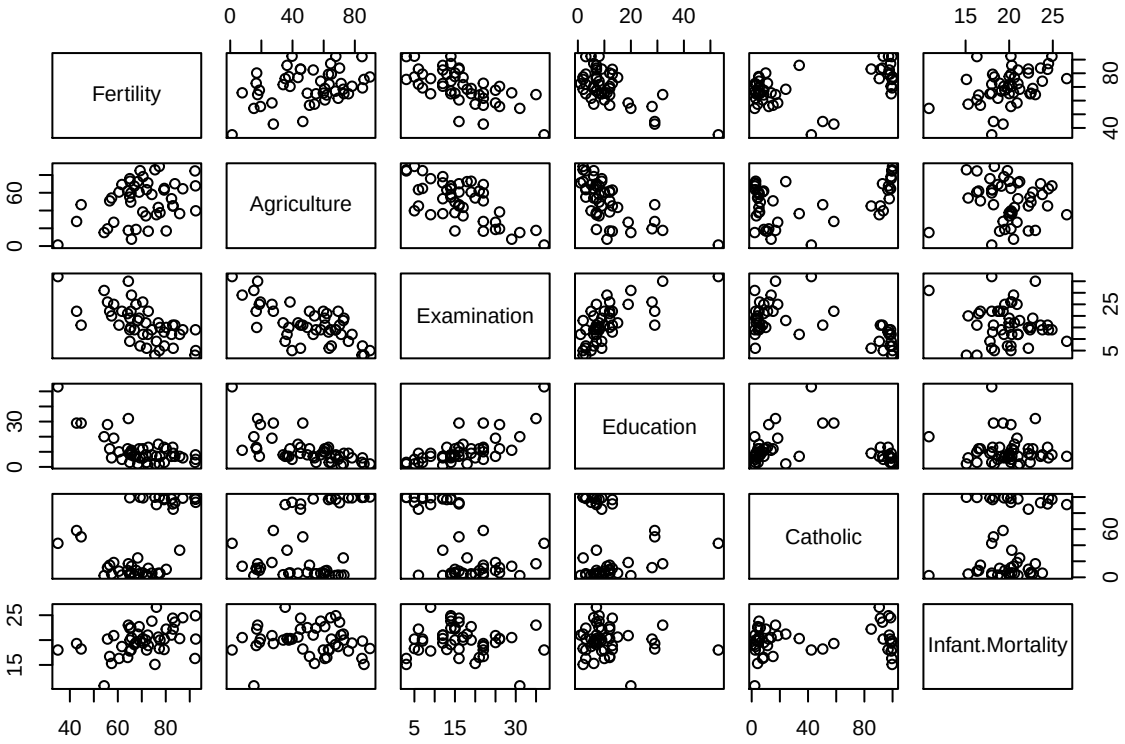
ex6

Exercise 6 (Page 37)

Look into the data set swiss of the Swiss socioeconomic survey from 1888. Determine the best model for describing education when ignoring the examination results. Use goodnees of fit, F -tests, AIC or BIC for argumentation.

http://jadianes.me/best-subset-model-selection-with-R

```
library(knitr)
library(datasets)
data(swiss)
pairs(swiss)
```



```
cor(swiss)
```

```
##                  Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility              0.41655603
## Agriculture           -0.06085861
## Examination           -0.11402160
## Education             -0.09932185
## Catholic               0.17549591
## Infant.Mortality       1.00000000
```

```
help("swiss")
```

Examination is the only category which is is positively correlated with Education. To all others it has a negativ correlation. According to the plot and the numbers Education seems strongly correlated with Fertility and Agriculture, and minor to Catholic. We will also use the library leaps with the best.subset function to confirm this.

```
library(leaps)
best.subset <- regsubsets(Education~., swiss, nvmax=5)
best.subset.summary <- summary(best.subset)
best.subset.summary$outmat
```

```
##          Fertility Agriculture Examination Catholic Infant.Mortality
## 1  ( 1 ) " "       " "         "*"         " "      " "
## 2  ( 1 ) "*"       "*"         " "         " "      " "
## 3  ( 1 ) "*"       "*"         " "         "*"      " "
## 4  ( 1 ) "*"       "*"         "*"         "*"      " "
## 5  ( 1 ) "*"       "*"         "*"         "*"      "*"
```

Full Model

Nevertheless we fit a full model that contains all categories exept Examination. We do that so we can later compare it to our reduced model.

```
fullfull_modell=lm(swiss[,4]~swiss[,1]+swiss[,2]+swiss[,5]+swiss[,6])
summary(fullfull_modell)
```

```
##
## Call:
## lm(formula = swiss[, 4] ~ swiss[, 1] + swiss[, 2] + swiss[, 5] +
##     swiss[, 6])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4029  -2.7803  -0.7571   2.4934  12.8590
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 49.99303    6.18641   8.081 4.31e-10 ***
## swiss[, 1]  -0.52070    0.07869  -6.617 5.14e-08 ***
## swiss[, 2]  -0.22880    0.03906  -5.857 6.37e-07 ***
## swiss[, 5]   0.08333    0.02179   3.825 0.000428 ***
## swiss[, 6]   0.28437    0.30040   0.947 0.349243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.224 on 42 degrees of freedom
## Multiple R-squared:  0.7305, Adjusted R-squared:  0.7048
## F-statistic: 28.46 on 4 and 42 DF,  p-value: 1.804e-11
```

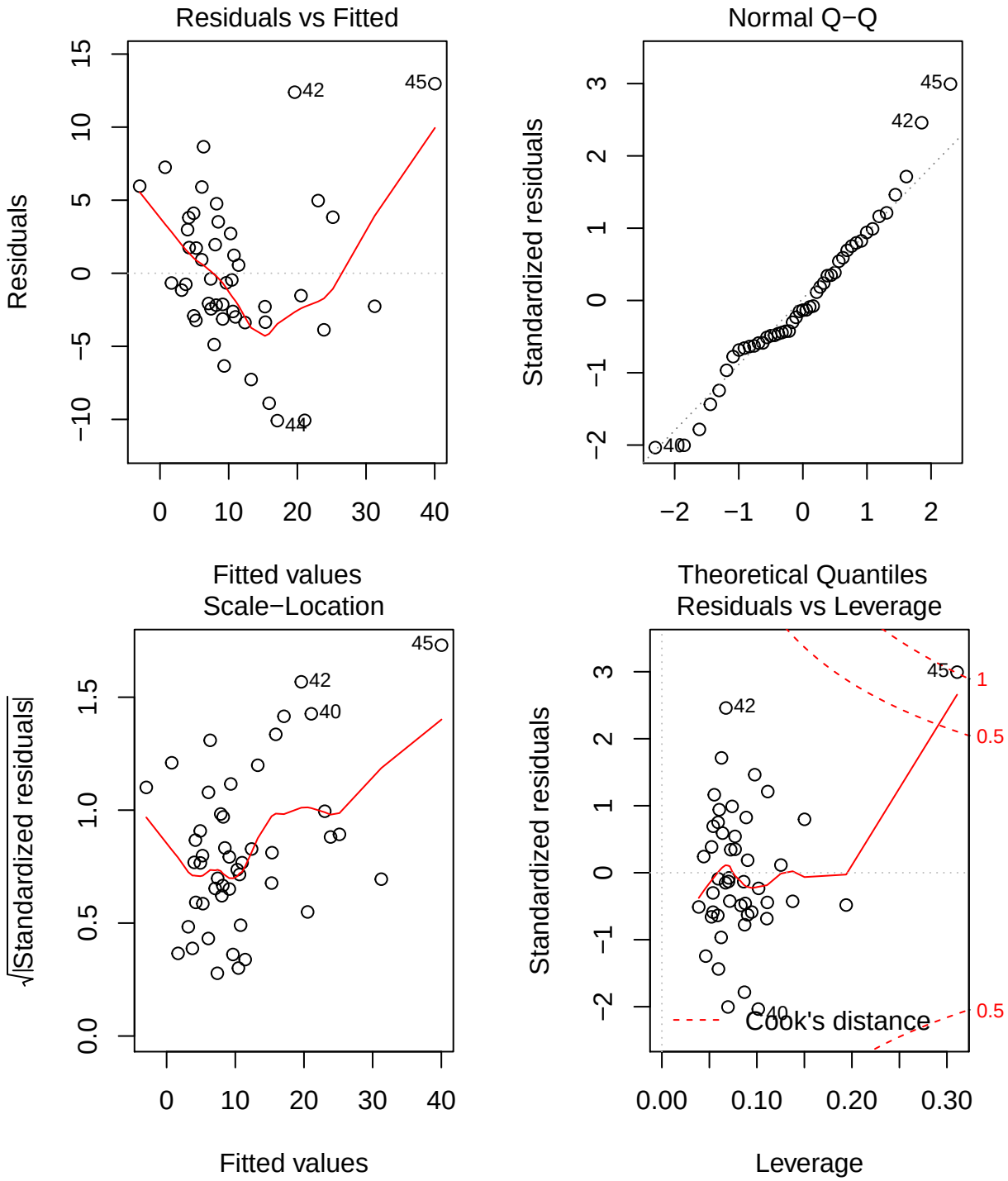
Reduced Model

Now we reduce our model to only two categories: Agriculture, Fertility and Catholic.

```
reduced_modell=lm(swiss[,4]~swiss[,1]+swiss[,2]++swiss[,5])
summary(reduced_modell)
```

```
##
## Call:
## lm(formula = swiss[, 4] ~ swiss[, 1] + swiss[, 2] + +swiss[,
##     5])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0852  -2.9521  -0.6678   3.2519  12.9706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.85051    4.64907  11.583 8.30e-15 ***
## swiss[, 1]  -0.48883    0.07104  -6.881 1.91e-08 ***
## swiss[, 2]  -0.23799    0.03779  -6.298 1.35e-07 ***
## swiss[, 5]   0.08440    0.02173   3.884  0.00035 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.218 on 43 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.7055
## F-statistic: 37.73 on 3 and 43 DF,  p-value: 4.123e-12
```

```
par(mfrow=c(1,2))
plot(reduced_modell)
```



Comparision We now use AIC and BIC to compare our two models. AS we can see our reduced model has a slightly lower score than the full one.

```
AIC(reduced_modell,fullfull_modell)
```

```
##                  df      AIC
## reduced_modell    5 294.4987
## fullfull_modell   6 295.5065
```

```r
BIC(reduced_modell,fullfull_modell)
```

```
##                  df      BIC
## reduced_modell    5 303.7495
## fullfull_modell   6 306.6074
```