Series 5

1. **Model diagnostics: simulation study** Assessing model diagnostic plots requires experience. Often it is difficult to decide whether a deviation from the theoretical centre 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.

   Simulate the following 4 models: one of them fulfils all model assumptions, other includes a systematic deviation from the linearity assumption,e.g., $\mathbb{E}[\epsilon_i] \neq 0$, and the two left include minor and major deviations from the constant variance assumption.

```
> 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)
```

   a) Decide which model has no violation of the model assumptions, minor deviation to non-constant variance, major deviation to non-constant variance and which model is non-linear.

   b) Plot each response `yy.[a,b,c,d]` versus `xx`. Fit a simple linear regression and plot the regression line into the corresponding scatter plot.

   c) Perform model diagnostics and have a look at the diagnostic plots. Where can we see the deviations? How large is the random variation within these plots?

   d) Repeat generating the random numbers a few times and study the variation in the resulting plots. You can also change the number of observations and track the changes in the plots.

   Assessing normal plots is equally difficult[1]. Even drawing samples from a normal distribution does not result in observations lying directly on the straight line. Now, we will use to following code to simulate new data and see how a skewed, a long-tailed and a short-tailed distribution look like in a Q-Q plot:

```
> qqnorm(rnorm(n), main=c("Normal distribution"))
> qqnorm(exp(rnorm(n)), main=c("Lognormal distribution"))
> qqnorm(rcauchy(n), main=c("Cauchy distribution"))
> qqnorm(runif(n), main=c("Uniform distribution"))
```

   e) Decide which random numbers are normal, skewed, short-tailed and long-tailed.

   f) Repeat generating the random numbers a few more times and study the variation in the resulting Q-Q plots. You can also change the number of observations and track the changes in the plots.

2. **a) Partial residual plots** Use the "prestige" data set from the package `library(car)`. Fit the following model

   `prestige` $\sim$ `income + education`.

   Generate the partial residual plots and perform a general residual analysis. Improve the model by transformation. Plot the resulting residuals versus the variables in the data set not used in the model so far. Considering these plots which variables do you expect to have a strong influence on the response? Add these variables in a stepwise manner as predictors to the model. Keep an eye on the summary output and the diagnostic plots to fit an optimal model.

[1] In the StatsNotes of the Department of Mathematics and Statistics at Murdoch University it reads: *A sufficiently trained statistician can read the vagaries of a Q-Q plot like a shaman can read a chicken's entrails, with a similar recourse to scientific principles. Interpreting Q-Q plots is more a visceral than an intellectual exercise. The uninitiated are often mystified by the process. Experience is the key here.*

b) Correlated errors
Use the "airquality" data set `library(faraway)`. Fit the model

$$\texttt{Ozone} \sim \texttt{Solar.R + Wind}.$$

Perform model diagnostics and check for correlated residuals. Plot the residuals versus the variable `Temp`. Improve the model to get an optimal fit.

3. Braking distance

The file bremsweg.rda contains measurements of braking distance (W, in feet) together with specific starting velocities (V, in mph). Perform a regression analysis.

a) Generate a scatter plot and solve any problems with the data if necessary.

b) Fit a suitable polynomial regression model.

c) Do you think this model is physically reasonable?

d) Perform a residual analysis. Which assumptions are violated?

e) Weighted regression Previously you have seen that the variance is not constant. Therefore, we fit a suitable weighted regression. Compare the results from the weighted and the not-weighted regression (e.g. summary, fitted values, plot fitted curves, residual analysis) and comment on the results.

f) Robust regression We use the data set `data(gala)` from the package `library(faraway)`. Fit a model with the following formula:

$$\texttt{Species} \sim \texttt{Area + Elevation + Scruz + Nearest + Adjacent}$$

Note that in this case the variables should be transformed. Take a look at the residual plots and fit a robust model. Compare the "blind fit" from the above formula with your best robust model fit using the transformed variables. Comment on your results. You can find additional information regarding the data set in the corresponding help file by using the command `?gala`.

Preliminary discussion: Monday, November 12.

Deadline: Monday, November 19.