Series 5

1. Assessing model diagnostic plots requires experience. Often it is difficult to decide whether a deviation is a systematic one (i.e. needing correction) or a random one (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 one predictor `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 the predictor.

a) For each model, create a scatter plot with the regression line, plot 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) Instead of `plot.lm()` use the function `resplot()` which is available from the course webpage (`https://stat.ethz.ch/education/semesters/as2015/asr/Uebungen/resplot.R`). The function `resplot()` uses resampling to visualize whether a model violation is present. How does the function perform for the four models?

c) Repeat generating the random numbers a few times (i.e. use different random seeds) and study the variation in the resulting plots. You can also change the number of observations and track the changes in the plots.

2. In this exercise, we would like to analyze how much savings differ between countries. The data set `savings.rda` contains 50 observations. For each country the values are averaged over the entire population and the 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`. Do a residual analysis.

b) Identify the three observations having the largest leverage and describe how these points differ from the remaining data points.

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

d) Now consider variable transformations. Plot the histograms of the individual variables and decide about suitable transformations. Fit the corresponding models and analyze the residuals. Finally, decide which model is most appropriate.

3. The data set `synthetisch.rda` contains the response `y` and the predictors `x1` and `x2`. Fit a multiple linear regression. We can assume that the errors E_i are independent and that the majority of the observations is distributed according to $\mathcal{N}(0, \sigma_E^2)$.

Perform a residual analysis and produce a 3D plot of the regression hyperplane. Does the 3D plot verify your conclusion from the residual analysis?

Hint: The 3D plot can be produced with the function `scatter3d()` from the package `car`. You will also need the package `rgl`.

Preliminary discussion: Monday, November 2.

Deadline: Monday, November 9.