**Advanced Topics in Data Analysis (MGMT 328 – Spring 2018)**

**Homework 2.**

**Problem 1. Data file (nels\_small.dta)**

This is data on school graduates in US.

1. Define a variable COLLEGE that equals one if a high school graduate chooses either a two-year or a four-year college, and zero otherwise. What percentage of the high school graduates attended college? (upon graduating from high school, either no college (PSECHOICE = 1), a two-yearcollege (PSECHOICE = 2), or a four-year college (PSECHOICE = 3).
2. Estimate a probit model explaining COLLEGE, using as explanatory variables GRADES, 13-point scale with 1 indicating the highest grade and 13 the lowest; FAMINC, gross family income in $1,000 increments; FAMSIZ, number of family members; PARCOLL = 1 if the most educated parent had a college degree; FEMALE = 1 if female; and BLACK = 1 if black. Are the signs of the estimated coefficients consistent with your expectations? Explain. Are the estimated coefficients statistically significant?
3. Using the estimates in (b), predict the probability of attending college for a black female with GRADES = 5, FAMINC = sample mean, from a household with five members, with a parent who attended college. Repeat this probability calculation with GRADES = 10.
4. Repeat the calculations in (c) for (i) a white female and (ii) a white male.
5. Re-estimate the model in (b), but omitting the variables PARCOLL, BLACK, and FEMALE. How are the signs and significance of the remaining coefficients affected?

Provide all answers directly in the jupyter notebook.

**Problem 2. Estimating the effect of training no wage earnings.**

In our classroom example on propensity score matching we have used NSW data to construct the control group. Now instead you are asked to use 2 additional control groups: CPS controls 15,992 observations and PSID controls (2490 observations). The purpose is still the same: estimate the impact of training on wages in 1978. The Data can be found on: <http://users.nber.org/~rdehejia/data/nswdata2.html>

Use the python notebook that we have used in the classroom as a basis for your analysis. For your convenience, I’m also attaching the slides I’ve prepared on this topic, where you can find some background info as well as results of estimations. You might get somewhat different outcomes, but conclusions should not defer materially.

Report the results directly in the jupyter notebook.

**Problem 3. Extramarital affairs in Deep Learning framework**

When studying the logistic regression, we covered a model that was trying to predict extramarital affairs for women. You can check that the prediction accuracy we were able to reach using conventional machine learning/econometric technique was 73%. Now, once you were introduced to abc of deep learning, you are invited to look at the same model through neural network prism. Purpose of this exercise is to check whether the accuracy of the model can be improved using deep learning tools. You are free to experiment with the number of layers or hidden units, but as a suggestion consider 2 intermediate layers and 64 hidden units. Again as a starting points try to run the model for 20 epochs.

*Note: Use the notebooks 2.1 2.5 and 3.7 from Deep Learning github depositary and first 3 chapters of Deep Learning book as a reference in case there unclear points on how to start.*

Report the results directly in the jupyter notebook.