Economics 142
Problem Set #3

1. Consider a regression model of the relationship between y_i and a vector of explanatory variables x_i: $y_i = x_i'\beta + u_i$. Let $\hat{\beta}$ denote the OLS estimates of the coefficients, which are assumed to satisfy the condition: $\sum_{i=1}^{N} x_i(y_i - x_i'\hat{\beta}) = 0$ where N is the sample size.

(a) Show that if x_i contains a constant, then $\bar{y} = \bar{x}'\hat{\beta}$, where
$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

(b) Show that if x_i contains a dummy variable for membership in group g (which has N_g observations in the sample) then $\bar{y}_g = \bar{x}_g'\hat{\beta}$, where
$\bar{y}_g = \frac{1}{N_g}\sum_{i\in g} y_i$, and $\bar{x}_g = \frac{1}{N_g}\sum_{i\in g} x_i$

(c) Complete the proof of the Frisch-Waugh theorem for the sample OLS regression coefficients by showing that the j^{th} row of $\hat{\beta}$ is:

$$\hat{\beta}_j = [\frac{1}{N}\sum_{i=1}^{N} \hat{\xi}_i^2]^{-1}[\frac{1}{N}\sum_{i=1}^{N} \hat{\xi}_i y_i]$$

where $\hat{\xi}_i$ is the *estimated residual* from an OLS regression of x_{ji} on all the other $x's$:

$$x_{ji} = x_{(\sim j)i}'\hat{\pi} + \hat{\xi}_i.$$

HINT: follow all the steps used in Lecture 4 for the corresponding properties of the population regression coefficients.

2. On the course web site you will find a .csv data set called ovb.csv (for "omitted variable bias"). This has 11,306 observations on men and 10,601 observations on women who are age 35-44, and worked in 2011, and were surveyed in the March 2012 Current Population Survey. The variables are: age (in years); female (0 or 1); imm (an indicator for immigrant status); hispanic (an indicator for hispanic ethnicity); black (an indicator for black race), asian (an indicator for asian race); educ (years of education, ranging from 0 to 20, with value of 12 for high school grads, 16 for people with a BA, etc); wagesal (total earnings last year); wage (average hourly wage earned last year); logwage (the log of wage); state (a categorical variable indicating state of residence, 93=California, 74=Texas, 21=NY, etc); and 3 indicators for government workers (fedwkr, statewkr, and localwkr, which are 1 if the person works for the Federal govt, a state government, or a local government). We are using the Census Bureau convention that people can be Hispanic ethnicity and of any race, so it is possible to be Hispanic and Asian.

As a way to check you have captured the data correctly, the last page of this problem set shows the means of the variables (plus their mins and maxes) for men (female=0) and women (female=1).

In lecture we presented a table showing regressions for **female** workers in this sample. There were 5 models:

1. logwage = constant, immigrant status
2. log wage = constant, education
3. immigrant status=constant, education
4. education = constant, immigrant status
5. logwage= constant, education, immigrant status

(a) Using the omitted variable formula, write out an expression for the OLS estimate of the coefficient on immigrant status from model (1), if the true model is model (5).

(b) Using a regression package, estimate the 5 models, and show the values of the terms for part (a), first for females, then for males. Your answers for females should be the same as the ones reported in the table in Lecture 5.

(c) Consider 3 groups of immigrants: ***Asian immigrants*** are those with (asian=1) and (hispanic=0) and (imm=1). ***Hispanic immigrants*** are those with (hispanic=1) and (imm=1). ***Other immigrants*** are those with (imm=1) who are not included in the first 2 groups. Redo the 5 models for females and for males, distinguishing the 3 groups of immigrants. So your models will have 3 separate dummies for the 3 immigrant groups, treating natives as the omitted group. Put your results in 2 new tables that are similar to the table in Lecture 5, and include these tables in your anwers.