Predictive Modeling

Series 3

Exercise 3.1

We consider a data set originating from a medical application which is contained in the data file `catheter.rda` . The data frame consists of measurements of the `height` in centimeter and of the `weight` in kilogram of patients. The response variable `catlength` refers to the optimal length of catheters used for the examination of the heart. In this exercise, we intend to predict the optimal catheter length on the basis of the available data set `catheter`.

a) Fit a simple linear regression model for both `catlength` $\sim$ `height` and `catlength` $\sim$ `weight`. Are the predictors significant?

b) Fit a multiple linear regression model `catlength` $\sim$ `height` + `weight`. Is there any influence of the predictors on the response variable ? Is it significant?

c) Test the null hypotheses $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. Compare the results with those from the two simple linear regressions. Comment and explain the differences if there are any.

d) For a child that is 120 cm tall and has a weight of 25 kg, compute the 95 % prediction interval

   - by means of the multiple regression model,
   - by means of the simple regression models(2x).

   In practice, a prediction error of ± 2 cm would be acceptable. Do the data and the models allow for a prediction of `catlength` that is sufficiently precise? Does it make sense to use both predictors?

Exercise 3.2

In this exercise, we would like to analyze to which extent savings differ between countries. The data set `savings.rda` contains data for 50 countries. For each country the values are averaged over the entire population and over the time period 1960 - 1970. The variables have the following meanings:

- `sr` : proportion of the available income that is saved

- `pop15` : proportion of the population that is younger than 15 years
- `pop75` : proportion of the population that is older than 75 years
- `dpi` : per capita income
- `ddpi` : growth rate of `dpi`

a) Fit the model `sr` $\sim$ `pop15` + `pop75` + `dpi` + `ddpi`. Carry out a residual analysis.

b) Identify the three countries having the largest leverage and describe how they differ from the remaining data points. **R**-Hints:

```
## Observations with the largest leverage in
## decreasing order
sort(hatvalues(fit), decreasing = TRUE)
```

c) Remove the data point (country) with the largest Cook's distance from the analysis. To what extent do the results change?

d) Now consider the following variable transformations:

   - Log-transform `sr`
   - Log-transform the predictor variables `dpi` and `ddpi`, but not the response variable `sr`
   - Log-transform `sr` and the predictor variables `dpi` and `ddpi`

   Fit the corresponding models and analyze the residuals. Finally, decide which model is most appropriate.

Exercise 3.3

The data set `synthetisch.rda` contains the response Y and the predictors X_1 and X_2. Fit a multiple linear regression model. We can assume that the errors ϵ_i are independent and that the majority of the observations are distributed according to $\mathcal{N}(0, \sigma^2)$.

Carry out a residual analysis and generate a 3D plot of the regression hyperplane. Does the 3D plot confirm your conclusion from the residual analysis?

R-Hints: The 3D plot can be generated by means of the **R**-function `scatter3d()` from the package `car`. You also required the package `rgl` which you can install as follows:

```r
install.packages("car")
install.packages("rgl")
```

Then use:

```r
## 3D plot
library(car)
scatter3d(y ~ x1 + x2, data = synthetisch)
```

The plot shows up in an interactive window.

Exercise 3.4

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. 20 measurements of the quality (`Q`), the speed (`S`) and the used cutting tool (`CTT`) are found in the file `lathe.dat`. (Source: Montgomery and Runger, *Applied Statistics and Probability for Engineers*).

a) Generate a scatter plot of `Q` versus `S`.

b) Fit a regression model for which the response variable `Q` depends linearly on the speed `S` of the lathe. The slope should be the same for both types of cutting tools, but the intercept should depend on the tool used. Write the linear regression model using a dummy variable for the cutting tool. Add the regression line to the plot in a).

c) Fit the following model:

```r
lm(Q ~ S * CTT, data = lathe)
```

What is the name of this model? Write the linear regression model using a dummy variable for the cutting tool.

d) Let us assume that the slope depends as well on the cutting tool used. Would a different slope be plausible according to the collected data? Answer the question with a hypothesis test at the 5 % level of significance. What is the probability of a type II error?

Exercise 3.5

The Australian Bureau of Agricultural and Resource Economics conducts an annual survey of the agroindustry. In 1991, 451 farms in New South Wales took part in this survey. The raw data is contained in the file `farm.rda`. The variables have the following meanings:

- `revenue` : response variable, total revenue of the farm
- `costs` : predictor variable, total costs of the farm
- `region` : predictor variable, code for different regions within New South Wales
- `industry` : predictor variable, code for cultivation (1 = wheat, 2=wheat, sheep, cattle, 3=sheep, 4=cattle, 5=sheep, cattle).

The aim is to fit a suitable regression model that predicts the `revenue` of a farm. You will need to perform the following steps:

a) Preprocess the data in such a way that factor variables are correctly stored in the data frame. **R**-Hints:

```
## Load data
load(".../farm.rda")
## Check properties of the data
str(farm)
```

If the data type of a variable needs to be transformed into a factor variable, use the **R**-function `factor(...,labels=...)`, e.g.

```
farm$industry <- factor(farm$industry, labels = c("wheat",
    "wheat_sheep_cattle", "sheep", "cattle", "sheep_cattle"))
```

b) Fit the complete regression model and check the residuals. Does the model fit? If not, carry out a log-transformation of all non-factor variables. Does the model fit now? Continue the analysis with the log-transformed variables.

c) For a cattle farm in `region` 111 with `costs` of 100 000, what is the predicted `revenue`?

d) Test whether `region` has a significant influence on revenue when the other predictors are given. **R-Hint**: Use `drop1(fit, test="F")`.

e) Add an interaction term between `region` and `industry`.

   - How many parameters are estimated in total?

- Is the interaction term significant? Carry out a suitable test.
- What is the intuitive meaning of the interaction term, how do we have to interpret it in the context of the model? What is the conclusion?

f) We now have several models of varying complexity. First, there is the model with all predictors and the interaction term between `region` and `industry`. Then, there is the model with the main effects followed by the one without the variable `region`. Last, the least complex model is the one that does not distinguish between `region` and `industry`. Which model is best suited for predicting the `revenue` of a farm?

Result Checker

E 3.2: b) Libya, United States, Japan

E 3.5: c) 154 914

Predictive Modeling

Solutions to Series 3

Solution 3.1

a) In both cases, the predictor variable in the simple linear model is highly significant:

```
## load data
load("Daten/catheter.rda")
## Simple linear regressions
fits1 <- lm(catlength ~ height, data = catheter)
fits2 <- lm(catlength ~ weight, data = catheter)
summary(fits1)

##
## Call:
## lm(formula = catlength ~ height, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0929 -0.7298 -0.2608  1.1652  6.6879
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.12706    4.24700   2.855 0.017090 *
## height       0.23774    0.04034   5.893 0.000152 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.009 on 10 degrees of freedom
## Multiple R-squared:  0.7764,Adjusted R-squared:  0.7541
## F-statistic: 34.73 on 1 and 10 DF,  p-value: 0.0001525
```

```
summary(fits2)

##
## Call:
## lm(formula = catlength ~ weight, data = catheter)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9676 -1.4963 -0.1386  2.0980  7.0205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.62631    2.00264  12.796 1.59e-07 ***
## weight       0.61613    0.09759   6.313 8.75e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.797 on 10 degrees of freedom
## Multiple R-squared:  0.7994,Adjusted R-squared:  0.7794
## F-statistic: 39.86 on 1 and 10 DF,  p-value: 8.755e-05
```

b)

```
## Multiple regression
fit <- lm(catlength ~ height + weight, data = catheter)
summary(fit)

##
## Call:
## lm(formula = catlength ~ height + weight, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0497 -1.2753 -0.2595  1.9095  6.9933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.08527    8.77037   2.404   0.0396 *
## height       0.07681    0.14412   0.533   0.6070
## weight       0.42752    0.36810   1.161   0.2753
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.94 on 9 degrees of freedom
## Multiple R-squared:  0.8056,Adjusted R-squared:  0.7624
## F-statistic: 18.65 on 2 and 9 DF,  p-value: 0.0006301
```

Yes, there is an influence of the predictor variables on the response variable.

This influence is measured by means of the global F-test. The corresponding p-value is smaller than 0.01 so that the null hypothesis is rejected at the 1 % level. At least one of the predictors is relevant.

c) As we can see from the summary output (see above), both null hypotheses are retained, i.e. the predictors are not significant. Is this a contradiction to the results obtained in the two simple linear regression models? The answer is no. In multiple linear regression, on the basis of the hypotheses tests we decide whether the predictor variable **`height`** is required when the value of the predictor **`weight`** is known. The answer is no and the same holds vice versa. On the other hand, the global F-test indicates that we need at least one of the two predictors. So we do not need to include both predictors simultaneously but we need one of them. This situation occurs when the predictors are strongly correlated. Due to the smaller p-value we would prefer to include the predictor **`weight`** in the regression model.

d)

```r
## prediction intervals
newdat <- data.frame(height = 120, weight = 25)
predict(fits1, newdata = newdat, interval = "prediction")

##          fit      lwr      upr
## 1 40.65609 31.20891 50.10327

predict(fits2, newdata = newdat, interval = "prediction")

##          fit      lwr      upr
## 1 41.02954 32.06162 49.99747

predict(fit, newdata = newdat, interval = "prediction")

##          fit      lwr      upr
## 1 40.99072 31.53989 50.44154
```

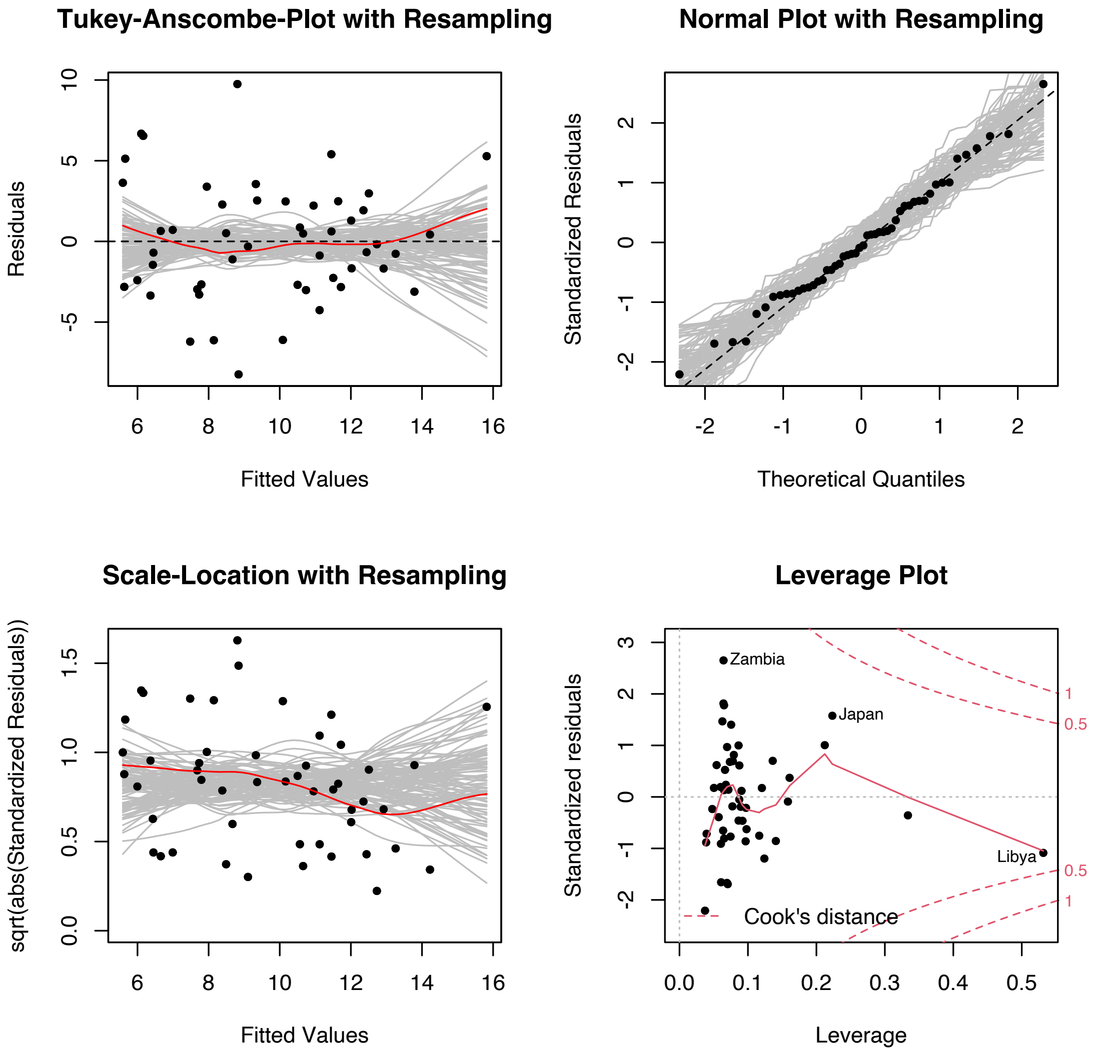
The prediction intervals differ slightly. We note that the prediction interval obtained on the basis of the multiple regression model is the largest among the three prediction intervals, contrary to what we might have expected. For, the multiple linear regression model includes most information. However, the multiple model requires the estimation of one additional parameter on the basis of the 12 available data points. This is associated with a larger estimation error of each single parameter.

In general, the prediction accuracy increases by including an additional parameter. But in this case the increase of the estimation error has a stronger effect, here a negative one. This is due to the fact that the two predictors are strongly correlated - adding the second predictor when the first one is already present does hardly yield additional information.

In practice, a prediction error of ± 2 cm would be acceptable. Thus, the data and the fitted models do not allow for a prediction of `catlength` that is sufficiently precise.

Solution 3.2

a)
```r
## Load data
load("Daten/savings.rda")
source("resplot.R")
## Model without transformations
fit <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
## Residuals and Cook's Distance
par(mfrow = c(2, 2))
resplot(fit)
```



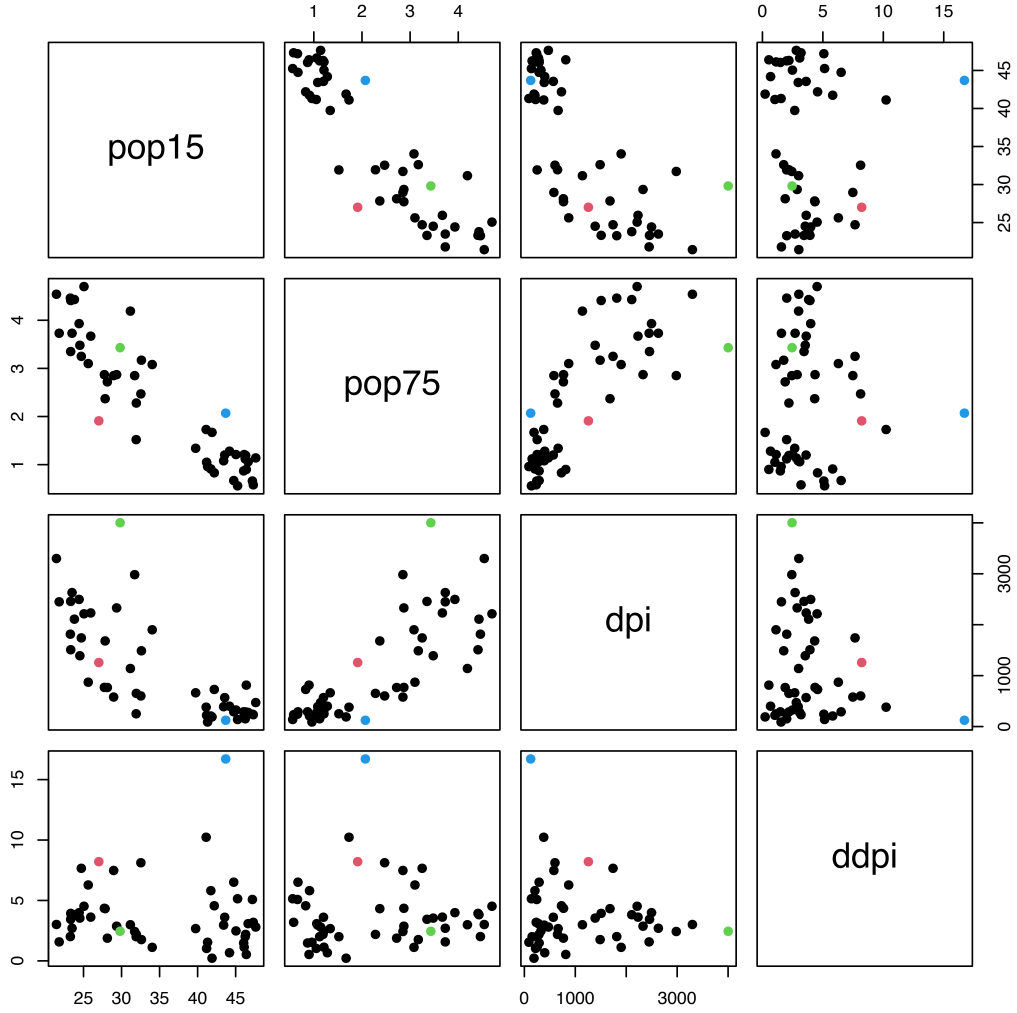
The residual plots don't indicate any serious violation of the model assumptions, hence the model seems appropriate. There are a few points with large leverage but none of these points is influential since Cook's distance is smaller than 0.5 for all points.

b)

```
## Observations with the largest leverage
sort(hatvalues(fit), decreasing = TRUE)[1:3]

##          Libya United States         Japan
##      0.5314568     0.3336880     0.2233099
```

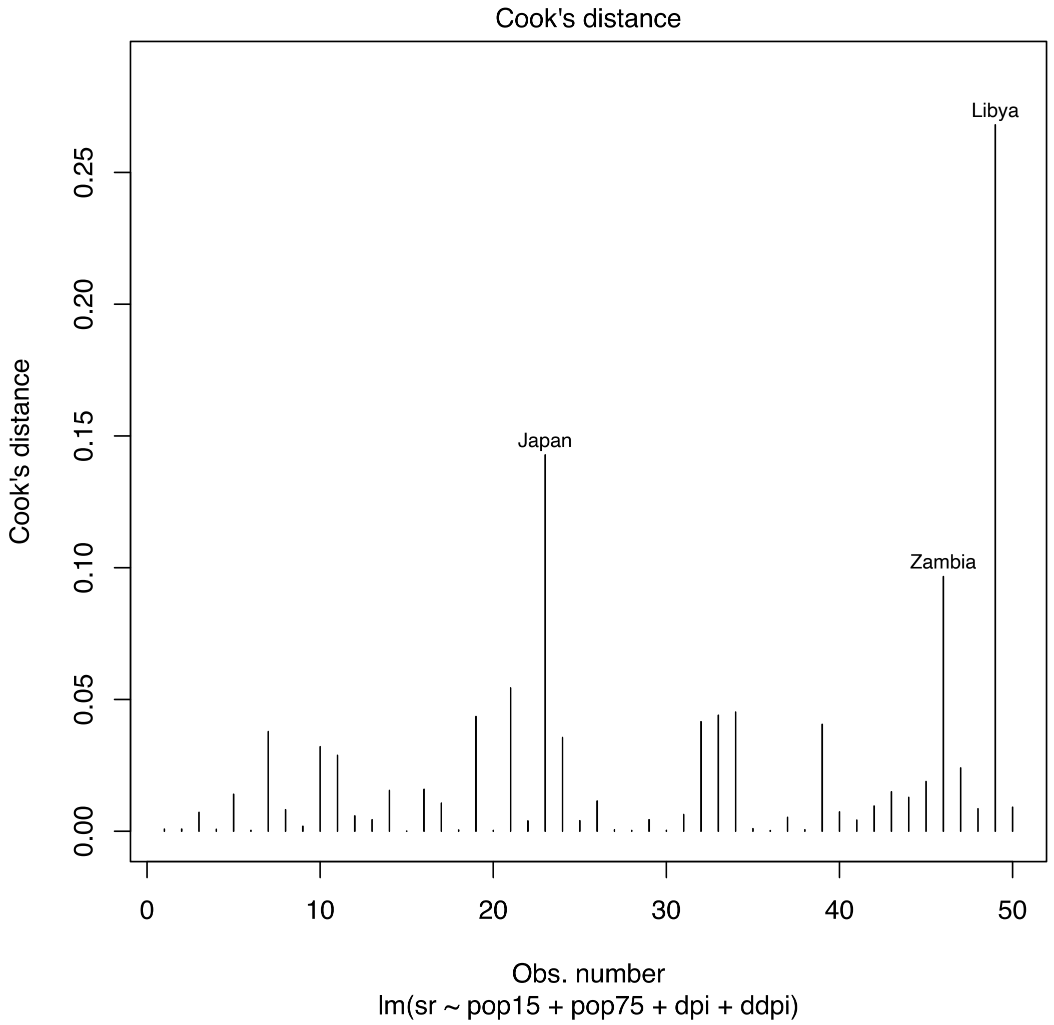
The three countries with the largest leverage are Libya, the USA, and Japan. The simplest way to see why these points have conspicuous predictor configurations is to plot pairwise scatter plots.

```r
weli <- which(rownames(savings) %in% c("Libya", "United States",
    "Japan"))
farb <- rep(1, nrow(savings))
farb[weli] <- c(2, 3, 4)
## Japan (red), USA (green), Libya (blue)
pairs(savings[, -1], pch = 19, col = farb)
```



In these plots, Japan is depicted as red point, USA as green point and Libya as blue point. The latter has a very low value of **dpi**, but a very large value of **ddpi**. The USA have the largest **dpi** value and a relatively large value of **pop75**. Japan, on the other hand, lies at the border in several scatter plots but is not extraordinary with respect to a single feature.

c)
```
## Analysis without data point with largest Cook's
## distance
plot(fit, which = 4)   ## exclude Libya
```



```
weli <- which(rownames(savings) == "Libya")
fit1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi, data = savings[-weli,
    ])
## Comparison of the estimated coefficient
coef(fit)

##   (Intercept)         pop15         pop75
## 28.5660865407 -0.4611931471 -1.6914976767
##           dpi          ddpi
## -0.0003369019  0.4096949279
```

```
coef(fit1)
```

```
##   (Intercept)         pop15         pop75
## 24.5240459788 -0.3914401268 -1.2808669233
##           dpi          ddpi
## -0.0003189001  0.6102790264
```

```
summary(fit)
```

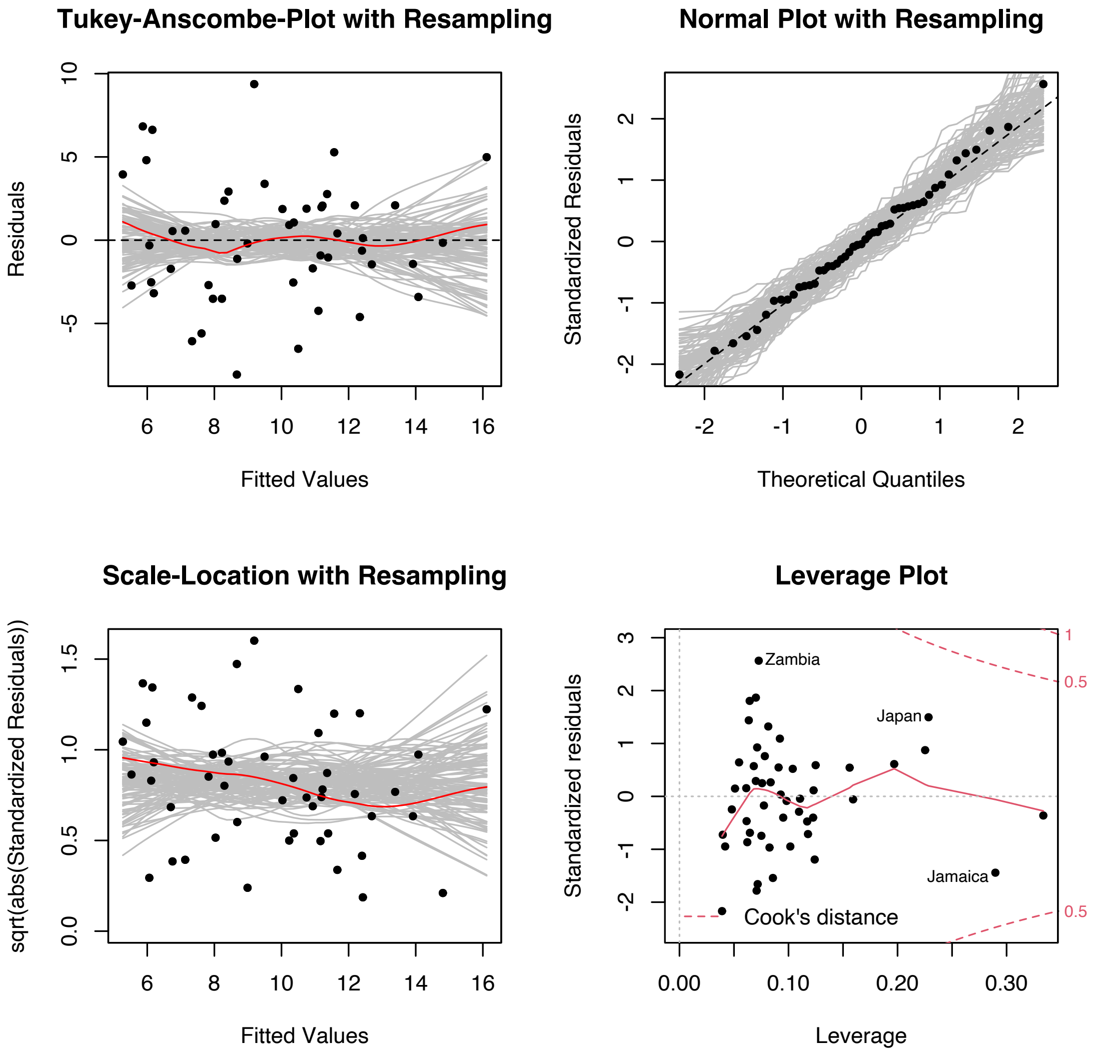
```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2422 -2.6857 -0.2488  2.4280  9.7509
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
## pop15       -0.4611931  0.1446422  -3.189 0.002603 **
## pop75       -1.6914977  1.0835989  -1.561 0.125530
## dpi         -0.0003369  0.0009311  -0.362 0.719173
## ddpi         0.4096949  0.1961971   2.088 0.042471 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.803 on 45 degrees of freedom
## Multiple R-squared:  0.3385,Adjusted R-squared:  0.2797
## F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

```
summary(fit1)
```

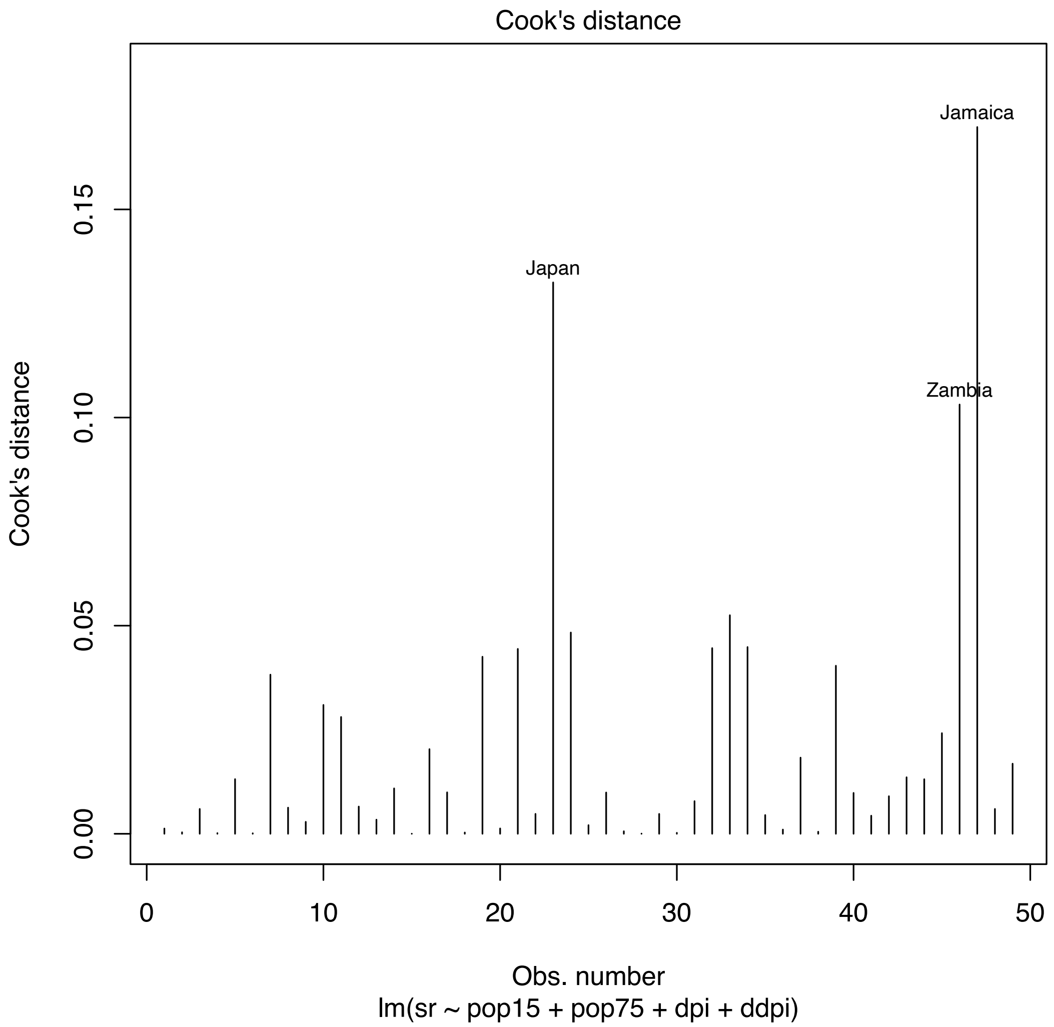
```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings[-weli
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0699 -2.5408 -0.1584  2.0934  9.3732
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.5240460  8.2240263   2.982  0.00465 **
## pop15       -0.3914401  0.1579095  -2.479  0.01708 *
## pop75       -1.2808669  1.1451821  -1.118  0.26943
## dpi         -0.0003189  0.0009293  -0.343  0.73312
## ddpi         0.6102790  0.2687784   2.271  0.02812 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.795 on 44 degrees of freedom
## Multiple R-squared:  0.3554,Adjusted R-squared:  0.2968
## F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

```
par(mfrow = c(2, 2))
resplot(fit1)
```



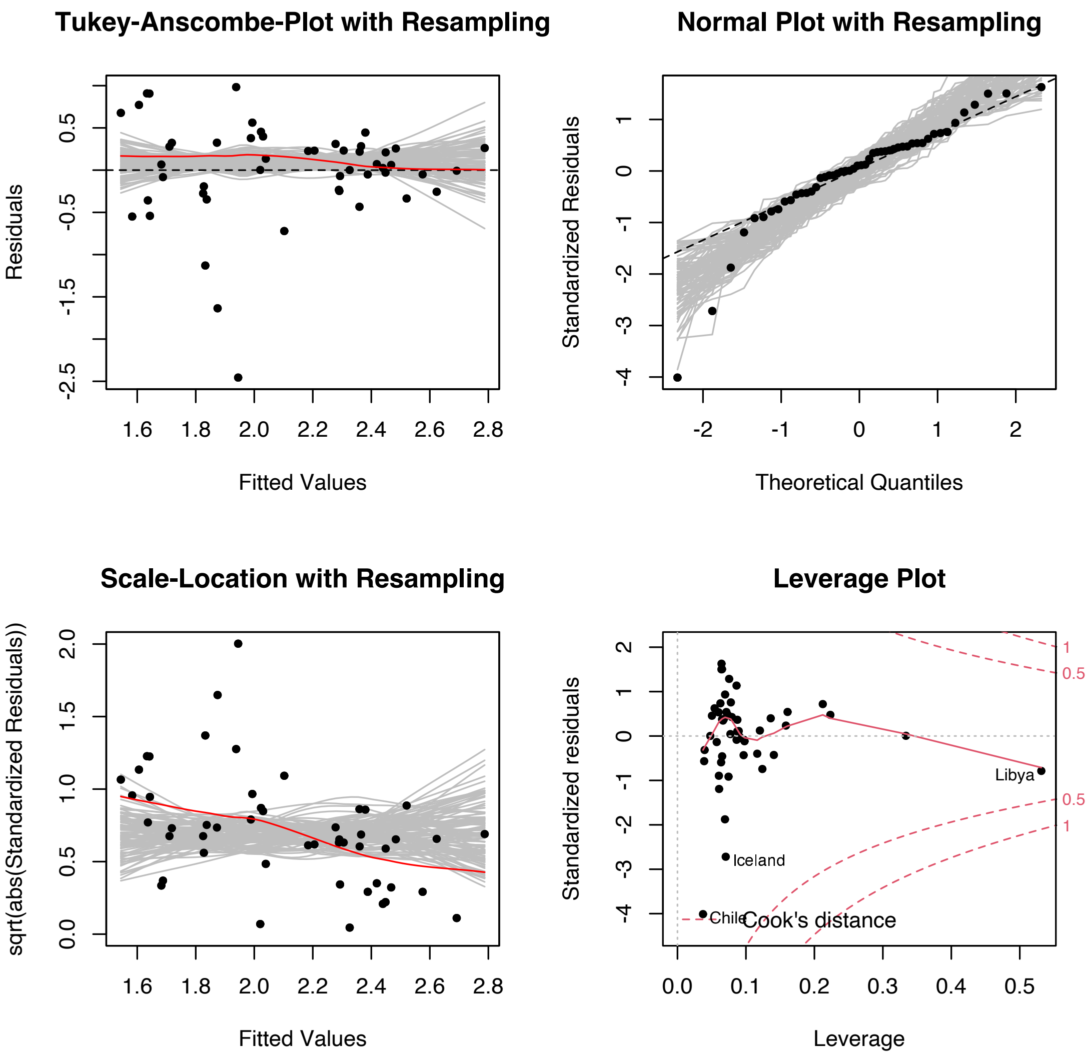
```r
par(mfrow = c(1, 1))
plot(fit1, 4, pch = 20)
```



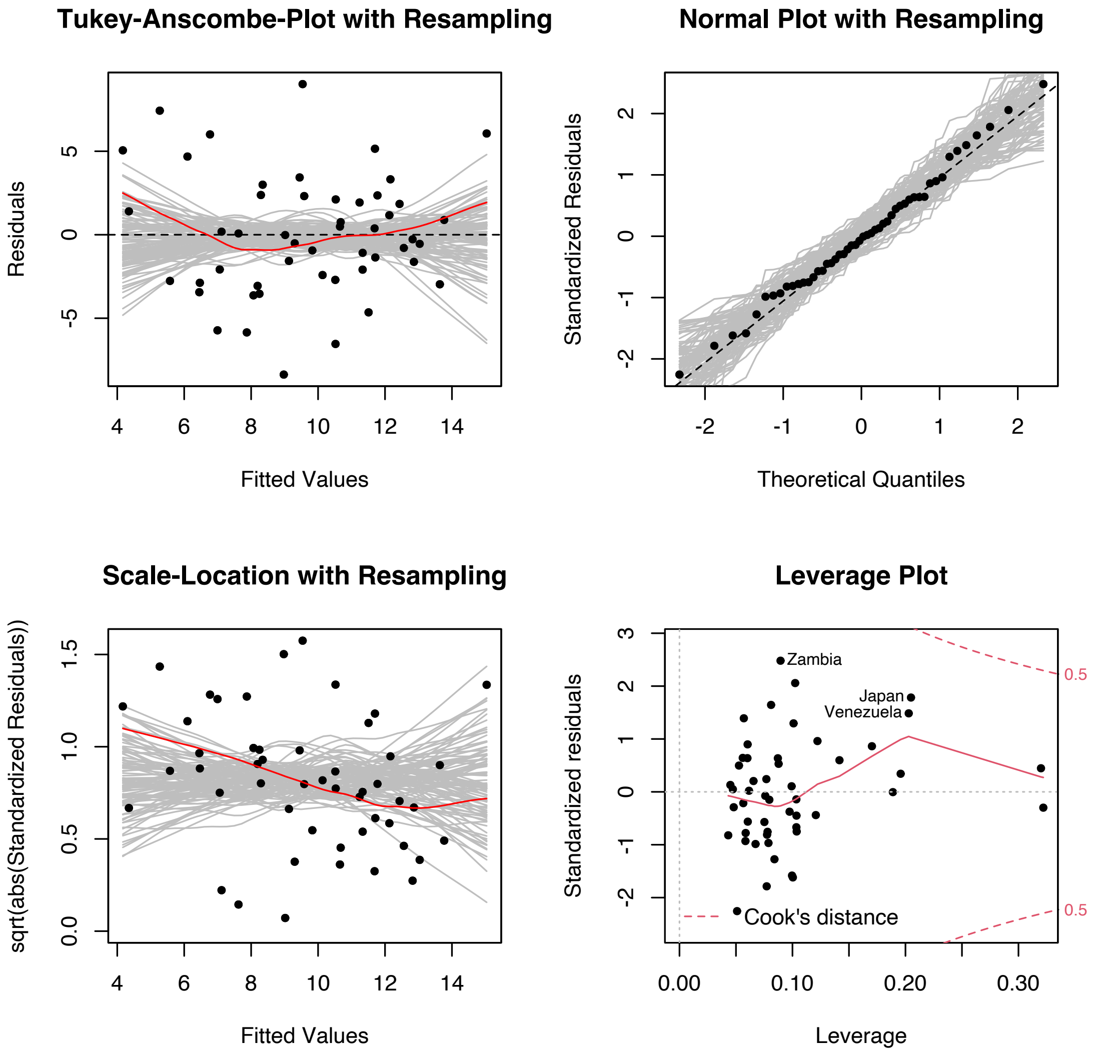
The results only change slightly. The coefficients have similar magnitudes and the same predictors are significant. Also the residual analysis does not yield entirely new insights. This is not surprising as we have seen that Libya does not have a large influence, even though its leverage is high.

d) A transformation may be appropriate for the response variable **sr** as well as for **dpi** and **ddpi**. We shall experiment with three different models. In the first one we transform the response but not the predictors. In the second one we transform both predictors but not the response. In the third model, we transform both the response and the predictors.

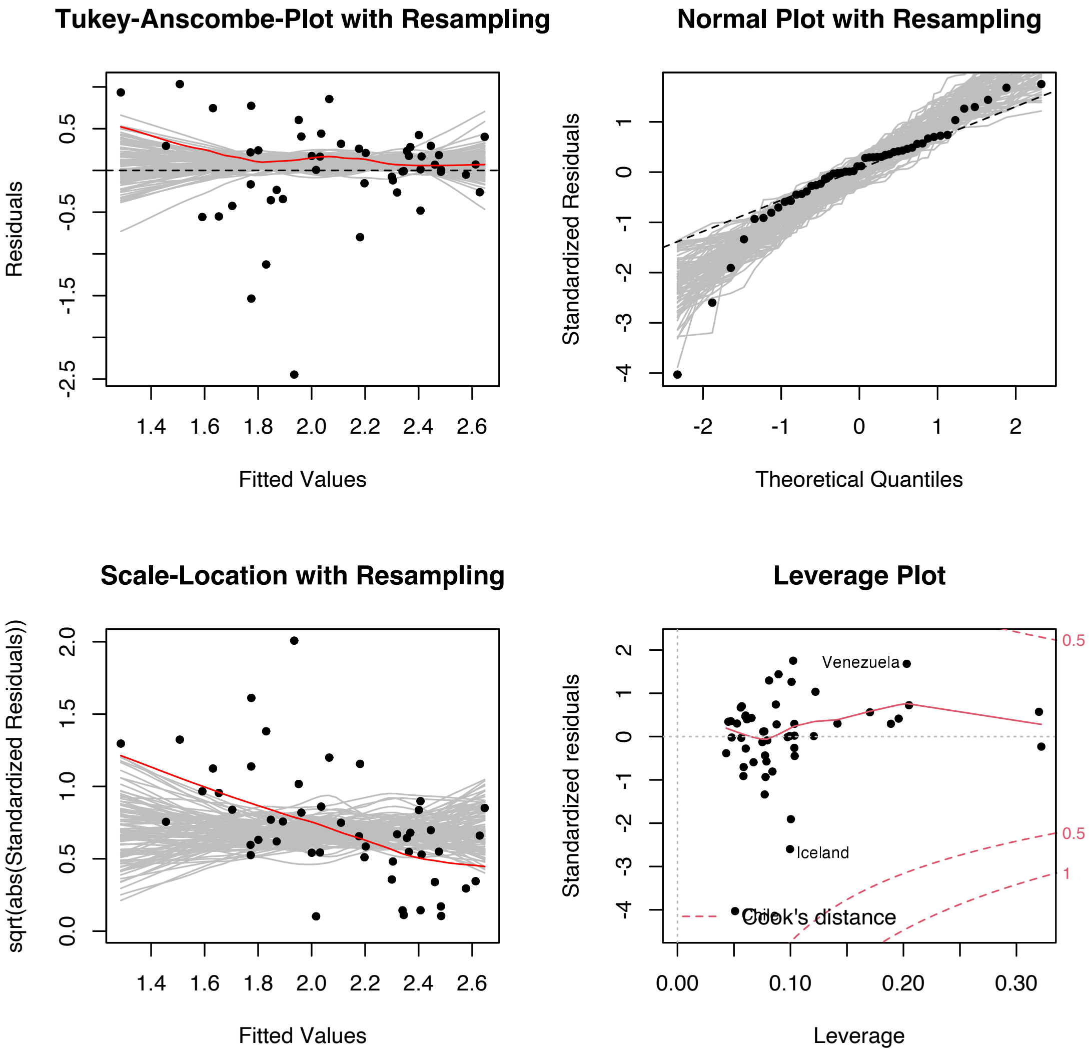
```
## Consider additional models : 1
fit2 <- lm(log(sr) ~ pop15 + pop75 + dpi + ddpi, data = savings)
resplot(fit2)
```



```
## Consider additional models : 2
fit3 <- lm(sr ~ pop15 + pop75 + log(dpi) + log(ddpi), data = savings)
resplot(fit3)
```



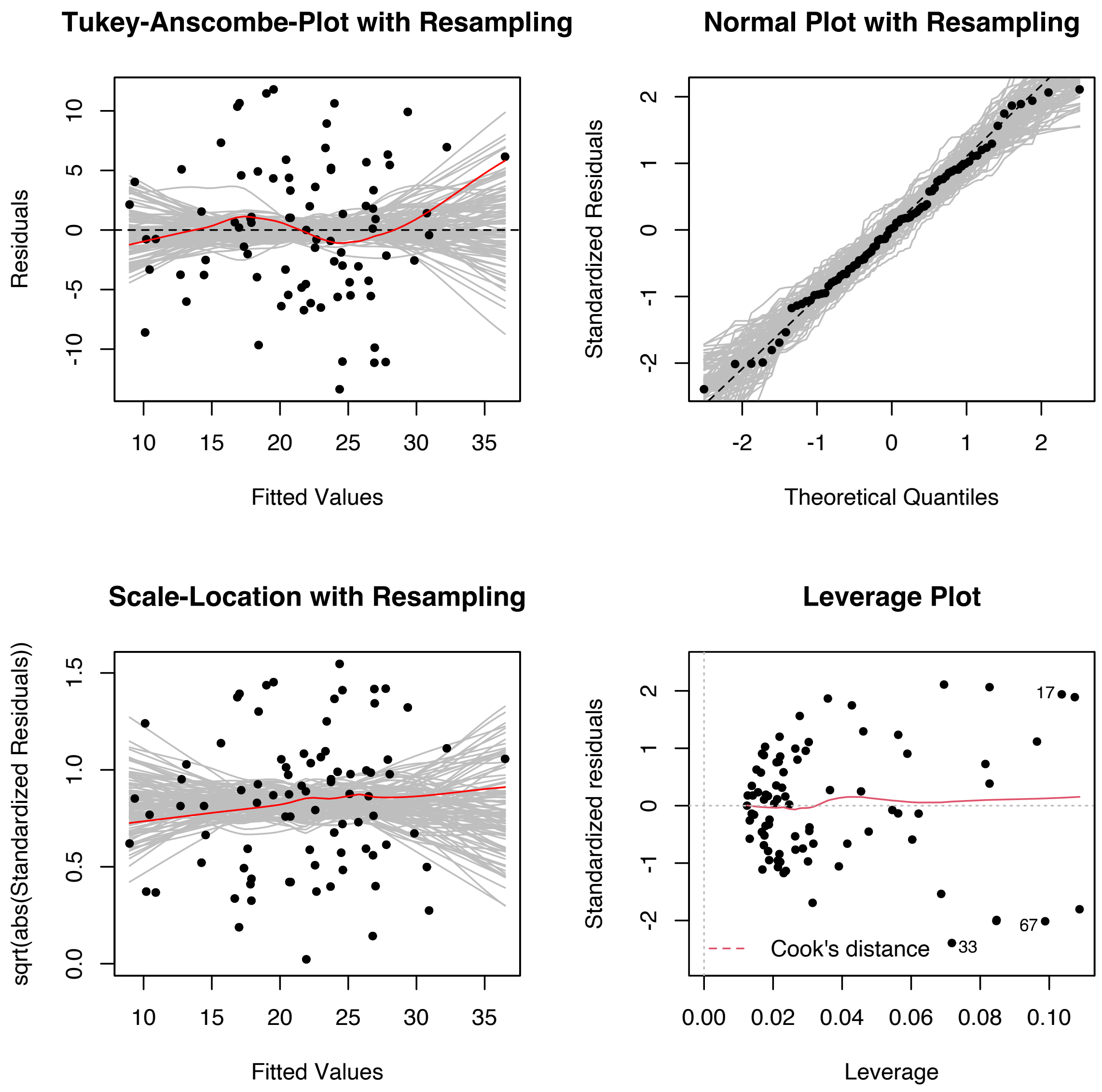
```
## Consider additional models : 3
fit4 <- lm(log(sr) ~ pop15 + pop75 + log(dpi) + log(ddpi),
    data = savings)
resplot(fit4)
```



The residual plots do not show any improvement with respect to the first model. Therefore, we will use the original model without transformations.

Solution 3.3

```
## load data
load("Daten/synthetisch.rda")
source("resplot.R")
## fit
fit <- lm(y ~ x1 + x2, data = synthetisch)
par(mfrow = c(2, 2))
resplot(fit)
```



```
summary(fit)

##
## Call:
## lm(formula = y ~ x1 + x2, data = synthetisch)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -13.3668  -3.8685   0.1167   4.3564  11.8021
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  72.9020     9.6482   7.556 5.96e-11 ***
## x1           -2.0837     0.4882  -4.268 5.37e-05 ***
## x2            1.4258     0.1828   7.802 1.98e-11 ***
## ---
## Signif. codes:  
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.799 on 80 degrees of freedom
## Multiple R-squared:  0.4963,Adjusted R-squared:  0.4837 
## F-statistic: 39.41 on 2 and 80 DF,  p-value: 1.226e-12
```

The residual plots look fine. There is a slight but acceptable deviation in the Tukey-Anscombe plot and a slight but as well acceptable indication of increasing variance in the scale-location plot.

The 3D plot contradicts the above conclusion. The majority of the data points scatter around a plane that is completely different from the least squares solution. In addition, there are a few high leverage points that lie outside the point cloud on both sides of the regression plane. The least squares fit tries to accommodate all observations as good as possible. This results in the residuals having a similar size, i.e. the least squares plane does not fit well anywhere but it does not fit badly anywhere, either. Robust methods could account for this and weigh influential observations less strongly.

This example shows that the least squares solution can yield bad results which are not detected by the model diagnostics tools. This situation can occur when outliers occur in groups instead of as single observations. In such a setting, Cook's distance does not work reliably as a measure for influential data points as leaving out a single observation does not change much (and Cook's distance is based on this change).

While this example has been constructed artificially and is certainly an extreme one, it does show that robust methods constitute an important additional tool in regression analysis. Unfortunately, studying these methods is beyond the scope of this course.

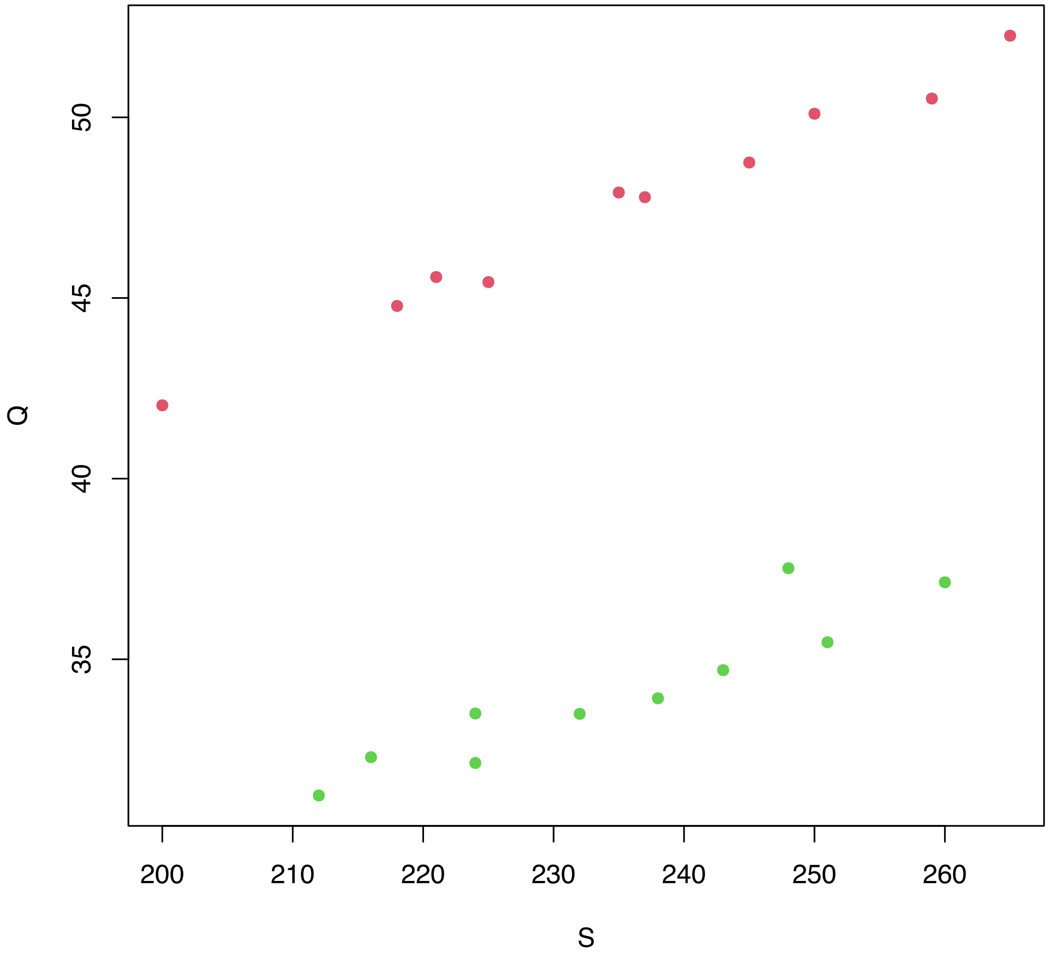
Solution 3.4

a)

```
data <- read.table("Daten/lathe.dat", sep = "", header = TRUE)
summary(data)
```

```
##        Q               S             CTT           
##  Min.   :31.23   Min.   :200.0   Length:20         
##  1st Qu.:33.81   1st Qu.:223.2   Class :character  
##  Median :39.77   Median :236.0   Mode  :character  
##  Mean   :40.83   Mean   :235.2                     
##   3rd Qu.:47.82    3rd Qu.:248.5  
##   Max.   :52.26    Max.   :265.0
```

```
plot(Q ~ S, data = data, col = as.integer(as.factor(CTT)) +
    1, pch = 16)
```



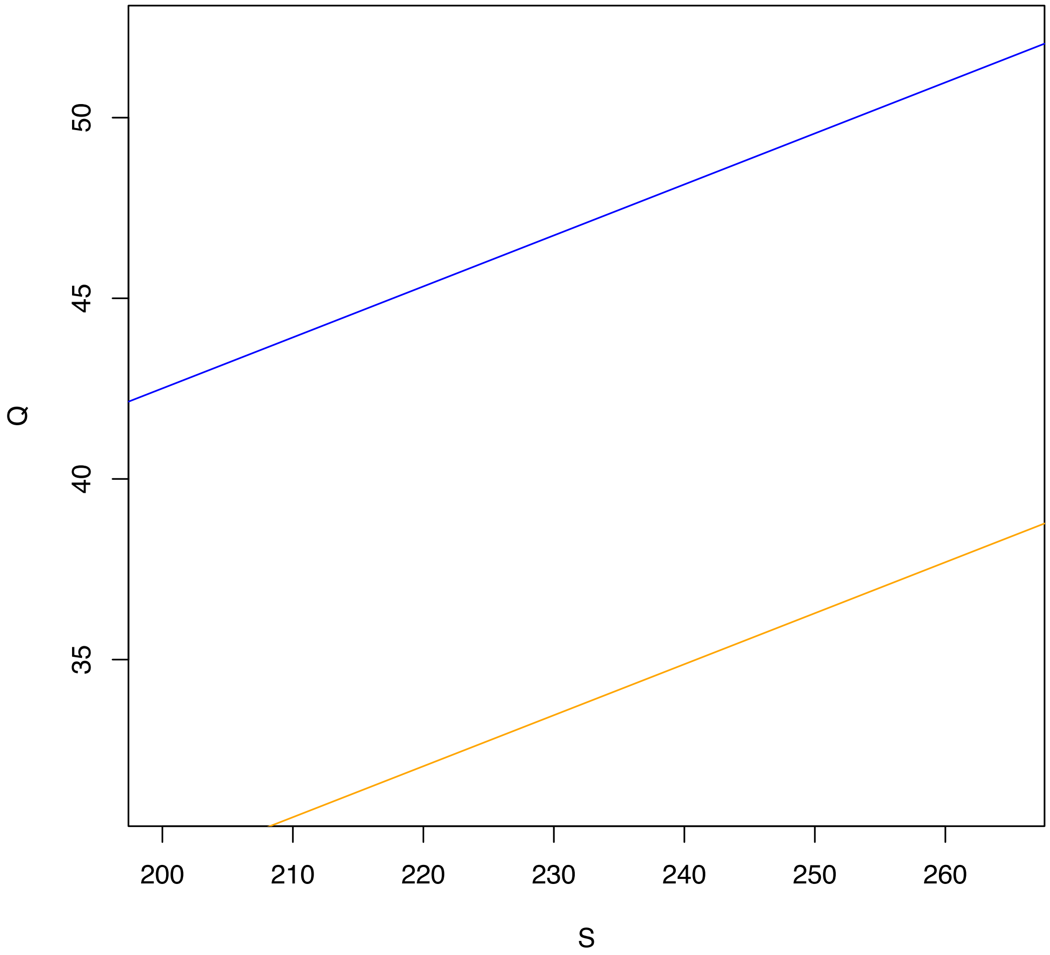
The data points scatter around two almost parallel straight lines, depending on the used cutting tool.

b)
```
data.lm1 <- lm(Q ~ S + CTT, data = data)
summary(data.lm1)
```

```
## 
## Call:
## lm(formula = Q ~ S + CTT, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9546 -0.5039 -0.1804  0.4893  1.5188
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  14.276196   2.091214   6.827 2.94e-06 ***
## S             0.141150   0.008833  15.979 1.13e-11 ***
## CTTDM416    -13.280195   0.302879 -43.847  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6771 on 17 degrees of freedom
## Multiple R-squared:  0.9924,Adjusted R-squared:  0.9915
## F-statistic:  1104 on 2 and 17 DF,  p-value: < 2.2e-16
```

```
plot(Q ~ S, data = data, col = as.integer(CTT) + 1, pch = 16)
abline(coef(data.lm1)[1:2], col = "blue")
abline(coef(data.lm1)[1:2] + c(coef(data.lm1)[3], 0), col = "orange")
```



Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where Y is the surface finish, X_1 is the lathe speed in revolutions per minute, and X_2 is an indicator variable denoting the type of cutting tool used; that is,

$$X_2 = \begin{cases} 0 & \text{for tool type 302} \\ 1 & \text{for tool type 416} \end{cases}$$

The parameters in this model may be easily interpreted. If $X_2 = 0$, the model becomes

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

which is a straight-line model with slope β_0 and intercept β_1. However, if $X_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416

c)

```
data.lm2 <- lm(Q ~ S * CTT, data = data)
summary(data.lm2)

##
## Call:
## lm(formula = Q ~ S * CTT, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68655 -0.44881 -0.07609  0.30171  1.76690
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.50294    2.50430   4.593   0.0003 ***
## S            0.15293    0.01060  14.428 1.37e-10 ***
## CTTDM416    -6.09423    4.02457  -1.514   0.1495
## S:CTTDM416  -0.03057    0.01708  -1.790   0.0924 .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6371 on 16 degrees of freedom
## Multiple R-squared:  0.9936,Adjusted R-squared:  0.9924
## F-statistic: 832.3 on 3 and 16 DF,  p-value: < 2.2e-16
```

The last part S:CTTDM416 is the so called interaction between the variables `S` and `CTT`. Since `CTT` is a dummy variable, it means that `S` has a -0.0306 times larger (which means a 0.0306 times smaller) slope for `CTT`=DM416.

It is also possible to use indicator variables to investigate whether tool type affects both the slope and intercept. Let the model be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \cdot X_2 + \epsilon$$

where X_2 is the indicator variable. Now if tool type 302 is used, $X_2 = 0$, and the model is

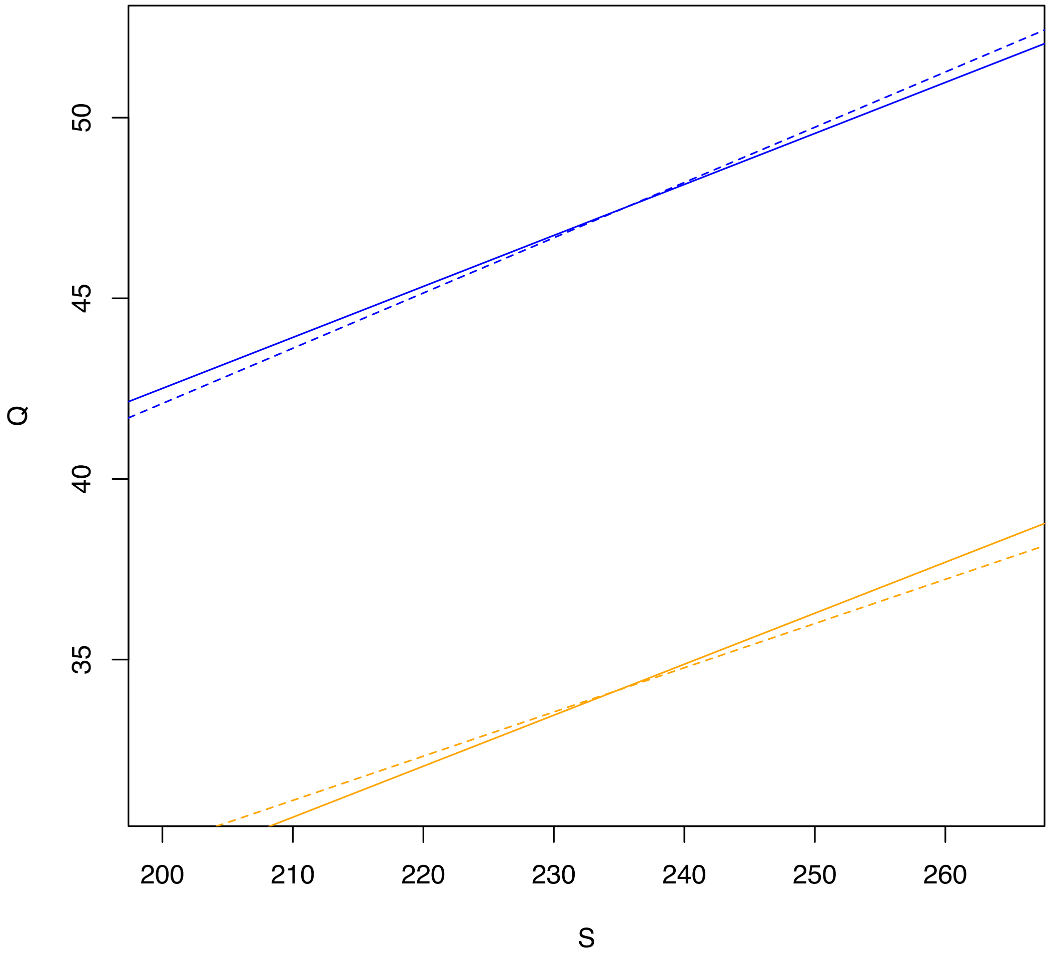
$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

If tool type 416 is used, $X_2 = 1$, and the model becomes

$$Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon$$

Note that β_2 is the change in the intercept and that β_3 is the change in slope produced by a change in tool type.

```r
plot(Q ~ S, data = data, col = as.integer(CTT) + 1, pch = 16)
abline(coef(data.lm1)[1:2], col = "blue")
abline(coef(data.lm1)[1:2] + c(coef(data.lm1)[3], 0), col = "orange")
abline(coef(data.lm2)[1:2], col = "blue", lty = 2)
abline(coef(data.lm2)[1:2] + coef(data.lm2)[3:4], col = "orange",
    lty = 2)
```



The differences are visible but not very large.

d) Since the p-value for the differences of the slopes is 0.0924, the difference is not significant at the 5 % level, although it would be significant at the 10 % level.

We may have commited a type II error, which means that the null hypothesis is retained although the alternative hypothesis is correct. Since we do not have any information about the alternative hypothesis (we only know the value is not equal 0), we cannot calculate the distribution of the test statistic assuming the alternative is true and that is why we cannot provide more information about the probability of a type II error than it lies between 0 and 1.

Solution 3.5

a) First, we check the structure of the data frame:

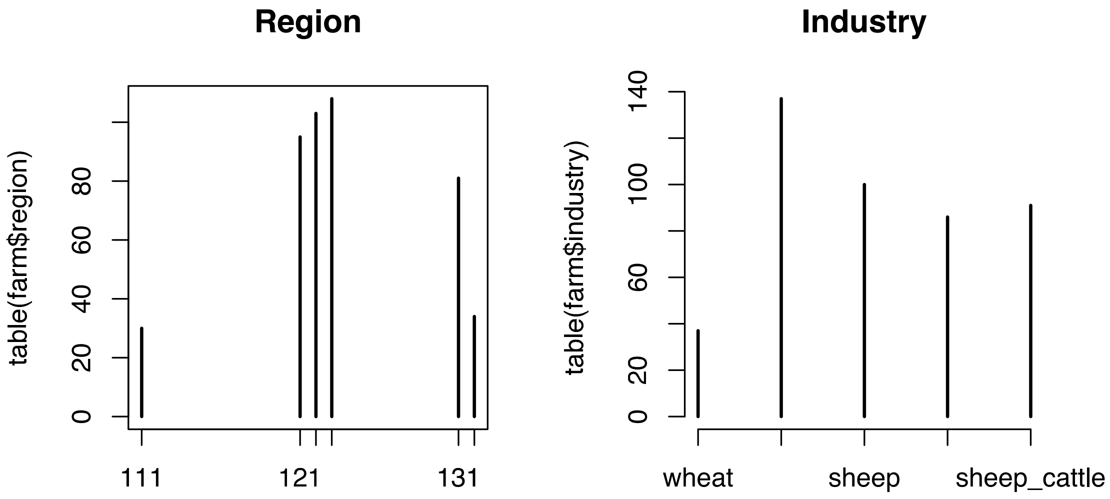
```
## Load data
load("Daten/farm.rda")
## Check properties of the data
str(farm)

## 'data.frame': 451 obs. of  4 variables:
##  $ region  : int  111 111 111 111 111 111 111 111 111 111 ...
##  $ industry: int  3 5 2 1 2 5 2 3 3 3 ...
##  $ costs   : int  115096 75443 378857 433590 347417 327745 714462
##  $ revenue : int  147652 82920 442726 649628 407836 472569 576372
```

All variables are of data type *int*. This is incorrect for the factor variables `region` and `industry` and would lead to incorrect regression results. We define the factor variables as follows:

```
farm$region <- factor(farm$region)
farm$industry <- factor(farm$industry, labels = c("wheat",
    "wheat_sheep_cattle", "sheep", "cattle", "sheep_cattle"))
```

For the advanced reader: We now check whether there are sufficiently many observations for all levels of the factor variables. It is recommended that there are at least five observations for each level.

```
## Visualization
par(mfrow = c(1, 2))
plot(table(farm$region), main = "Region")
plot(table(farm$industry), main = "Industry")
```

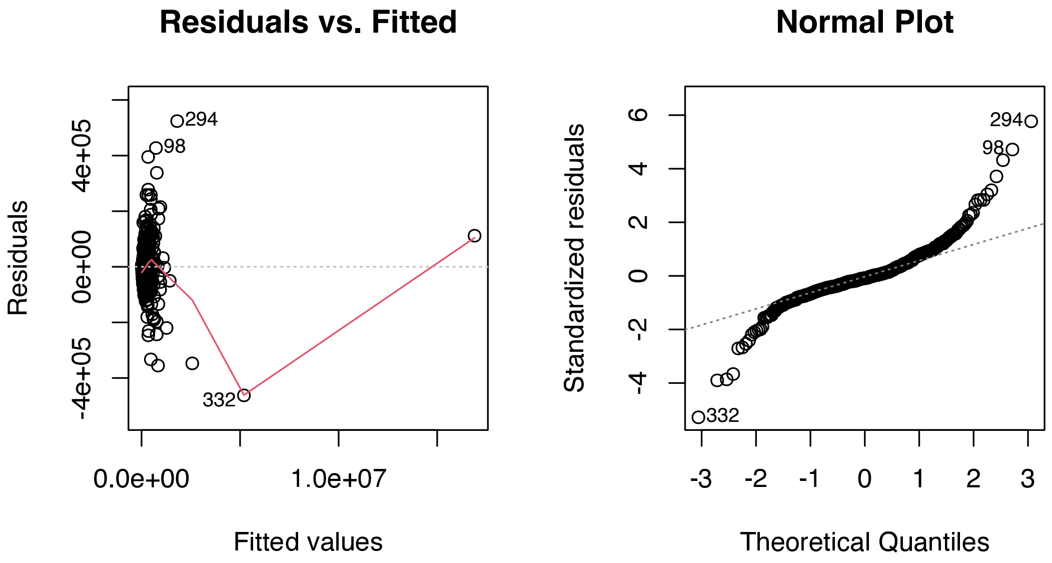


The number of observations are sufficient for all levels of the factor variables.

b)

```
## Fit the complete regression model
fit1 <- lm(revenue ~ costs + region + industry, data = farm)
```

```
## Residual analysis
par(mfrow = c(1, 2))
plot(fit1, which = 1, caption = "", main = "Residuals vs. Fitted")
plot(fit1, which = 2, caption = "", main = "Normal Plot")
```



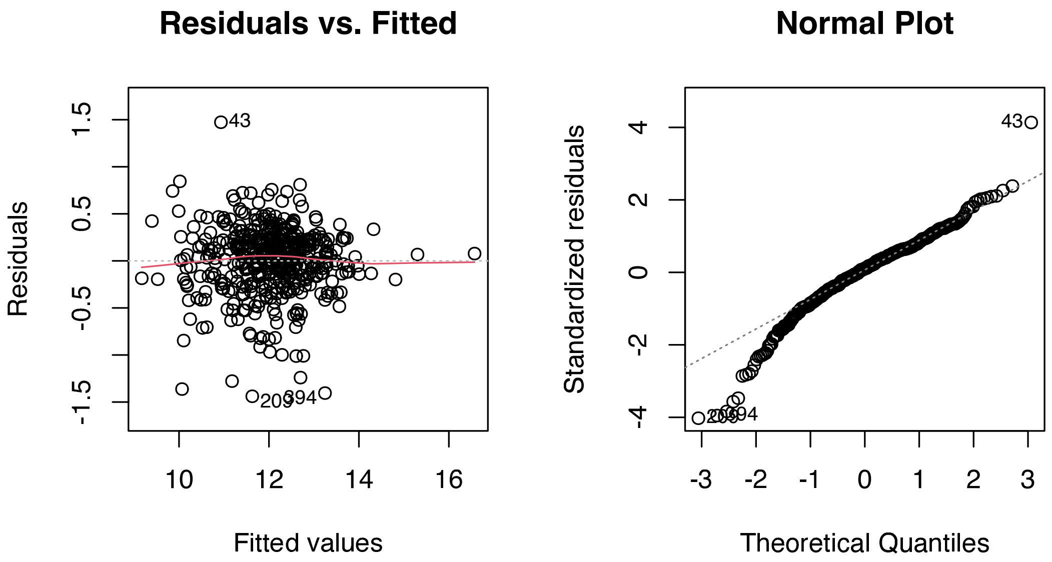
The Tukey-Anscombe plot and the normal plot clearly show that the model assumptions are violated. A variable transformation is absolutely necessary.

We log-transform the predictor variable **costs** and the response variable **revenue**.

```
## Fit the complete regression model
fit <- lm(log(revenue) ~ log(costs) + region + industry,
    data = farm)
summary(fit)

##
## Call:
## lm(formula = log(revenue) ~ log(costs) + region + industry, data =
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43881 -0.17143  0.03773  0.22168  1.47317
```

```
##
## Coefficients:
##                              Estimate Std. Error
## (Intercept)                  1.379636   0.248432
## log(costs)                   0.917954   0.018617
## region121                   -0.076883   0.077353
## region122                   -0.082997   0.076912
## region123                   -0.036680   0.076151
## region131                   -0.003855   0.079775
## region132                   -0.243938   0.100536
## industrywheat_sheep_cattle -0.155614   0.068023
## industrysheep               -0.222879   0.071421
## industrycattle               0.002649   0.075844
## industrysheep_cattle        -0.171106   0.072947
##                            t value Pr(>|t|)
## (Intercept)                  5.553 4.86e-08 ***
## log(costs)                  49.306  < 2e-16 ***
## region121                   -0.994  0.32081
## region122                   -1.079  0.28113
## region123                   -0.482  0.63027
## region131                   -0.048  0.96148
## region132                   -2.426  0.01565 *
## industrywheat_sheep_cattle  -2.288  0.02263 *
## industrysheep               -3.121  0.00192 **
## industrycattle               0.035  0.97215
## industrysheep_cattle        -2.346  0.01944 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3612 on 440 degrees of freedom
## Multiple R-squared:  0.8712,Adjusted R-squared:  0.8683
## F-statistic: 297.7 on 10 and 440 DF,  p-value: < 2.2e-16
```

```
## Residual analysis
par(mfrow = c(1, 2))
plot(fit, which = 1, caption = "", main = "Residuals vs. Fitted")
plot(fit, which = 2, caption = "", main = "Normal Plot")
```



The Tukey-Anscombe plot does not indicate the presence of any systematic error. The normal plot shows that the distribution of the residuals is skewed to the left and there is one large positive outlier no. 43). In summary, the assumptions seem to be fulfilled to a sufficient degree but not entirely.

c) Using `predict()` we obtain the prediction on the log scale. We thus need to transform the value back to the original scale. So the predicted `revenue` is 154 914.2 Dollar.

```
## predict
newdat <- data.frame(costs = 10^5, region = "111", industry = "cattle
predi <- predict(fit, newdata = newdat)
exp(predi)

##        1
## 154914.2
```

For the advanced reader: In order to obtain the expected `revenue`, we need the following correction:

```
## predict
newdat <- data.frame(costs = 10^5, region = "111", industry = "cattle
predi <- predict(fit, newdata = newdat)
exp(predi)

##        1
## 154914.2
```

```
exp(predi + 0.5 * summary(fit)$sigma^2)

##        1
## 165357.7
```

So the expected **revenue** is 165 357.7 Dollar.

d)

```
drop1(fit, test = "F")

## Single term deletions
##
## Model:
## log(revenue) ~ log(costs) + region + industry
##            Df Sum of Sq    RSS     AIC   F value
## <none>                   57.41 -907.62
## log(costs)  1    317.21 374.62  -63.69 2431.0923
## region      5      1.36  58.77 -907.03    2.0906
## industry    4      2.77  60.18 -894.39    5.3007
##               Pr(>F)
## <none>
## log(costs) < 2.2e-16 ***
## region     0.0655074 .
## industry   0.0003542 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The predictor variable **region** is not significant as can be seen from the p-value 0.0655 of the partial F-test.

e)

- 31 parameters are estimated since the model has 420 degrees of freedom and there are 451 observations.
- We have sufficiently many observations since there are more than five observations for every estimated parameter.
- To test the interaction term we need to do a partial F-test. We could do this explicitly with the **R**-function **anova()** but using **drop1()** is more convenient:

```
## Option 1
f.big <- lm(log(revenue) ~ log(costs) + region + industry +
    region * industry, data = farm)
f.small <- lm(log(revenue) ~ log(costs) + region + industry,
    data = farm)
```

```
anova(f.small, f.big)

## Analysis of Variance Table
##
## Model 1: log(revenue) ~ log(costs) + region + industry
## Model 2: log(revenue) ~ log(costs) + region + industry + regio
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    440 57.411
## 2    420 54.540 20    2.8706 1.1053 0.3404
```

```
# Option 2
drop1(f.big, test = "F")

## Single term deletions
##
## Model:
## log(revenue) ~ log(costs) + region + industry + region * indus
##                 Df Sum of Sq    RSS     AIC   F value
## <none>                         54.54 -890.75
## log(costs)        1   303.467 358.01  -44.14 2336.9109
## region:industry 20     2.871  57.41 -907.62    1.1053
##                 Pr(>F)
## <none>
## log(costs)      <2e-16 ***
## region:industry 0.3404
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction term is not significant and can be excluded from the model.

- The interaction term can be interpreted in an intuitive way as follows: **`region`** and **`industry`** do not influence **`revenue`** independently and additively, but the influence of **`industry`** differs between regions. However, as we have seen this is not the case for this data set.

f) The interaction term is not significant as we have seen above. So we exclude it and are left with the main effects model (model without interaction term). As we have concluded for the main effects model, the predictor variable **`region`** is not significant, so we will exclude it as well. Thus, the log-transformed response variable **`revenue`** is predicted by the log-transformed predictor variable **`costs`** and by the predictor **`industry`**. In this model both predictors are significant. Hence, we select this model.