Lab 10- Zadanie domowe

Mateusz Kapusta

2022-05-24

Zadanie nr 2

```
model<-glm(class~.,data=biopsy, family="binomial")
t<-stepAIC(model,trace=TRUE,direction="both",scope = list(upper = ~., lower = ~1))
```

```
## Start:  AIC=122.89
## class ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9
##
##          Df Deviance    AIC
## - V2      1   102.89 120.89
## - V5      1   103.27 121.27
## - V3      1   104.74 122.74
## <none>        102.89 122.89
## - V9      1   106.61 124.61
## - V8      1   106.66 124.66
## - V4      1   110.31 128.31
## - V7      1   110.33 128.33
## - V1      1   120.72 138.72
## - V6      1   122.07 140.07
##
## Step:  AIC=120.89
## class ~ V1 + V3 + V4 + V5 + V6 + V7 + V8 + V9
##
##          Df Deviance    AIC
## - V5      1   103.27 119.27
## <none>        102.89 120.89
## - V9      1   106.66 122.66
## - V3      1   106.66 122.66
## - V8      1   106.76 122.76
## + V2      1   102.89 122.89
## - V4      1   110.64 126.64
## - V7      1   110.70 126.70
## - V1      1   121.10 137.10
## - V6      1   122.07 138.07
##
## Step:  AIC=119.27
## class ~ V1 + V3 + V4 + V6 + V7 + V8 + V9
##
##          Df Deviance    AIC
## <none>        103.27 119.27
## + V5      1   102.89 120.89
## - V9      1   107.14 121.14
```

```
## + V2     1   103.27 121.27
## - V8     1   107.72 121.72
## - V3     1   107.90 121.90
## - V7     1   111.69 125.69
## - V4     1   112.17 126.17
## - V1     1   121.55 135.55
## - V6     1   123.15 137.15
```

```
tp<-stepAIC(model,trace=TRUE,direction="both",scope = list(upper = ~., lower = ~1),k=log(nrow(biopsy)))
```

```
## Start:  AIC=168.15
## class ~ V1 + V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9
##
##        Df Deviance    AIC
## - V2    1   102.89 161.63
## - V5    1   103.27 162.00
## - V3    1   104.74 163.48
## - V9    1   106.61 165.35
## - V8    1   106.66 165.40
## <none>      102.89 168.15
## - V4    1   110.31 169.05
## - V7    1   110.33 169.07
## - V1    1   120.72 179.46
## - V6    1   122.07 180.81
##
## Step:  AIC=161.63
## class ~ V1 + V3 + V4 + V5 + V6 + V7 + V8 + V9
##
##        Df Deviance    AIC
## - V5    1   103.27 155.48
## - V9    1   106.66 158.87
## - V3    1   106.66 158.88
## - V8    1   106.76 158.97
## <none>      102.89 161.63
## - V4    1   110.64 162.86
## - V7    1   110.70 162.91
## + V2    1   102.89 168.15
## - V1    1   121.10 173.31
## - V6    1   122.07 174.28
##
## Step:  AIC=155.48
## class ~ V1 + V3 + V4 + V6 + V7 + V8 + V9
##
##        Df Deviance    AIC
## - V9    1   107.14 152.83
## - V8    1   107.72 153.41
## - V3    1   107.90 153.59
## <none>      103.27 155.48
## - V7    1   111.69 157.37
## - V4    1   112.17 157.86
## + V5    1   102.89 161.63
## + V2    1   103.27 162.00
## - V1    1   121.55 167.24
## - V6    1   123.15 168.84
##
```

```
## Step:  AIC=152.83
## class ~ V1 + V3 + V4 + V6 + V7 + V8
##
##        Df Deviance    AIC
## - V3    1   112.26 151.42
## - V8    1   112.57 151.72
## <none>      107.14 152.83
## - V7    1   116.29 155.45
## + V9    1   103.27 155.48
## - V4    1   116.57 155.73
## + V5    1   106.66 158.87
## + V2    1   107.02 159.23
## - V6    1   126.32 165.48
## - V1    1   134.01 173.17
##
## Step:  AIC=151.42
## class ~ V1 + V4 + V6 + V7 + V8
##
##        Df Deviance    AIC
## <none>      112.26 151.42
## + V3    1   107.14 152.83
## + V9    1   107.90 153.59
## + V2    1   108.97 154.65
## + V5    1   110.79 156.48
## - V8    1   125.78 158.41
## - V4    1   125.96 158.59
## - V7    1   128.36 160.99
## - V6    1   145.91 178.54
## - V1    1   162.83 195.47
```

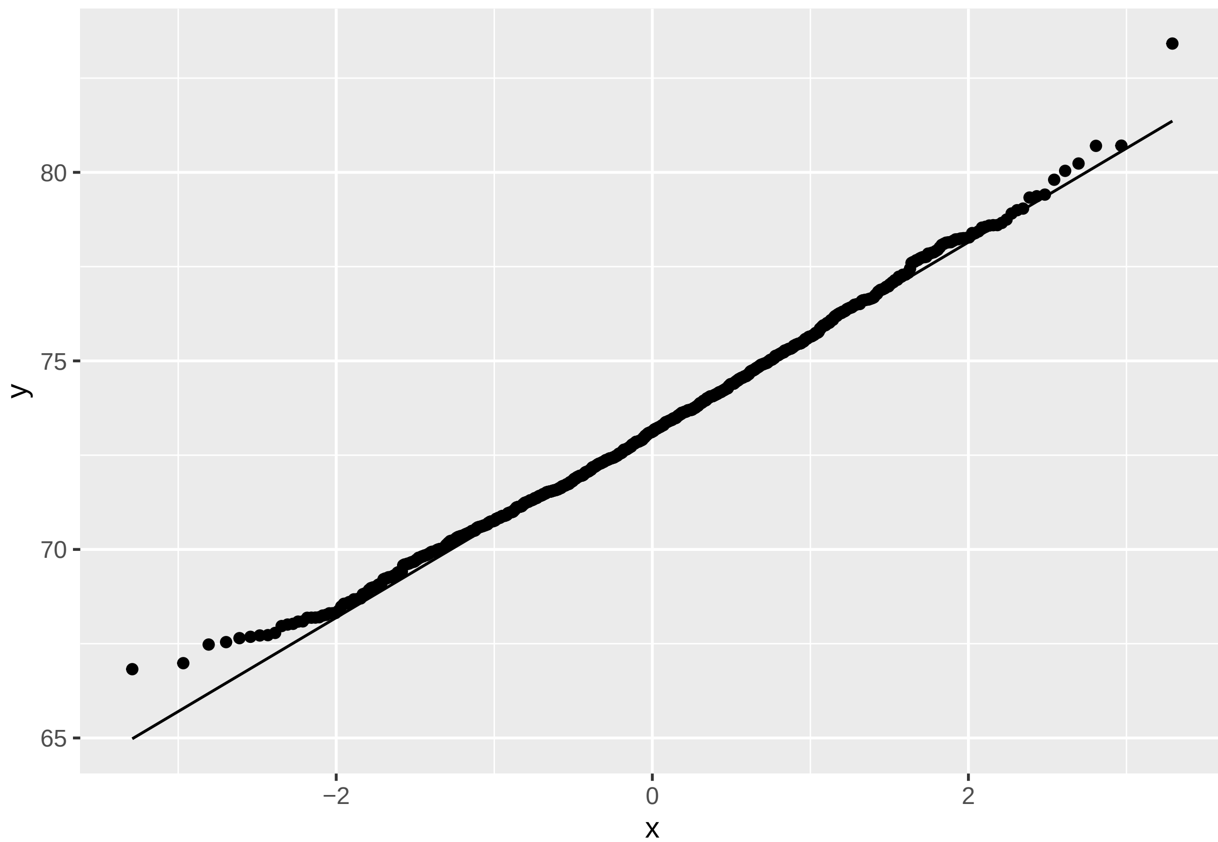
```
k=15
name<-paste0(names(tp$coefficients)[2:length(tp$coefficients)],sep="+",collapse = "")
name<-substr(name,1,nchar(name)-1)
train_control <- trainControl(method='cv', number=k)
tppred <- train(as.formula(paste("class~",name)), data=biopsy,
                     method='glm', family=binomial, trControl=train_control)
wyntp<-tppred$results["Accuracy"]
name<-paste0(names(t$coefficients)[2:length(t$coefficients)],sep="+",collapse = "")
name<-substr(name,1,nchar(name)-1)
train_control <- trainControl(method='cv', number=k)
tpred <- train(as.formula(paste("class~",name)), data=biopsy,
                     method='glm', family=binomial, trControl=train_control)
wynt<-tpred$results["Accuracy"]
```

Średnia dokładnośc modelu wytrenowanego sugerując się maksymalizacją AIC to 0.967568438003221 natomiast BIC 0.969306104523496 czyli oba modele różnią się tylko minimalnie.

Zadanie nr 3

```
fun<-function(dane,indeksy)
{
  sum(sapply(indeksy, \(x) dane[x,]/length(indeksy)))
}
```

```r
wyn<-numeric(1000)
d<-read.csv("walter.csv")
for (i in 1:1000)
{
  x<-sample(1:nrow(d),replace=TRUE)
  wyn[i]<-fun(d,x)
}
ggplot(data=data.frame("mean"=wyn))+stat_qq(aes(sample=mean))+stat_qq_line(aes(sample=mean))
```

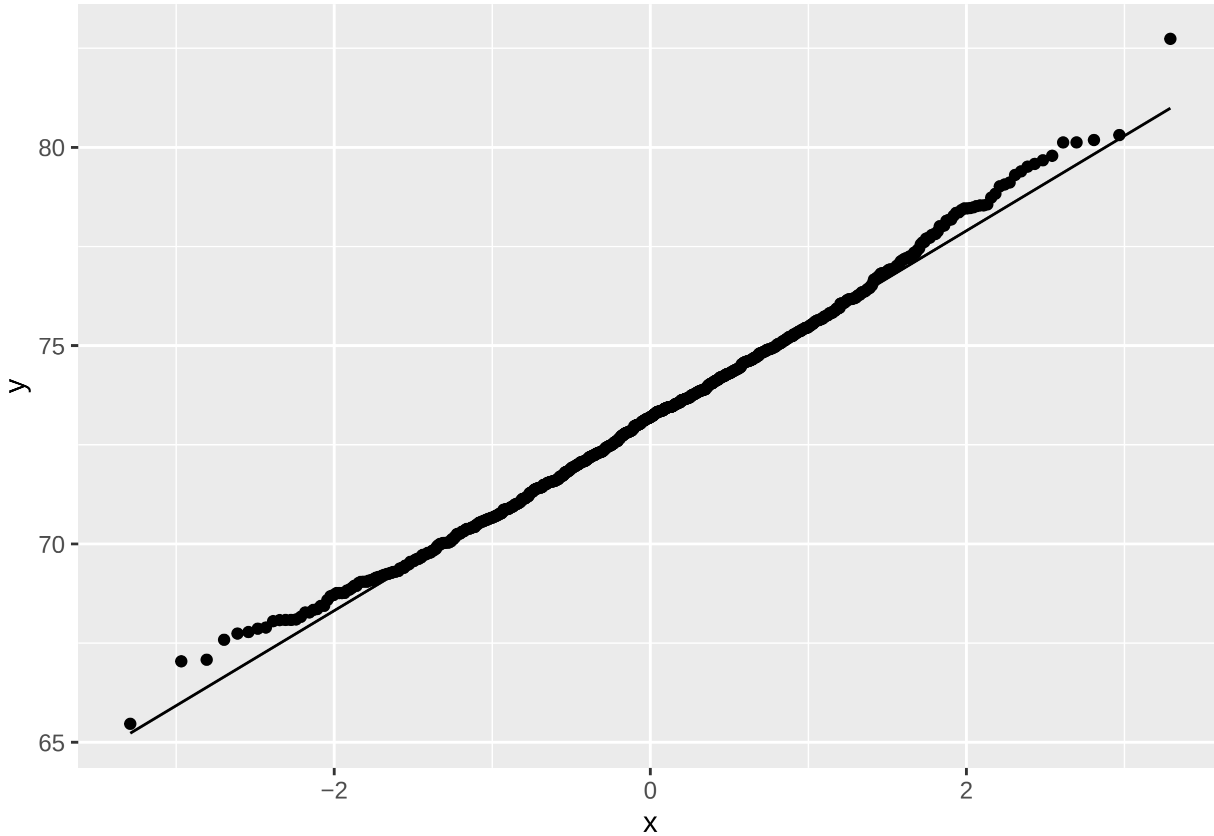


Widzimy, że rozkład średniej jest normalny albowiem dane dobrze leżą na lini prostej na wykresie kwantylowym. Teraz porównujemy to z funkcją boot.

```r
wynp<-boot(d,fun,1000)
wynp
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d, statistic = fun, R = 1000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 73.24421 -0.08187895     2.42163
```

```
ggplot(data=data.frame("mean"=wynp$t))+stat_qq(aes(sample=mean))+stat_qq_line(aes(sample=mean))
```



Tak jak w przypadku ręcznego bootstrapu widizmy, że dane pochodza z grubsza z rozkładu normalnego i przypominają te wykonane ręcznie.