November 2, 2010
Morten Frydenberg

POSTGRADUATE COURSE IN
LINEAR AND LOGISTIC REGRESSION
Day 1
Afternoon exercises:

You will start by considering the data set used as the main example in chapter 2 in the book. The data consist of *plasma glycolate* (mg/dl) and *ph* measured on 18 patients with ethylene glycol poisoning. The data is found in the files `glyco.dta` and `glyco.sav`

1. Make a scatter plot of `glyco` versus `ph`.
2. Fit the simple linear regression of `glyco` on `ph`.
   Make a short summary of the results.
3. Find the predicted values, standardized residuals and leverages.
4. Make diagnostic plots to evaluate the validity of the model.
5. Make diagnostics plot to see if any data points need extra attention (large leverage and/or large residual.)
6. Find the expected value of `glyco` (with CI's) for a person with ph=7.4.
   Comment on what you found.
7. Explain why the simple linear regression model you have looked at is invalid.

In the second part you will work on the GFR and Creatinine example from the lecture. The data is found in `gfrdata.dta` and `gfrdata.sav`.
Here we will focus on the simple linear regression of ln(`gfr`) on ln(`cr`)

8. Make a scatter plot of ln(`gfr`) versus ln(`cr`).
   Hint generate two new variables ln(`gfr`) and ln(`cr`) first.
9. Fit the simple linear regression of ln(`gfr`) on ln(`cr`).
10. Find the predicted values, standardized residuals and leverages.
    Make diagnostic plots to evaluate the validity of the model.
    Make diagnostics plot to see if any data points need extra attention (large leverage and/or large residual).
11. Write down the equation for the fitted line, based on the estimates you found in 9. Translate this into an equation involving `gfr` and `cr`. That is, translate the model on the log-scale back to the original scale.
12. Test the hypothesis that the slope in the regression is -1.
    What is the interpretation of the hypothesis in terms of `gfr` and `cr` ?