Instructions:

1. This problem set is due at the beginning of class on October, 29.
2. You must hand-in hard copy of your homework.
3. If your homework is completely typed (including equations), then you will receive up to 10% of extra credit. You cannot earn above 100% on a homework.
4. Each part of every single question is worth 10pts.
5. **You MUST show ALL of your work to receive full credit.**
6. **You MUST turn PART I separately from the PART II.**

PROBLEM SET 5
Due October 29

PART I.

1. A researcher plans to study the causal effect of police on crime, using data from a random sample of U.S. counties. He plans to regress the county's size of the police force on county's crime rate (per capita). Define the variables as following: cRate=crime rate per capita and PoliceSize=size of the county's police force.

   a) Write down the population model.
   b) Would you expect the effect of β_1 to be positive or negative?
   c) How would you interpret the coefficients from your population regression?
   d) Name another **3 factors** that at the same time affect crime rate and size of the county's police force besides population of the county (**variable** Population), and explain your reasoning.
   e) Suppose that the only factor besides cRate affecting PoliceSize is the population of the county. Also, assume that you do not have access to data on the population of the county. What kind of bias would you expect to see in your estimate of β_1 because you are unable to control for the size of the population? Why?
   f) Name one irrelevant variable, i.e. a variable that does not belong to the model. Explain why that variable is irrelevant.
   g) Name one variable that is purely a control variable, i.e. variable that affects the dependent variable but is not correlated with the independent variable of interest.

2. Table 1 reports the estimates of the effects of personal characteristics on a monthly wage. The variable birth order measures order of birth, i.e. 1st child, 2nd child etc. Use this table to answer the following questions:
    a) How many different regressions are estimated in Table 1?
    b) Write down the estimated regression for column 6 and make sure to name variables in a meaningful way.
    c) Construct 95% confidence intervals for all coefficients in column 6.
    d) Interpret each coefficient in column 6. (Hint: *Do not forget to state if the effect is statistically different than 0).*
    e) **The total variation** in monthly wage is 138787905. Calculate adjusted R^2 for each regression.
    f) What is predicted wage for an individual who has 11 years of experience, has 12 years of education, works 40 hours per week, has 2 years of experience with current employer, is a 2nd child in family, has two siblings, is married, has a father who has 8 years of education, is not black, does not live in south, and lives in metropolitan area?
    g) Compute residual for this individual given that the observed monthly wage is $759.
    h) What is the difference in wages between single individual with 10 years of experience and a married individual with 8 years of experience, all else equal?
    i) In the column 6, are the effects of being black, living in the south and living in metropolitan jointly important variables in determining monthly wage? (Hint: You can use R^2 to test this hypothesis).
    j) In the column 5, are the effects of being black and living in south jointly important variables in determining monthly wage? (Hint: You can use R^2 to test this hypothesis).
    k) Do explanatory (independent) variables in column 6 jointly explain monthly wages? (Hint: You can compute RSS from the unrestricted model using R^2 and TSS from part e.)

Table 1. Regression results of the effects of personal characteristics on monthly earnings

	(1) monthly earnings Coeff./Std. err.	(2) monthly earnings Coeff./Std. err.	(3) monthly earnings Coeff./Std. err.	(4) monthly earnings Coeff./Std. err.	(5) monthly earnings Coeff./Std. err.	(6) monthly earnings Coeff./Std. err.
years of work	14.86***	13.22***	12.54***	16.80***	16.26***	15.60***
experience	(3.25)	(3.41)	(3.38)	(3.89)	(3.88)	(3.79)
years of education	74.89***	70.91***	72.08***	62.81***	62.37***	59.64***
	(6.30)	(6.78)	(6.72)	(7.75)	(7.70)	(7.54)
average weekly hours	-1.73	-1.32	-1.60	-3.58	-3.85	-4.13*
	(1.70)	(1.80)	(1.79)	(2.00)	(2.00)	(1.95)
years with current	8.13**	7.56**	7.00**	3.98	3.31	4.15
employer	(2.50)	(2.59)	(2.57)	(2.91)	(2.91)	(2.84)
birth order		-12.62	-11.59	-15.55	-15.08	-18.84
		(10.01)	(9.92)	(11.73)	(11.78)	(11.52)
number of siblings		-8.15	-8.33	-0.81	2.79	5.65
		(7.03)	(6.97)	(7.77)	(8.00)	(7.83)
=1 if married			180.88***	179.09***	173.90***	179.80***
			(43.69)	(47.40)	(47.22)	(46.13)
father's education				17.58***	14.93**	12.62**
in years				(4.85)	(4.90)	(4.81)
=1 if black					-123.12*	-163.81**
					(54.63)	(53.83)
=1 if live in south					-55.12	-33.42
					(31.12)	(30.63)
=1 if live in						179.56***
Metropolitan area						(31.19)
Constant	-205.47	-84.83	-242.33	-237.75	-160.03	-223.27
	(127.51)	(142.52)	(146.21)	(164.27)	(165.38)	(161.91)
R-squared	0.1468	0.1529	0.1698	0.1828	0.1944	0.2325
N. of cases	935.0000	852.0000	852.0000	680.0000	680.0000	680.0000

4. Use the data set named "Earnings_and_Height.dta" under Files\Problem Set\Problem Set 5\ on Carmen to answer following question.

   These data are taken from the US National Health Interview Survey for 1994. They are a subset of the data used in Anne Case and Christina Paxson's paper "Stature and Status: Height, Ability, and Labor Market Outcomes," Journal of Political Economy, 2008, 116(3): 499-532.

   The dataset contains information on 17,870 workers. The table on the last page describes the variables.

   a. Generate summary statistics table for variables: earnings, height, weight, race, educ, age, sex.
      i. Notice that you have categorical variables here which means that you will need to create binary variables. (In PS 3, you have learned commands to generate new variable. Make sure to give meaningful names to your new variables).
      ii. When you create new variables, you will need to calculate basic statistics number of observations, mean, standard deviation, min, and max. (You have used command that generates this information in PS3.)
      iii. These two commands will produce a table of basic statistics in a word document:

```
estpost sum (write names of your variables here)
esttab using "mystats.rtf", replace cells(" count mean(fmt(2)) sd(fmt(2)) min(fmt(2)) max(fmt(2))")  noobs
```

- After you run the command, in the command window there will a link: (output written to basicStats.rtf). Click on this link for the file to open.
- You can copy-paste this table into another document if necessary. Make sure to edited table.
- Your table should have a name and be nicely organized. See example below.

Example of a good table:

Table 1. Summary Statistics

Variable	Observations	Mean	Standard Deviation	Min	Max
Earnings	17870	46875.32	26923.29	4726.39	84054.75
Education	17870	13.54	2.64	0.00	19.00
Northeast	17870	0.20	0.40	0.00	1.00
Midwest	17870	0.26	0.44	0.00	1.00
South	17870	0.32	0.47	0.00	1.00
West	17870	0.22	0.41	0.00	1.00

PART II: PROJECT

1. At this point you have your question and two variables of interest: outcome variable y, and main explanatory variable x.
    a. Now, it is time to think about **all other factors** that might affect your dependent and independent variable (IMPORTANT VARIABLES) and about other variables that you might want to introduce into your regression to reduce variation in your outcome variable (CONTROL VARIABLES).

STEP 1: You should **list at least 5 other factors** that affect your x and y **at the same time**.

STEP 2: When you list each factor make sure to describe how each factor is affecting both your dependent(y) and independent (x) variables at the same time, i.e. here you will be talking about your expectations about the correlation of these other variables with your y and main x.

STEP 3: Check if any of the 5 variables you listed are in the data set.

- The factors that you think are important but you cannot measure since there are not available in the data set, you want to keep for later discussion. You will need them at the later stage when you discuss omitted factors that could impact your results.

STEP 4: Think about what CONTROL VARIABLES you want to include in the regression. See if these variables are present in the data set.

STEP 5: If you do not have 5 variables, go to step 1 and do step 2-4 until you get 5 variables. You must have at least one variable that is IMPORTANT VARIABLE.

b. After you identify the variables of interest (dependent and independent variables, and other factors), it is time to prepare data set for estimation.
    - After you finish, your data set should have **at least 7** variables.
    - Your finalized data set should have only variables of interest.
        - Use **command** *keep* to save variables that you will be using for your project.
    - Each variable should have appropriate values (i.e. if you are using wage in your project, then you cannot have for example wage being negative or 0).
        - You can use **command** keep if var_name<= or keep if var_name>= to drop observations which do not make sense.
        - If you have categorical variables, then you need to create binary variables (Make sure to name your binary variable is the same as your base group. For example, Sex=1 if female and Sex=2 if male. I want to make a dummy variable female=1 if female and 0 otherwise. My base group are females. Thus, I will name variable female).
        - When you are done "cleaning" data set, you should have all variables having the same number of observations

b. The last step is to create a table that looks like the table in PART I question 3 by using similar commands as in question 3, and then write a few paragraphs describing this table.