Series 6

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

$$\texttt{prestige} \sim \texttt{income + education}.$$

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.
Hint. Other variable transformations you may use in this exercise.

- **Count data, where** $y \geq 0$:
  square-root trantsformation: $y' = \sqrt{y}$
- **Proportions** $(0 \leq y \leq 1)$:
  arcsine transformation: $y' = sin^{-1}(\sqrt{y})$
  (if the variable y is a percentage, first perform $y' = y/100$ and then apply the proportions transformation on y').

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.
Hint. When plotting residiuals against the variable `Temp` be careful of the missing values. To test for autocorrelation use the Durbin-Watson test in `R`. For example: `dwtest(fit, alternative="two.sided")`.

2. 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.
Hint. To do a weighted regression in `R` add a parameter `weight` when fitting a linear model. For example: `fit <- lm(y ~ x + z, weights=...)`.
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`.
Hint. To fit a robust regression model in `R` use the function `rlm()`.

3. Which of the following statements are false and why?

a) It is always best to leave unusual observations out of the model.

b) A data point can be both an outlier and an influential observation.

c) In a partial residual plot we plot the response variable y vs. predictor x_k.

d) If the residuals indicate a zero expectation and a non-constant variance, a recommendable model to use would be a weighted regression model.

e) If the residuals indicate a long-tailed and a non-constant variance it is better to use a robust regression model rather than attempt some variable transformations.

Preliminary discussion: Monday, November 24.

Deadline: Monday, December 01.