Problem Set 3

This problem set is designed to perform predictive tasks using linear models.

You are allowed to work in groups of up to three students, but you must disclose the members of your group. Individual submissions are required. The code you submit may be identical to the one of the other group members, but we expect the comments and answers to the questions to be your own.

Your submission should consist of: (i) a R markdown document with code, figures and comments, and (ii) a zip folder containing all files needed for replication. You may upload these materials via the course's Canvas website (please do not email us with your homework submission).

Each of the four questions below summarize the required tasks, with the individual bullet points detailing the steps needed to complete them.

0) Download the market-level and market-airline-level datasets,

`"market_level.R"`

and

`"market_airline_level.R".`

Set the seed to 0 and randomly allocate 1,000 rows of the market-level data to a test set, to be used only in (7). Use the rest to do the following.

(1) Estimate a linear probability model, predicting whether American Airlines enters a market as a function of the number of competitors. Note: American Airlines' ticket carrier id is "AA".

2) Repeat (1) using a logit model instead of a linear probability model.

3) Repeat (1) using a probit model instead of a linear probability model.

4) Compute non-parametric estimates of the conditional probabilities of entering. (ie compute the conditional probability of entering conditional on each number of competitors directly from the data).

5) Plot the fitted values of each regression in one graph (i.e. estimated probabilities on the y-axis and the number of competitors on the x-axis). In words, explain the coefficients of the first three models. How do the estimated relationships compare? Should we interpret these relationships causally? Are the estimates for the probit and logit similar? Should we have expected this ex ante?

6) Obviously other covariates matter in predicting whether or not American will enter a particular route. In addition to the number of competitors, add the average market distance, market size, hub route indicator, vacation route indicator, slot controlled indicator, and market income to the set of predictors. Fit to the data L_1 regularized logistic regression (ie Lasso for logit, pg 125-126 ESL) where the full model includes all squared terms and second-order cross terms. Using a 10-fold cross validation procedure, find the optimal value of lambda.

7) Calculate the sum of squared prediction errors on the test set for each of your 5 models and put them in a table. Explain your results.