Solution to Series 5

1. **a)** From the plots below we can derive the following:

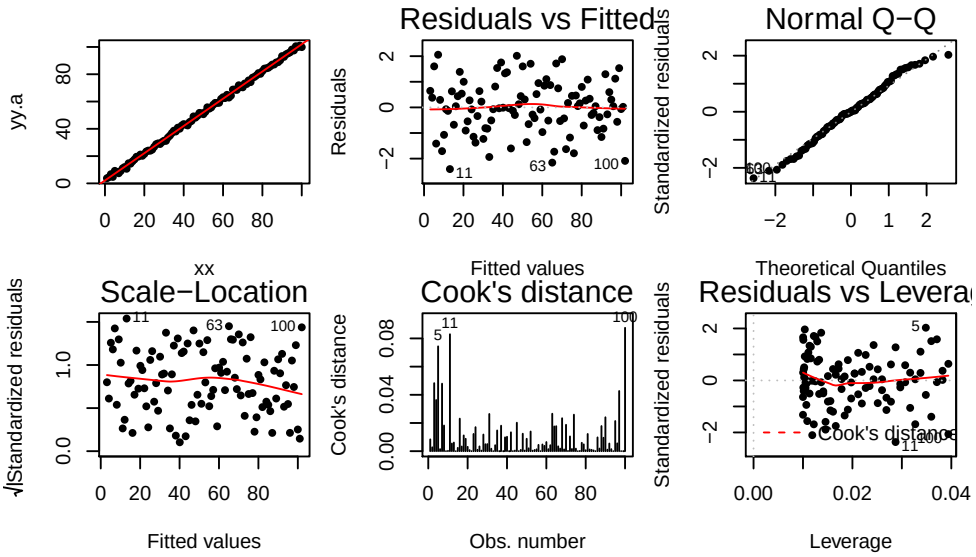
.a Model assumptions valid.

.b Model contains strong non-constant variance.

.c Variance slightly non-constant.

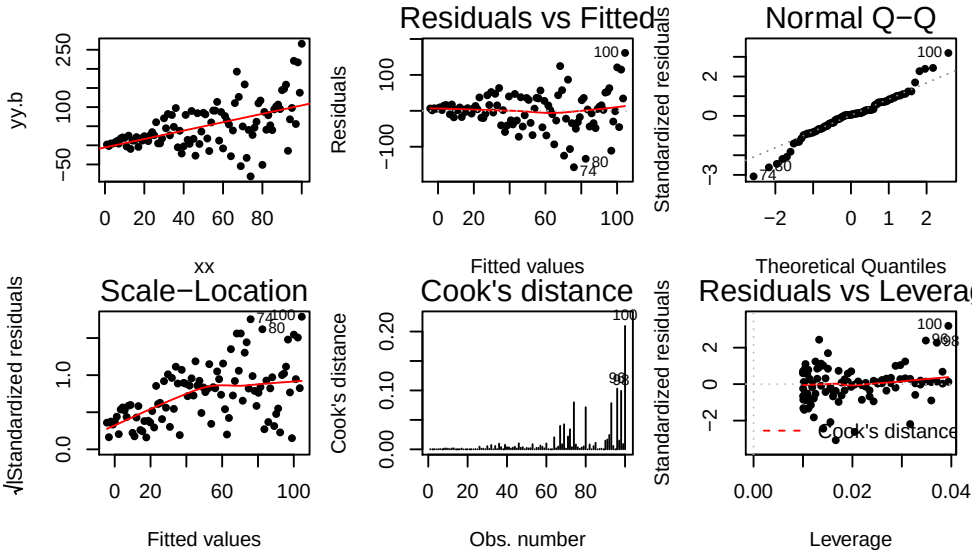
.d Non-linear model (linear model shows systematic error).

```
> ## yy.a: scatter plots, residuals and Cook's Distance
> par(mfrow=c(2,3))
> plot(yy.a ~ xx, pch=20)
> abline(fit <- lm(yy.a ~ xx), col="red")
> plot(fit,1:5,pch=20)
```



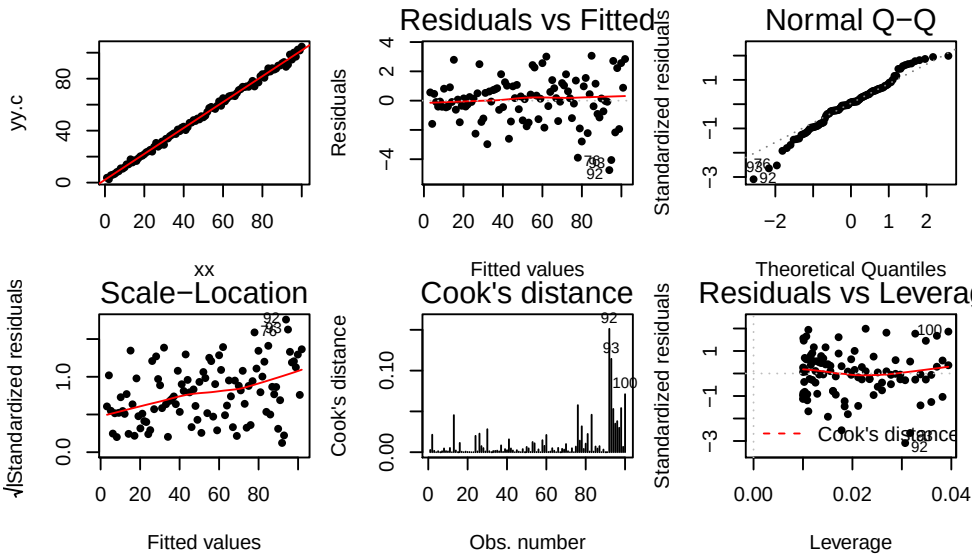
yy.a: For the first model the residual plots look perfect. Only in the plot containing Cook's distance, there are a few values that are slightly larger than the rest. These are the observations with the smallest/largest x-values. However, since those values are far from 0.5, there is no problem.

```
> ## yy.b: scatter plots, residuals and Cook's Distance
> par(mfrow=c(2,3))
> plot(yy.b ~ xx, pch=20)
> abline(fit <- lm(yy.b ~ xx), col="red")
> plot(fit,1:5,pch=20)
```



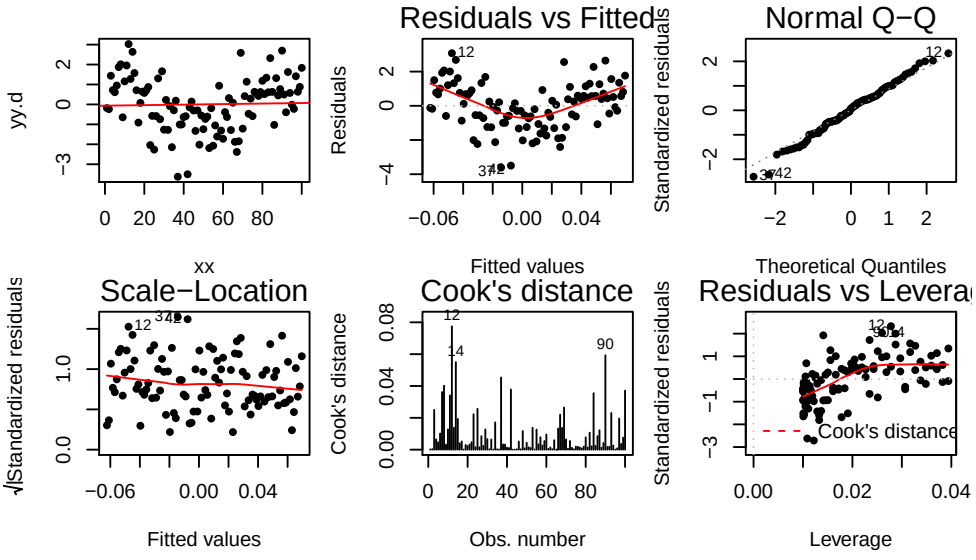
`yy.b`: In case of the second model, we see the increasing variance with the magnitude of the fitted values in the Tukey-Anscombe-Plot. The Normal plot shows a violation of the normality assumption, even though the errors do follow a Normal distribution per definition. However, the variance is not constant which also needs to be fulfilled for the Normal plot (so that the points follow a straight line). So the violation stems from the fact that the variance is not constant. In the scale-location plot we can also see the increase in the variance. There are no leverage points nor influential data points – even though the points with large observation numbers have larger values of Cook's distance.

```
> ## yy.c: scatter plots, residuals and Cook's Distance
> par(mfrow=c(2,3))
> plot(yy.c ~ xx, pch=20)
> abline(fit <- lm(yy.c ~ xx), col="red")
> plot(fit,1:5,pch=20)
```



`yy.c`: For the third model, the analysis is similar as in case of the second model. This is the case because the model violations are similar. The model violation is less accentuated than in the previous example.

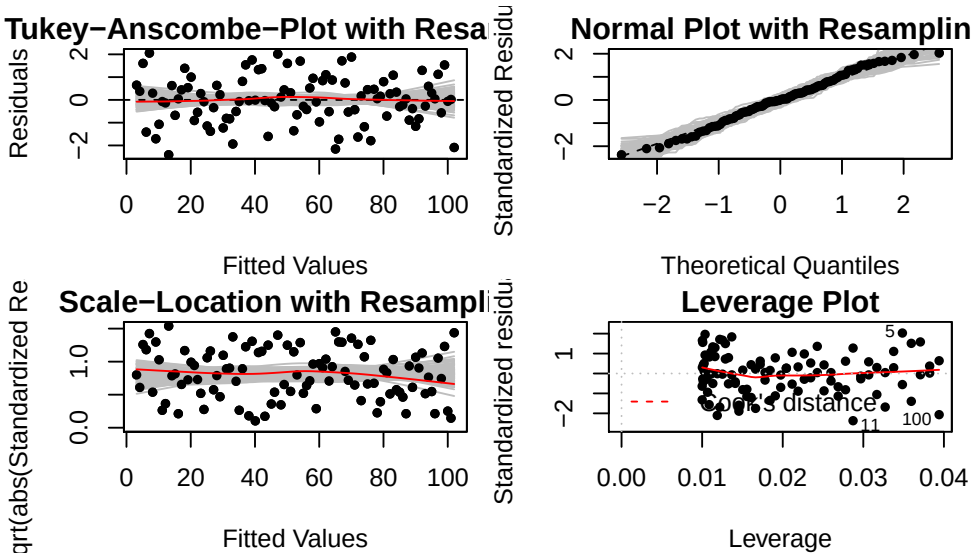
```
> ## yy.d: scatter plots, residuals and Cook's Distance
> par(mfrow=c(2,3))
> plot(yy.d ~ xx, pch=20)
> abline(fit <- lm(yy.d ~ xx), col="red")
> plot(fit,1:5,pch=20)
```



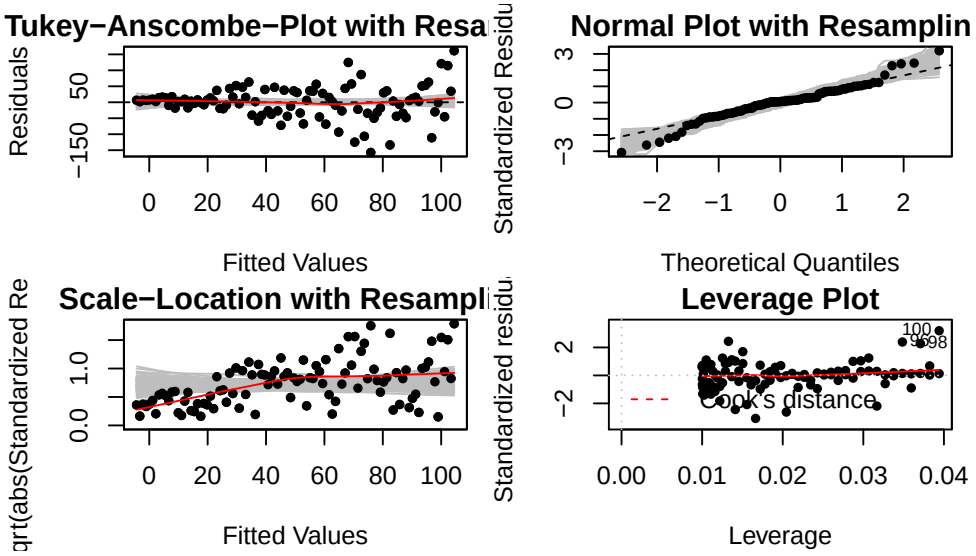
`yy.d`: In case of the fourth model, the systematic error can be easily detected in the Tukey-Anscombe plot since it exhibits a U-shaped pattern. The Normal plot and the scale-location plot do not show any abnormalities. There are no influential data points but the smoother deviates from the horizon in

the leverage plot. This is the case because the points with large leverage (i.e. points at the border of this simple regression) have systematically positive residuals.

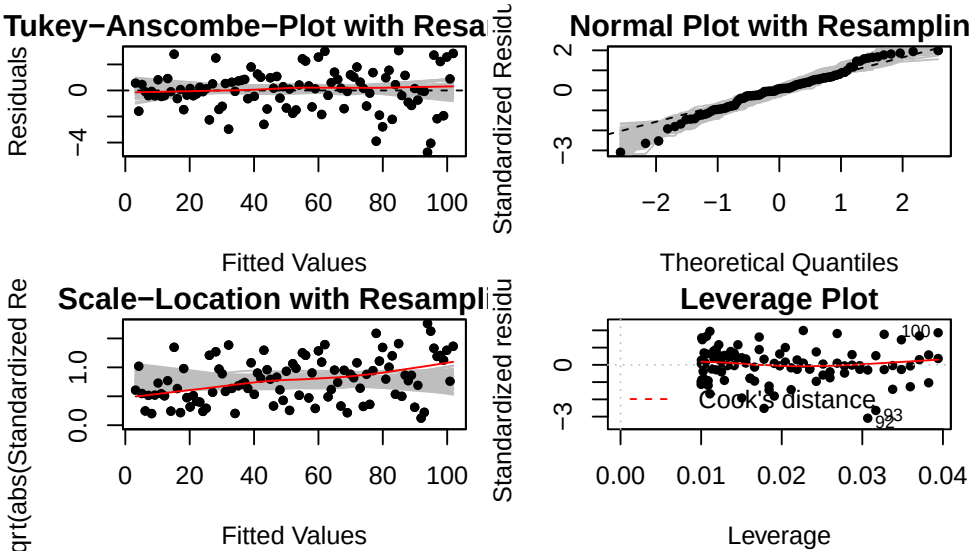
b)
```
> ## source function (needs to be in your working directory)
> source("resplot.R")
> ## yy.a: residual plots with resampling
> par(mfrow=c(2,2))
> fit <- lm(yy.a ~ xx)
> resplot(fit)
```



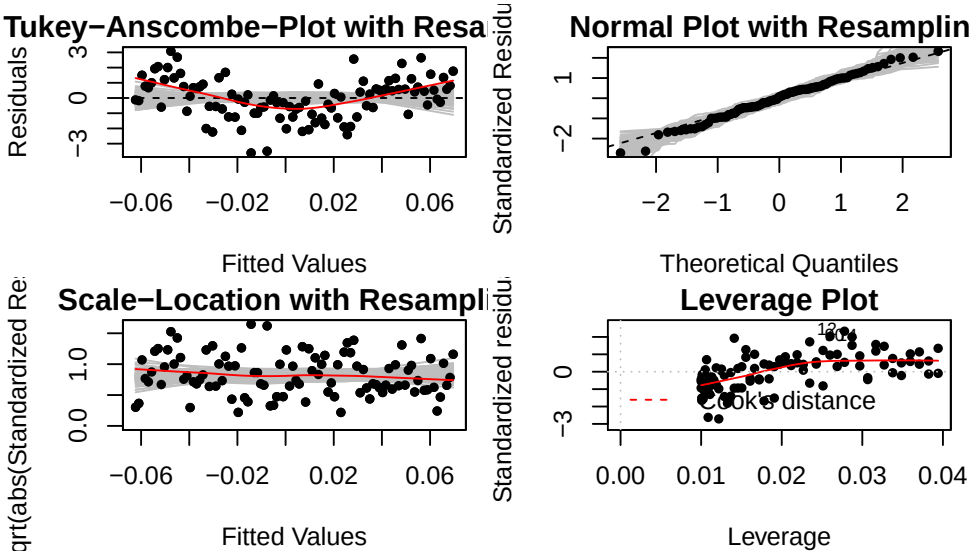

```
> ## yy.b: residual plots with resampling
> fit <- lm(yy.b ~ xx)
> resplot(fit)
```




```
> ## yy.c: residual plots with resampling
> fit <- lm(yy.c ~ xx)
> resplot(fit)
```




```
> ## yy.d: residual plots with resampling
> fit <- lm(yy.d ~ xx)
> resplot(fit)
```

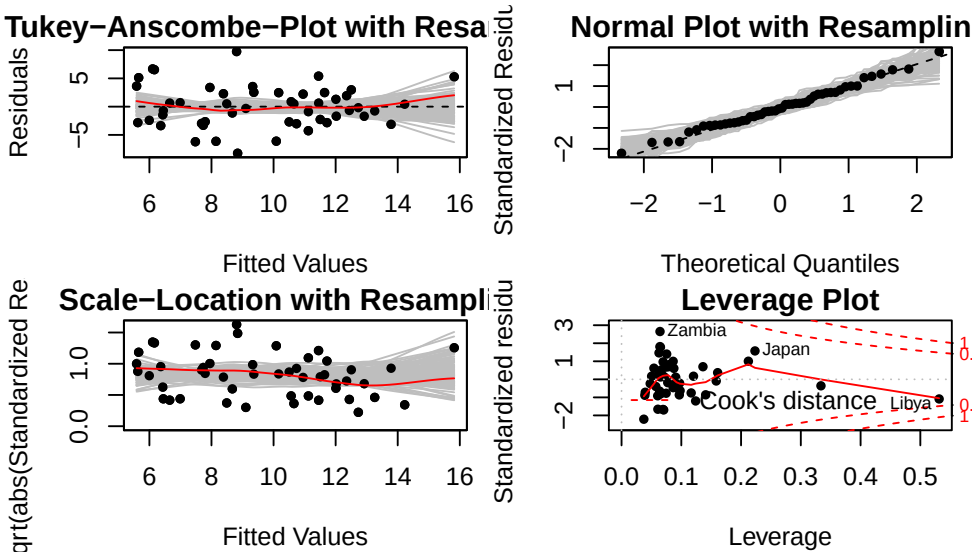


As you can see from the plots, the function does a good job in detecting the three model violations. Additionally, it does not make a mistake "in the other direction", either. I.e. the smoother does not lie outside of the gray area in cases where the model assumptions are fulfilled. In other words, there are no "false positives".

c) The exercise should be repeated generating new random numbers (remember to change the argument of `set.seed` or just eliminate it). Manipulating the number of observations is also instructive. However, the above described structures are of general nature and will largely remain on the repetitions. Detecting model violations is more difficult the fewer observations are present.

2. a)

```
> ## load data
> load("savings.rda")
> source("../ex1/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)
```



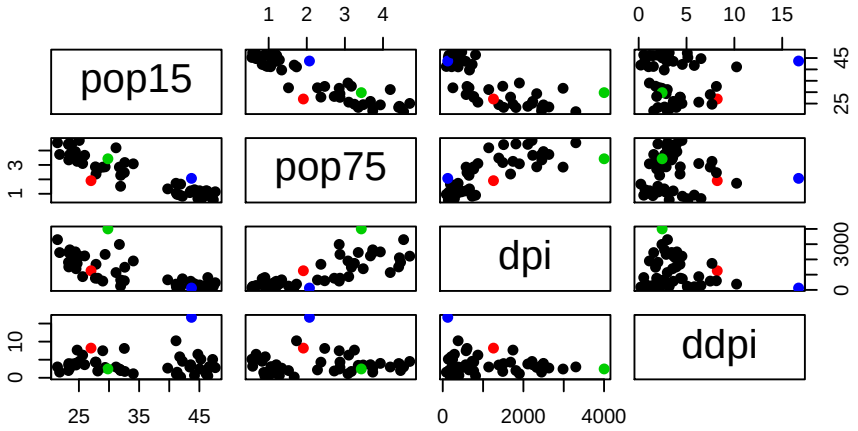
The residuals look ok and the model can be used. There are a few points with large leverage but none of these points is influential as Cook's distance is smaller than 0.5 for all points.

b)

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

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

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

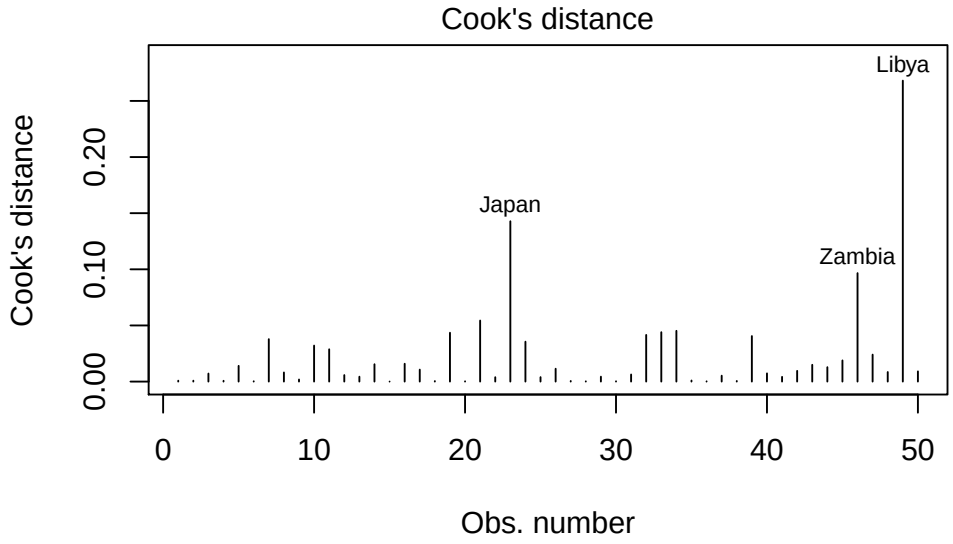


The three countries with the largest leverage are Libya, the USA, and Japan. To simplest way to see why these points have extraordinary predictor configurations is to plot pairwise scatter plots.
In the plots, Japan corresponds to red, USA to green and Libya to blue. The latter has a very low value of `dpi` but a very large value of `ddpi`. The USA have the largest `dpi` value and a relatively large proportion for pop75. Japan, on the other hand, lies at the border in several scatter plots but is not extraordinary with respect to a single feature.

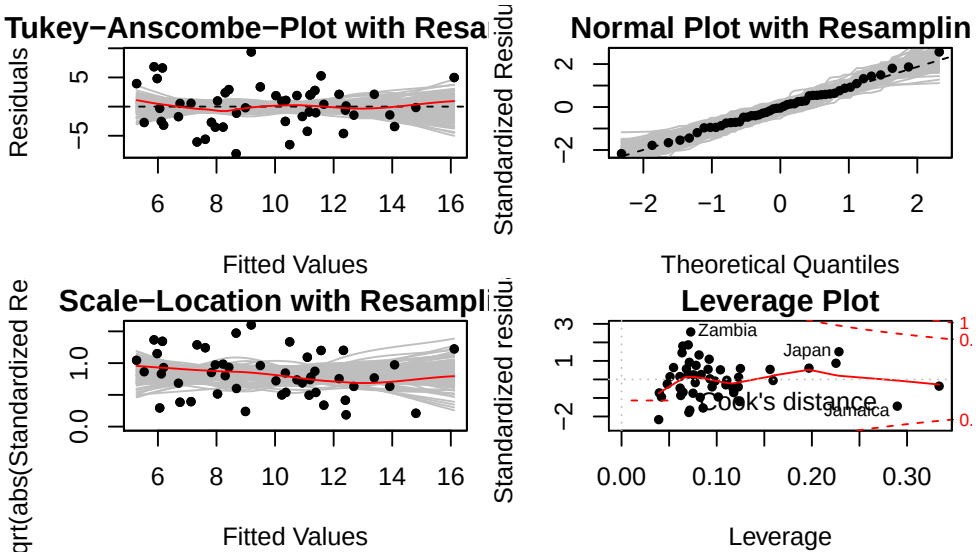
c)

```
> ## 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); coef(fit1)
```

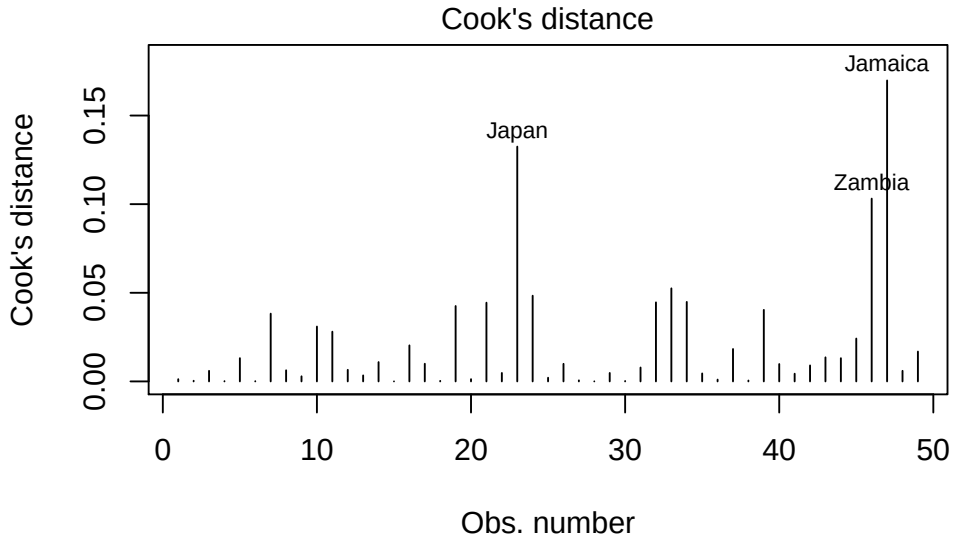
```
  (Intercept)         pop15         pop75           dpi          ddpi
28.5660865407 -0.4611931471 -1.6914976767 -0.0003369019  0.4096949279
  (Intercept)         pop15         pop75           dpi          ddpi
24.5240459788 -0.3914401268 -1.2808669233 -0.0003189001  0.6102790264
> summary(fit); summary(fit1)

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

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)
```




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

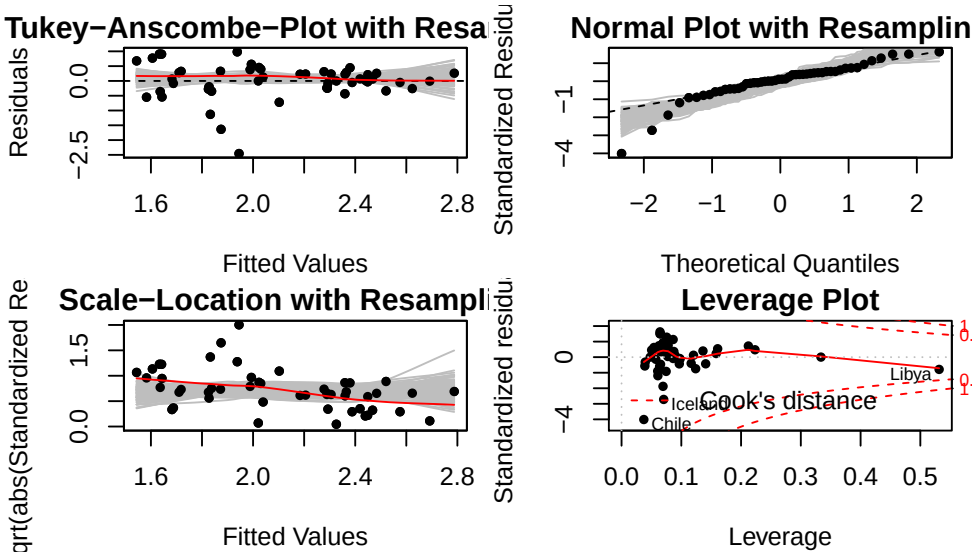


The results only change slightly. The coefficients have similar magnitudes and the same predictors are significant. Also the residual analysis does not yield entirely new insights. This is not surprising as we have seen that Libya does not have a large influence, even though its leverage is high.

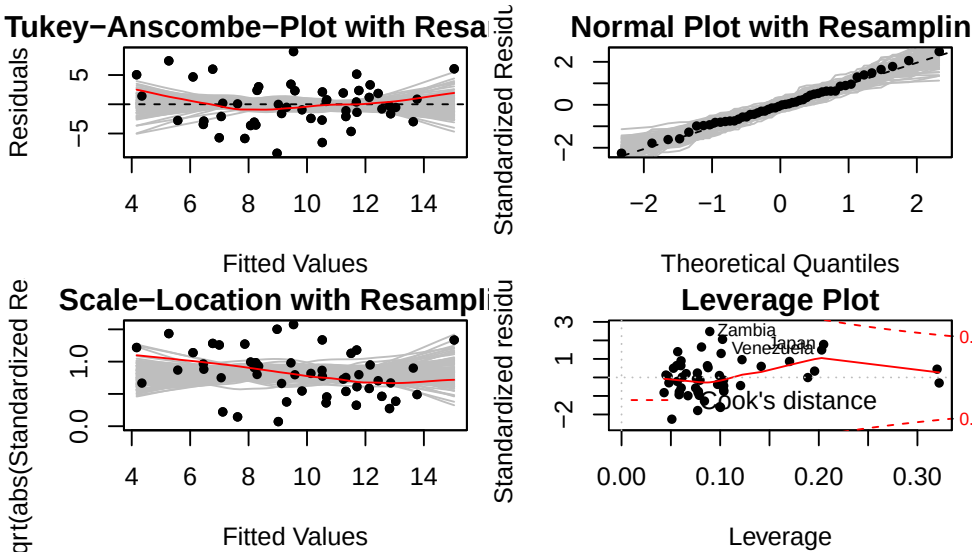
d)
```
> ## consider additional models
> par(mfrow=c(2,3))
> hist(savings$sr, col="limegreen")
> hist(savings$pop15, col="limegreen")
> hist(savings$pop75, col="limegreen")
> hist(savings$dpi, col="limegreen")
> hist(savings$ddpi, col="limegreen")
```

A transformation makes sense for the targest `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.

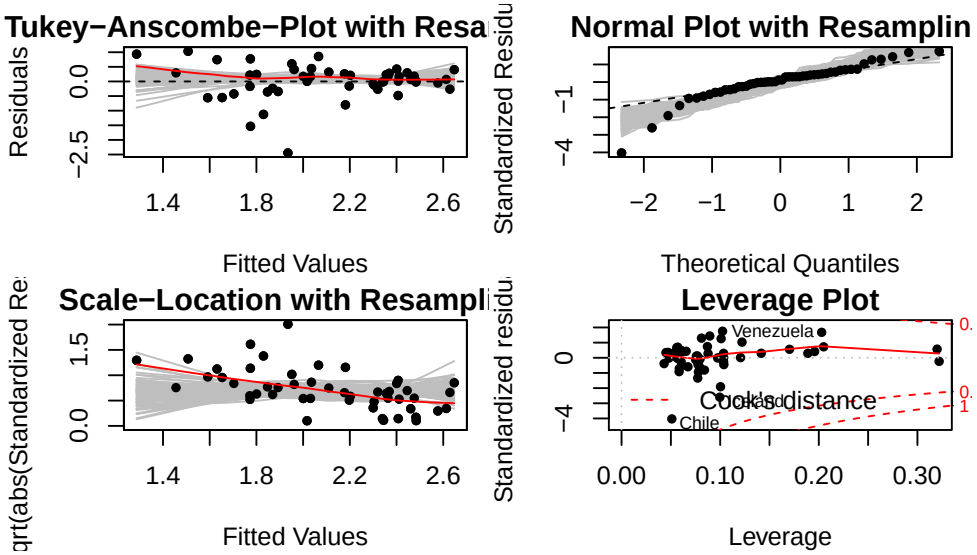
```
> ## consider additional models : 1
> fit2 <- lm(log(sr) ~ pop15 + pop75 + dpi + ddpi, data=savings)
> resplot(fit2)
```




```
> ## 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)
```

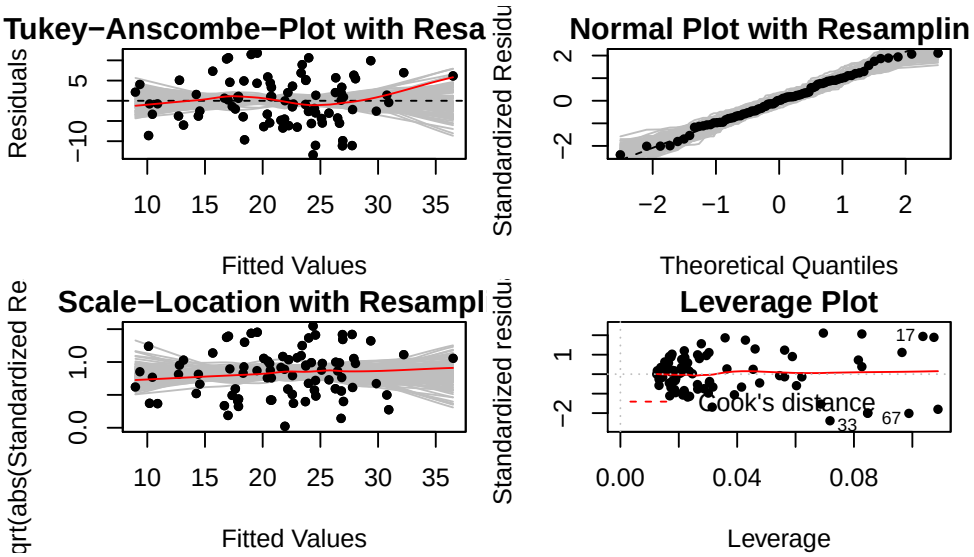


The residual plots look worse than in the first model. Therefore, we will use the orignal model without transformations.

3.

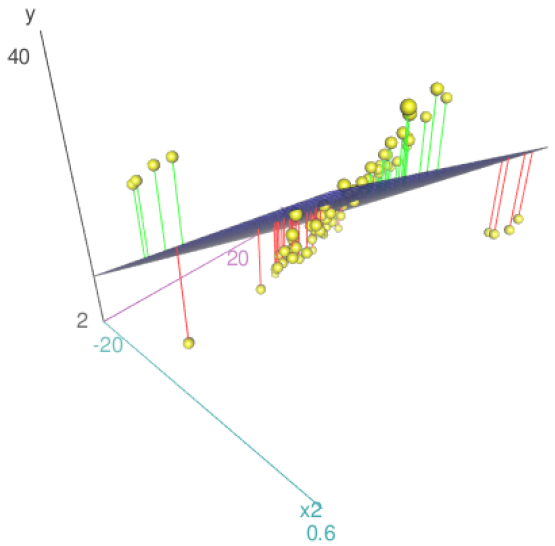
```
> ## load data
> load("synthetisch.rda")
> source("../ex1/resplot.R")
> ## fit
> fit <- lm(y ~ x1 + x2, data=synthetisch)
> par(mfrow=c(2,2))
> resplot(fit)
> summary(fit)

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 tolerable deviation in the Tukey-Anscombe plot and a slight but also tolerable sign of increasing variance in the scale-location plot.

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



The 3D plot contradict the above conclusion. The majority of the data points scatters around a plane that is completly different from the OLS solution. Additionally, there are a few high leverage points that lies outside the point cloud on both sides of the regression plane. The OLS fit tries to accomodate all

observations as good as possible. This results in the residuals having a similar size, i.e. the OLS 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 OLS 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 artifically and is certainly an extreme one, it does show that robust methods constitute an important additional tool in regression analysis. Unfortunately, studying these methods lies beyond the scope of this course.