Predictive Modeling

Series 2

Exercise 2.1

The following exercise is based on the `windmill` data, found in the data file `windmill.dat`. We consider the following three regression models:

a) Naive:

$$\text{current} = \beta_0 + \beta_1(\text{wind speed}) + \epsilon$$

b) First-Aid Transformation:

$$\log(\text{current}) = \beta_0 + \beta_1 \log(\text{wind speed}) + \epsilon$$

c) Transformation according to expert knowledge:

$$\text{current} = \beta_0 + \beta_1 \frac{1}{\text{wind speed}} + \epsilon$$

Check by means of residual plots, for which of the three models the assumptions of the linear regression model are fulfilled.

Exercise 2.2

We consider the model

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

where

$$x = 100 \cdot \log(\mathit{pressure})$$

to fit the `Forbes` data. Check with the help of residual plots, whether the assumptions of the linear regression model are fulfilled.

Exercise 2.3

While fitting and visualizing simple linear regression models becomes a routine after a while, assessing whether a model fits the data remains a challenging task. We will practice this aspect of linear regression analysis with two additional data sets:

a) The file **gas.rda** contains the gas **consumption** (in kWh) and the difference of **temperature** (in Celcius) inside and outside of 15 houses which are heated with gas. The measurements were collected over a long time period and then averaged. The goal is to predict the gas consumption on the basis of the temperature difference. Plot the regression line and perform a residual analysis.

b) The file **antique_clocks.rda** contains the **age** and the **price** of antique clocks that are auctioned. The goal is to predict the price on the basis of the age of the clock. Plot the regression line and perform a residual analysis.

R-Hints: To load the data files, use the following **R**-function

```
load("../gas.rda")
```

Exercise 2.4

The article *Characterization of Highway Runoff in Austin, Texas, Area* was based on a data set with $x =$ **rainfall volume** and $y =$ **runoff volume** for a particular location. The values are:

rainfall volume	5	12	14	17	23	30	40	47	55	67	72	81	96	112	127
runoff volume	4	10	13	15	15	25	27	46	38	46	53	70	82	99	100

a) Generate a scatter plot of **runoff volume** versus **rainfall volume**. Fit a simple linear regression model, add the regression line to the plot and generate the **R** summary output for the regression model.

b) How much of the observed variation in **runoff volume** can be explained by the simple linear regression model with response variable **runoff volume** and predictor variable **rainfall volume**?

c) Is there a significant linear association between **runoff volume** and **rainfall volume**? Provide an illustrative interpretation of the regression coefficients.

d) Use the regression fit to predict the `runoff volume` when the `rainfall volume` takes on the value 50. Also compute the 95 % prediction interval for a `rainfall volume` value of 50.

e) Assess the model assumptions by means of the model diagnostics tools we have discussed.

f) Fit a new regression model on the basis of the log-transformed variables and add it to the scatter plot.

g) Assess the strength and the significance of the linear relationship between the log-transformed variables `runoff volume` and `rainfall volume`. Interpret the regression coefficients.

h) Predict the expected `runoff volume` when rainfall takes on a value of 50. Generate a 95 % prediction interval and compare it to the original solution. Add the prediction interval for arbitrary x to the scatter plot.

i) Perform a residual analysis. Which model is more appropriate and why?

Exercise 2.5

In an experiment marine bacteria were exposed to x-rays during 15 intervals of six minutes. The following table contains the number of `surviving bacteria` after each `interval`:

`interval`	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
`surv. bact.`	255	211	197	166	NA	106	104	60	56	38	36	32	21	19	15

a) Show the relation between the number of `surviving bacteria` and the number of radiation `intervals`. Does it make sense to fit a least squares regression model to the data?

b) Fit a simple linear regression model and check the model assumptions.

c) Improve the model by transforming the response variable or/and the predictor. **Hint**: Theory suggests that per radiation interval the proportion of bacteria that is killed remains constant.

d) Predict the missing value for the fifth interval and compute a 95 % prediction interval. In addition, compute the estimate for the relative decrease in the number of surviving bacteria, together with a 95 % confidence interval. Last, determine the expected number of bacteria at the beginning, i.e. before the first radiation interval. Also compute a 95 % confidence interval for this value.

Exercise 2.6

Assessing model diagnostic plots requires experience. Often it is difficult to decide whether a deviation is systematic (i.e. needing correction) or due to random variations (i.e. just variability in the data). Experience can be gained by performing model diagnostics on problems where it is known whether the model assumptions hold or do not hold. This allows us to identify the naturally occurring variability in the results.

In the following we simulate a predictor variable xx and four responses $yy.a$, $yy.b$, $yy.c$, and $yy.d$.

```
set.seed(21)
n <- 100
xx <- 1:n
yy.a <- 2 + 1 * xx + rnorm(n)
yy.b <- 2 + 1 * xx + rnorm(n) * (xx)
yy.c <- 2 + 1 * xx + rnorm(n) * (1 + xx/n)
yy.d <- cos(xx * pi/(n/2)) + rnorm(n)
```

Fit four simple linear regression models using xx as predictor variable.

a) For each model, generate a scatter plot with the regression line, generate the four standard residual plots and the plot containing Cook's distance. Decide for each model which of the assumptions are fulfilled and which ones are violated. Verify your claims with the construction of the responses.

b) Use the function **`resplot()`** which is available on moodle. The function **`resplot()`** uses resampling to visualize whether a model violation is present. How does the function perform for the four models?

Result Checker

E 2.4: b) $R^2 = 0.98$

c) $\hat{\beta}_0 = -1.12$ and $\hat{\beta}_1 = 0.827$

d) $[28.53, 51.92]$

h) 39.88 and $[29.86744, 53.25903]$

E 2.5: d) Prediction and Prediction interval for interval 5 : 124.0672 and $[96.29711, 159.8]$
Prediction and Prediction interval for interval 0 : 352.2568 $[307.3775, 403.6888]$

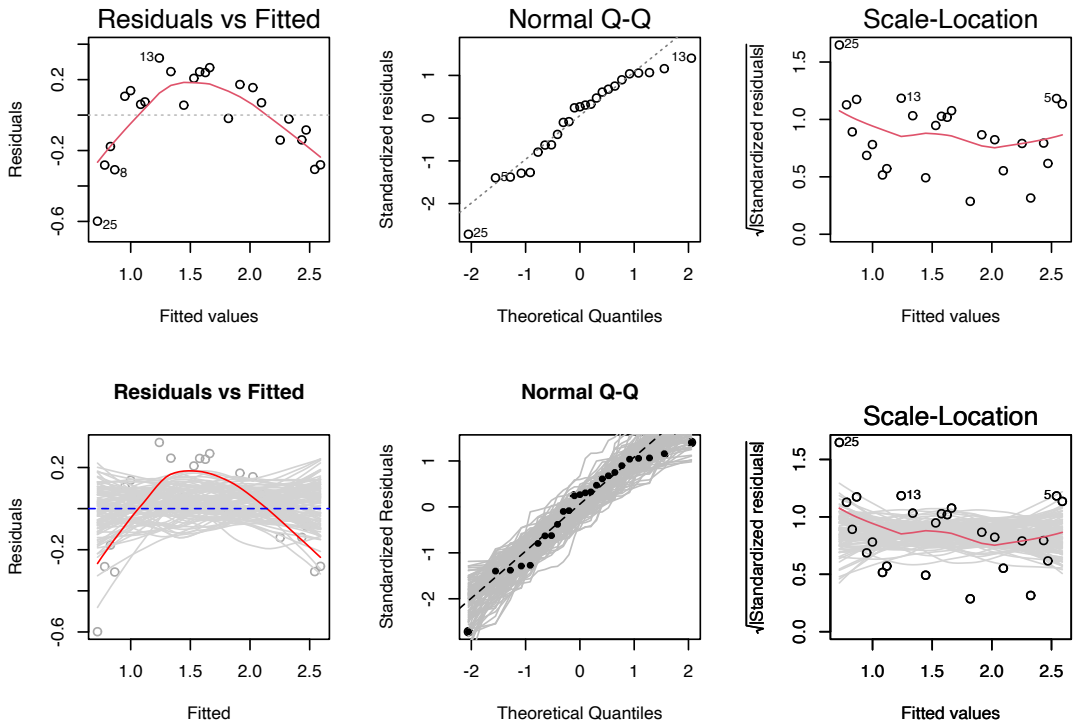
Predictive Modeling

Solutions to Series 2

Solution 2.1

a)
```
path <- "Daten/"
windmill <- read.table(paste(path, "windmill.dat", sep = ""),
    header = T)
fit.lm1 <- lm(current ~ wind_speed, data = windmill)

# Plot standard residual plots
par(mfrow = c(2, 3))
plot(fit.lm1, which = 1:3)

# Simulations for Tukey-Anscombe plot
plot(fitted(fit.lm1), resid(fit.lm1), col = "darkgrey",
    main = "Residuals vs Fitted", xlab = "Fitted", ylab = "Residuals"
set.seed(111)
for (i in 1:100) {
    sresid <- sample(resid(fit.lm1), replace = TRUE)
    lines(loess.smooth(fitted(fit.lm1), sresid), col = "lightgrey",
        lwd = 1)
}
lines(loess.smooth(fitted(fit.lm1), resid(fit.lm1)), col = "red",
    lwd = 1)
abline(h = 0, col = "blue", lty = "dashed")

# Simulation for QQ-plot

qq <- qqnorm(rstandard(fit.lm1), main = "Normal Q-Q", ylab = "Standar
for (i in 1:100) {
    sresid <- rnorm(length(qq$x), mean(qq$y), sd(qq$y))
    lines(sort(qq$x), sort(sresid), col = "grey")
}
points(qq$x, qq$y, pch = 20)
qqline(rstandard(fit.lm1), lty = 2)

# Simulation for scale location plot

plot(fit.lm1, which = 3)
```

```
for (i in 1:100) {
    sresid <- sqrt(abs(sample(rstandard(fit.lm1), replace = TRUE)))
    lines(loess.smooth(fitted(fit.lm1), sresid), col = "lightgrey",
        lwd = 1)
}
par(new = T)
plot(fit.lm1, which = 3)
```



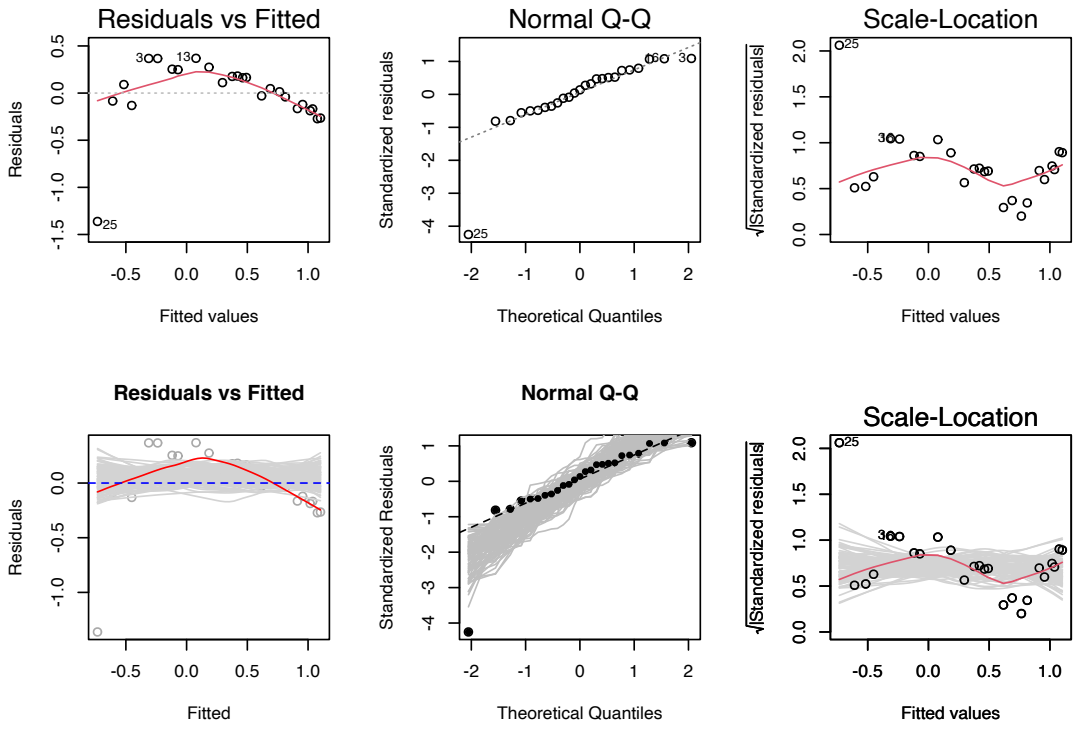
- *Tukey-Anscombe*: The banana-like shape of the smoothing line suggests that the expected value is not constant zero. This structure is as well visible in the graphics containing the simulated smoothing lines: the (red) smoothing line is not entirely contained in the grey band of simulated curves.
- *Normal Plot*: The data points scatter next to the straight line. All data points fall into the band of curves which may arise due to statistical fluctuations. There is no evidence to doubt the assumption that the residuals follow a normal distribution.
- *Scale-Location*: The smoothing line tends to decrease, but it is still part of the grey band of simulated curves. Thus, there is no evidence against the assumption that the variance is constant.

- *Time Correlation:* Since we do not dispose of any information concerning the time order of the measurements, we cannot analyze any time correlations. Such a correlation would represent a (clear) violation of the independence of the residuals.

Conclusion: The fitted model is insufficient, because systematic deviations of the expected value show up. A transformation of the predictor variables may solve the problem.

b) The simulation is performed in analogy to a), but we need to adjust the model:

```
windmill$lwind_speed <- log(windmill$wind_speed)
windmill$lcurrent <- log(windmill$current)
fit.lm2 <- lm(lcurrent ~ lwind_speed, data = windmill)
```



- *Tukey-Anscombe*: The banana-like shape of the smoothing line suggests again that the expected value is not constant zero. This structure is as well visible in the graphics containing the simulated curves, where the (red) smoothing line is not entirely part of the band of simulated curves.
- *Normal Plot*: The data points scatter near to the straight line, with the exception of observation 25, which shows a strong deviation from the straight

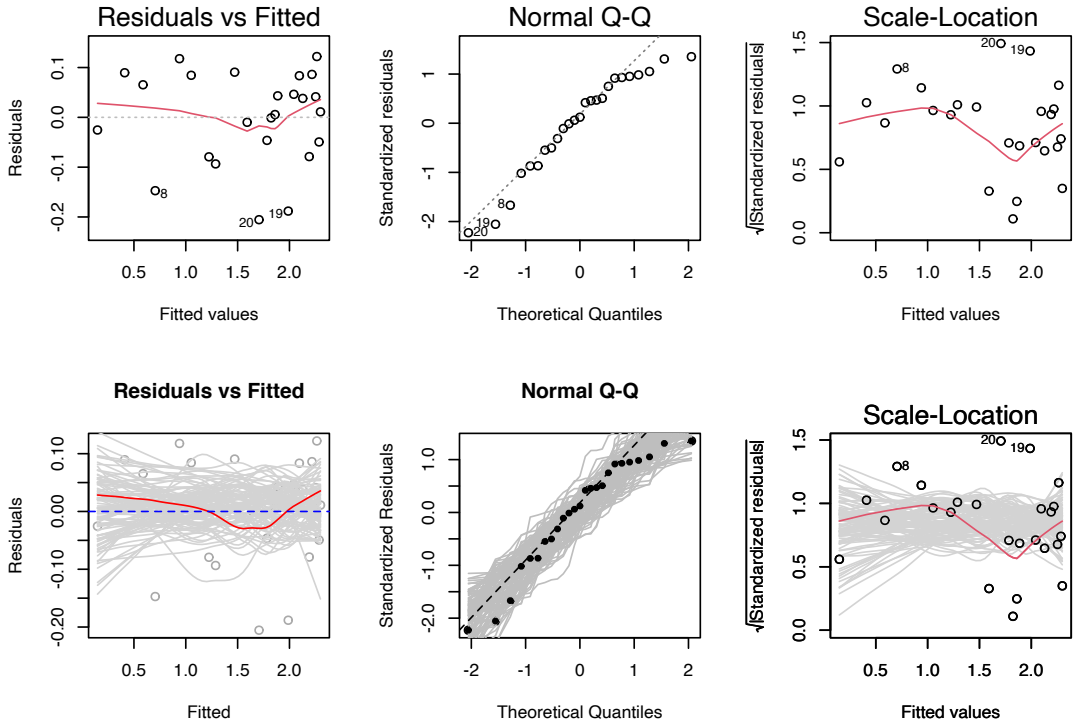
line and falls outside of the band of simulated curves. This outlier seems to violate the normal distribution assumption.

- *Scale-Location*: The smoothing line follows a wave-like curve and, in addition to that, touches the border of the grey region with the simulated curves.

Conclusion: This model fit is as well insufficient and shows peculiarities in all three diagnosis plots.

c) The simulations are carried out in analogy to a), however we need to adjust the model:

```
windmill$x <- 1/windmill$wind_speed
fit.lm3 <- lm(current ~ x, data = windmill)
```



- *Tukey-Anscombe*: The smoothing line shows a slight banana-like shape, however it does not seem to be problematic according to the graphics with the simulated curves, because the red curve lies inside the grey band of curves.
- *Normal Plot*: With the exception of the points at the right margin, the data points scatter nicely around the straight line. The right margin points indicate short-tailedness, which does not represent, however, a problem for the least-squares method.

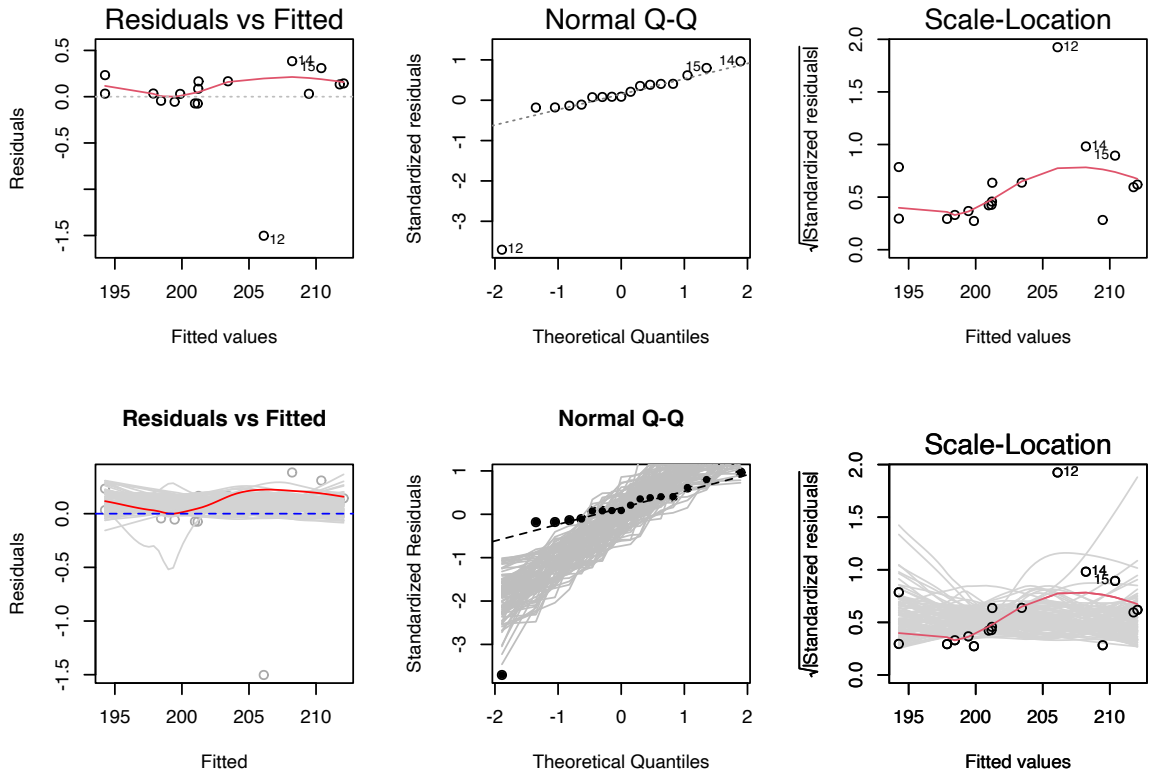
- *Scale-Location*: The smoothing line shows a dent-like pattern, which however does not represent a problem according to the simulated curves.

Conclusion: The residual analysis for this model fit does not show any evidence of violating the assumptions of the linear model.

Solution 2.2

The simulation plots are generated in analogy to exercise 1. Use the model:

```
path <- "Daten/"
Forbes <- read.table(paste(path, "Forbes.dat", sep = ""),
    header = T)
Forbes$x <- 100 * log(Forbes$pressure)
Forbes.lm <- lm(y ~ x, data = Forbes)
```



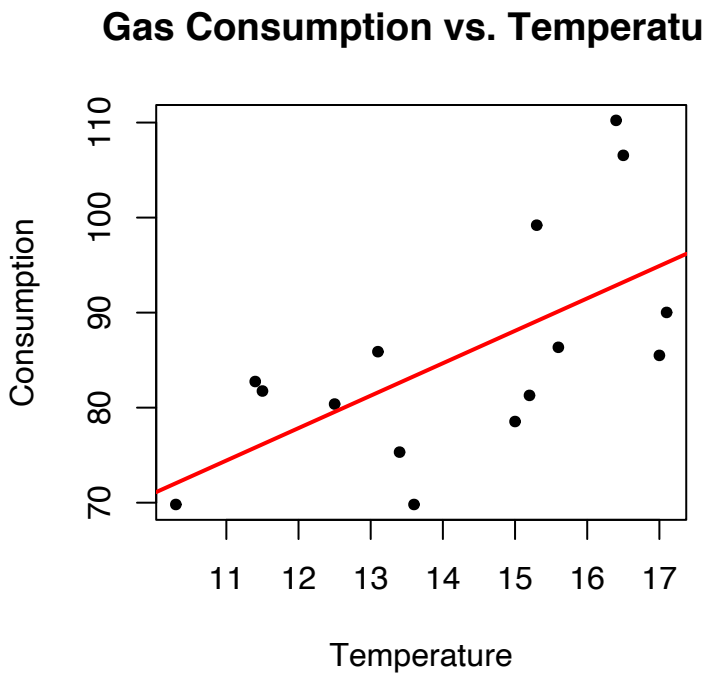
- *Tukey-Anscombe*: The smoothing line shows a slightly suspicious pattern. Since the smoother however falls entirely into the grey band of simulated curves, this pattern does not represent a serious problem.

- *Normal Plot*: The data points scatter nicely around the straight line, with the exception of observation 12, which deviates clearly from the straight line. It lies outside of the grey region. Because of data point 12, the normal distribution assumption is violated.
- *Scale-Location*: The smoother exhibits a suspicious pattern : it departs from the grey band consisting of simulated curves. The normal distribution assumption seems to be violated.
- *Time Correlation*: Because we do not dispose of any information concerning the time order of the measurements, we are not in the position to analyze any time correlations. Such correlations would represent very strong evidence against the independence of the residuals.

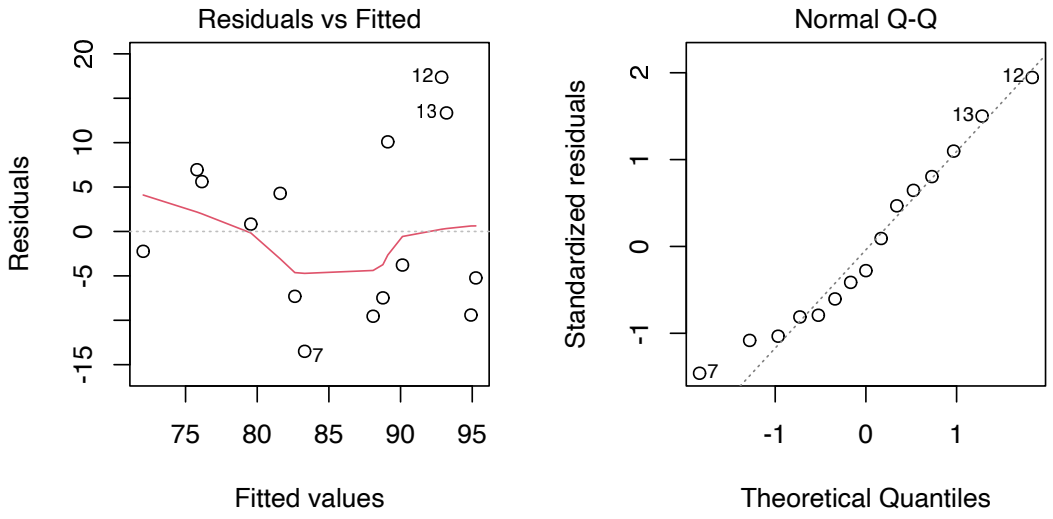
Conclusion: The model fit is insufficient and shows rather strong evidence of violation of the assumptions in all three diagnosis plots.

Solution 2.3

a)
```r
## load data
load("Daten/gas.rda")
## analysis of the gas consumption
par(mfrow = c(1, 1))
plot(consumption ~ temperature, data = gas, pch = 20, ylab = "Consump
    xlab = "Temperature")
title("Gas Consumption vs. Temperature")
fit <- lm(consumption ~ temperature, data = gas)
abline(fit, col = "red", lwd = 2)
```




```
par(mfrow = c(1, 2))
plot(fit, which = 1:2)
```

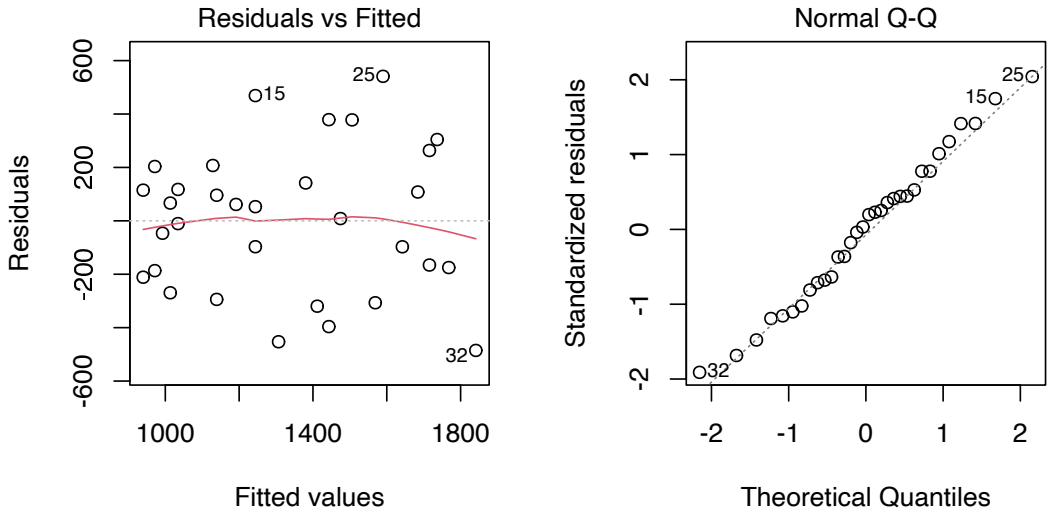


At first sight the scatter plot seems to suggest that the regression line fits well the data. However, the plot *Residuals vs. Fitted* gives another impression. The smoother shows a strong deviation towards the bottom which indicates the pre-

sence of a systematic error. Similarly, the variance seems to be larger for large fitted values. The normal plot does not show any abnormalities. Whether the observations are correlated cannot be determined based on these two plots. We would have to know whether the observations were recorded in a temporal order and whether residuals of observations close in time show abnormalities. In summary, the two assumptions $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_\epsilon^2$ might be violated. A log-transformation would yield a much better fit and should be applied.

b)
```
## load data
load("Daten/antique_clocks.rda")
## analysis of the gas consumption
par(mfrow = c(1, 1))
plot(price ~ age, data = antique_clocks, pch = 20, ylab = "price",
    xlab = "age")
title("Price vs. Age")
fit <- lm(price ~ age, data = antique_clocks)
abline(fit, col = "red", lwd = 2)
```


```
par(mfrow = c(1, 2))
plot(fit, which = 1:2)
```



You can generate the residual plots as well by using

```
plot(fitted(fit), resid(fit), pch = 20)
lines(loess.smooth(fitted(fit), resid(fit)), col = "red")
title("Residuals vs. Fitted")
abline(h = 0, lty = 2)
qqnorm(resid(fit), pch = 20)
qqline(resid(fit))
```

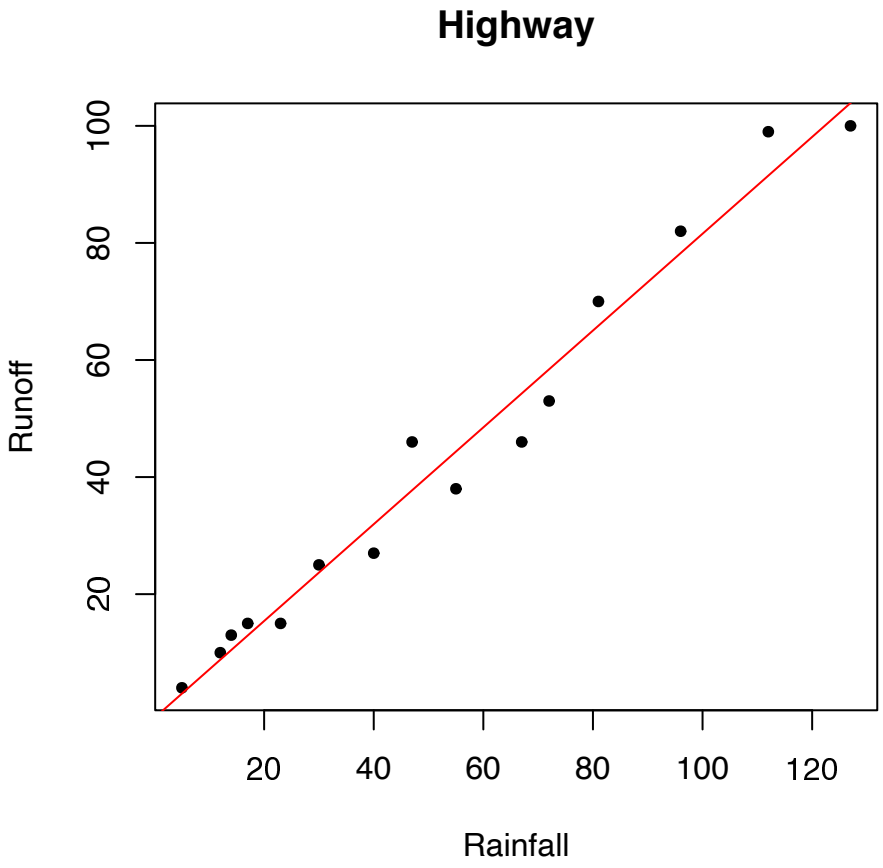
This model shows a good fit. The smoother in the plot *Residuals vs. Fitted* is almost horizontal and does not show systematic deviations from the x-axis. The variance of the data points is approximately constant. It is only slightly smaller on the left which is not a significant problem. The Normal plot does not show any abnormalities. Whether the observations are correlated cannot be determined based on these two plots. These could occur if the clocks were sold at different auctions with systematically larger or smaller prices. This would cause a correlation of the corresponding residuals. In summary, we may consider this model as fitting the data well.

Solution 2.4

a) First we type in the data. The scatter plot of **runoff** versus **rainfall** suggests that a linear relationship applies.

```
rainfall <- c(5, 12, 14, 17, 23, 30, 40, 47, 55, 67, 72,
    81, 96, 112, 127)
runoff <- c(4, 10, 13, 15, 15, 25, 27, 46, 38, 46, 53, 70,
```

```
    82, 99, 100)
hwy.runoff <- data.frame(rainfall = rainfall, runoff = runoff)
plot(hwy.runoff$runoff ~ hwy.runoff$rainfall, pch = 20,
    xlab = "Rainfall", ylab = "Runoff", main = "Highway")
## fit a simple linear regression
fit <- lm(runoff ~ rainfall, data = hwy.runoff)
abline(fit, col = "red")
```




```
## Summary
summary(fit)

##
## Call:
## lm(formula = runoff ~ rainfall, data = hwy.runoff)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.279 -4.424  1.205  3.145  8.261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.12830    2.36778  -0.477    0.642
## rainfall     0.82697    0.03652  22.642  7.9e-12 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.24 on 13 degrees of freedom
## Multiple R-squared:  0.9753,Adjusted R-squared:  0.9734
## F-statistic: 512.7 on 1 and 13 DF,  p-value: 7.896e-12
```

b) An R^2 of 0.98 is very high, i.e. a large part of the variation in the data can be explained by the linear regression model for the association between **runoff volume** and **rainfall volume**.

c) There is a significant linear association between **runoff volume** and **rainfall volume**, since the null hypothesis $\beta_1 = 0$ can be rejected very clearly.

As estimates we obtain $\hat{\beta}_0 = -1.12$, i.e. when it does not rain, the **runoff volume** is negative. Obviously, this is not possible. Note as well that we do not dispose of any observations at $x = 0$. Thus, interpreting the intercept represents an extrapolation. On the other hand, this shows as well that this model might not be the best. Last, we observe that the estimate of the intercept is not significant.

For the slope we obtain $\hat{\beta}_1 = 0.827$. This value is smaller than 1 (even significantly smaller as the 95 % confidence interval for the slope indicates). Thus, it is statistically shown that not all the rain runs off via the canalization. It is plausible that part of the rain evaporates or trickles away.

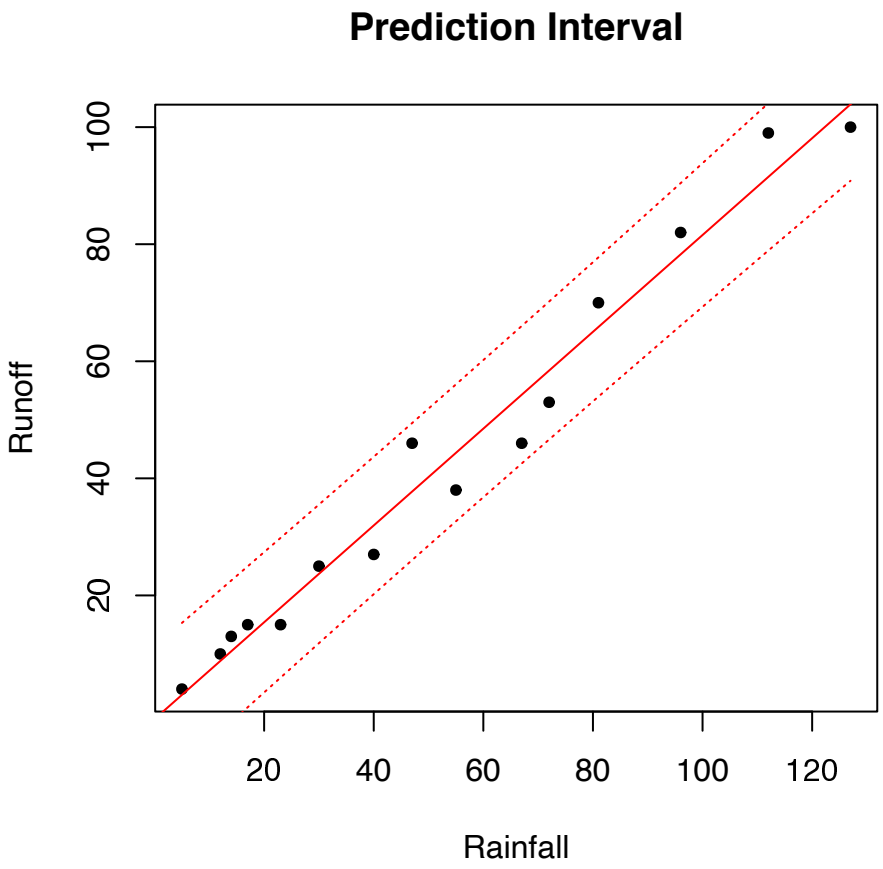
d)
```
pred <- predict(fit, newdata = data.frame(rainfall = 50),
    interval = "prediction")
```

If the **rainfall volume** takes on a value of 50 we find a **runoff volume** of 40.22 with a 95 % prediction interval of $[28.53, 51.92]$.

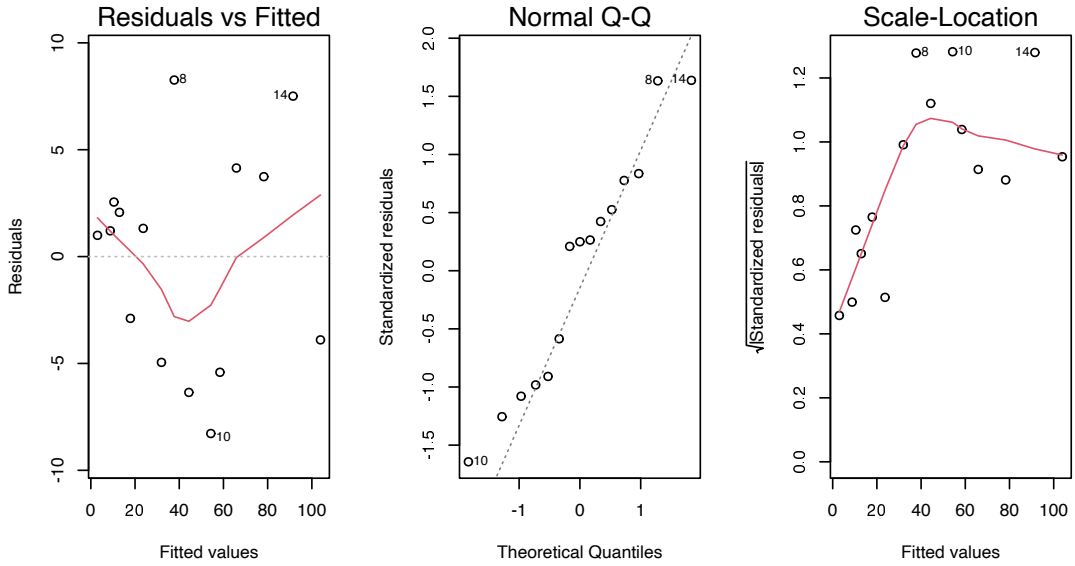
We can also plot the 95 % prediction interval to the data.

```
plot(hwy.runoff$runoff ~ hwy.runoff$rainfall, pch = 20,
    xlab = "Rainfall", ylab = "Runoff", main = "Prediction Interval")
abline(fit, col = "red")
```

```
interval <- predict(fit, interval = "prediction")
lines(hwy.runoff$rainfall, interval[, 2], lty = 3, col = "red")
lines(hwy.runoff$rainfall, interval[, 3], lty = 3, col = "red")
```



e)
```
par(mfrow = c(1, 3))
plot(fit, which = 1:3)
```



We check the following model assumptions

a) Expected value of the errors is zero

b) Constant error variance

c) Normal distribution of errors

with

a) a Tukey-Anscombe plot (residuals vs. fitted values) to assess assumption (i)

b) a scale-location plot to assess assumption (ii)

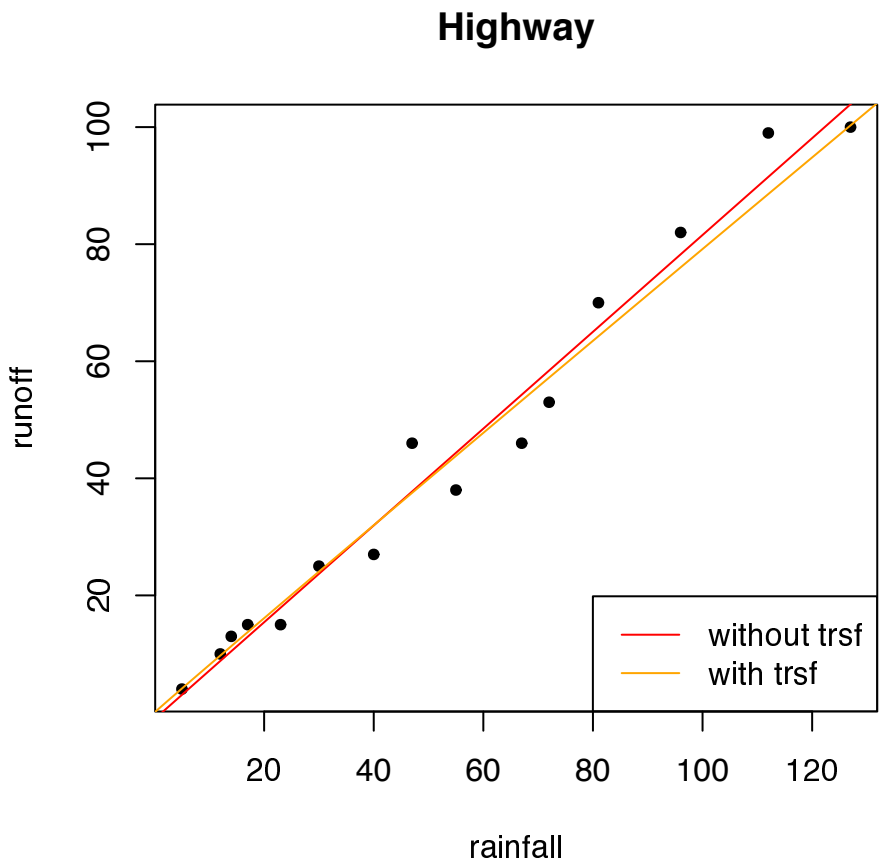
c) a normal plot to assess assumption (iii)

For the assumption of uncorrelated errors there is no suitable plot in this case. However, this is not a reason to assume that the errors are uncorrelated.

While the normal distribution assumption of the errors seems to be satisfied, the Tukey-Anscombe plot (residuals vs. fitted values) raises some doubts. Even though there is a large R^2 and also the test for the slope is highly significant, the expected value of the errors does not seem to be zero. The smoother deviates systematically from the horizontal line. Similarly, it seems like the variance of the residuals is larger at large rainfall values. Thus, the assumption of constant error variance might be violated. As we shall see, we can describe the relation between runoff and rainfall more accurately with a different simple linear regression model.

f) After fitting the new regression model, we need to exponentiate the fitted values in order to add them to the original plot.

```
par(mfrow = c(1, 1))
fit.loglog <- lm(log(runoff) ~ log(rainfall), data = hwy.runoff)
plot(runoff ~ rainfall, data = hwy.runoff, pch = 20, main = "Highway"
abline(fit, col = "red")
xx <- data.frame(rainfall = 0:150)
yy <- predict(fit.loglog, newdata = xx)
lines(xx$rainfall, exp(yy), col = "orange")
legend("bottomright", lty = 1, col = c("red", "orange"),
    legend = c("without trsf", "with trsf"))
```



g)

```
summary(fit.loglog)

##
## Call:
## lm(formula = log(runoff) ~ log(rainfall), data = hwy.runoff)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20980 -0.10952  0.02828  0.08727  0.20388
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.18369    0.13803  -1.331    0.206
## log(rainfall)  0.98917    0.03676  26.908 8.75e-13 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1293 on 13 degrees of freedom
## Multiple R-squared:  0.9824,Adjusted R-squared:  0.981
## F-statistic:   724 on 1 and 13 DF,  p-value: 8.748e-13
```

There has already been a strong relation between **rainfall** and **runoff** without any variable transformation. We found in this case a p-value of 10^{-12} for the null hypothesis $\beta_1 = 0$ and an R^2 of 0.975. The variable transformation leads to a p-value of 10^{-13} for the null hypothesis $\beta_1 = 0$ and an R^2 of 0.982.

In addition, the model with transformed variables should be prefered from a practical point of view since it cannot yield negative runoff values. Further, the coefficient β_1 is easier to interpret. Without transformation, $\hat{\beta}_1 = 0.827$ means that for an additional unit of rain, there are 0.827 additional units of runoff.

If we transform the variables, that is, $Y' = \log(Y)$ and $X' = \log(X)$, then the straight line is fitted on the log-log-scale:

$$Y' = \beta_0' + \beta_1' X' + \epsilon'$$

We can derive the relation on the original scale by taking the exponential function on both sides:

$$Y = \exp(\beta_0') \cdot x^{\beta_1'} \cdot \exp(\epsilon') = \beta_0 \cdot x^{\beta_1} \cdot \epsilon$$

where $\exp(\beta_0') = \beta_0$ and $\beta_1' = \beta_1$. The slope from the log-log-scale is the exponent to x on the original scale. Moreover, we have a multiplicative rather than an additive model, where the error term follows a log-normal distribution. Hence, the errors will scatter more the larger X is, and are skewed towards the right, i.e. larger values. The interesting aspect is the interpretation of the model equation.

If X increases by 1 %, then Y increases by β_1%:

$$\begin{aligned}\tilde{Y} &= \beta_0 \cdot (x(1+0.01))^{\beta_1} \cdot \epsilon \\ &\approx \beta_0 \cdot (x^{\beta_1} + 0.01 \cdot \beta_1) \cdot \epsilon \\ &= Y + \beta_1 \cdot 0.01 \cdot Y\end{aligned}$$

With the transformation, the interpretation is that for 1 % of additional amount of rain, there is 0.989 % of additional runoff. In other words, 98.9 % of the rain runs off via the canalization, the rest evaporates or trickles away.

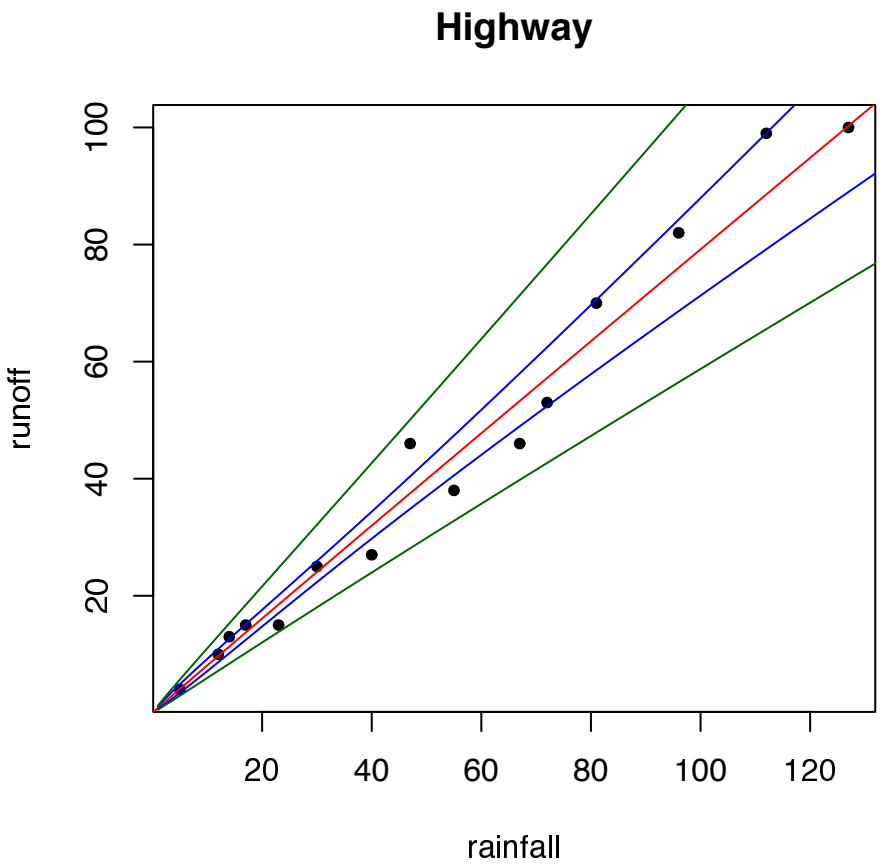
h)
```
## prediction
new.x <- data.frame(rainfall = 50)
pred.y <- predict(fit.loglog, new.x, interval = "prediction")
exp(pred.y)

##        fit      lwr      upr
## 1 39.88372 29.86744 53.25903
```

Remark for the advanced reader: Note that the point prediction is not the expected value of the response variable but its median. If we want to compute the expected value, we need to adjust the model as follows:

```
exp(pred.y[1] + (summary(fit.loglog)$sigma^2)/2)

## [1] 40.21832
```

The predicted value is very close to the one from the model without log-transformations. This however is not true in general. In some cases, the difference can be large. In addition, note that the 95 % prediction interval is no longer symmetric:

```
## prediction and confidence interval
plot(runoff ~ rainfall, data = hwy.runoff, pch = 20, main = "Highway'
xx <- data.frame(rainfall = 0:150)
yy <- predict(fit.loglog, newdata = xx, interval = "confidence")
lines(xx$rainfall, exp(yy[, 1]), col = "red")
lines(xx$rainfall, exp(yy[, 2]), col = "blue")
lines(xx$rainfall, exp(yy[, 3]), col = "blue")
yy <- predict(fit.loglog, newdata = xx, interval = "prediction")
lines(xx$rainfall, exp(yy[, 2]), col = "darkgreen")
lines(xx$rainfall, exp(yy[, 3]), col = "darkgreen")
```

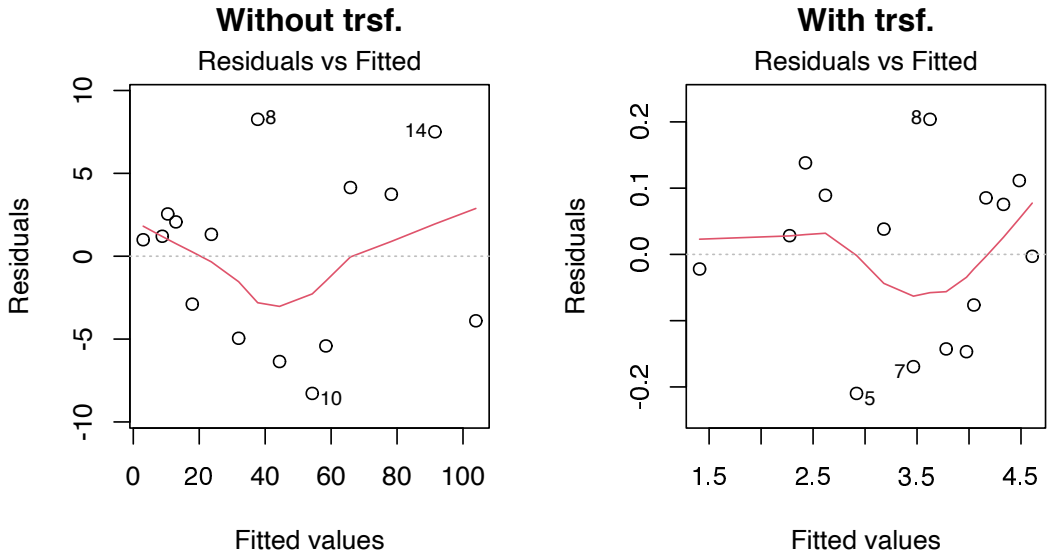


The asymmetry is not very pronounced in this example. In other cases, it can be much stronger.

i) The residual plot for the regression model based on transformed variables seems to be more suited than the model on the basis of the original variables.

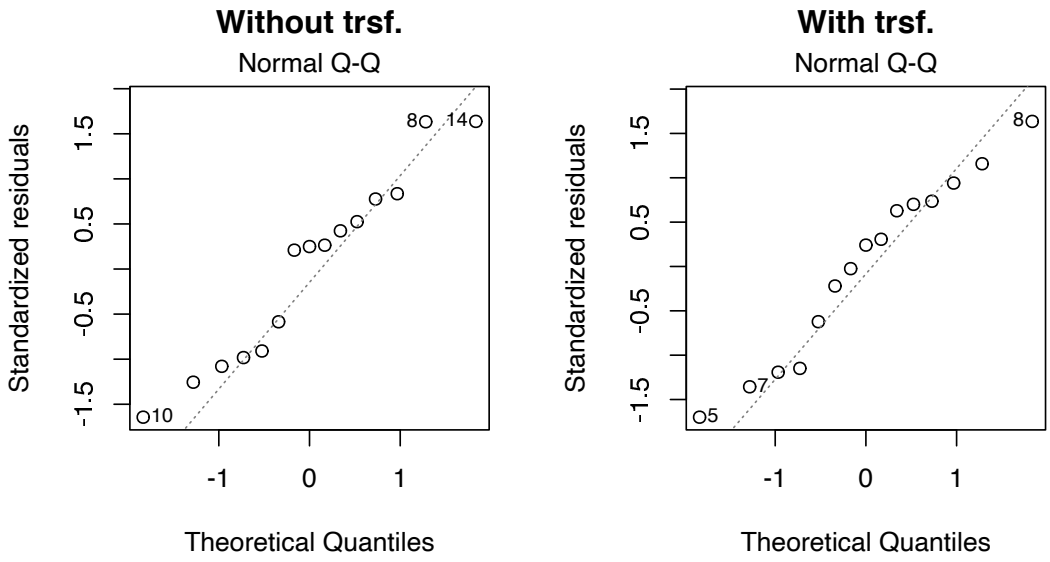
The improvement is not significant, but the model on the basis of transformed variables seems to have more positive properties. Another advantage of the model with transformed variables is its property of not yielding negative values - neither fitted values nor in the prediction interval. The transformed model predicts a runoff of 0 for a rainfall of 0 which is another desirable property. In summary, there are only small differences between the two models but the model with the transformed variables is more appropriate.

```
## residuals plots
par(mfrow = c(1, 2))
plot(fit, which = 1, main = "Without trsf.")
plot(fit.loglog, which = 1, main = "With trsf.")
```



Both of the normal plots do not show any abnormalities:

```
## Normal plots
par(mfrow = c(1, 2))
plot(fit, which = 2, main = "Without trsf.")
plot(fit.loglog, which = 2, main = "With trsf.")
```

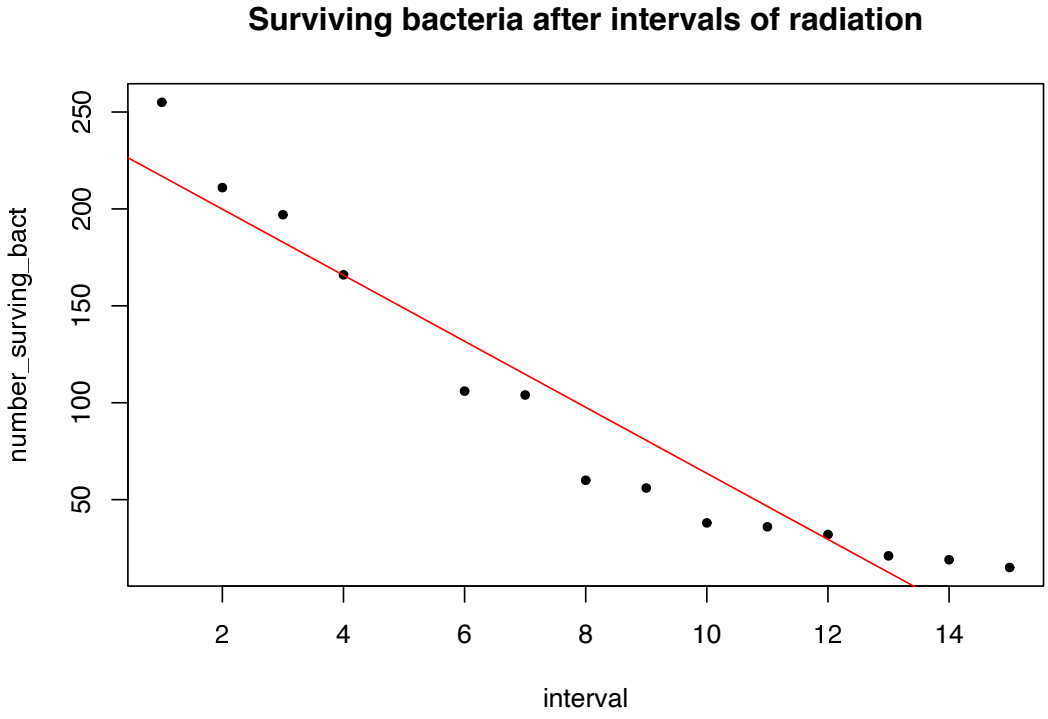


Solution 2.5

a) The scatter plot shows that a least squares regression line does not describe the

relation between the two quantities appropriately. We need to apply variable transformations.

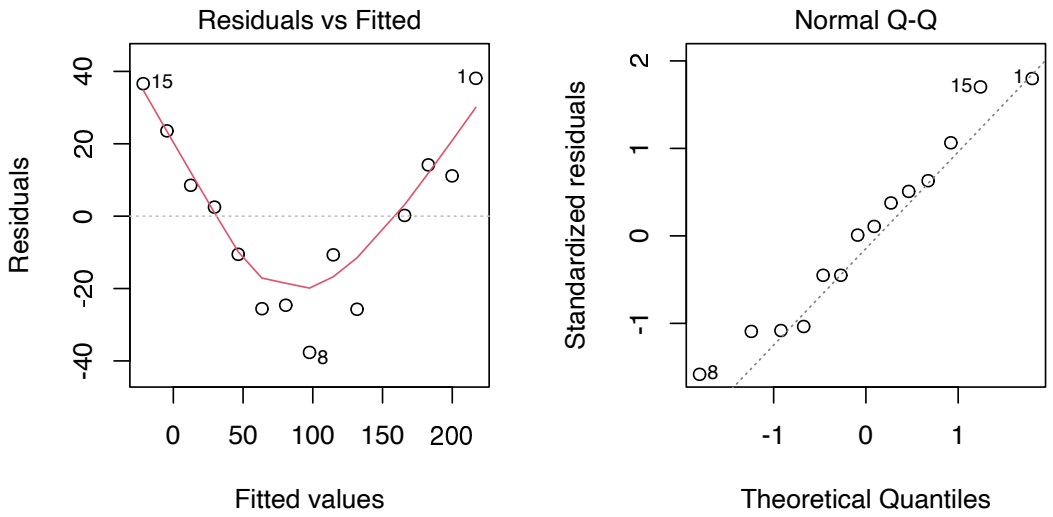
```
## load data
interval <- 1:15
number_surving_bact <- c(255, 211, 197, 166, NA, 106, 104,
    60, 56, 38, 36, 32, 21, 19, 15)
## scatter plot
par(mfrow = c(1, 1))
plot(interval, number_surving_bact, pch = 20)
title("Surviving bacteria after intervals of radiation")
## regression
fit.orig <- lm(number_surving_bact ~ interval)
## add regressions line to plot
abline(fit.orig, col = "red")
```



b) As the *Residuals vs. Predictor* plot shows, the expected value of the errors is not equal to zero. Instead there is a systematic deviation. The picture is typical for situations where a transformation should be applied. The normal plot does not show any conspicuous pattern. Please note, however, that it does not make much sense to study the distribution of the residuals carefully since they result from an incorrect model. The residuals corresponding to the correct model will

be different and may have a different distribution.

```
## diagnostics plots
par(mfrow = c(1, 2))
plot(fit.orig, which = 1)
plot(fit.orig, which = 2)
```



c) A log-transformation should be applied to the response variable number of **surviving bacteria**. This variable can only take on positive values. For the predictor, we do not need to apply a transformation.

The scale of the response is arbitrary, negative values could also be used. In addition, the log-response model fulfills the hint given in the problem formulation: per radiation interval the proportion of bacteria that is killed remains constant. Let us denote by $x(t)$ the number of surviving bacteria after time interval t . Then, the proportion of bacteria killed per time interval Δt is given by

$$\frac{x(t+\Delta t) - x(t)}{x(t) \cdot \Delta t}$$

If this proportion is supposed to be constant, then this expression approaches for $\Delta t \to 0$

$$\frac{dx}{dt} = \text{const.} \cdot x(t)$$

The solution to this differential equation is $x(t) = \text{const.} \cdot \exp(t)$, thus the relation between time interval t and number of surviving bacteria x is given by $\log(x(t)) = \log(\text{const.}) + t$.

We thus will fit the following regression model:

$$\log(X) = \beta_0 + \beta_1 \cdot t + \varepsilon$$

The estimate of the coefficient β_1 is

```
fit.log <- lm(log(number_surving_bact) ~ interval)
coef(fit.log)[2]

##   interval
## -0.2087074
```

We interpret $\hat{\beta}_1$ as the amount by which $\log(X)$ is increased (or decreased) after each time interval.

At time $t+1$, X is changed according to

$$X(t+1) = \exp(\beta_0) \cdot \exp(\beta_1 \cdot t + \beta_1) \cdot \exp(\varepsilon) = X(t) \cdot \exp(\beta_1)$$

Therefore, X is increased (or decreased) by the factor $\exp(\hat{\beta}_1)$ after each time interval, that is

```
fit.log <- lm(log(number_surving_bact) ~ interval)
exp(coef(fit.log)[2])

##  interval
## 0.8116327
```

After each interval, 81.16 % of the bacteria remain alive. In other words, on average 18.84 % of the bacteria are killed per interval.

d) The prediction on the original scale for the interval 5 can be obtained as follows:

```
## predictions and intervals
new.x <- data.frame(interval = c(5))
predi <- predict(fit.log, newdata = new.x, interval = "prediction")
## prediction for the median of the conditional
## distribution when interval=5
exp(predi)

##        fit      lwr      upr
## 1 124.0672 96.29711 159.8456
```

Remark for the advanced reader: This value is the median of the conditional distribution. The expected value is slightly larger, namely

```
## prediction for the expected value of the
## conditional distribution when interval=5
exp(predi + (summary(fit.log)$sigma^2)/2)[1]

## [1] 124.8256
```

The estimate for the relative decrease in the number of surviving bacteria was already discussed above:

```
## proportional change after one interval
exp(-0.208707)

## [1] 0.811633
```

The 95 % confidence interval is given by:

```
## confidence interval
exp(confint(fit.log, "interval"))

##              2.5 %    97.5 %
## interval 0.7998456 0.8235935
```

To estimate the expected number of bacteria at the beginning, we predict for the interval 0, together with the confidence interval:

```
## starting value (i.e. interval=0), including
## confidence interval
new.x <- data.frame(interval = c(0))
predi <- predict(fit.log, newdata = new.x, interval = "confidence")
# prediction for the median of the conditional
# distribution when interval=0
exp(predi)

##        fit      lwr      upr
## 1 352.2568 307.3775 403.6888

# prediction for the expected value of the conditional
# distribution when interval=0
exp(predi + (summary(fit.log)$sigma^2)/2)[1]

## [1] 354.41
```

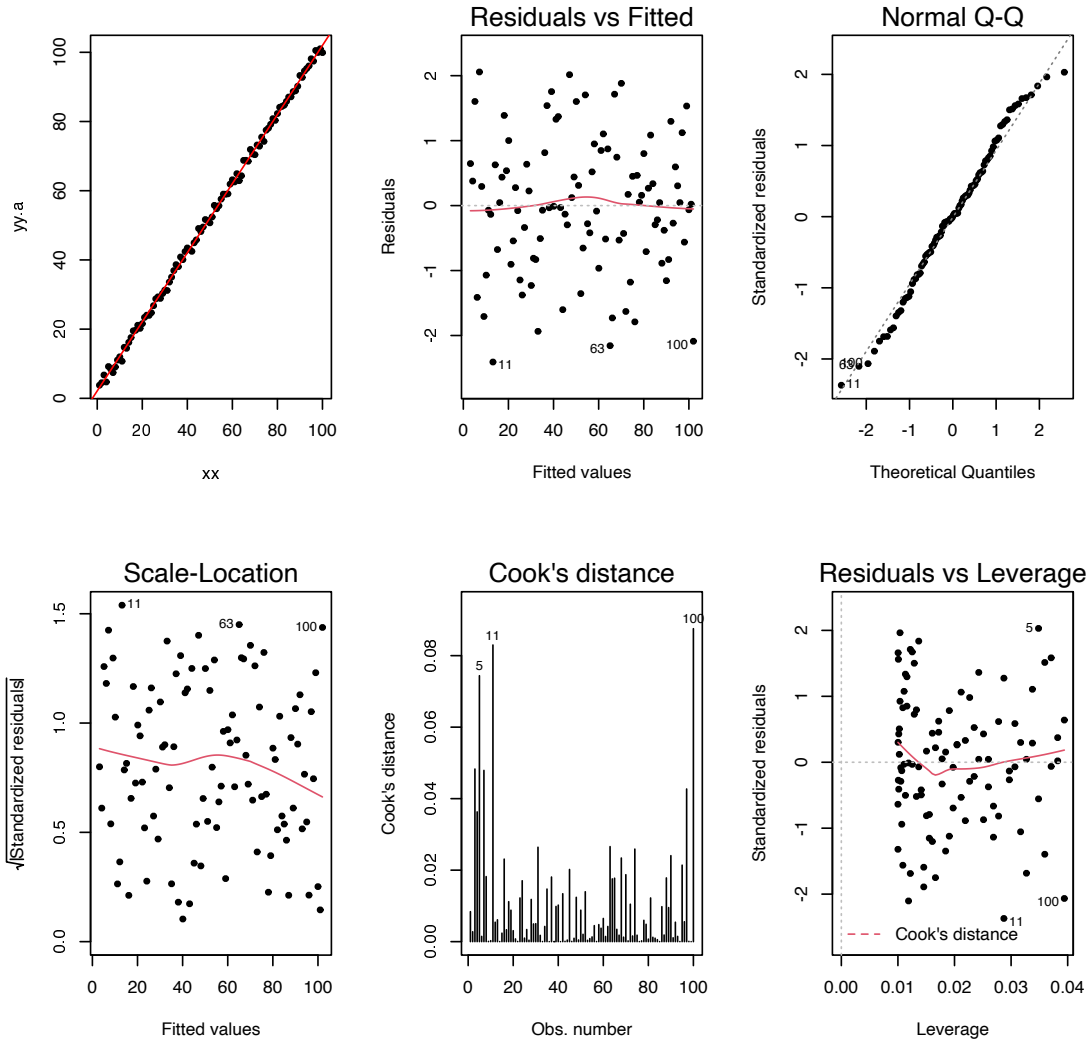
Solution 2.6

a) From the plots below we conclude:

- Model assumptions valid

- Model contains strong non-constant variance
- Variance slightly non-constant
- 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)
```

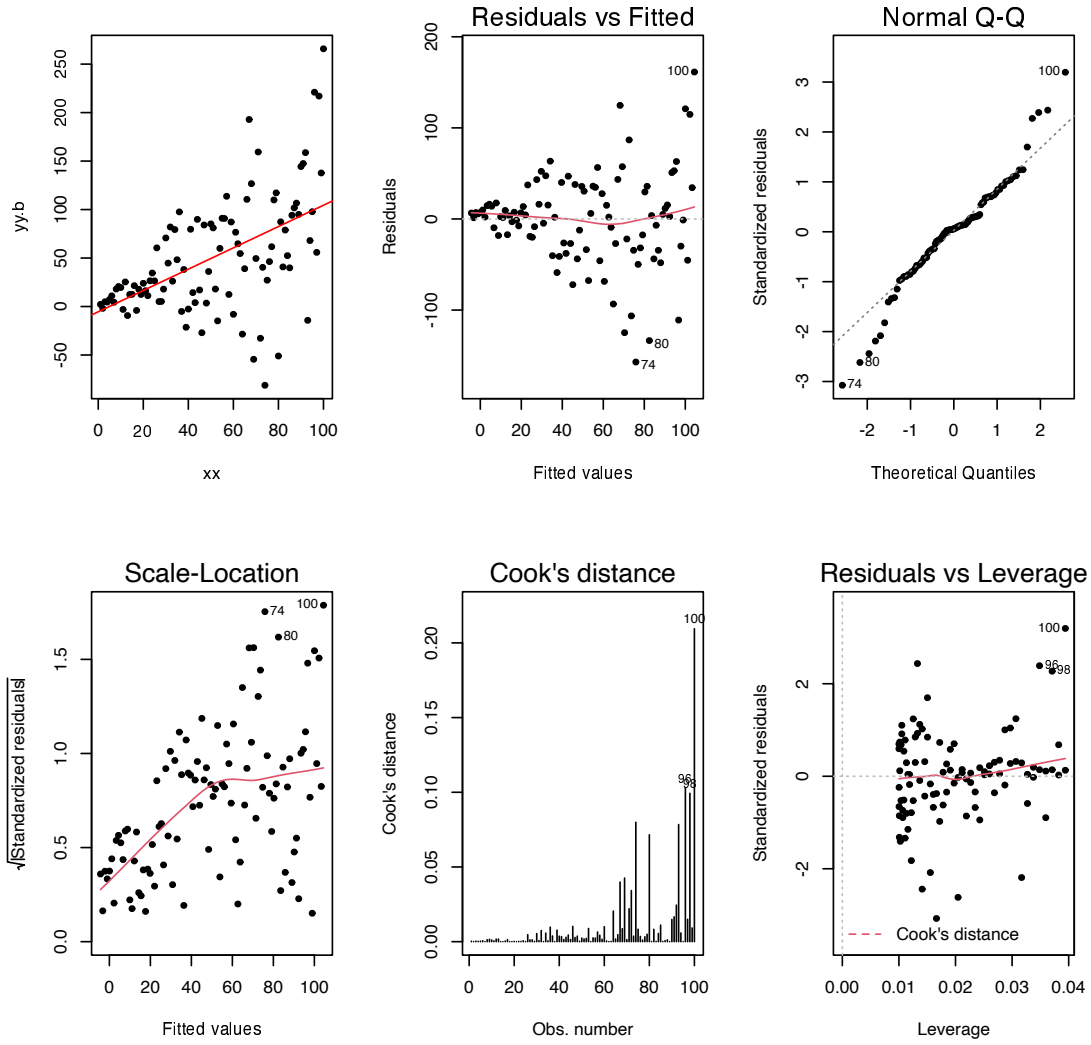


yy.a:

For the first model the residual plots look perfect. Only in the plot displaying 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)
```

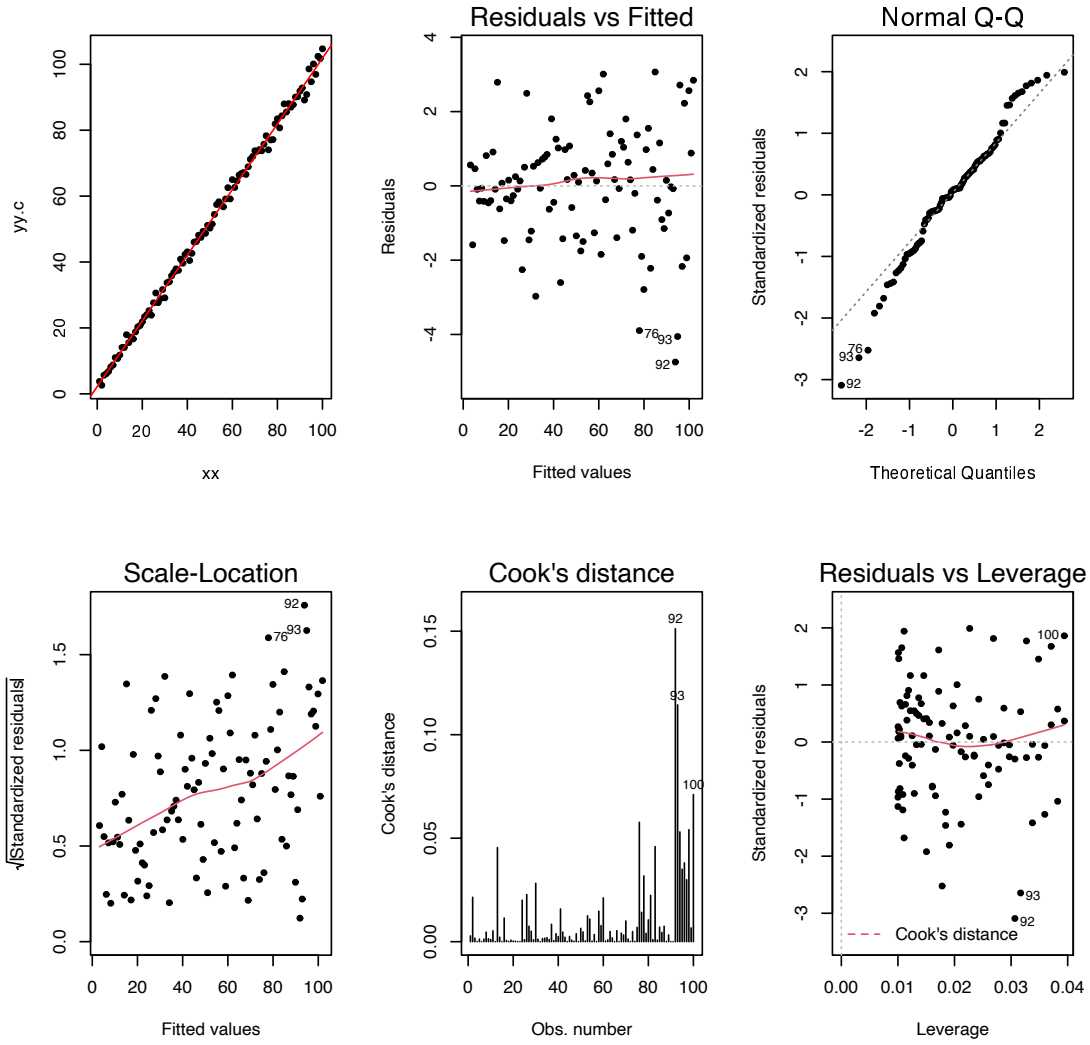


yy.b:

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 originates 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.

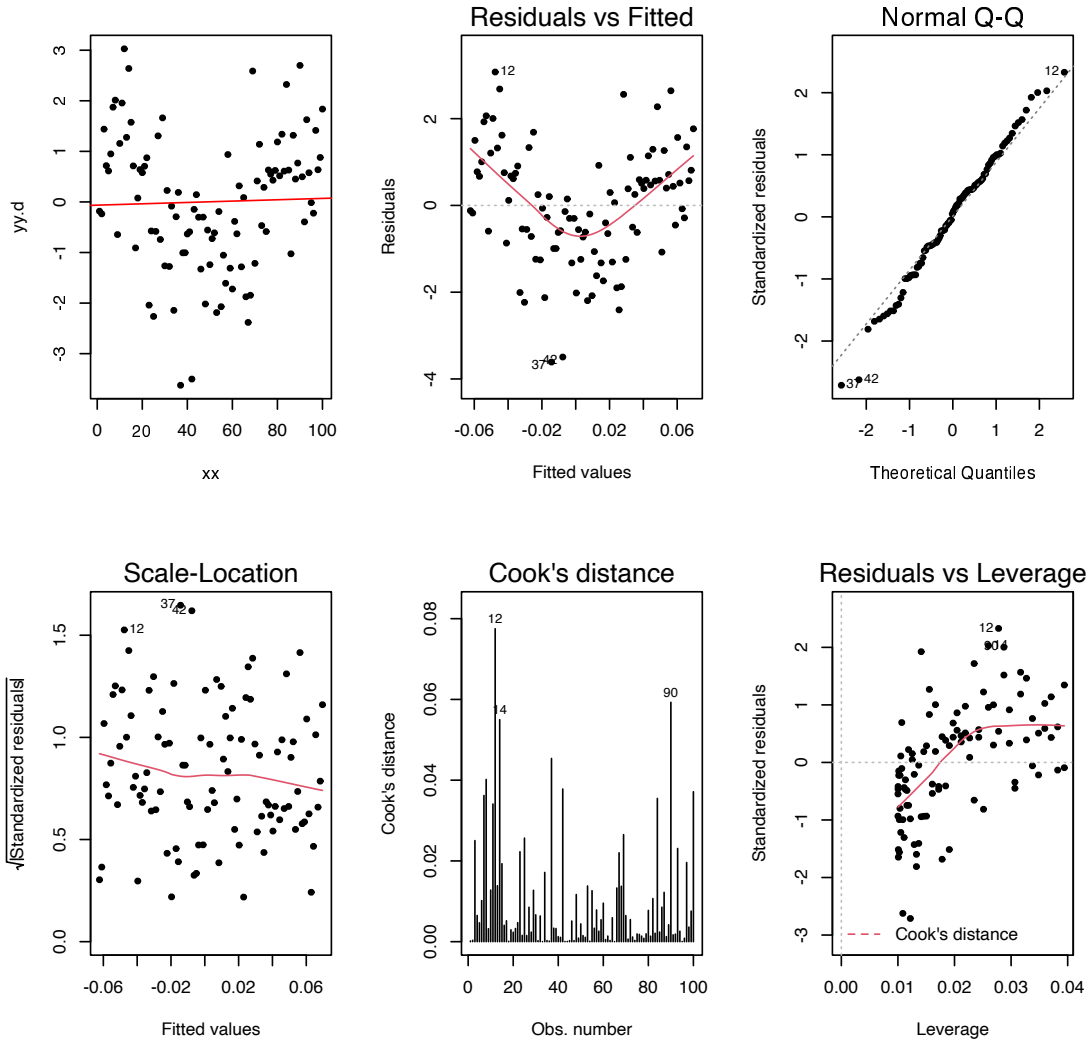
```
## yy.c: scatter plots, residuals and Cook's Distance
par(mfrow = c(2, 3))
plot(yy.c ~ xx, pch = 20)
abline(fit <- lm(yy.c ~ xx), col = "red")
plot(fit, 1:5, pch = 20)
```



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 pronounced than in the previous example.

```r
## 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)
```

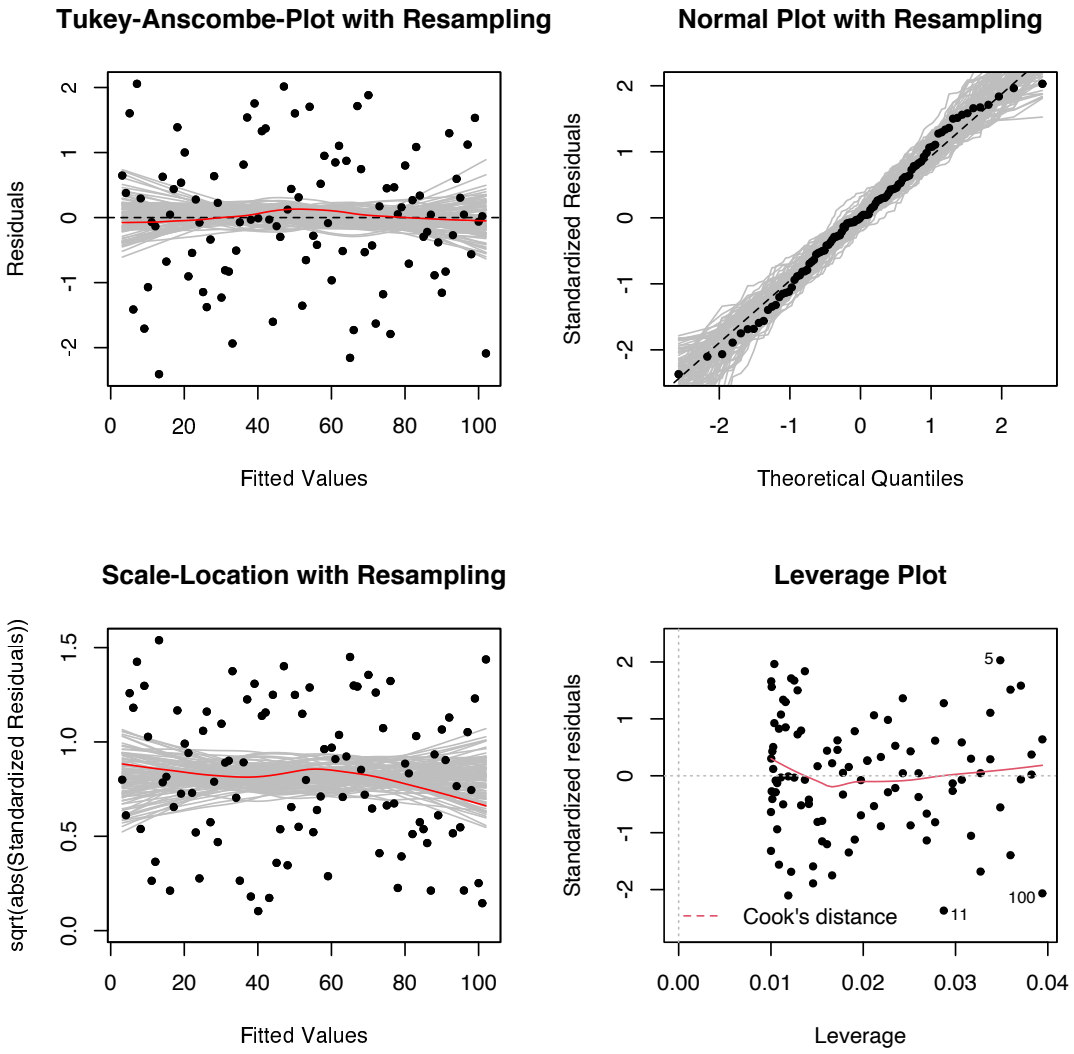


yy.d:

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 horizontal line in the leverage plot. This may be explained by the points with large leverage (i.e. points at the border of this simple regression) have systematically positive residuals.

b) The procedure is shown for the first model:

```
## 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)
```



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