Labb 11

Problem 20

1a) It seems reasonable that the income will decrease with illiteracy and that the amount of murders will increase with illiteracy. Also, it very unlikely that this kind of relationship would be anything else but linear. Otherwise it would mean that the amount of murders/income will increase/decrease very fast with illiteracy. Below we see the two plots with regression lines. The confidence intervals show that we can reject the null hypothesis (on 5% significane level) that the slope is 0, i.e there is no linear relationship between murder/income and illiteracy.

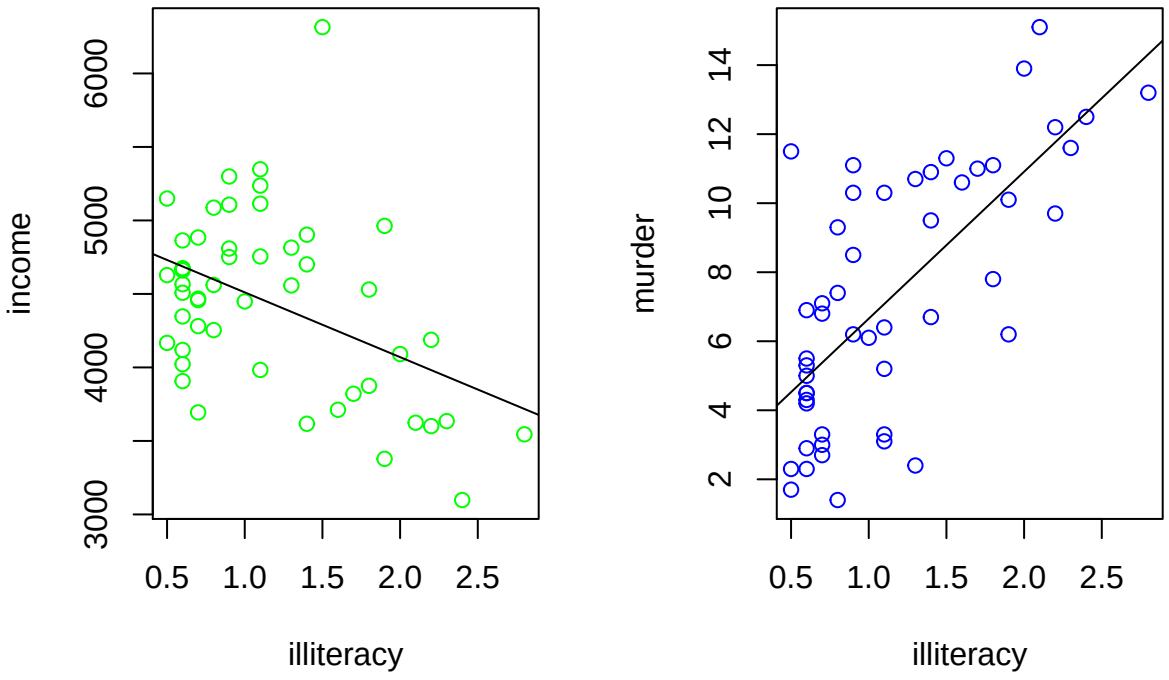
Furthermore, with these regression lines we can do prediction. I.e, if we get a new illiteracy we can construct a 95% confidence interval for the amount of murders and income.

```
states <- as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy", "Income", "Frost")])

illiteracy = states$Illiteracy
murder = states$Murder
income = states$Income


par(mfrow = c(1,2))
plot(illiteracy,income,col="green")
ill_inc.lm = lm(income ~ illiteracy )
abline(ill_inc.lm)


plot(illiteracy,murder,col="blue")
ill_murdl.lm = lm(murder ~ illiteracy)
abline(ill_murdl.lm)
```



```r
confint(ill_inc.lm,level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 4604.9180 5297.7216
## illiteracy  -703.7516 -177.4789
```

```r
print("****")
```

```
## [1] "****"
```

```r
confint(ill_murdl.lm,level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.7511819 4.042369
## illiteracy  3.0074184 5.507495
```

1b)

```r
states <- as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy", "Income", "Frost")])

illiteracy = states$Illiteracy
murder = states$Murder
income = states$Income
frost = states$Frost
population = states$Population

lm = lm(murder ~ illiteracy+income+frost+population)
lm
```

```
##
## Call:
## lm(formula = murder ~ illiteracy + income + frost + population)
##
## Coefficients:
## (Intercept)   illiteracy       income        frost   population
##   1.235e+00    4.143e+00    6.442e-05    5.813e-04    2.237e-04
```

```r
confint(lm)
```

```
##                     2.5 %       97.5 %
## (Intercept) -6.552191e+00 9.0213182149
## illiteracy   2.381799e+00 5.9038743192
## income      -1.312611e-03 0.0014414600
## frost       -1.966781e-02 0.0208304170
## population   4.136397e-05 0.0004059867
```

```r
X = matrix(c(rep(1,length(murder)),population,illiteracy,income,frost),nrow=length(murder))

y = murder
est = solve(t(X)%*%X)%*%(t(X)%*%y)
est
```

```
##              [,1]
## [1,] 1.2345634112
## [2,] 0.0002236754
## [3,] 4.1428365903
## [4,] 0.0000644247
## [5,] 0.0005813055
```

The coefficients should be interpreted as how much the murder variable increases when keeping everything fixed and increasing a predictor variable by 1 unit. It seems as if illiteracy has the biggest impact on murders and income the smallest. The confidence interval shows that we can reject the null hypothesis that (on 5% significance level) that there is no linear relationship between murder and illiteracy. We can also note that we can not reject the null hypothesis that there is no relationship between murder and income. So it could be the case that the income does not affect murder rate.

We see that the coefficients are almost the same as the ones obtained using R's lm-function.

Problem 21

```
library(MASS)
states <- as.data.frame(state.x77[, c("Murder", "Population", "Illiteracy", "Income", "Frost")])

illiteracy = states$Illiteracy
murder = states$Murder
income = states$Income
frost = states$Frost
population = states$Population
n = length(murder)

lm4 = lm(murder ~  population+illiteracy+income+frost)

## AIC with 4 predicots
X = matrix(c(rep(1,n),population,illiteracy,income,frost),nrow=length(murder))

y = murder
beta = solve(t(X)%*%X)%*%(t(X)%*%y)
sigma2 = 1/n*t(y-X%*%beta)%*%(y-X%*%beta)

AIC4 = 2*n*log(sqrt(2*pi))+2*n*log(sqrt(sigma2))+1/sigma2*t(y-X%*%beta)%*%(y-X%*%beta)+2*(5+1)
AIC4
```

```
##          [,1]
## [1,] 241.6429
```

```
AIC(lm4)
```

```
## [1] 241.6429
```

```
## AIC with 2 predictors

X = matrix(c(rep(1,length(murder)),population,illiteracy),nrow=length(murder))
lm2 = lm(murder ~  population+illiteracy)

y = murder
beta = solve(t(X)%*%X)%*%(t(X)%*%y)
sigma2 = 1/n*t(y-X%*%beta)%*%(y-X%*%beta)

AIC2 = 2*n*log(sqrt(2*pi))+2*n*log(sqrt(sigma2))+1/sigma2*t(y-X%*%beta)%*%(y-X%*%beta)+2*(3+1)
AIC2
```

```
##          [,1]
## [1,] 237.6565
```

```
AIC(lm2)
```

```
## [1] 237.6565
```

```
###

k = stepAIC(lm4,direction="backward")
```

```
## Start:  AIC=97.75
## murder ~ population + illiteracy + income + frost
##
##              Df Sum of Sq    RSS     AIC
## - frost       1     0.021 289.19  95.753
## - income      1     0.057 289.22  95.759
## <none>                    289.17  97.749
## - population  1    39.238 328.41 102.111
## - illiteracy  1   144.264 433.43 115.986
##
## Step:  AIC=95.75
## murder ~ population + illiteracy + income
##
##              Df Sum of Sq    RSS     AIC
## - income      1     0.057 289.25  93.763
## <none>                    289.19  95.753
## - population  1    43.658 332.85 100.783
## - illiteracy  1   236.196 525.38 123.605
##
## Step:  AIC=93.76
## murder ~ population + illiteracy
##
##              Df Sum of Sq    RSS     AIC
## <none>                    289.25  93.763
## - population  1    48.517 337.76  99.516
## - illiteracy  1   299.646 588.89 127.311
```

We see that the AIC is lower when using only population and illiteracy as predictors compared to the AIC when using population,illiteracy,income and frost as predictors. This means that we should go for the model that uses only two predictors since it fits our data better.

The stepAIC function starts with all variables as predictors and then removes variables and finds the model that has the the smallest AIC. This turns out to be the one when we have only population and illiteracy as predictors.