STA 2201S Homework 3. *due Wednesday April 1 2015 at 11.59pm on Blackboard*

Usual preamble (see HW 2)

1. *Faraway, Exercise 10.1.* The `ohio` data concern 536 children from Steubenville, Ohio, and were taken as part of a study on the health effects of air pollution. Children were in the study for four years from age seven to ten. The response was whether they wheezed or not. The variables are:

   `resp`: an indicator of wheeze status (1 = yes, 0 = no)
   `id`: an identifier for the child
   `age`: 7 years = −2, 8 years = −1, 9 years = 0, 10 yrs = 1
   `smoke`: an indicator of maternal smoking at the first year of the study
   (1 = smoker, 0 = nonsmoker)

   (a) Fit an appropriate GEE model and determine the effects of age and maternal smoking on wheezing.

   (b) What is the predicted probability that a 7 year-old with a smoking mother wheezes?

   (c) Analyse the data using a GLM, with a separate intercept for each subject. Then analyze the data using a GLM but treating all the observations as independent (i.e. ignoring `id`.) Indicate how the conclusions change, and which set of results seem most appropriate for assessing the effect of maternal smoking on wheeze.

   (d) Sum the number of times wheezing is recorded for a child over the four measurements and model this as a function of the smoking status of the mother.[1] Now determine the effect of smoking on the response. Compare this result to the previous analyses and discuss which is preferable.

2. *Faraway, Exercise 11.2.* The dataset `uswages` was drawn as a sample from the Current Population Survey in 1988. The response is `wage`, weekly wages in dollars (adjusted for inflation), and the other variables in the data set are `educ`, years of education, `exper`, years of experience, `race` (1 Black/ 0 White), `smsa` (1 if living in a standard metropolitan statistical area/ 0 if not), a set of dummy variables to indicate region of employment (`ne, mw, we, so`), and a dummy variable to indicate part-time work (`pt`). Of interest is how years of education are associated with wages, and whether this effect is different in different subgroups determined by the other variables.

   (a) Using `wage` as the response, fit one or more semi-parametric regression models, with smooth functions of education (and possibly experience), and including relevant dummy variables as appropriate. Choose the method, and model for that method, that you prefer, and summarize the results.

   (b) Fit a parametric model to `wage` or a transformation of `wage` and compare the results to those of the semi-parametric model that you chose in (a).

[1] `R` code for this is provided in the textbook.

(c) Faraway suggests smoothing the square root of the absolute value of the residuals from the semi-parametric model fit, and smoothing these as a function of `educ`. Why do you think he suggested this?

3. The article "Cognitive control in media multitaskers" by Ophir, Nass and Wagner (*PNAS*, 2009) is widely quoted as evidence that we are not as good at multi-tasking as we think. The authors report on a set of experiments involving different tasks: a filtering task, a continuous performance task, and a set of 'two- or three- back tasks'. This question concerns only the first, filtering task.

   (a) The authors' state "Repeated-measures ANOVA revealed a group×distractor level interaction, $F(1, 39) = 4.61, P < 0.04$". Write down an algebraic form for a linear model for the response, K^2, as a function of the relevant covariates, and describe how you think the F-statistic was computed. Why is the denominator degrees of freedom for the F-test equal to 39?

   (b) Figure 1B is presented as evidence of the interaction effect. How does this plot indicate interaction? Is the model and associated degrees of freedom in (a) consistent with this plot?

   (c) In the Discussion the authors write "The present research suggests that individuals who frequently use multiple media approach fundamental information processing activities differently than do those who consume multiple media streams much less frequently." Do you think they have demonstrated this convincingly? If not, what further information do you think would be useful?

4. *Measurement error in regression:* Suppose y depends on x in a simple linear regression :

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_j, \quad i = 1, \ldots, n; \; \epsilon_i i.i.d. \sim (0, \sigma_e^2), \; x_i i.i.d. \sim (\mu_x, \sigma_x^2).$$

   We assume that x and ϵ are independent. Instead of observing x_i, we are only able to observe a corrupted value $w_i = x_i + u_i$, where u_i is independent of x_i, and $u_i i.i.d. \sim (0, \sigma_u^2)$. The least squares estimator from this regression

$$\hat{\beta}_1 = \Sigma(y_i - \bar{y})(w_i - \bar{w})/\Sigma(w_i - \bar{w})^2.^3$$

   Find an expression for the limit in probability of $\hat{\beta}_1$ and thus deduce that it will normally be an under-estimate of the true regression coefficient β_1. In what special circumstance will it be consistent for β_1?

   Usually this result is relied on to argue that if there is uncertainty in the x's used in a given regression, the association with the response will be attenuated, i.e. less likely to be significantly different from zero. Diggle, Liang & Zeger[4] show for logistic regression that the estimate from the marginal model of GEE is approximately $(c^2\nu^2 + 1)^{-1/2}$ times the estimate from a random effects model, where ν^2 is the variance of the random effect, and c^2 is approximately 0.346.

[2] a measure of performance described in the Materials and Methods supplement

[3] you don't need to show this

[4] Analysis of Longitudinal Data (2002) Oxford, Ch. 7.4