Economics 142
Problem Set #5

Part 1: The goal of this part of the problem set is to analyze the effect of field of degree on the gender pay gap for younger workers with a bachelor's degree in the 2010 American Community Survey (ACS). The data set bydegree.csv contains 70,079 observations on people age 26-40 who have a bachelor degree (and no other degrees) and were working in the previous year, and worked full time for the whole year.

agep=age of the person
race = 1 if white, 2 if black, 3 if Asian, 4 if other
logwage=log weekly wage
female = 1 if female
hispanic = 1 if hispanic
field = 2-digit major field of degree
mfield = 1 digit grouped field of degree
st=state of residence.
The mfield categories are:
1=communications and journalism
2=computer science and statistics
3=education
4=engineering
5=social science and psychology
6=business
7=natural sciences
8=humanities, arts and history
9=medical-related
10=all other

1. Construct a 5 group race/ethnicity variable for hispanics (any race), white-non-Hispanics, black-non-Hispanics, Asian-non-Hispanics, and other-non-Hispanics. Show the distributions of males and females in these 5 categories.

2. a) Show the fractions of males and females in each category of "mfield" and the mean log weekly wages (logwage) of both groups in each category.

b) Calculate the weighted average mean log wage females and males across all categories. What is the male-female log wage gap? Compute a counterfactual mean for females if they had the same distribution across categories as males. Using this counterfactual, explain what fraction of the female-male log wage gap is "explained" by mfield.

c) Calculate a counter-factual mean for males if they had the same distribution across categories as females. Using this counterfactual, explain what fraction of the female-male log wage gap is "explained" by mfield. Why is it different from your answer for part b?

3. In this part we will try to account for use differences in major field, race, and age between males and females using a reweighting model estimated using a logit.

e) Fit a *logistic model* for the probability of being a male, as a function of age race, and mfield. Specifically define the outcome $m_i = 1[male]$, and fit the logistic model with a constant, a linear and quadratic age term, dummies for the ethnicity/race categories, dummies for each of the possible categories of mfield, and **interactions** of age with the dummies for the categories of mfield. Save the predicted probability $\hat{p}_i$, which is the predicted probability individual i is male. Form the weight $w_i = \hat{p}_i/(1-\hat{p}_i)$ if person i is female and $w_i = 1$ if i is male.

a) Check that your weights are working by calculating the weighted average fractions of females in each category of mfield, and comparing this to the unweighted fraction of females in the category, and the unweighted fraction of males in the category. The weighted mean should be close to the mean for males.

b) regress log weekly wage on a female dummy to find the the "unadjusted" male-female wage gap. Check that this is equal to the gap you calculated in question 2a.

c) Re-run the regression using weighted least squares with weights w_i as calculated above. You should find that the female coefficient is reduced by about 25% in magnitude.

d) As an alternative to reweighting you could also run an unweighted regression of log weekly wages on a female dummy, linear and quadratic age term, dummies for the ethnicity/race categories, and dummies for each of the possible categories of mfield. You should find that the female coefficient is reduced by about 25% in magnitude relative to the unweighted OLS regression without other controls.

Part 2: In this part of the problem set we will fit a few models to the Princeton Twins Survey data.

The data set is called twins142.csv. The variables are:

famid = family id variable

t=1,2 for twin #1 or twin #2

age = age (some observations have coded in part year values)

educ = education

oeduc = education of twin

lw = log wage

married = dummy 1 if married 0 if not

omarried = dummy if twin is married

female = 1 if female

ofemale = 1 if twin is female.

exp = "labor market experience" = age - educ - 6

oexp = experience of twin

NOTE: all twins are one of two identical twins in the data set. So female=ofemale in all cases.

1. Estimate a simple model relating log wages to: education, experience, experience-squared, married status, and female.

2. Fit the same model separately for men and women. Does the model look different?

3. Construct the mean family marriage rate for each person in the data set (i.e., the fraction of the twins that is married, which can be 0, 1/2, or 1).

a) verify that when you regress marriage of a twin on the average fraction of the siblings who are married, you get a coefficient of 1.

b) add mean fraction of siblings married to your gender-specific wage models from part 2. How does the addition of this variable affect the estimated coefficient on marriage. Give an interpretation of the patterns and how they differ between men and women.