Problem Set 04
Linear Models – Fall 2015
Due date: 2015-10-28

Problems sets are due at the start of class on the due date. Please hand write or type up and print the solutions; we will not accept e-mail solution sets except in exceptional circumstances. You may discuss problem sets with others, but must write up your own solutions. This means that you should have no need to look at other's final written solutions. Many of these problems come from a variety of textbooks, which are referenced in the problems. These are for citation purposes and not because you will need to consult the text itself (though you may feel free to do so).

I. Binomial regression

This question explores and additional dimension of the airline dataset using binomial regression. I have produced a further downsampled version of the data, available at (`euler.stat.yale.edu/~tba3/stat612/class_data/airline2007_pset04.Rds`).

1. Construct a binary variable `late` that indicated whether a flight arrives over one hour late. Over the entire data dataset, fit the following model using both ordinary least squares and generalized least squares with a logistic link function:

$$late \sim dest. \tag{1}$$

Calculate the predicted values from both of these models, and find the maximum absolute difference between the two sets of predictions. What is going on here? Why?

2. Now, using the same data calculate the local scheduled departure hour (Hint: the data gives the actual departure time; you need to adjust this with the departure delay). Fit the following model on the entire dataset using both ordinary least squares and generalized least squares with a logistic link function:

$$late \sim hour. \tag{2}$$

Again, calculate the predictions from both models and determine the maximum absolute difference between the two sets of predictions. How (and why) does this differ from the answer in question 1?

Plot the predicted probabilities from both models as a function of hour; describe what is going on here and the difference between the two models.

3. Now calculate the model given in Equation 2 for all five binomial link functions available within R: “logit”, “probit”, “cloglog”, “cauchit”, “log”. Produce the same plot as before. Is there a natural ordering to these links?

4. Finally, fit the following model using only data in Group I for both ordinary least squares and generalized least squares with a logistic regression (logit link):

$$late \sim hour + dest + schedTotTime. \tag{3}$$

Now, predict values for each of these on Group II. Calculate the mean squared errors for each and comment on how close or far away from one another they are.

Calculate the centiles for the predictions of both models and plot them against one another. Add vertical and horizontal lines at 0 and a line with a slope of 1 and intercept of 0. Describe the difference between the predicitons of both models.

II. Simulations

1. Construct a random matrix X with 100 columns and 1000 rows filled with independent random uniform variables. Generate a random β vector of length 100 by generating independent trials from a standard random normal distribution.

Simulate the following model for 2000 trials with a fixed X and β:

$$y = X\beta + \epsilon, \quad \epsilon_i \sim_{i.i.d.} \mathcal{N}(0, 5^2) \tag{4}$$

Compute both the ordinary least squares $\widehat{\beta}_{ols}$ and the following modified estimator for α in 0.9, 0.8, 0.5, 0.2, and 0.1:

$$\widetilde{\beta}_{\alpha} = \alpha \cdot \widehat{\beta}_{ols} \tag{5}$$

saving the squared parameter error ($||\beta - \widehat{\beta}||_2^2$) for all 6 models. Which has the best average error rate? Try to explain what is going on (hint: you may want to save other results from each simulation to try to answer this).

2. Redo the previous question with each of the following tweaks (that is, do them one by one) and comment on if, how, and why the results change:

A. use Cauchy errors (i.e., ϵ_i)

B. set p equal to 10

C. set n equal to 10000