Prof. Andreas Orth
High Integrity Systems, 2nd semester: Data Mining
Assignment 6 week 6

Look at the data in the table Ass6-dm-data.xls. They consist of cell-count data for different leukocyte-subtypes of healthy children. The first column is age in months. The counts are given in number of cells per µl of blood. They have been measured using flow-cytometry, in which cells are marked and coloured according cluster molecules that they present on their surface and then counted using powerful microscopes and software.

Leukocyte is another name for "white blood cells". Subtypes are Lymphocytes, Monocytes and Granulocytes (the latter have been excluded from this study). The Abbreviation CD means Cluster Designation and is a state of the art way to classify cells in the immune system. In fact CD14+ are monocytes, Lymphocytes have three subtypes: CD3+ are T-cells (cells that have been trainied in the thymus); CD19+ are B-cells (grow up in the bone marrow and are not trained outside); CD56+ are natural killer cells. T-cells are again differentiated into CD3+CD4+, so called T-helper cells and CD3+CD8+, so called cyto-toxic cells (details can be found at www.med.uni-duesseldorf.de/praedimm/immu.html)

Exercise 1:
Do the following "silly" experiment: Try to predict age from the blood counts. Is this possible? Use a model thats linear in all eight markers. (Don't forget to take out line 95, which for some reason contains no data.) Study residuals. Are there outliers? Look at the plot "residuals vs predicted" values. What do you notice? Look at the histogram of residuals. Do you think the residuals are normally distributed?
Predict age for the children in the table "prediction data".

Exercise 2:
Instead of taking the raw marker data, use transformed x-data. This means: take log(x+1) for all data. (You can either do this in R or in MS-Excel resp. Open-Office). Remodel the y-data. Answer the same questions as above. Has the model improved?

Exercise 3:
(a) Do the analysis above using square root of age as the dependent variable (and the logs as in exercise 2). Again analyse residuals. Answer the same questions as above. Predict age for the children in the table "prediction data".
(b) Do the analysis above using log of age as the dependent variable. Again analyse residuals. Answer the same questions as above. Predict age for the children in the table "prediction data".

Which one of the models would you suggest using for further analysis? Why?