Exercise 5 solutions - ST2304

Christoffer Høyvik Hilde

4/24/2018

Problem 1

```
rawdata<-read.csv("https://www.math.ntnu.no/emner/ST2304/2018v/LifeExpectancy.csv")
NoSA <- rawdata[rawdata$Country!= "South Africa",] ## Remove South Africa from the dataset

mod0 <- lm(Life.exectancy~1, data=NoSA)
mod1 <- update(mod0, .~ . + Health.Spending.per.capita)
mod2 <- update(mod1, .~ . + I(Health.Spending.per.capita^2))
mod3 <- update(mod2, .~ . + I(Health.Spending.per.capita^3))
mod4 <- update(mod3, .~ . + I(Health.Spending.per.capita^4))
mod5 <- update(mod4, .~ . + I(Health.Spending.per.capita^5))
mod6 <- update(mod5, .~ . + I(Health.Spending.per.capita^6))
mod7 <- update(mod6, .~ . + I(Health.Spending.per.capita^7))
mod8 <- update(mod7, .~ . + I(Health.Spending.per.capita^8))
mod9 <- update(mod8, .~ . + I(Health.Spending.per.capita^9))
mod10 <- update(mod9, .~ . + I(Health.Spending.per.capita^10))

AIC(mod0,mod1,mod2,mod3,mod4,mod5,mod6,mod7,mod8,mod9,mod10)
```

```
##       df      AIC
## mod0   2 236.9086
## mod1   3 215.9586
## mod2   4 187.3340
## mod3   5 179.6972
## mod4   6 180.6243
## mod5   7 181.0067
## mod6   8 180.9642
## mod7   9 179.1760
## mod8  10 178.6553
## mod9  11 179.7920
## mod10 12 181.7918
```

Model 8 has the lowst AIC, but keep in mind that when you have several models within 2 AIC values of the best model, these models are equivalent. However, often when we have several equivalent models within 2 AIC of each other we prefer the simplest of them (Principle of parsimony), in this case that would be model 3.

```
BIC(mod0,mod1,mod2,mod3,mod4,mod5,mod6,mod7,mod8,mod9,mod10)
```

```
##       df      BIC
## mod0   2 240.3840
## mod1   3 221.1716
## mod2   4 194.2847
## mod3   5 188.3855
## mod4   6 191.0503
## mod5   7 193.1703
## mod6   8 194.8656
## mod7   9 194.8150
```

```
## mod8   10 196.0320
## mod9   11 198.9063
## mod10 12 202.6439
```

Using BIC the best model is model 3. BIC (+ log(n)p) will add more per parameter than AIC (+ 2p), and will in most cases come prefer a simpler model. In this case AIC and BIC seem to agree.

Problem 2

This is the code used for simulating the data:

```r
N <- 50
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.4.3
```

```r
muX <- c(0,0) # mean of bivariate distribution
Corr <- 0.5 # correlation
sigmaX <- matrix(c(1,Corr,Corr,1), nrow=2) # covariance matrix
x <- mvrnorm(N, muX, Sigma=sigmaX) # 2 columns: x[,1] & x[,2]
# Simulate from a different model
N <- 50; alpha <- 0; sigma <- 1 # same throughout
beta1 <- -20
beta2 <- 5000
mu <- alpha + beta1*x[,1] + beta2*x[,2]
y <- rnorm(N, mu, sigma)
mod <- lm(y ~ x[,1] + x[,2])
```

1. No correlation, $\beta_1 = 1$ and $\beta_2 = 1$

```r
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 1]     1 34.240  34.240  35.271 3.329e-07 ***
## x[, 2]     1 67.473  67.473  69.506 8.037e-11 ***
## Residuals 47 45.626   0.971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 2]     1 36.280  36.280  37.373 1.813e-07 ***
## x[, 1]     1 65.433  65.433  67.404 1.240e-10 ***
## Residuals 47 45.626   0.971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The order of the variables in the model doesnt affect which model is selected as the best.

2. Correlation = 0.7, $\beta_1 = 1$ and $\beta_2 = 0$

```r
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 1]     1 34.461  34.461 38.8810 1.184e-07 ***
## x[, 2]     1  0.318   0.318  0.3591    0.5519
## Residuals 47 41.657   0.886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 2]     1 15.648 15.6476  17.655 0.0001174 ***
## x[, 1]     1 19.131 19.1314  21.585 2.754e-05 ***
## Residuals 47 41.657  0.8863
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we see that the order of the variables do affect which model is selected. Since $\beta_2 = 0$, the correct model here is $y = \beta_1 x_1$. The anova of mod1 (lm(y ~ x[,1] + x[,2])) shows that adding x_2 to the model y~x_1 doesnt make the model better, thus it correctly show that the best model is $y = \beta_1 x_1$. The anova of mod2 (lm(y ~ x[,2] + x[,1])) shows that adding x_1 to the model y ~ x_1 makes the model better, thus the best model would be $y = \beta_1 x_1 + \beta_2 x_2$. The reason for this is the correlation between x_1 and x_2. The anova funtion tests the models against each other in the order specified, so in the case of mod2 it will first test whether $y = \beta_2 x_2$ is better than $y = 1$, which it is because x_2 alone will have an indirect effect on y, through its correlation with x_1. Then it tests whether $y = \beta_1 x_1 + \beta_2 x_2$ is better than $y = \beta_2 x_2$, which it also is because x_1 has a direct effect on y.

3. Correlation = 0.7, $\beta_1 = 1$ and $\beta_2 = 1$

```r
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x[, 1]     1 135.654 135.654 108.327 8.671e-14 ***
## x[, 2]     1  31.800  31.800  25.394 7.366e-06 ***
## Residuals 47  58.857   1.252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x[, 2]     1 137.006 137.006 109.407 7.358e-14 ***
## x[, 1]     1  30.448  30.448  24.314 1.062e-05 ***
## Residuals 47  58.857   1.252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The order of the variables in the model doesnt affect which model is selected as the best.

4. Correlation = -0.8, $\beta_1 = 5$ and $\beta_2 = 5$

```r
anova(mod1)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 1]     1  51.95   51.95  50.083 6.248e-09 ***
## x[, 2]     1 522.03  522.03 503.315 < 2.2e-16 ***
## Residuals 47  48.75    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x[, 2]     1  48.03   48.03   46.31 1.615e-08 ***
## x[, 1]     1 525.95  525.95  507.09 < 2.2e-16 ***
## Residuals 47  48.75    1.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The order of the variables in the model doesnt affect which model is selected as the best.