Problem 2

Suppose you work at the U.S. Bureau of Labor Statistics and your task is to analyze the effect of education on the wages.

Data Set

First you need to obtain the data for the analysis. Recommended data set would be National Longitudinal Survey of Youth 1997, which you can access through the following LINK.

Once you enter in NLS Investigator, make sure to select following options:
Select the study you want to work with:
NLSY97 (National Longitudinal Survey of Youth 1997)
Select a substudy:
NLSY97 1997-2017 (round 1 - 18)

In the **Review Selected Variables**' tab, you can see 6 basic demographic variables already pre-selected:

NLS Data - Required ID Variable (PUBID) and Recommended Demographic Variables

R0000100	PUBID	Pubid, youth case identification code
R0536300	KEY!SEX	Key!Sex, rs gender (symbol)
R0536401	KEY!BDATE_M	Key!Bdate, rs birthdate month/year (symbol)
R0536402	KEY!BDATE_Y	Key!Bdate, rs birthdate month/year (symbol)
R1235800	CV_SAMPLE_TYPE	Sample type. Cross-sectional or oversample
R1482600	KEY!RACE_ETHNICITY	Combined race and ethnicity (symbol)

Additionally, you might find following variables important to include in the analysis (please choose wisely between suggested variables to avoid co-linearity between regressors, do not include very similar variables together):

NLS Data - additionally suggested variable candidates (not exhaustive list)

Education degree/test results		
Z9083900	CVC_HIGHEST_DEGREE_EVER	Highest degree received
R9829600	ASVAB_MATH_VERBAL_SCORE_PCT	ASVAB MATH_VERBAL score percent
Z9033700	CVC_SAT_MATH_SCORE_2007	Highest SAT math score 2007
Z9034100	CVC_ACT_SCORE_2007	Highest ACT score 2007
Experience –*this is a cumulative measure capturing labor market experience*		
Z9065401	CVC_WKSWK_ADULT2_ALL	# Weeks all jobs from age 20
Marital Status	*take for the most recent year*	
E7013810	MAR_STATUS_2018.10	2018 Marital: marital status in month 10
E7023102	MAR_COHABITATION_2011.02	2011 Marital: cohabitation status in month 02
Children – *take for the most recent year*		
U1852600	CV_BIO_CHILD_HH	# Bio children r has in household
Parental Education - Mother		
R1302500	CV_HGC_BIO_MOM	Biological mothers highest grade completed
R1302400	CV_HGC_BIO_DAD	Biological fathers highest grade completed
R1302700	CV_HGC_RES_MOM	Residential mothers highest grade completed
R1302600	CV_HGC_RES_DAD	Residential fathers highest grade completed
Income – *take for the most recent year*		
U2857200	YINC-1700	Total income from wages and salary in past year

If you find other variables that you consider relevant and important in the data set, feel free to include them in your analysis as well.

Please see the detailed descriptions of the variable on website.

Analysis

Let us simplify the task and check how individual's income changes with obtaining college degree. As a starting point, consider the following simple wage determination model, but feel free to add more explanatory variables if you find them relevant:

$$lnWage = B_1 + B_2Degree + B_3Exper + B_4Exper^2 + u_i \tag{1}$$

Present you work in a form of a short (maximum 3-page) report that will summarize your empirical approach, the data used, results, comments, and conclusions. Treat it as a real report that would be distributed among the management of the institution. Treat the points below as a suggestive checklist.

a. Data preparation:

- Download data from the webpage; keep only males in the dataset.
- prepare a table with summary statistics. Remember that this should be the statistics for your final estimation sample, not for all the downloaded data.
- Finally explain your rational behind choosing them for your analysis.

b. Think about the relationship between college degree and wages:

- Write down the econometric model you will use to estimate the relationship between education and wages.
- Estimate your model and explain the results
- Do you expect any of the explanatory variables in your model to be endogenous? Explain your intuition.
- If you suspect endogeneity, then are your OLS results reliable? Are they unbiased? Efficient? If you expect a bias - in which direction it goes?

c. When an explanatory variable is endogeneous, one can deal with the problem caused by endogeneity by using a proxy variable. Is it possible to use this approach here? If yes, than which variable you would use as a proxy. Explain. Run the relevant estimation if possible.

d. Another way of dealing with endogeneity is using instrumental variables. Can you find proper instrument/instruments in the data? Explain why the chosen variable/varibles might be a good instrument/instruments. Use proper statistical tests whenever possible.

e. Present estimation results using all methods that you have tried, preferably in one table so that regression coefficients can be compared. Comment on similarities and differences and explain which model is the most reliable and why. Do not forget to reply to the main question of this task.