UWEO StatR201 – Winter 2013: Homework 3

Due: Thursday February 14, before Class (grace period 1 additional week, with notification)

Assaf Oron, assaf@uw.edu

Reading related to this assignment: Lecture 5; Dobson and Barnett Sections 3.1, 3.2, 3.4; Hastie et al., Sections 4.4.1, 4.4.2, 5.1, 5.2.

Instructions:

- Please submit online in the class dropbox. Please submit either ***.pdf is accepted as the main submission (with code pasted verbatim), or *.rmd.** (tip: to save time when using knitr, use `'cache=TRUE'` in the chunk header for any code chunks that run big simulations – this will prevent them from re-running each time you recompile the code).
- **Starred (*) questions and question-parts are not required.** You may submit them if you choose, or do any part of them without submitting.
- **Grading is determined chiefly by effort, not by correctness. If your submission shows evidence of independent, honest effort commensurate with the amount of homework assigned – you will receive full credit.**

0. If you haven't yet submitted HW2 Question 3, please submit it together with HW3.

1. Logistic Regression interpretation.

Download the 'HosmerLemeshowHeart.csv' dataset, one of the many accompanying the Hosmer and Lemeshow logistic regression textbook. It documents incidence of heart disease for 100 individuals, vs. their age. Follow the steps done in class for the Challenger dataset – this time trying to quantify the prevalence of heart disease as a function of age.

a. Make the usual checks (missing data, outliers, visualize the age distribution, etc.). Create a descriptive table of prevalence vs. age, by cutting age into 5 “bins” using 'quantile'.

b. Run a logistic regression for disease status as a function of age. Make sure to center age in a meaningful way. Interpret the intercept and the age-effect coefficient (the latter, including 95% CI's). Calculate the estimated prevalence of heart disease at age 60, based on this rather modest dataset.

c*. Try to examine whether the effect might be nonlinear. Compare the estimated values at midpoints of the 5 'bins' from (a) above, to the logit of the observed frequencies. If there is a suspected nonlinearity, try to replace the linear age covariate with a polynomial or spline.

2. Splines

Choose another variable from the Boston dataset (but *not* one with many 'funny' values), and repeat the exercise we did in class for distance from job centers: plot vs. the residuals from a model with poverty-rate and # of rooms, identify the regions of strongest nonlinearity, and implement a strategy of adding knots to a natural-spline basis. Decide how many d.f. are needed, and plot the fitted curves.