Task p. 52 / Anwendung S. 52

FK

automatic

Working directory

```
> setwd("D:/kronthafranz/Documents/01Lehre/06Quantitative Forschungsmethoden
dt en")
```

Load data

```
> load("D:/kronthafranz/Documents/01Lehre/06Quantitative Forschungsmethoden
dt en/06Regression/abortion.RData")
```

Descriptive statistics

```
> summary(abortion)
```

```
       State       Abortion        Religion         Price
 ALABAMA   : 1   Min.   : 4.30   Min.   : 9.80   Min.   :228.0
 ALASKA    : 1   1st Qu.:13.43   1st Qu.:23.80   1st Qu.:271.2
 ARIZONA   : 1   Median :18.40   Median :29.65   Median :294.5
 ARKANSAS  : 1   Mean   :20.58   Mean   :32.65   Mean   :305.1
 CALIFORNIA: 1   3rd Qu.:25.35   3rd Qu.:38.67   3rd Qu.:329.8
 COLORADO  : 1   Max.   :46.20   Max.   :76.70   Max.   :461.0
 (Other)   :44
     Laws          Funds           Educ           Income
 Min.   :0.00   Min.   :0.00   Min.   :64.30   Min.   :14082
 1st Qu.:0.00   1st Qu.:0.00   1st Qu.:72.03   1st Qu.:17086
 Median :0.00   Median :0.00   Median :76.70   Median :18881
 Mean   :0.36   Mean   :0.24   Mean   :75.93   Mean   :19216
 3rd Qu.:1.00   3rd Qu.:0.00   3rd Qu.:80.10   3rd Qu.:20843
 Max.   :1.00   Max.   :1.00   Max.   :86.60   Max.   :27150


    Picket
 Min.   :  0.00
 1st Qu.: 39.25
 Median : 50.00
 Mean   : 52.34
 3rd Qu.: 67.00
 Max.   :100.00
```

Correlation coefficients

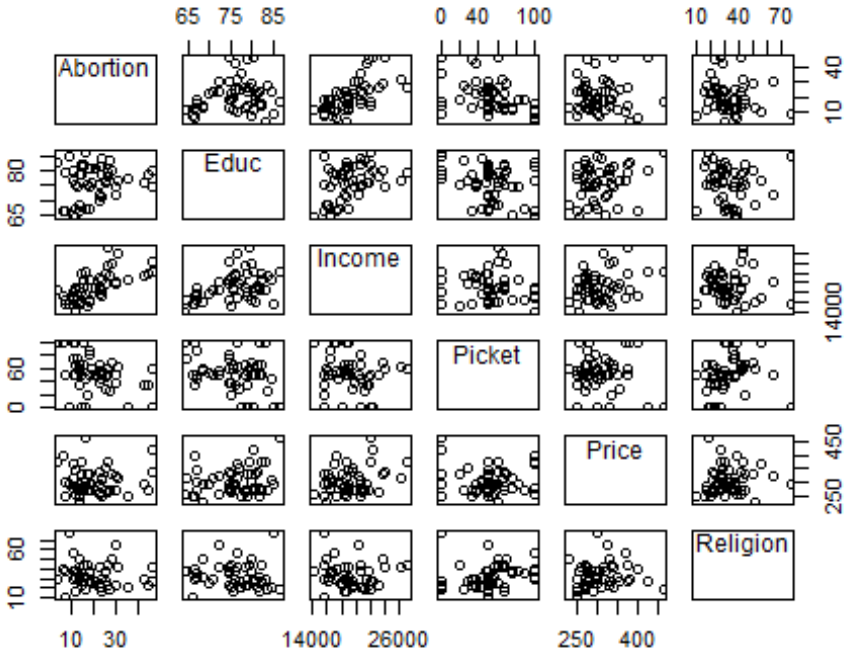
```
> cor(abortion[,c("Abortion","Educ","Income","Picket","Price","Religion")],
+   use="complete")

             Abortion        Educ      Income      Picket        Price
Abortion  1.000000000  0.19487479  0.64720351 -0.37742264  0.003097758
Educ      0.194874794  1.00000000  0.44139524 -0.30962113  0.248312375
Income    0.647203508  0.44139524  1.00000000 -0.16067340  0.302700617
Picket   -0.377422643 -0.30962113 -0.16067340  1.00000000 -0.070030556
Price     0.003097758  0.24831237  0.30270062 -0.07003056  1.000000000
Religion -0.125183848 -0.07988469 -0.07117385  0.20730929  0.086685910
            Religion
Abortion -0.12518385
Educ     -0.07988469
Income   -0.07117385
Picket    0.20730929
Price     0.08668591
Religion  1.00000000
```

Scatterplots

(not for the two dummy-variables)

```
> scatterplotMatrix(~Abortion+Educ+Income+Picket+Price+Religion,
reg.line=FALSE,
+    smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   id.n=0, diagonal = 'none', data=abortion)
```



--> Linear relationship between abortion and income

--> No relationship between abortion and picet, price and religion

--> Is there a non-linear relationship between abortion and educ? It is hard to say

Estimate the model

```
> RegModel.1 <- lm(Abortion~Educ+Funds+Income+Laws+Picket+Price+Religion,
+   data=abortion)
> summary(RegModel.1)


Call:
lm(formula = Abortion ~ Educ + Funds + Income + Laws + Picket +
    Price + Religion, data = abortion)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-11.6110  -4.6493  -0.6696   4.5253  15.9514

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.2839573 15.0776294   0.947   0.3489
Educ        -0.2872551  0.1995545  -1.439   0.1574
Funds        2.8200030  2.7834747   1.013   0.3168
Income       0.0024007  0.0004552   5.274 4.35e-06 ***
Laws        -0.8731018  2.3765662  -0.367   0.7152
Picket      -0.1168712  0.0421799  -2.771   0.0083 **
Price       -0.0423631  0.0222232  -1.906   0.0635 .
Religion     0.0200709  0.0863805   0.232   0.8174
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.063 on 42 degrees of freedom
Multiple R-squared:  0.5774,    Adjusted R-squared:  0.507
F-statistic: 8.199 on 7 and 42 DF,  p-value: 2.847e-06
```

--> Model is significant

--> R2 is 57.8%

--> Income, picket and price (10%) are significant

--> Educ, funds, laws abd religion are not significant

Evaluate GM assumptions

Add regression statistics

```
> abortion<- within(abortion, {
+   fitted.RegModel.1 <- fitted(RegModel.1)
+   residuals.RegModel.1 <- residuals(RegModel.1)
+   rstudent.RegModel.1 <- rstudent(RegModel.1)
+   hatvalues.RegModel.1 <- hatvalues(RegModel.1)
+   cooks.distance.RegModel.1 <- cooks.distance(RegModel.1)
+   obsNumber <- 1:nrow(abortion)
+ })
```

GM1: Linearity and complete specification

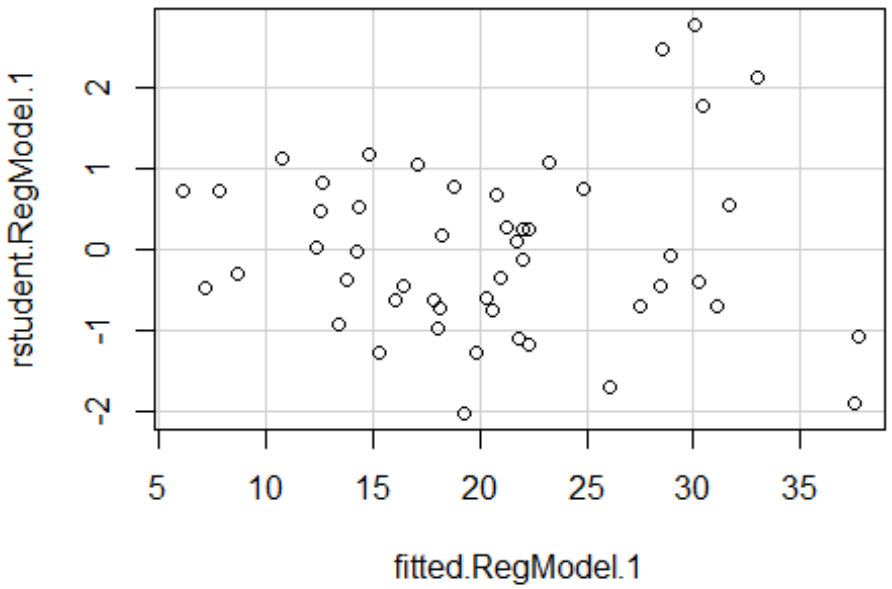
--> Specification is a matter of theory

--> Linearity is already considered (but there is uncertainty for the relationship between abortion and educ)

--> The interested student can try to model an inverse u-shaped relationship

GM2: Expected value = 0

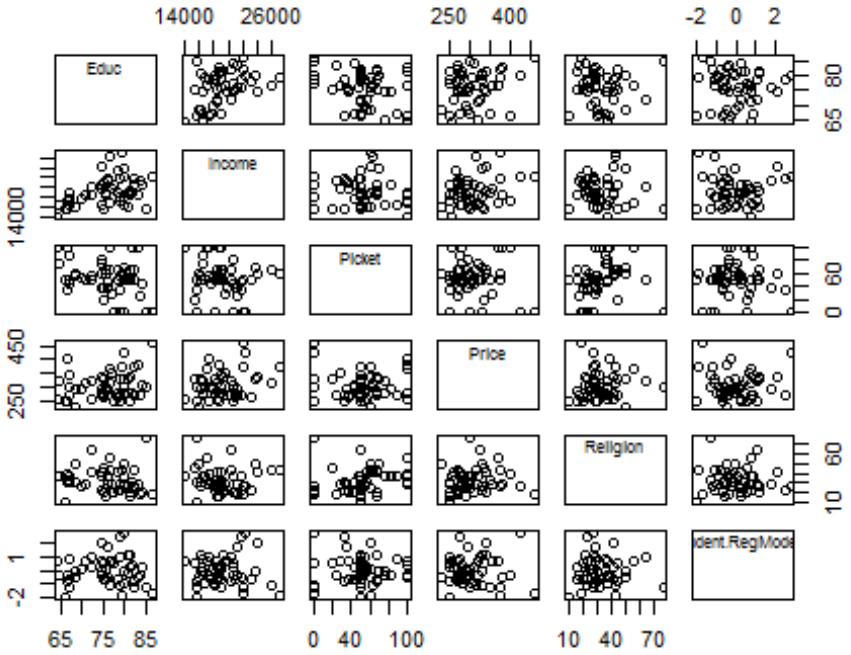
```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=abortion)
```



--> Looks good, but not for intervalls at the end of the fitted values

GM3: Error term is correlated with independent variables?

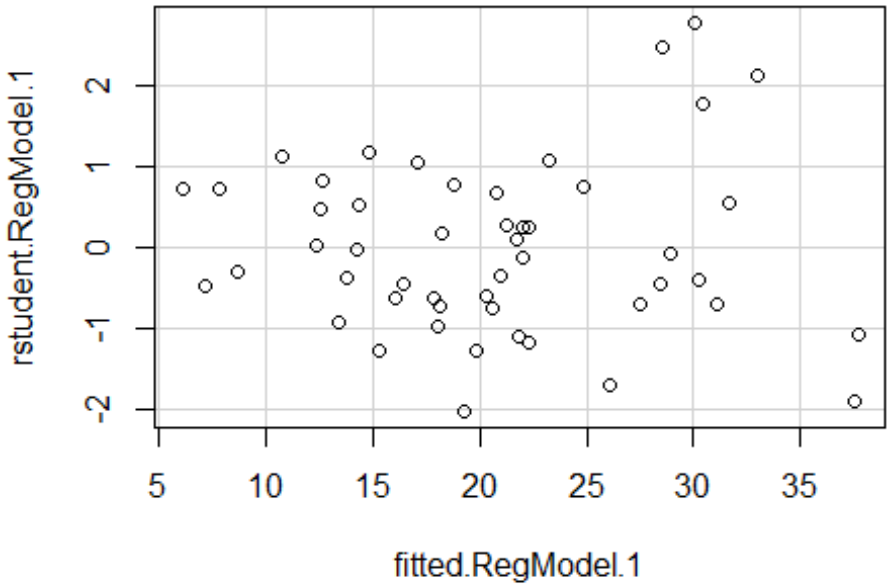
```
> scatterplotMatrix(~Educ+Income+Picket+Price+Religion+rstudent.RegModel.1,
+   reg.line=FALSE, smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), id.n=0, diagonal = 'none', data=abortion)
```



--> Looks good

GM4: Heteroscedasticity?

```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=abortion)
```

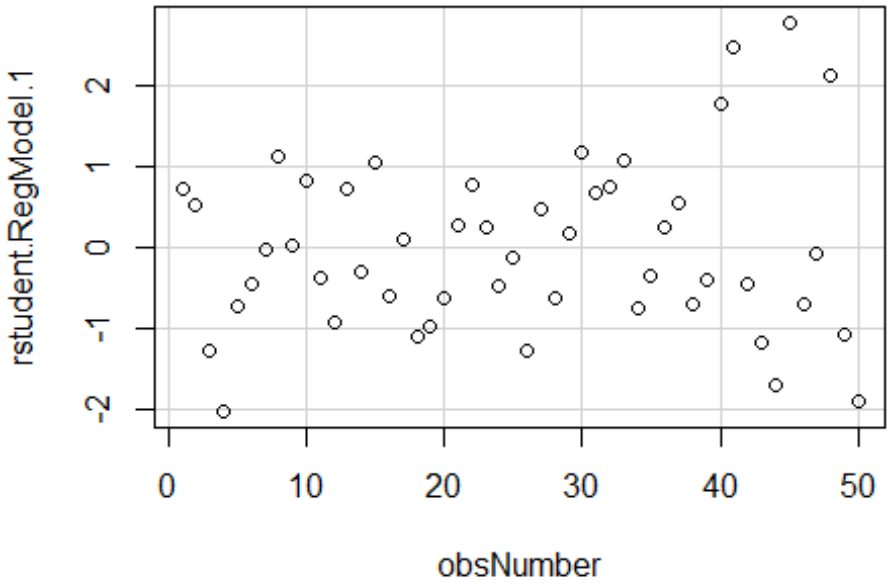


--> Seems to be the case that the variance increases

--> Heteroscedasticity might be a problem

GM5: Autocorrelation?

```
> scatterplot(rstudent.RegModel.1~obsNumber, reg.line=FALSE, smooth=FALSE,
+   spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   data=abortion)
```



--> No pattern

GM6: Multicollinearity

correlation coefficients

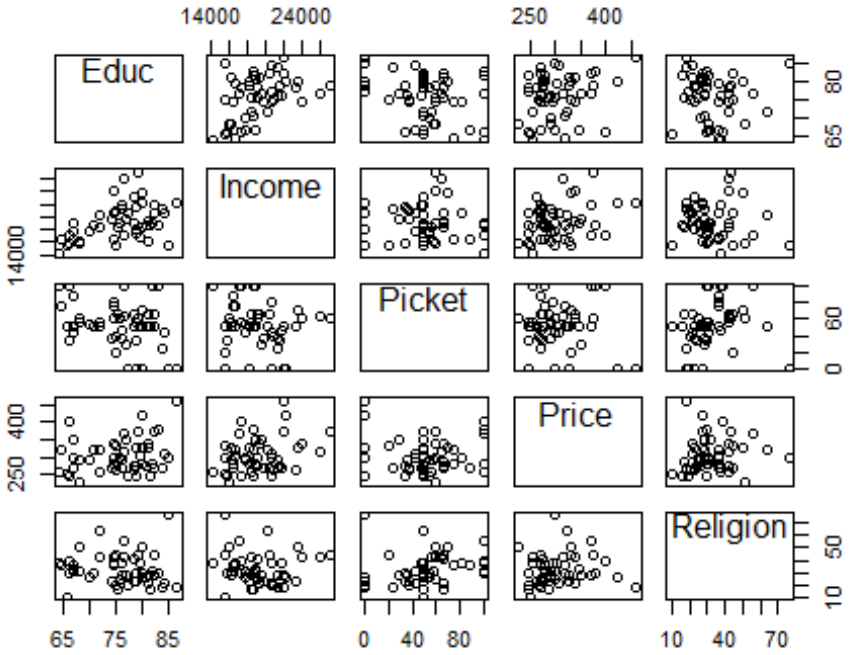
```
> cor(abortion[,c("Educ","Income","Picket","Price","Religion")],
+   use="complete")

               Educ      Income      Picket       Price    Religion
Educ      1.00000000  0.44139524 -0.30962113  0.24831237 -0.07988469
Income    0.44139524  1.00000000 -0.16067340  0.30270062 -0.07117385
Picket   -0.30962113 -0.16067340  1.00000000 -0.07003056  0.20730929
Price     0.24831237  0.30270062 -0.07003056  1.00000000  0.08668591
Religion -0.07988469 -0.07117385  0.20730929  0.08668591  1.00000000
```

--> No high correlations

Scatterplots

```
> scatterplotMatrix(~Educ+Income+Picket+Price+Religion, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   id.n=0, diagonal = 'none', data=abortion)
```



--> Confirms impression

Variance inflation factors

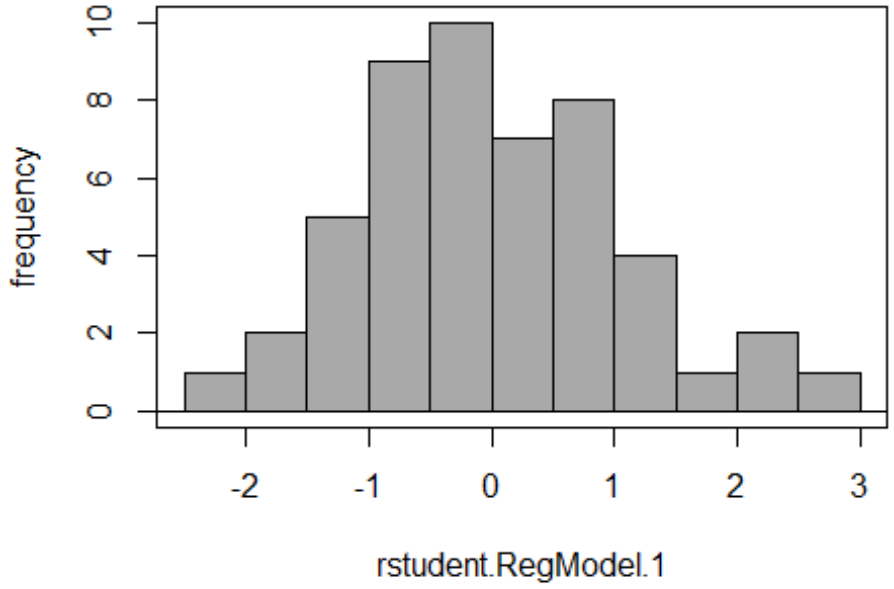
```
> vif(RegModel.1)

    Educ    Funds   Income     Laws   Picket    Price Religion
1.380153 1.416584 1.606933 1.304444 1.214623 1.153023 1.175216
```

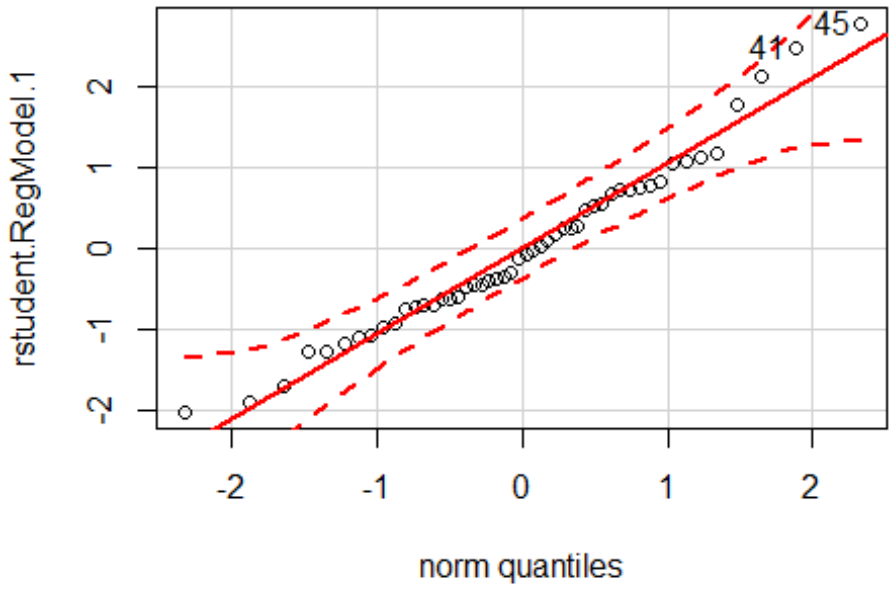
--> No VIF close to or higher than 4

GM7: Normal distribution?

```
> with(abortion, Hist(rstudent.RegModel.1, scale="frequency",
+   breaks="Sturges", col="darkgray"))
```




```
> with(abortion, qqPlot(rstudent.RegModel.1, dist="norm", id.method="y",
+   id.n=2, labels=rownames(abortion)))
```




```
45 41
50 49
```

```
> library(nortest, pos=15)

> with(abortion, shapiro.test(rstudent.RegModel.1))


    Shapiro-Wilk normality test

data:  rstudent.RegModel.1
W = 0.9742, p-value = 0.3399
```

--> No problem

Course of action

--> There seems to be a problem with GM2

--> There seems to be a problem with GM4

--> A first approach would be to solve the problem with heteroscedasticity (for example transform the dependent variable as described in the lecture)

--> Estimate the model again

--> Evaluate the new model