Problem Set 2
Machine Learning
PPHA 30545
Due: Monday, February 7

1. *The objective of this question is to stress the point that the best linear predictor, although we have seen it is the best linear approximation to the conditional expectation, may not be a good approximation and that interpretation of the coefficients must be done with great care.*

a. Suppose that the true model relating y, x_1 and x_2 is

$$y_i = \beta_0 + f_1(x_{i,1})\beta_1 + f_2(x_{i,2})\beta_2 + \epsilon_i,\ i = 1, ..., n,$$

and that $\beta_2 > \beta_1$. Suppose, however, that we run the following model $y_i = b_0 + x_{i,1}b_1 + x_{i,2}b_2 + \epsilon_i$. Come up with β_0, β_1, β_2, f_1 and f_2 such that when estimating the true model on data simulated from the true model we obtain $\hat{\beta}_2 > \hat{\beta}_1$, but estimating the model

$$y_i = b_0 + x_{i,1}b_1 + x_{i,2}b_2 + \epsilon_i, i = 1, ..., n,$$

we obtain $\hat{b}_2 < \hat{b}_1$.

b. Can you provide regression diagnostics that would suggest that the linear model $y_i = b_0 + x_{i,1}b_1 + x_{i,2}b_2 + \epsilon_i$ is the wrong model for the data?

2. *The point of this question is to investigate the different constraints making factorial models interpretable.*

We use the dataset catholic from the paper "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling," by Altonji, J.G., T.E. Elder, and C.R. Taber (2005). We model reading test results using covariates on gender, high school attendance and family income. We consider different specifications of fix effects and interacted models.
In order to get the dataset you have to install the package wooldrige from CRAN and create the dataset with factors using the code in the attached file create_factors_pset_2.

a. Run and interpret the following model: lm(read12 ~ gender).

b. Run and interpret the following model: lm(read12 ~ gender - 1).

c. Run and interpret the following model: lm(read12 ~ gender*hs)

d. Run and interpret the following model: lm(read12 ~ gender*lfaminc)

e. *bonus question*: suppose you run the regression lm(read12 ~ factor1*factor2), where factor1 has m levels and factor2 has n levels. How many of the interaction terms have to be set to zero so that the remaining ones are identified?

3. Do the following questions from Chapter 5 of your class texbook Introduction to Statistical Learning.

a. Question 2

b. Question 3

c. Question 8

d. Question 9