Due: Feb. 1 by midnight

(1) Use the dataset "heights.dta" which is saved in Canvas.
   (a) Fit a linear regression model predicting earnings from height. What transformation should you perform in order to interpret the intercept from this model as average earnings for people with average height?
   (b) Fit some regression models with the goal of predicting earnings from some combination of sex, height, and education. Be sure to try various transformations and interactions that might make sense. Choose your preferred model and justify.
   (c) Interpret all model coefficients in the preferred model.

(2) Use the dataset "pollution.dta" which is saved in Canvas. We will create a model of mortality that is an extreme oversimplification as it does not adjust for crucial factors such as age and smoking.
   (a) Create a scatterplot of mortality rate versus level of nitric oxides. Do you think linear regression will fit these data well? Fit the regression and evaluate a residual plot from the regression.
   (b) Find an appropriate transformation that will result in data more appropriate for linear regression. Fit a regression to the transformed data and evaluate the new residual plot.
   (c) Interpret the slope coefficient from the model you chose in (b).
   (d) Now fit a model predicting mortality rate using levels of nitric oxides, sulfur dioxide, and hydrocarbons as inputs. Use appropriate transformations when helpful. Plot the fitted regression model and interpret the coefficients.
   (e) Cross-validate: fit the model you chose above to the first half of the data and then predict for the second half. Explain why the procedure you just followed was not really cross-validation.

(3) Use the dataset "ProfEvaltnsBeautyPublic.csv" which is saved in Canvas.
   (a) Fit regression models predicting evaluations ("courseevaluation") given many of the inputs in the dataset. Consider interactions, combinations of predictors, and transformations as appropriate.
   (b) Select a model that you think is best and explain why you chose it over the other models

(4) You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine the logistic regression model that is consistent with this information. (For simplicity you may want to assume that income is measured in units of $10,000).

(5) Use the dataset "hvs02_sorted.csv" which is saved in Canvas and contains data on rodents in a sample of New York City apartments.

(a) Build a logistic regression model to predict the presence of rodents (the variable rodent2 in the dataset) given indicators for the ethnic groups (race). Combine categories as appropriate. Interpret the estimated coefficients.

(b) Add to your model some other potentially relevant predictors describing the apartment, building and community district. Build your model using the general principles explained in Section 4.6. Interpret and explain the coefficients for the ethnicity indicators in your model.

(6) Use the dataset "wells.dat" which is saved in Canvas.

(a) Fit a logistic regression for the probability of switching using as predictors: distance, log(arsenic), and their interaction. Interpret the estimated coefficients and their standard errors.

(b) Make graphs as in Figure 5.12 to show the relation between probability of switching, distance, and aresenic level.

(c) Following the procedure described in Section 5.7, compute the average predictive differences corresponding to the following and discuss the results:

- A comparison of dist = 0 to dist = 100, with arsenic held constant.
- A comparison of dist = 100 to dist = 200, with arsenic held constant.
- A comparison of arsenic = 0.5 to arsenic =1.0, with dist held constant.
- A comparison of arsenic = 1.0 to arsenic = 2.0, with dist held constant.