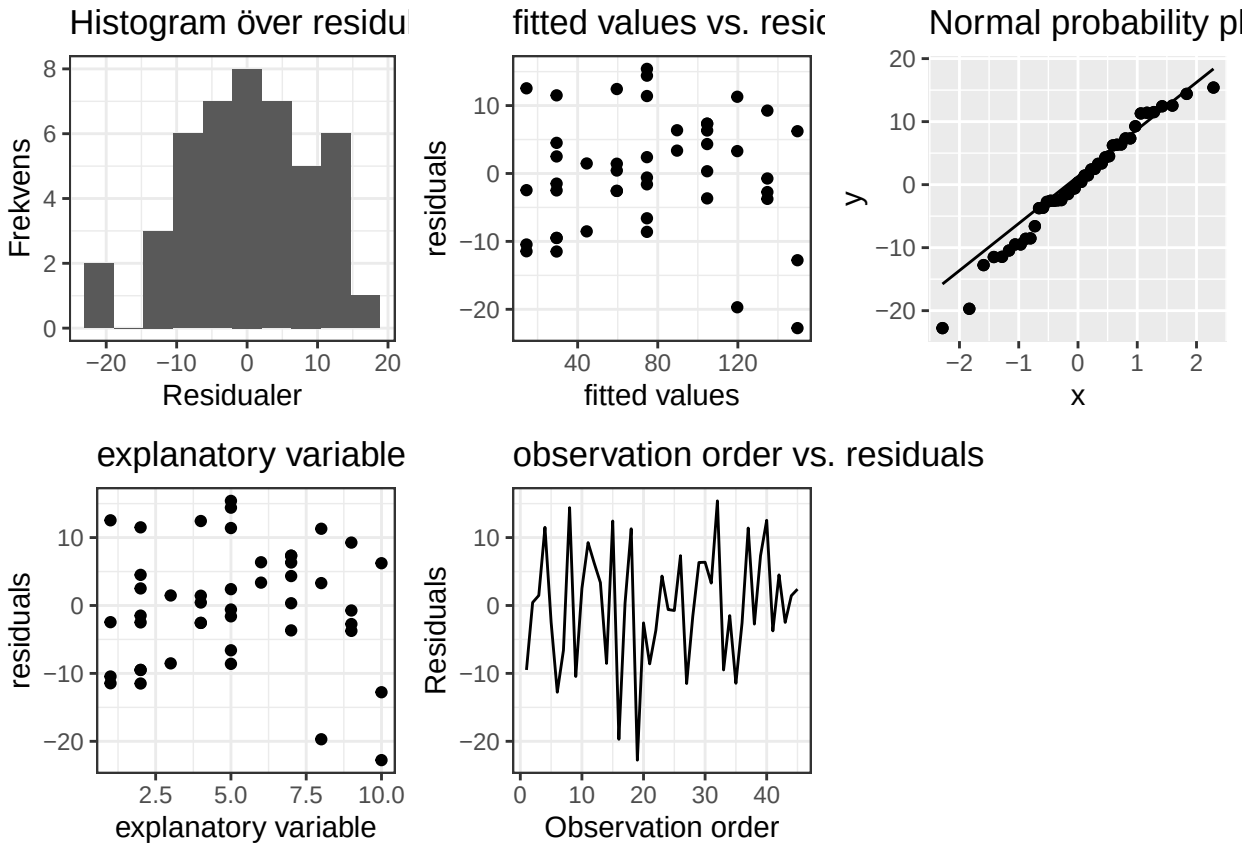
```
##
## Call:
## lm(formula = Minuter ~ Maskiner, data = Uppgiftetttjugo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## Maskiner     15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

```
myanova <- anova(linjar_modell)
cbind(myanova, 'CriticalValue' = qf(1-.05, myanova[1,1], myanova[2,1]))
```

```
##           Df    Sum Sq     Mean Sq  F value       Pr(>F) CriticalValue
## Maskiner   1 76960.423 76960.42298 968.6572 4.009032e-31      4.067047
## Residuals 43  3416.377    79.45063       NA           NA      4.067047
```

Hypotesen som testas i F-testet är hur antalet kopieringsmaskiner påverkar antalet minuter servicepersonalen spenderar. Då vår F-statistiska (968.7) är större än det kritiska värdet på 4.067, vilket innebär att vi kan förkasta nollhypotesen om att antal kopieringsmaskiner inte påverkar antal minuter servicepersonalen spenderar.

2.2 Uppgift 2



I normalfördelningstabellen syns extremvärden som bäst, de som framgår främst är de två observationerna längst ned samt den högsta, bortsett från det följer observationerna linjen någorlunda bra, däremot inte tillräckligt bra för att kunna klassas som normalfördelade.

Det syns trender i observationsordningen, däremot är dessa trender inte enhetliga, vilket tyder på att variansen för residualerna inte är konstant.

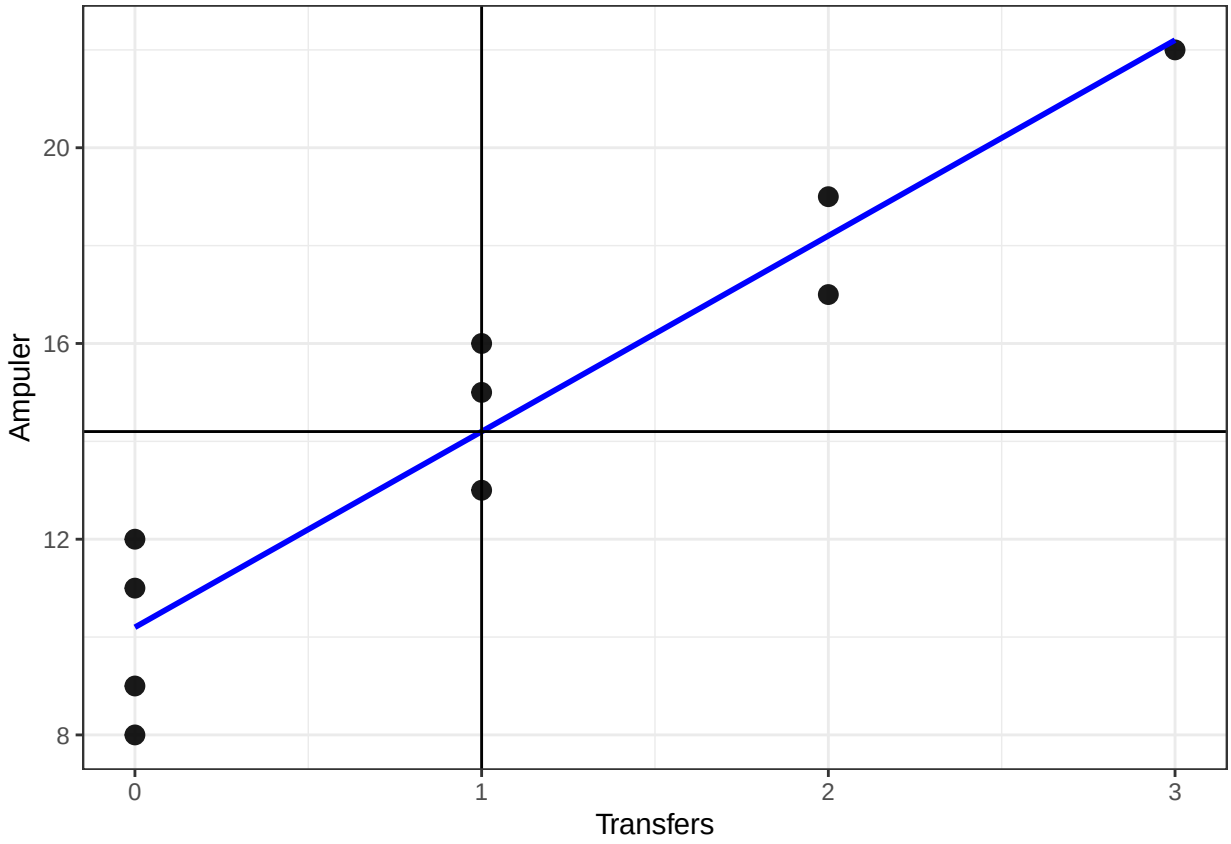
I explanatory vs Residuals är punkterna relativt utspridda, däremot är det nedre delen av diagrammet tom. Punkterna centraliserar sig kring 0.

Histogrammet visar upphov till att tillvisso vara normalfördelat med uppfyller inte fullständigt kravet för att betraktas som normalfördelat.

2.2.1 Uppgift 1.21 Airfreight Breakage

```
##
## Call:
## lm(formula = Ampuler ~ Transfers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## Transfers     4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

2.2.2 a) Obtain the estimated regression function. Plot the regression function.Does a linear regression function appear to give a good fit here?



Den linjära regressionslinjen passar bra samtidigt som att den i en punkt till och med går igenom punkten.

b) Obtain a point estimate of the expected number of broken ampules when X = 1 transfer is made

$\hat{y} = b_0 + b_1 x_0$

$\hat{y} = 10.2 + 4 * 1 = 14.2$

Det förväntade värdet ampuler som är sönder vid en frakt är 14.2

c) Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer

$\hat{y} = 10.2 + 4 * 2 = 18.2$

Vid två försändelser ökar det förväntade värdet av ampuler som går sönder med 4 enheter.

d) Verify that your fitted regression line goes through the point(Xbar, Ybar)

Detta svaras på i grafen som visas på a) uppgiften.

2.3 Uppgift 2.6

2.3.1 a) Estimate B1 with a 95 % CI, interpret your interval estimate

$$\beta_0 \pm t_{\alpha/2}^{(n-2)} \cdot \frac{s}{\sqrt{SS_{xx}}}$$
$$4 \pm 1.08$$
$$[2.92 : 5.08]$$

Med 95 % konfidens går det att säga att B1 kommer ligga mellan 2.92 och 5.08.

2.3.2 b) Conduct a t-test to decide whether or not there is a linear association between the numbers of times a carton is transferred (X) and number of broken ampules (Y)

```
t.test.lm = lm(Ampuler ~ Transfers, data=Flygdata)
summary(t.test.lm)
```

```
##
## Call:
## lm(formula = Ampuler ~ Transfers, data = Flygdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## Transfers     4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

Då t-värdet som returneras (8.52) är högre än vårt kritiska värde 2.3060041 kan vi förkasta nollhypotesen om att det inte råder något statistiskt signifkant samband mellan Antalet Ampuler som går sönder och Transfers.

2.3.3 c) Obtain a 95 percent confidence interval for β_0 and interpret it

$$\beta_0 \pm t_{\alpha/2}^{(n-2)} \cdot \frac{s}{\sqrt{SS_{xx}}}$$
$$10.2 \pm 1.52963$$
$$8.67 : 11.73$$

Alltså kommer antalet ampuler som går sönder vid 0 tranfers att med ett 95 % konfidens intervall ligga mellan 8.67 och 11.73

2.3.4 d) A consultant has suggested that the mean number of broken ampuels should not exceed 9.0 when no transfers are made. Conduct a appropriate test.

```
# H0: B0 är större eller lika med 9
# H1: B0 är större än 9
B0 <-10.2
u <-9
s_b_0 <-0.6633
t_tezt <-(B0-u)/s_b_0
t_tezt
```

```
## [1] 1.809136
```

```
t_krit <-qt(0.025,df=8,lower.tail = FALSE)
t_krit
```

```
## [1] 2.306004
```

Eftersom att teststatistiskan från t-testet understiger det kritiska kan vi inte förkasta H0. Alltså går det inte att säga om medelvärdet av trasiga Ampuler överstiger 9.

2.4 2.15

2.4.1 a) Estimate the mean breakage for 2 transfers, and 4, use separate 99 % confidence intervals

För X=2

$$18.2 \pm 2.23$$

$$[15.97 : 20.43]$$

Med 99% konfidens ligger antalet förstörda ampuller när det är 2 tranporter mellan 16 (15.97) och 21 (20.43).

För X =4

$$26.2 \pm 5$$

$$[21.2 : 31.2]$$

Med 99% konfidens ligger antalet förstörda ampuller när det är 4 tranporter mellan 16 (15.97) och 21 (20.43).

2.4.2 b) Obtain a 99% prediction interval for the mean number of ampules broken in this shipment(X=2)

$$18.2 \pm 5.5$$

$$[12.7 : 23.7]$$

När två transfers sker kommer det med 99 % predictionssäkerhet att var mella 12.7 och 23.7 ampuler som går sönder.

2.5 Uppgift 2.25

2.5.1 a Set up the Anova table. Which elements are additive?

```
anova(Lm_Flygdata)
```

```
## Analysis of Variance Table
##
## Response: Ampuler
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Transfers  1  160.0   160.0  72.727 2.749e-05 ***
## Residuals  8   17.6     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I anova utskriften som syns är det Sum Sq och Df (Degrees of freedom) som är dom additiva elementen.

2.5.2 b) Conduct an f-test to decide where or not there is a liner association between the number of times a carton is transferred and the number of broken ampules.

```
summary(Lm_Flygdata)
```

```
##
## Call:
## lm(formula = Ampuler ~ Transfers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## Transfers     4.0000     0.4690   8.528 2.75e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

```
myanova1 <- anova(Lm_Flygdata)
cbind(myanova1, 'CriticalValue' = qf(1-.05, myanova1[1,1], myanova1[2,1]))
```

```
##           Df Sum Sq Mean Sq  F value       Pr(>F) CriticalValue
## Transfers  1  160.0   160.0 72.72727 2.748669e-05      5.317655
## Residuals  8   17.6     2.2       NA           NA      5.317655
```

Teststatistikan som ges från F-test (72.73) är betydligt mycket högre än det kritiska värdet (5.317) vilket innebär att vi kan förkasta nollhypotesen om att antalet transfers inte har en påverkan på hur många ampuler som går sönder.

3 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.

4 Referenser