Task p. 51-2 / Anwendung S. 51-2

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/reg_country.RData")
```

Descriptive statistics

```
> summary(data_country_2)

       country        expect          urban             doc
 Afghanistan:  1   Min.   :41.00   Min.   :  5.00   Min.   : 0.188
 Albania    :  1   1st Qu.:56.00   1st Qu.: 28.25   1st Qu.: 1.185
 Algeria    :  1   Median :68.00   Median : 48.00   Median : 6.305
 Angola     :  1   Mean   :66.31   Mean   : 48.78   Mean   :10.521
 Argentina  :  1   3rd Qu.:76.00   3rd Qu.: 69.50   3rd Qu.:16.667
 Australia  :  1   Max.   :83.00   Max.   :100.00   Max.   :42.918
 (Other)    :116                                    NA's   :1
      bed               gdp              rad
 Min.   :  2.525   Min.   :  120   Min.   :  1.562
 1st Qu.: 11.887   1st Qu.:  400   1st Qu.: 11.746
 Median : 22.051   Median : 1110   Median : 21.277
 Mean   : 34.875   Mean   : 4158   Mean   : 31.186
 3rd Qu.: 50.315   3rd Qu.: 4375   3rd Qu.: 40.000
 Max.   :135.135   Max.   :22470   Max.   :200.000
 NA's   :6
```

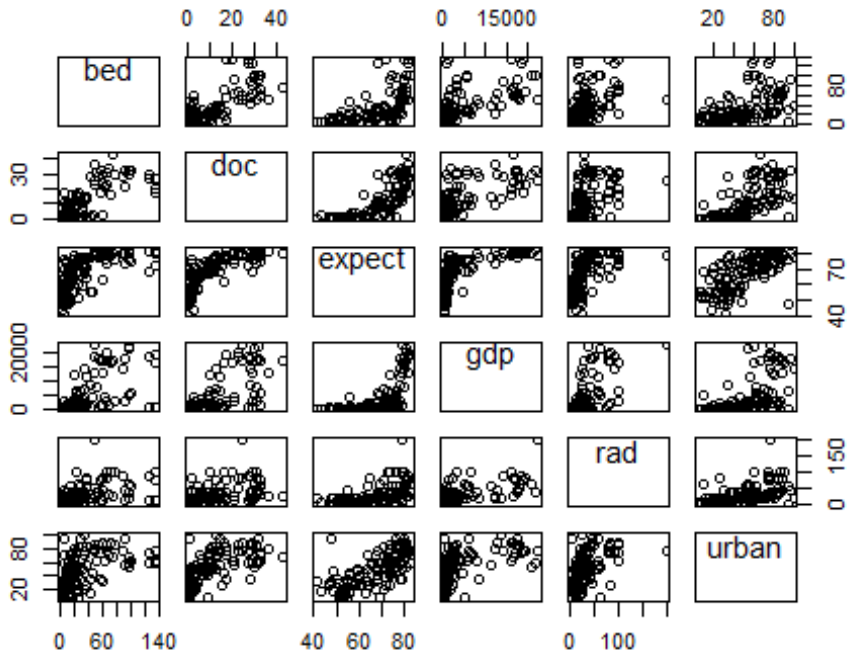
Correlation coefficients

```
> cor(data_country_2[,c("bed","doc","expect","gdp","rad","urban")],
use="complete")

             bed       doc    expect       gdp       rad     urban
bed    1.0000000 0.7704197 0.6251971 0.6499957 0.4986648 0.5005619
doc    0.7704197 1.0000000 0.7818952 0.7142052 0.5280187 0.6842779
expect 0.6251971 0.7818952 1.0000000 0.6679829 0.5636072 0.6966961
```

```
gdp    0.6499957 0.7142052 0.6679829 1.0000000 0.6620972 0.6104549
rad    0.4986648 0.5280187 0.5636072 0.6620972 1.0000000 0.5339263
urban  0.5005619 0.6842779 0.6966961 0.6104549 0.5339263 1.0000000
```

Scatterplot

```
> scatterplotMatrix(~bed+doc+expect+gdp+rad+urban, reg.line=FALSE,
smooth=FALSE,
+   spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9), id.n=0, diagonal
= 'none',
+    data=data_country_2)
```



--> non-linearity for nearly all relationships between dependent and independent variables

--> linearity only between expect and urban

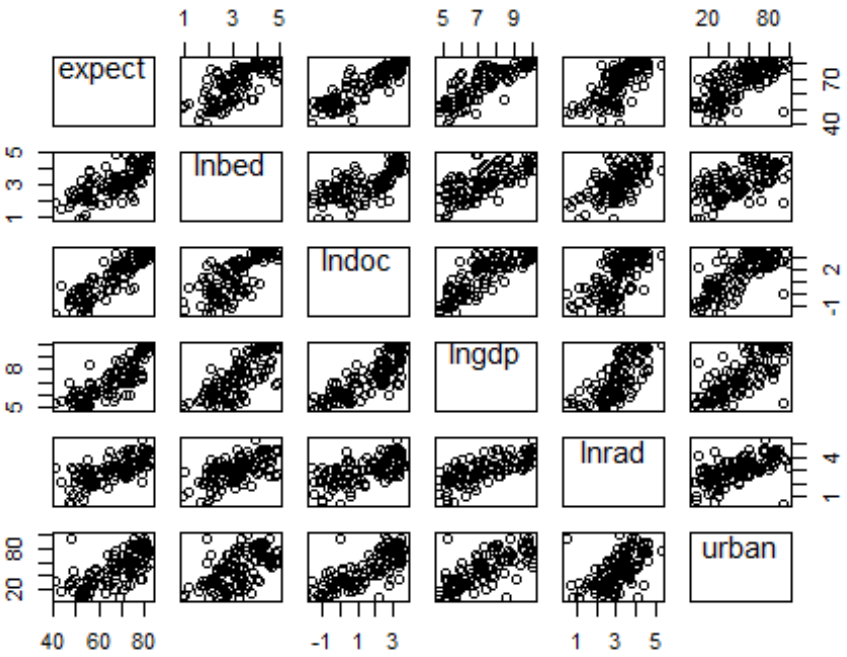
Generate new variables: lnbed, lndoc, lngdp, lnrad

```
> data_country_2$lnbed <- with(data_country_2, log(bed))

> data_country_2$lndoc <- with(data_country_2, log(doc))

> data_country_2$lngdp <- with(data_country_2, log(gdp))

> data_country_2$lnrad <- with(data_country_2, log(rad))
```

--> to generate linearity there are more than one possibility, one possibility that works is the log

Consider scatterplot again

```
> scatterplotMatrix(~expect+lnbed+lndoc+lngdp+lnrad+urban, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
id.n=0,
+   diagonal = 'none', data=data_country_2)
```



--> relationships between expect and independent variables are more or less linear

Estimate model

```
> RegModel.1 <- lm(expect~lnbed+lndoc+lngdp+lnrad+urban, data=data_country_2)
> summary(RegModel.1)


Call:
lm(formula = expect ~ lnbed + lndoc + lngdp + lnrad + urban,
    data = data_country_2)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8341  -2.6580   0.0932   2.9010  14.0639

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.76703    3.17385  12.845  < 2e-16 ***
lnbed        1.14739    0.74880   1.532  0.12832
lndoc        4.06873    0.56290   7.228 6.85e-11 ***
```

```
lngdp         1.70932    0.61574   2.776  0.00647 **
lnrad         1.54173    0.68607   2.247  0.02662 *
urban        -0.02002    0.02917  -0.686  0.49397
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.742 on 110 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8272,    Adjusted R-squared:  0.8194
F-statistic: 105.3 on 5 and 110 DF,  p-value: < 2.2e-16
```

--> Model is significant

--> R2 is 82.7%

--> lndoc, lngdp and lnrad (5%) are significant

--> lnbed and urban are not significant

Evaluate GM assumptions

Add regression statistics

```
> data_country_2<- within(data_country_2, {
+   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(data_country_2)
+ })
```

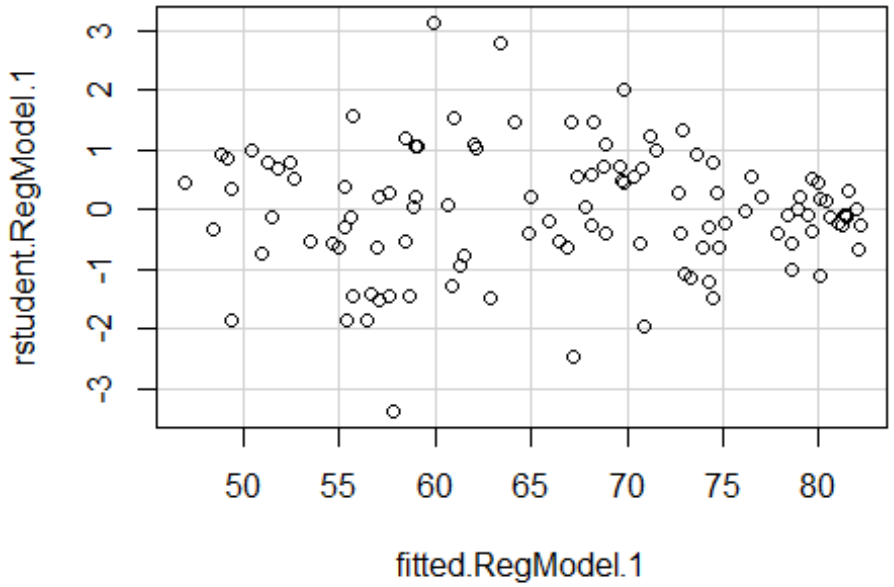
GM1: Linearity and complete specification

--> Complete specification is a matter of theory

--> Linearity is already considered

GM2: Expected value = 0

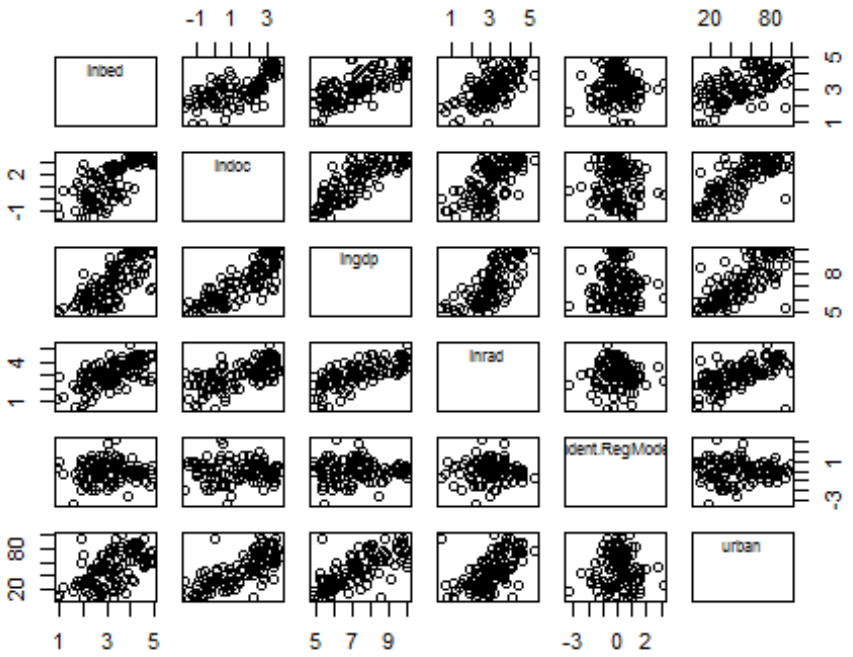
```
> scatterplot(rstudent.RegModel.1~fitted.RegModel.1, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, boxplots=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), data=data_country_2)
```



--> Looks good

GM3: Error term is correlated with independent variables?

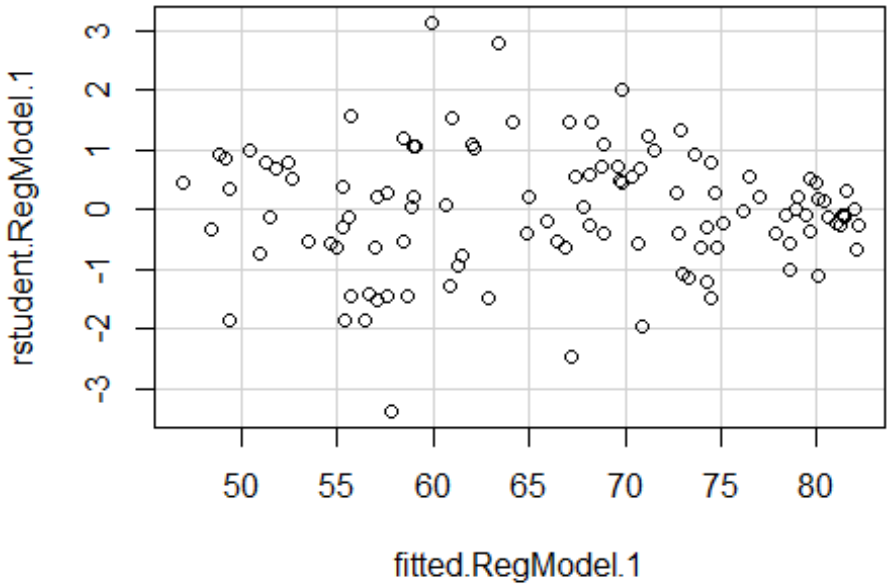
```
> scatterplotMatrix(~lnbed+lndoc+lngdp+lnrad+rstudent.RegModel.1+urban,
+   reg.line=FALSE, smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE,
+   levels=c(.5, .9), id.n=0, diagonal = 'none', data=data_country_2)
```



--> 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=data_country_2)
```



--> Looks not so bad

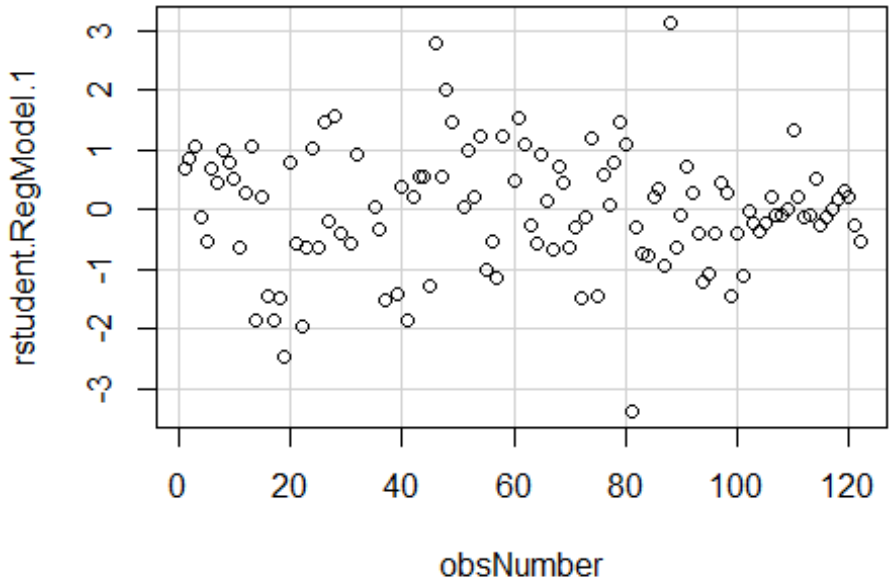
--> In the midst the variance seems larger

--> In the end the variance seems smaller

--> Might be a problem of outliers in the midst

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=data_country_2)
```



--> Looks fine, no pattern

GM6: Multicollinearity?

Correlation coefficients

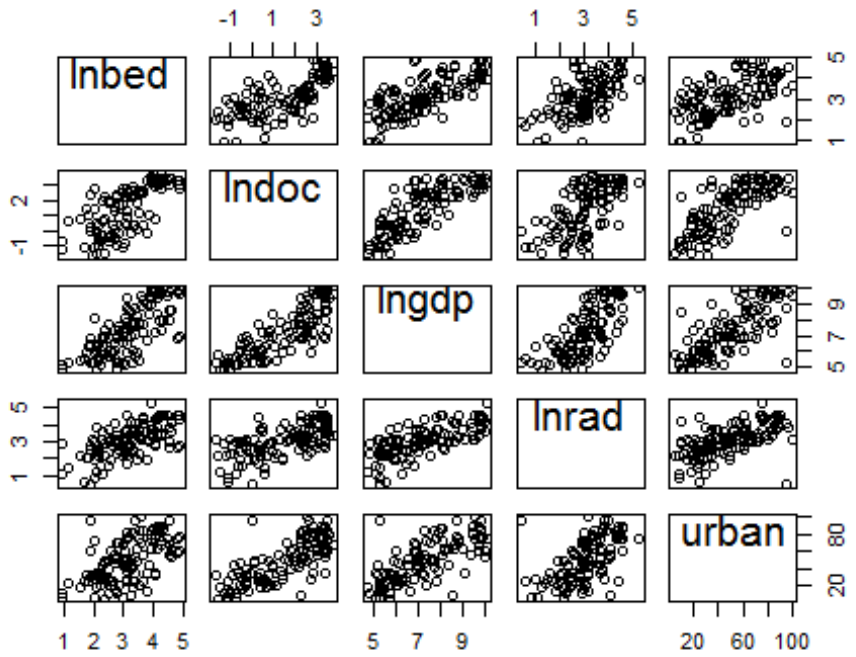
```
> cor(data_country_2[,c("lnbed","lndoc","lngdp","lnrad","urban")],
+   use="complete")

          lnbed     lndoc     lngdp     lnrad     urban
lnbed 1.0000000 0.7105475 0.7414043 0.6164773 0.5756946
lndoc 0.7105475 1.0000000 0.8236441 0.6332845 0.7628680
lngdp 0.7414043 0.8236441 1.0000000 0.7157287 0.7478097
lnrad 0.6164773 0.6332845 0.7157287 1.0000000 0.5792965
urban 0.5756946 0.7628680 0.7478097 0.5792965 1.0000000
```

--> high correlations between independent variables

--> especially between lndoc and lngdp

Scatterplot

```
> scatterplotMatrix(~lnbed+lndoc+lngdp+lnrad+urban, reg.line=FALSE,
+   smooth=FALSE, spread=FALSE, span=0.5, ellipse=FALSE, levels=c(.5, .9),
+   id.n=0, diagonal = 'none', data=data_country_2)
```



--> Scatterplot confirms impression

Variance inflation factor

```
> vif(RegModel.1)

   lnbed    lndoc    lngdp    lnrad    urban
2.460880 3.950080 4.614049 2.140492 2.698857
```

--> High variance inflation factors for lndoc and lngdp

--> Problem with multicollinearity

--> Leave a variable out that causes multicollinearity

--> In this case we can leave out either lndoc or lngdp (both are highly correlated)

GM7: Normal distribution?

do not forget to evaluate the normal distribution assumption

Course of action

--> leave out the variable responsible for multicollinearity

--> estimate the model again

--> evaluate the new model