STA 2201S 2014 Assignment 4. *due Friday, April 11 at 2 pm.*

Note: Do Q3 or Q4, not both

1. *Measurement error in regression:* Suppose we have a simple linear regression model

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \quad j = 1, \ldots, n$$

but the x_j are independently $N(\mu_x, \sigma_x^2)$. We assume further that $\epsilon_j \sim N(0, \sigma_e^2)$, independently of the x's. Suppose further that x_j is in fact measured with error, so the observations are (y_j, w_j), where

$$w_j = x_j + u_j, \quad u_j \sim N(0, \sigma_u^2),$$

and u is independent of both x and ϵ.

   (a) Show that $E(y_j \mid w_j) = \alpha_0 + \alpha_1 w_j$, and give expressions for α_0 and α_1 as functions of $\beta_0, \beta_1, \sigma_x^2, \sigma_u^2$ and μ_x.

   (b) Give an expression for the variance of the residual $y_j - E(y_j \mid w_j)$.

   (c) Deduce that the ordinary least squares estimator obtained by regressing y on w has expected value $\lambda\beta_1$, where $\lambda < 1$, and give an explicit expression for λ. This is sometimes referred to as *attenuation* due to measurement error.

   (d) Illustrate this on a set of simulated data. Your answer will be a plot of the data, the regression line for the regression of y on x, and the regression line for the regression of y on w.

   (e) Using simulated data, assess whether or not the same attenuation appears to hold in the logistic regression setting.

2. *SM Exercise 10.8.4*: Suppose that survival data consist of independent observations $(Y_j, C_j), j = 1, \ldots, n,$, where Y_j follows an exponential distributions with mean $\exp(x_j^{\mathrm{T}}\beta)$, and C_j is an indicator variable which equals 0 if Y_j is censored and 1 if Y_j is uncensored. Show that the likelihood for these data is the same as if the counts C_j followed Poisson distributions with means $y_j \exp(-x_j^{\mathrm{T}}\beta)$. Hence show that maximum likelihood estimates for the censored data model, and their standard errors based on observed information, can be obtained by modelling the censoring variable as Poisson with log link function and offset $\log y_j$.

3. *Semi-parametric regression. Do this question or Q4*
The dataset `teengamb` in the package `faraway` gives data on annual gambling expenditure per year (in pounds) (gamble), with several covariates: sex (0 = M, 1 = F), status (a score reflecting socio-economic status), income (pounds per week), verbal (a score from 0 -12 on a test of verbal ability), for a sample of 47 teenagers. Of interest is how covariates are associated with gambling expenditure, and whether there are differences between males and females.

(a) Using log(gamble + .1) as a response, fit one or more semi-parametric regression models, with smooth functions for log(income), status, and verbal, including sex as a categorical variable. Choose the method, and model for that method, that you prefer, and provide a plot of the estimate of the sex coefficient, as a function of the smoothing parameter(s) for the other explanatory variables; Fig. 3 in Peng et al. (2006) does this for the coefficient of PM10, relative to the smoothing parameter for time.

(b) Compare the results from your semi-parametric model chosen in (a) with those from a standard linear regression of log(gamble + .1) on all the variables (with a log-transformation for income).

4. *from Faraway CH. 11, Q5. Do this question or the preceding one.* The dataset `aatemp` in the package `faraway` gives the annual mean temperatures in Ann Arbor, Michigan, from 1854 - 2000. Faraway suggests fitting a smooth function to temperature as a function of year, to see whether this helps determine whether the temperature is changing over time.

   Try several different smoothing methods, choose your preferred method, and model for that method, and summarize the conclusions. Include trace plots of any needed tuning parameters and describe how you chose the value for your final model. Use the model to predict average mean temperature in 2010, and compare to the true value.

   It would be preferable to use a data set on a Canadian city, if you can find it. I would start at `climate.weather.gc.ca`, but there may be better sources.