4 October 2016
Erik Parner

PhD Course in Basic Biostatistics
Exercises, day 5

You are supposed to complete exercises 5.1and 5.2 at the supervised exercises corresponding to day 5. Exercise 5.3 is homework for the next week.

Exercise 5.1
Consider the data on *PEFR*, *height* and, *sex* we looked at during today's lecture, *PEFR.dta*. We will start by looking at the men only (i.e. discard the women) and mimic the analysis done for the women presented in the lectures.

1. Make a scatter plot of *PEFR* against *height* with a fitted line.
   Comment on what you see.

2. Fit the regression line of *PEFR* on *height*. What is the estimated slope, intercept and the standard deviation of the unexplained variation?

3. Generate a new variable *height170=height*-170 and fit the regression line of *PEFR* on *height170*. What is the estimated slope, intercept and the standard deviation of the unexplained variation?
   Compare with what you found in 2.
   What is the interpretation of the intercept?
   Is the slope statistically significant different from zero?

4. Generate the predicted values and the residuals under this model.
   Make a plot of the residuals against the predicted values and the residuals against *height170*. Comment on what you see.

5. Make a QQ-plot of the residuals. Comment on what you see.

We will now look at both sexes.

6. Run the commands
   ```
   regress PEFR b1.sex##c.height170
   regress PEFR b2.sex##c.height170
   ```
   and compare with model 2 on page 34 in the notes.

7. Run the commands:
   ```
   regress PEFR b1.sex c.height170
   lincom _cons + 2.sex
   ```
   and compare with model 3 on page 35 in the notes.

Exercise 5.2

To explore the possible role of a reduced physical fitness in connection with the diminished insulin sensitivity in first-degree relatives (FDR) of NIDDM patients, 21 relatives were examined. Glucose disposal (Rd) was measured during a hyperinsulinaemia euglycaemic clamp. In this exercise we will try to describe the association between Rd and VO2 max. The data are found in *fdr.dta*.

1. Make a scatter plot of Rd against VO2.

2. Compute the regression of Rd on VO2.

3. Check the assumptions.

4. What is the (Pearson) correlation between Rd and VO2.
   Is it statistically significant?
   Discuss the interpretation of and assumption behind this analysis.
   Test the hypothesis of no association between Rd and VO2 using the Spearman rank correlation.

5. Do you think a regression model of log(Rd) on log(VO2) gives a better description of the association between Rd and VO2 than the untransformed data?

6. Make a prediction of Rd for a relative in the study population with a VO2 max of 30, using the model you think fits best.

Exercise 5.3

The rate of evaporation of water from the surface of the skin (transepidermal water loss, TEWL) may be used as an indirect measurement of skin permeability and barrier function. A field study among workers in the fish processing industry was performed to obtain information about skin temperature and transepidermal water loss (TEWL) and their relationship during work. A similar study was performed among a control group. In both groups skin temperature and TEWL were measured at the tip of the 3rd finger at the volar aspect. Differences between the two groups might indicate damage to the skin barrier caused by contact with different irritants or difference in the environmental factors during the measurements. The data can be found in *tewl.dta.*

Previous experience with statistical analysis of a larger data set has shown that the logarithm of TEWL can be described by a normal distribution and $\log_{10}$ is commonly used in the litterature.

1. Compare the distributions of $\log_{10}$(TEWL) in the two groups.

It is well-known that TEWL increases with increasing skin temperature, and that $\log_{10}$(TEWL) depends approximately linearly of the skin temperature. Hence some of the variation in TEWL may be explained by the variation in skin temperature.

2. Compare the distributions of skin temperature in the two groups.

3. Compute the regression of $\log_{10}$(TEWL) on skin temperature in each group. Check and discuss the assumptions.

4. Compare the slopes in the two regression lines.

5. Assume the slopes to be identical and estimate an "adjusted" difference between groups.

6. In the literature it has been shown that an increase in the skin temperature of 5° corresponds to an increase in $\log_{10}$(TEWL) of 0.175. Is this finding compatible with the present data?

7. Write a summary including estimates and their associated 95% confidence intervals for the parameters in resulting model. The summary should include a short discussion of the validity of the model.