Labbrapport i Statistik

Laboration 5

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-10-03

Innehåll

1 Introduktion

I denna laboration kommer ett dataset användas.
I 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. Det som skiljer denna laboration från den tidigare laborationen är att samtliga 113 observationer inte kommer att användas samtidigt.

2 Databehandling

```
DS_C1 <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/Data%20set%20C.1") #Laddar
cols2 <-c("NMR","length","Age","infectionRisk","RoutineCR","RoutineChXrR","Nmbrofbeds",
          "Medschal","Reg","ADC","NmbrofNurs","AvaFAS") #Namnger kolumnerna
colnames(DS_C1) <-cols2
 DS_C1 <- subset (DS_C1, select = -c(NMR,Reg,Medschal))
```

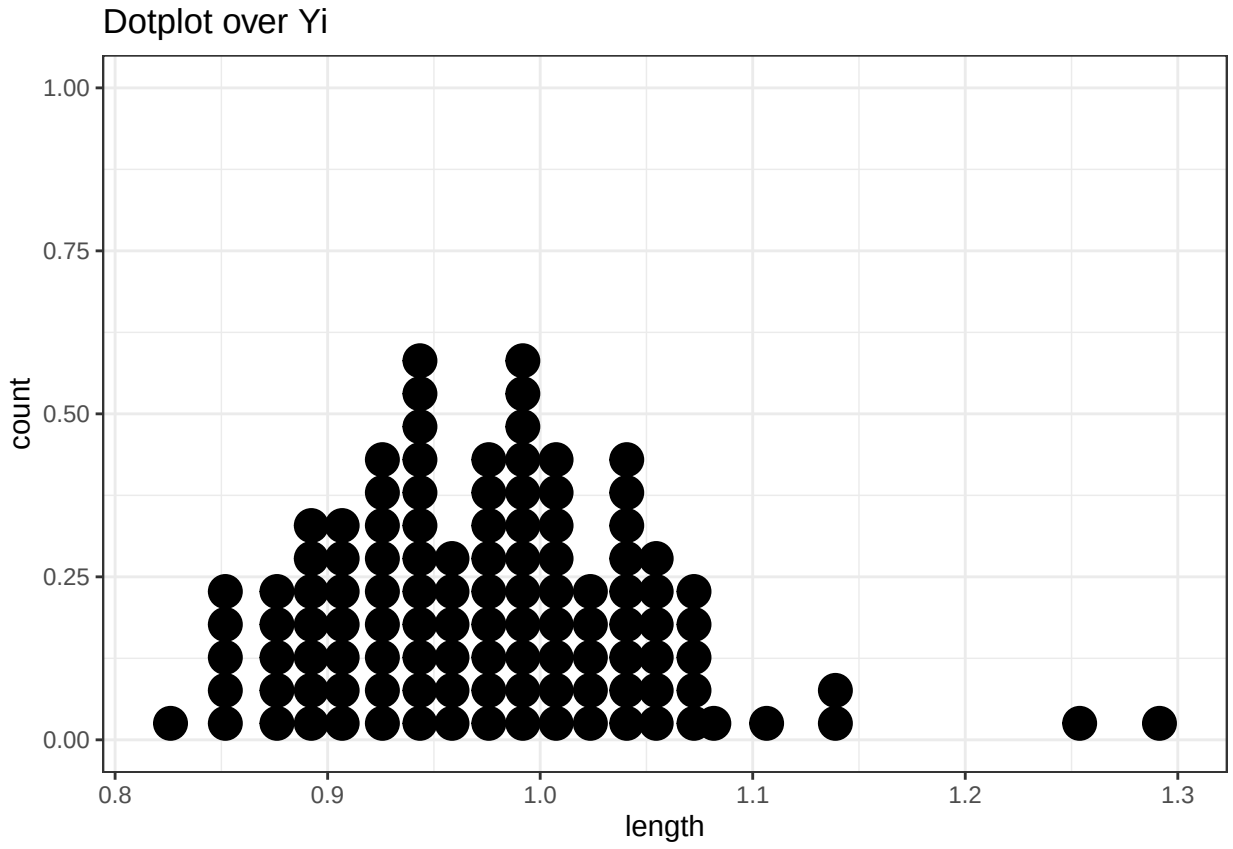
3 Uppgifter

3.1 Uppgift 1

3.1.1 9.25

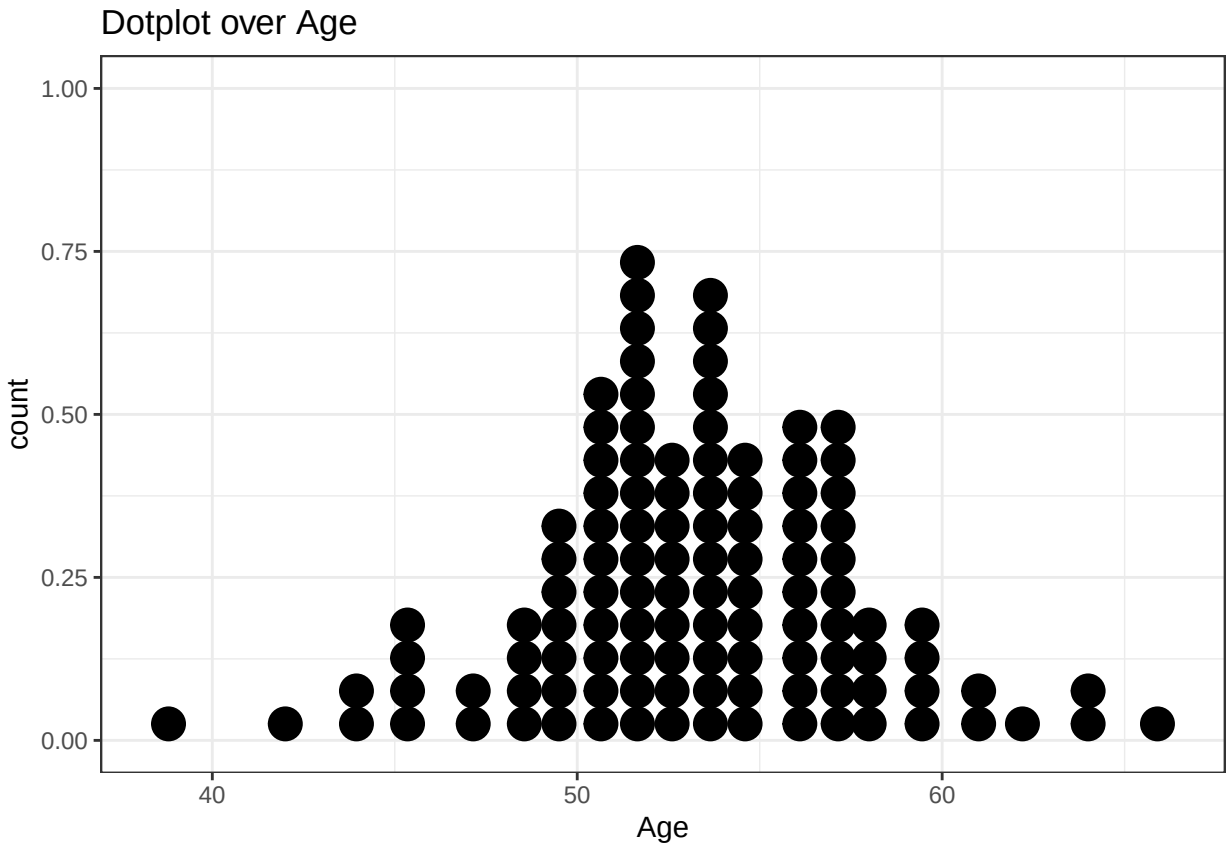
```
#Databehandling
DS_C1$length <-log(DS_C1$length,base = 10) #Logarithmerar responsvariabeln
DS_C2 <-DS_C1 #Skapar det andra datatsetet för den senare uppgiften
DS_C1_1 <-DS_C1 #Skapar datasetet som kommer användas i första delen av labben.
DS_C1_1 <-DS_C1_1[57:113,] #Det första datasetet får  57:113 raderna
DS_C2 <- DS_C2[1:56,] #Det andra datasetet får 1:56 raderna
```

```
ggplot(data = DS_C1, aes(x=length))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Yi")
```



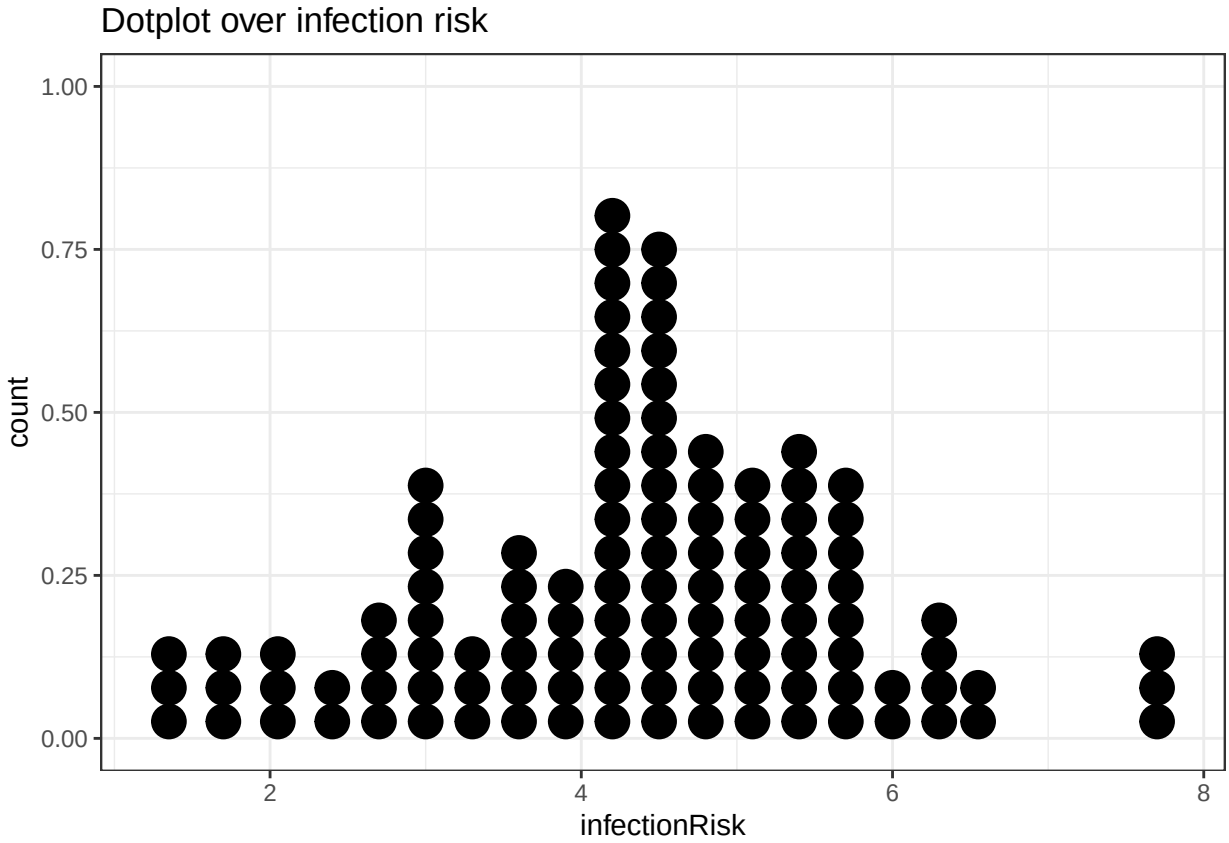
Vad gäller längden på sjukhusbesöken så koncentreras majoriteten till mellan 0.85 och 1.05 på grund av logarithmeringen, bortsett från det ser fördelningen relativt bra ut trots den positiva skevheten.

```
ggplot(data = DS_C1, aes(x=Age))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Age")
```



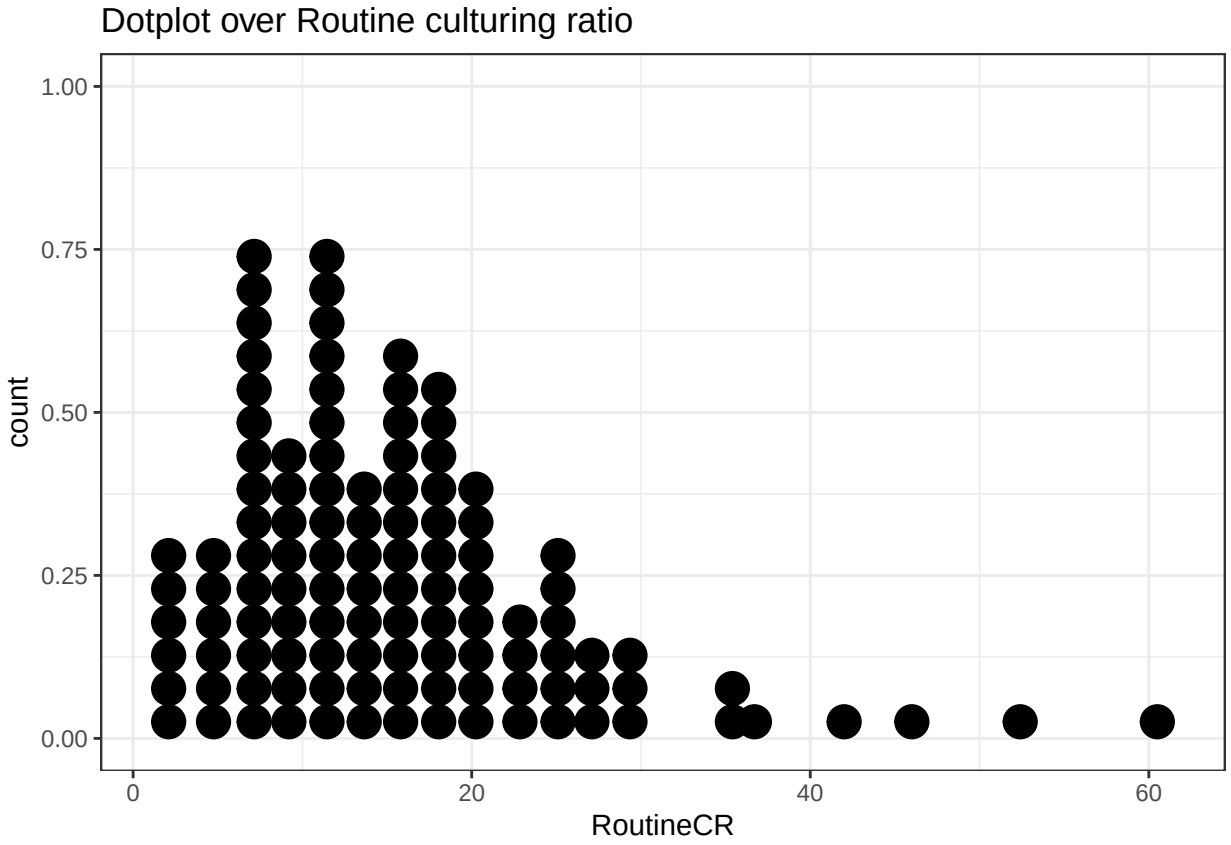
Vad gäller åldern på patienterna så returneras en fin fördelning med en median strax över 50 år.

```
ggplot(data = DS_C1, aes(x=infectionRisk))+
 geom_dotplot()+
 theme_bw()+
 labs(title = "Dotplot over infection risk")
```



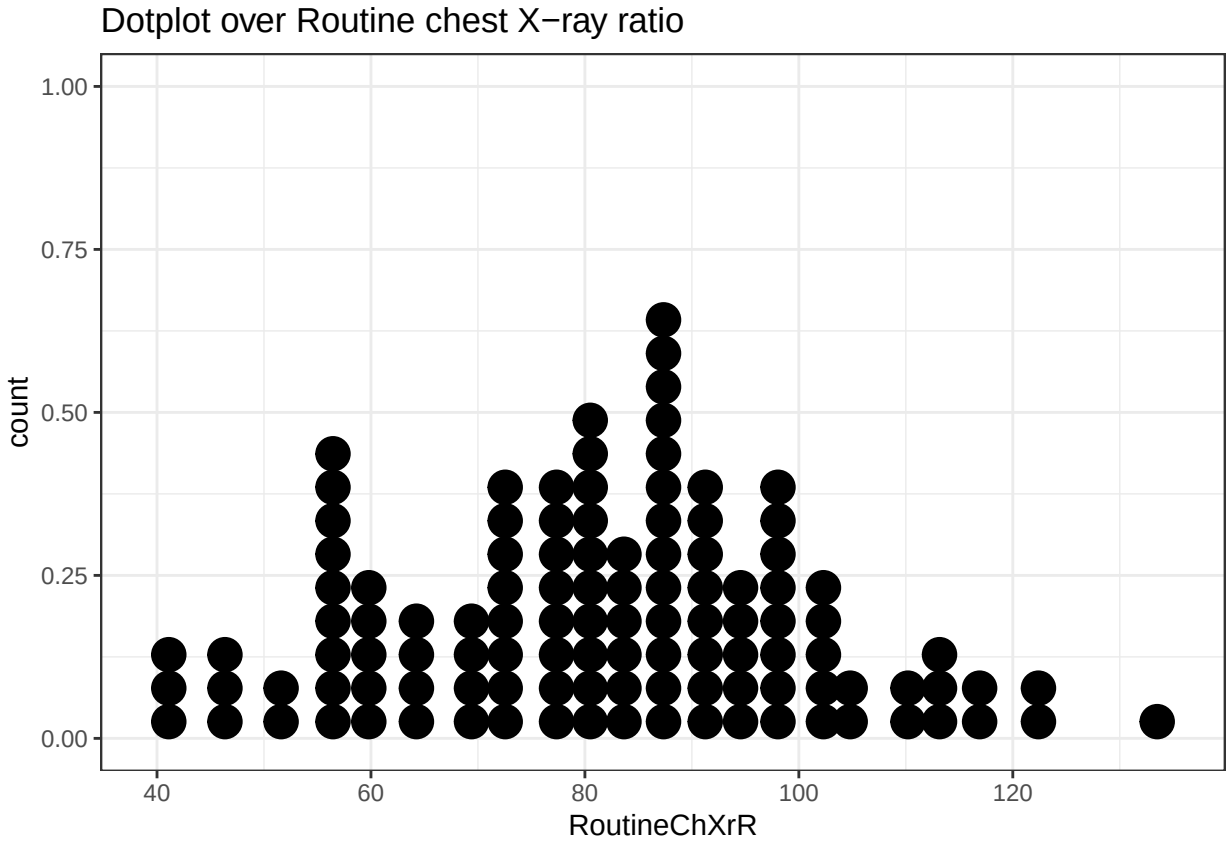
Gällande infection risk varierar variabeln väldigt mycket men har en koncentration mellan 3 och 6.

```
ggplot(data = DS_C1, aes(x=RoutineCR))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Routine culturing ratio")
```



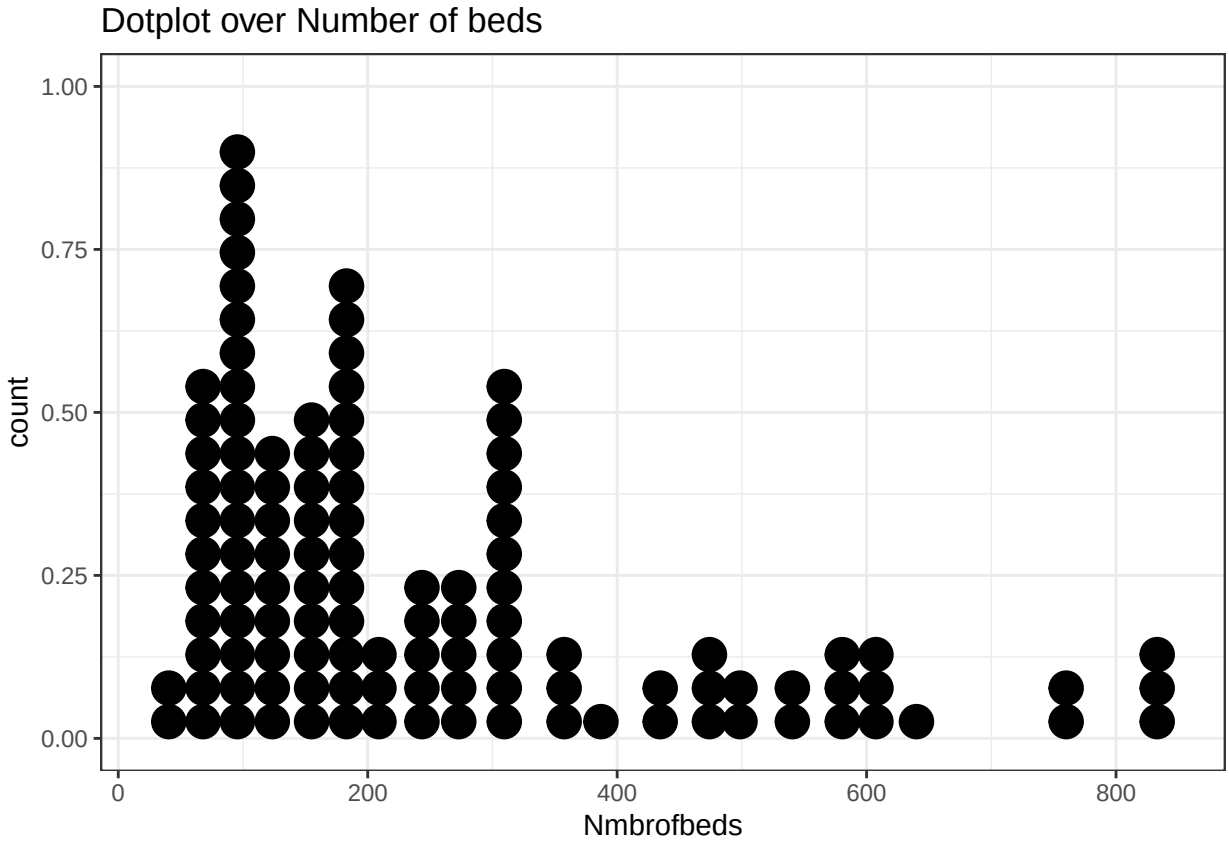
Koncentration kring 2-25 för att därefter vara väldigt låg runt 37-40, återigen en variabel som är positivt skev.

```
ggplot(data = DS_C1, aes(x=RoutineChXrR))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Routine chest X-ray ratio")
```



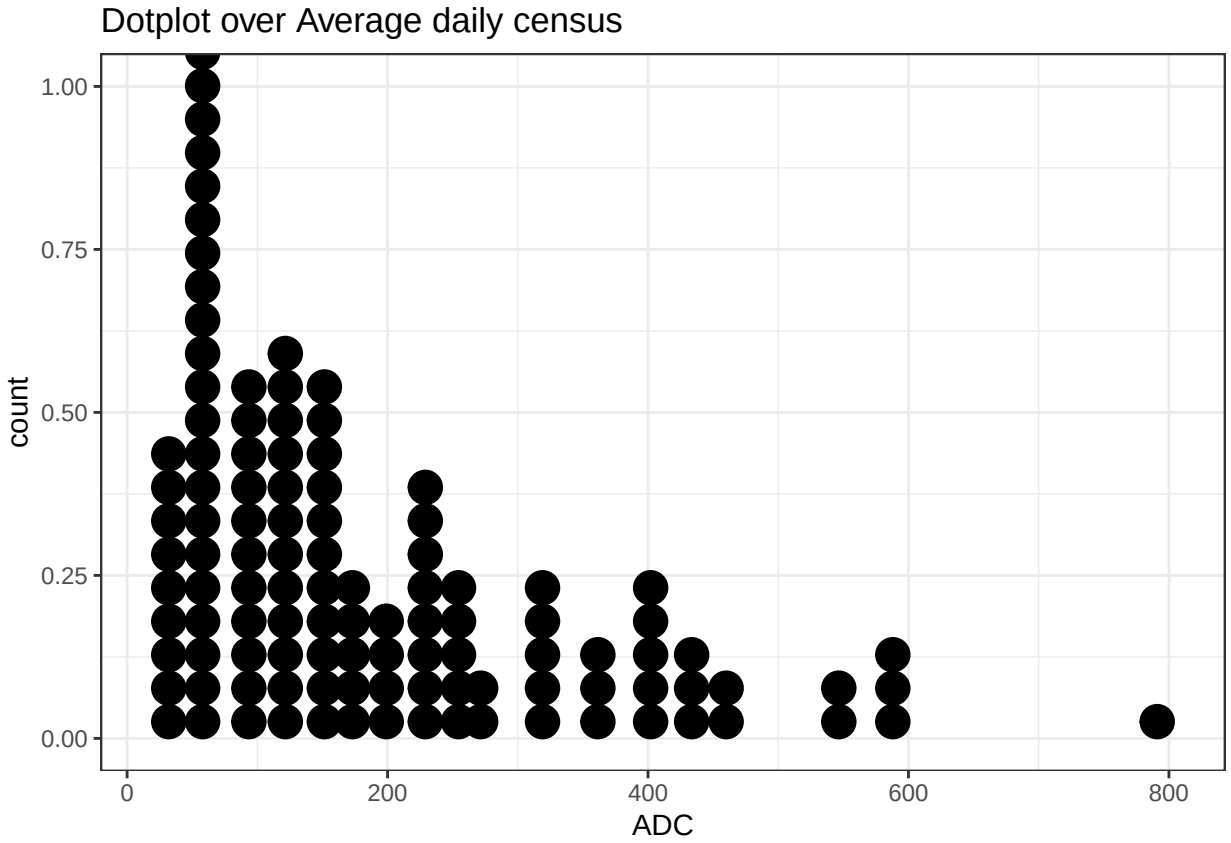
Antalet X-ray ratios koncentrerar sig mellan 70 och 100. Fördelningen är väldigt ojämn.

```
ggplot(data = DS_C1, aes(x=Nmbrofbeds))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Number of beds" )
```



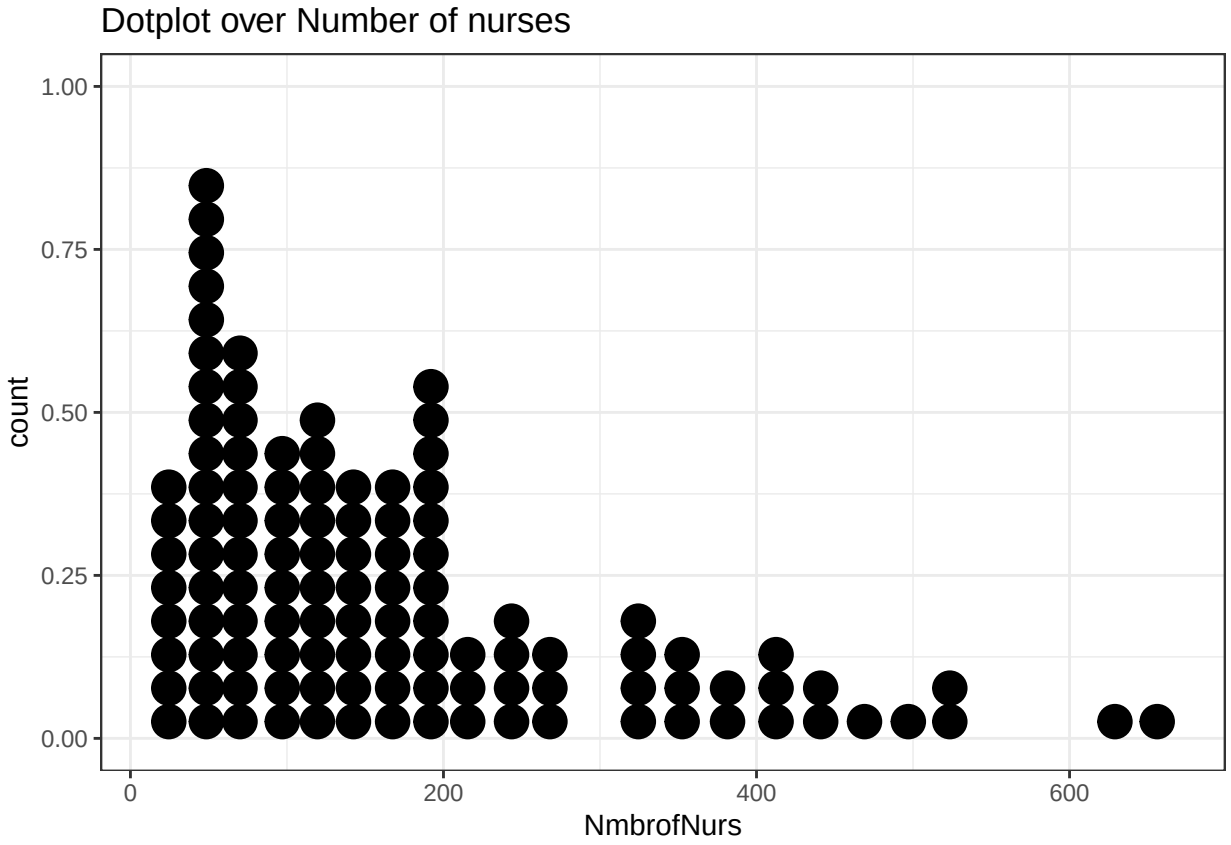
Väldigt positivt skev och majoriteten av sjukhusen har mellan 90 och 300 sängar.

```
ggplot(data = DS_C1, aes(x=ADC))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Average daily census")
```



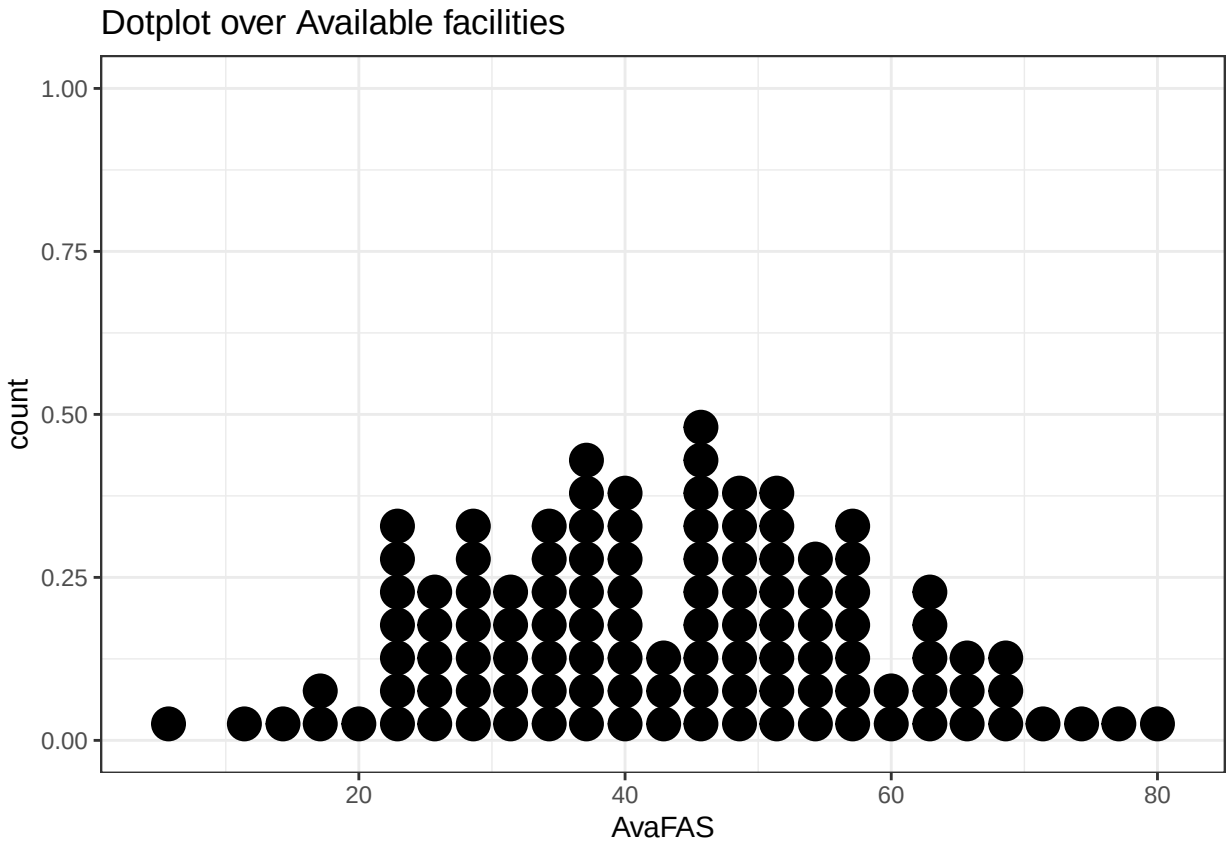
Gällande antalet människor på sjukhus är majoriteten mellan 20 till 180 dagar och med en stark topp vid 50.

```
ggplot(data = DS_C1, aes(x=NmbrofNurs))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Number of nurses" )
```



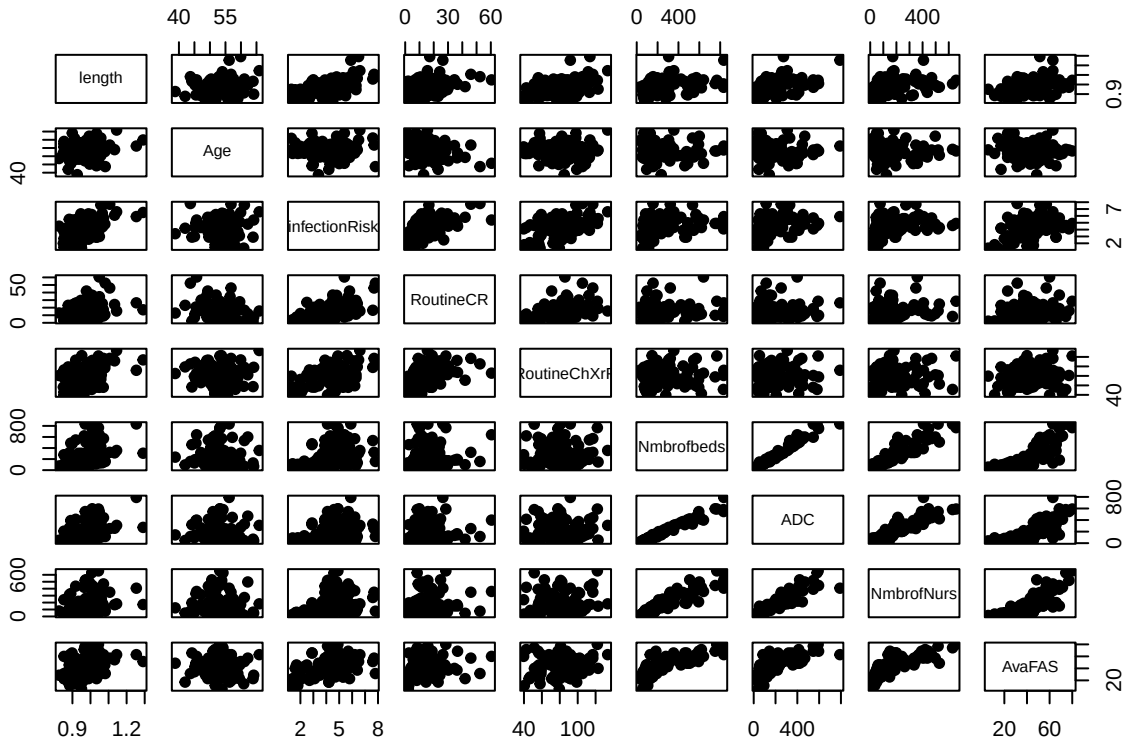
Vad det gäller antalatel sjuksköterskor är den också väldigt positiv skev med majoritet som är mellan 50-200 stycken.

```
ggplot(data = DS_C1, aes(x=AvaFAS))+
  geom_dotplot()+
  theme_bw()+
  labs(title = "Dotplot over Available facilities" )
```



Antalet tillgängliga anläggningar är en av de variablerna med bäst fördelning, variabeln följer normalfördelningskurvan bra men observationer fattas i mitten.

```r
#Scatterplot Matrix
pairs(DS_C1, pch = 19)
```



Bland de fyra nedersta variablerna uppvisas ett stundtals starkt parvis linjärt samband, bortsett från de variablerna ser fördelning sådär ut.

```r
#Correlation Matrix
cor(DS_C1[,2:8])
```

```
##                        Age infectionRisk  RoutineCR RoutineChXrR  Nmbrofbeds
## Age            1.000000000   0.001093166 -0.2258468  -0.01885490 -0.05882316
## infectionRisk  0.001093166   1.000000000  0.5591589   0.45339156  0.35977000
## RoutineCR     -0.225846789   0.559158869  1.0000000   0.42496204  0.13972495
## RoutineChXrR  -0.018854897   0.453391557  0.4249620   1.00000000  0.04581997
## Nmbrofbeds    -0.058823160   0.359770000  0.1397249   0.04581997  1.00000000
## ADC           -0.054774667   0.381411081  0.1429482   0.06291352  0.98099774
## NmbrofNurs    -0.082944616   0.393981340  0.1988998   0.07738133  0.91550415
##                        ADC  NmbrofNurs
## Age           -0.05477467 -0.08294462
## infectionRisk  0.38141108  0.39398134
## RoutineCR      0.14294821  0.19889983
## RoutineChXrR   0.06291352  0.07738133
## Nmbrofbeds     0.98099774  0.91550415
## ADC            1.00000000  0.90789698
## NmbrofNurs     0.90789698  1.00000000
```

Dom starkare korrelationerna (+-0.70+) finns mellan Average daily census och number of beds samt number of nurses och Available facilites and services då dessa variabler har ett högt samband mellan varandra. Däremot ska inte variabler såsom infection risk, och age tas bort för det.

3.1.2 Selection of regression models

```
formula <-length ~ Age + infectionRisk + RoutineCR + RoutineChXrR +
    Nmbrofbeds + ADC + NmbrofNurs + AvaFAS
model1 <- ols(formula = formula,data = DS_C1_1)

models <- regsubsets(length ~ Age + infectionRisk + RoutineCR + RoutineChXrR +
    Nmbrofbeds + ADC + NmbrofNurs + AvaFAS, data = DS_C1_1, nvmax = 9)
res.sum <- summary(models)
data.frame(
  Adj.R2 = which.max(res.sum$adjr2),
  CP = which.min(res.sum$cp),
  AIC= which.min(res.sum$bic)
)
```

```
##   Adj.R2 CP AIC
## 1      5  3   3
```

Från den ovanstående koden som returneras nämns det att i modellen som baseras på Ads.R2 kommer 5 variabler att vara med och i den som baseras på CP,criterion kommer 3 variabler att vara med. För AIC kommer också 3 variabler vara med.

```
res.sum
```

```
## Subset selection object
## Call: regsubsets.formula(length ~ Age + infectionRisk + RoutineCR +
##     RoutineChXrR + Nmbrofbeds + ADC + NmbrofNurs + AvaFAS, data = DS_C1_1,
##     nvmax = 9)
## 8 Variables  (and intercept)
##               Forced in Forced out
## Age               FALSE      FALSE
## infectionRisk     FALSE      FALSE
## RoutineCR         FALSE      FALSE
## RoutineChXrR      FALSE      FALSE
## Nmbrofbeds        FALSE      FALSE
## ADC               FALSE      FALSE
## NmbrofNurs        FALSE      FALSE
## AvaFAS            FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Age infectionRisk RoutineCR RoutineChXrR Nmbrofbeds ADC NmbrofNurs
## 1  ( 1 ) " " " "           " "       " "           " "        "*" " "
## 2  ( 1 ) " " " "           " "       "*"           " "        "*" " "
## 3  ( 1 ) "*" " "           " "       "*"           " "        "*" " "
## 4  ( 1 ) "*" " "           " "       "*"           " "        "*" "*"
## 5  ( 1 ) "*" " "           "*"       "*"           " "        "*" "*"
## 6  ( 1 ) "*" "*"           " "       "*"           "*"        "*" "*"
## 7  ( 1 ) "*" "*"           "*"       "*"           " "        "*" "*"
```

```
## 8  ( 1 ) "*" "*"           "*"       "*"          "*"        "*" "*"
##           AvaFAS
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

Vad gäller den första modellen kan variablerna som kommer att vara med hittas på femte raden för den andra och tredje som både har tre variabler återfinns variablerna på den tredje raden.

3.1.3 1. Best subset method, using the R2.adj criterion

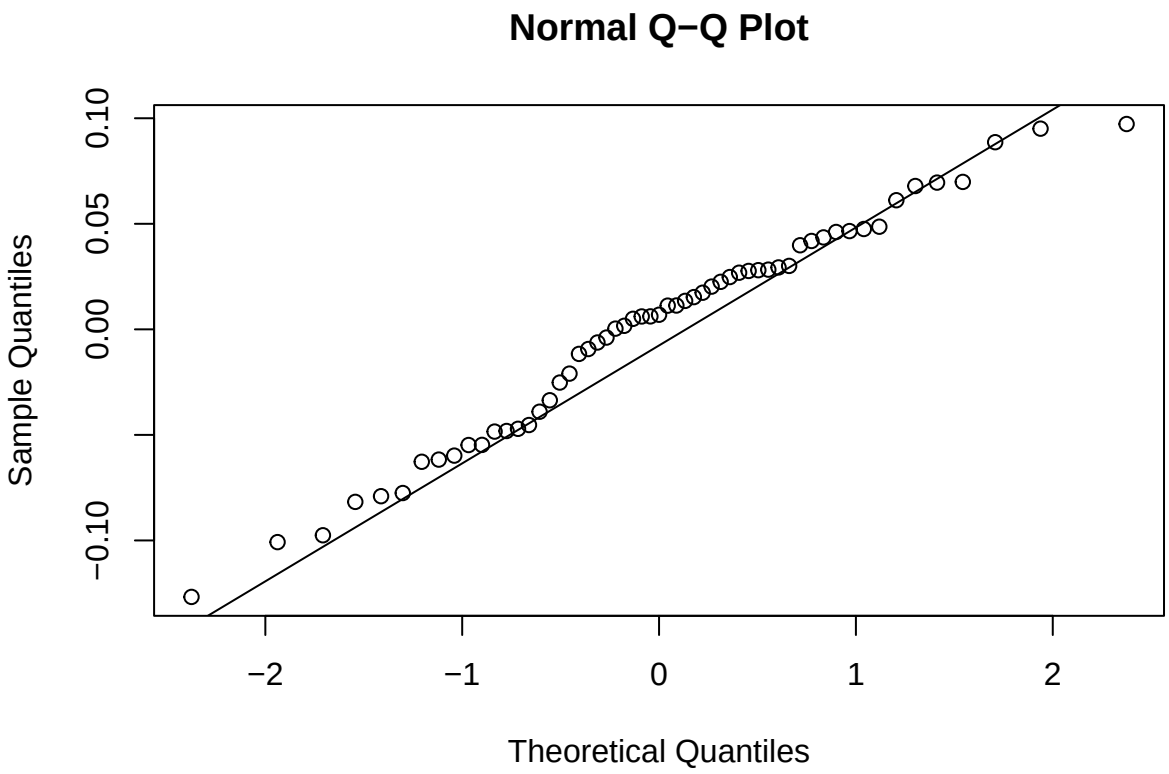
```
coef(models, 5)
```

```
##   (Intercept)           Age     RoutineCR  RoutineChXrR           ADC
##  0.6259771134  0.0036135298  0.0011788592  0.0010636217  0.0004296093
##    NmbrofNurs
## -0.0002146857
```

```
model1 <-lm(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C1_1)
```

Alltså kommer det i fallet där man jagar högst justerad förklaringsgrad vara de ovanstående fem variablerna som är med i modellen.

```
resmod1 =resid(model1)
qqnorm(resmod1)
  qqline(resmod1) #add a straight diagonal line to the plot
```



Residualerna följer linjen väldigt bra med ett par enstaka extremvärden, ett i nedre delen och ett i övre delen.

3.1.4 2. Best subset method, using the Cp criterion

```
model1_2 <- ols(formula = formula,data = DS_C1_1)
# with p-values, sls = significance level for staying in a model
model2 <- fastbw(fit=model1_2, rule="p", sls=0.05)
coef(models,2)
```

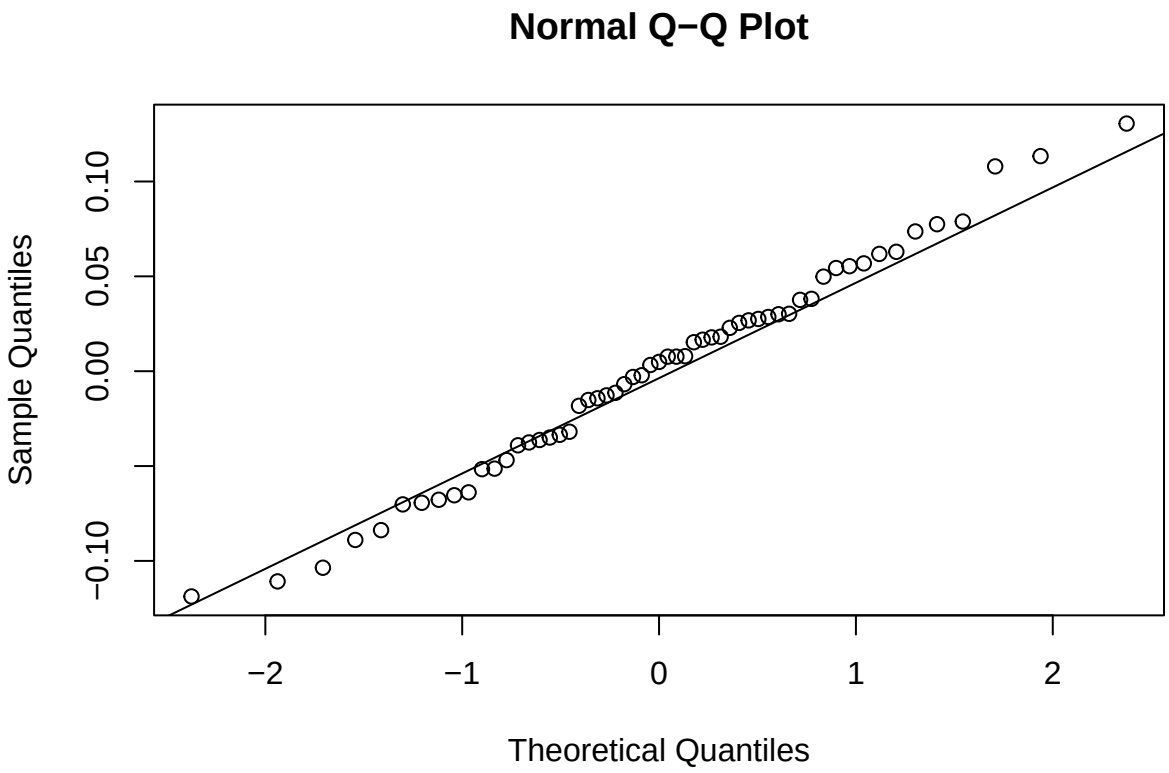
```
##  (Intercept) RoutineChXrR          ADC
## 0.8064182597 0.0013536224 0.0002720532
```

```
model2$coefficients
```

```
##    Intercept RoutineChXrR          ADC
## 0.8064182597 0.0013536224 0.0002720532
```

```
model2 <-lm(length~RoutineChXrR+ADC,data = DS_C1_1)
```

```
resmod2 =resid(model2)
qqnorm(resmod2)
  qqline(resmod2) #add a straight diagonal line to the plot
```



Residualerna följer linjen fint och uppvisar inte några extremvärden.

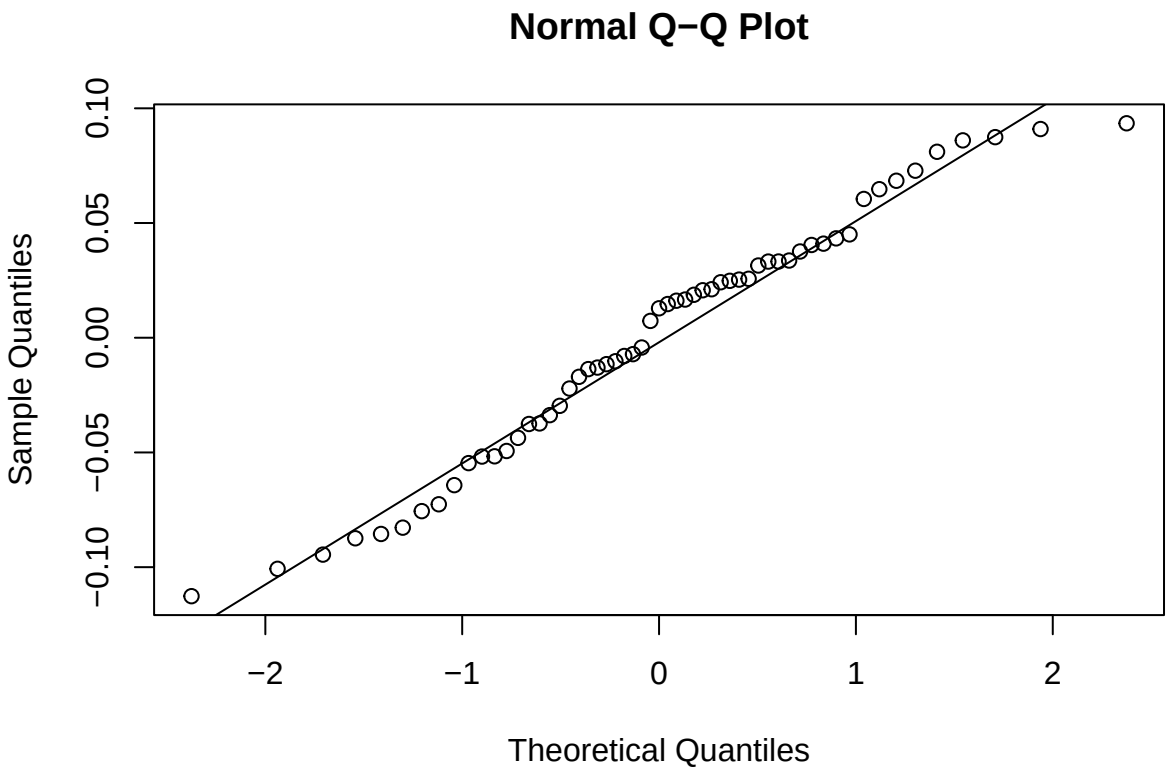
3.1.5 3. Backward elimination using the AIC

```
model1_0 <-ols(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C1_1)
# With aic
model3 <- fastbw(fit=model1_0, rule="aic")
model3$coefficients
```

```
##    Intercept           Age RoutineChXrR           ADC
## 0.6104338715 0.0038800967 0.0011747874 0.0002926124
```

```
model3 <-lm(length~Age+RoutineChXrR+ADC,data =DS_C1_1)
```

```
resmod3 =resid(model3)
qqnorm(resmod3)
qqline(resmod3) #add a straight diagonal line to the plot
```



Residualerna följer linjen fint och uppvisar enbart ett extremvärde mot slutet.

3.2 Uppgift 2: External Validation

3.2.1 Cross-validation

```
PRESS1 <- function(model) {
    i <- residuals(model)/(1 - lm.influence(model)$hat)
    sum(i^2)
}
cbind(PRESS1(model1),PRESS1(model2),PRESS1(model3))
```

```
##           [,1]      [,2]     [,3]
## [1,] 0.1895577 0.2070168 0.193714
```

I detta fall har vi jämfört de tre modellernas olika Press-statistiska och den modell som fått lägst värde är Modell1 som för övrigt också har flest variabler med.

3.2.2 External validation set

3.2.3 1. Do predictions on the validation set

```
pred_mod1 <-predict.lm(model1,newdata = DS_C2)
pred_mod2 <-predict.lm(model2,newdata = DS_C2)
pred_mod3 <-predict.lm(model3,newdata = DS_C2)
```

```
MSPR <- function(model) {
    (1/length(model))*sum(model-DS_C2$length)^2
}
cbind(MSPR(pred_mod1),MSPR(pred_mod2),MSPR(pred_mod3))
```

3.2.3.1 2. Calculate the mean squared prediction error, MSPR

```
##           [,1]       [,2]       [,3]
## [1,] 0.0161745 0.01547956 0.01586965
```

Jämförelsen är återigen väldigt lik jämförelsen vad gällde Press-statistikorna, samtliga tre modeller har väldigt litet MSPR, däremot skiljer sig den första modellen i form av ett extremt litet MSPR.

3.3 3. Select the set of explanatory variables that had the lowset MSPR in step 2). You will use those in the following exerices:

```
model_another <-lm(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C2)
smry1 <- summary(model_another)
smry2 <- summary(model1)
cbind(smry2$coefficients[ ,1:2],smry1$coefficients[,1:2])
```

```
##                   Estimate   Std. Error      Estimate   Std. Error
## (Intercept)   0.6259771134 0.0905330996  0.5529813398 0.1203562383
## Age           0.0036135298 0.0016364085  0.0053805577 0.0020809107
## RoutineCR     0.0011788592 0.0009219591  0.0023585415 0.0009094498
## RoutineChXrR  0.0010636217 0.0004209460  0.0009373458 0.0004805877
## ADC           0.0004296093 0.0001034021  0.0004339260 0.0001700745
## NmbrofNurs   -0.0002146857 0.0001370050 -0.0002978266 0.0001616502
```

I de två första kolumnerna syns koefficenterna och dess standardfel för första modellen och i de två senare kolumnerna syns samma värden för den senare modellen. Det som tydligt visas är hur koefficenterna i majoriteten av fallen får ett mer precis värde som både är högre och lägre men främst har lägre standardfel.

```
#MSE
mse <-function(model){
  mean(model$residuals)^2
}
cbind(mse(model_another),mse(model1))
```

```
##              [,1]         [,2]
## [1,] 2.258012e-37 6.770847e-36
```

Här ser man att den nya modellen har ett lägre värde och att den tidigare modellen har ett tydligt högre värde på MSE.

```
#R2
cbind(Rsq(model_another),Rsq(model1))
```

```
##           [,1]      [,2]
## [1,] 0.4017075 0.5512792
```

Den första modellen (model1) har i detta fall 16 procentenheter högre förklaringsgrad vilket kan förklaras med att modellen lämpar sig lämpar sig bättre för de 57 senare observationerna.

3.3.1 Take all data and create a new validation

```
set.seed(500)
DS_C3 <- DS_C1[sample(nrow(DS_C1), 56), ]
model_4 <-lm(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C3)
```

```
set.seed(300)
DS_C4 <- DS_C1[sample(nrow(DS_C1), 56), ]
model_5 <-lm(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C4)
```

```
set.seed(200)
DS_C5 <- DS_C1[sample(nrow(DS_C1), 56), ]
model_6 <-lm(length~Age+RoutineCR+RoutineChXrR+ADC+NmbrofNurs,data = DS_C5)
```

```
pred_mod4 <-predict.lm(model_4,newdata = DS_C3)
pred_mod5 <-predict.lm(model_5,newdata = DS_C4)
pred_mod6 <-predict.lm(model_6,newdata = DS_C5)
cbind(MSPR(pred_mod4),MSPR(pred_mod5),MSPR(pred_mod6))
```

```
##             [,1]       [,2]       [,3]
## [1,] 0.02330269 0.01507676 0.01984795
```

Samtliga tre modeller visar på en väldigt låg MSPR, första värdet visar på högst MSPR bland de tre nya modellerna och andra värdet är tydligt det lägsta.

4 Lärdomar

Målen med laborationen var att förstå och använda sig av olika kriterier samt krav under valet av modeller. Använda automatiska sökfunktioner för modelval och validera de utvalda modellerna.

Efter laborationen upplever vi förbättrade kunskaper kring ovanstående ämne, än en gång förstår vi även vikten att läsa laborationen noggrant så det är tydligt vad som ska göras.