Department of Economics **UN3412**
Columbia University **Fall 2020**

Midterm Exam
Introduction to Econometrics
for both sections
(Best and Erden)

Instructions:

This exam has three questions and an academic integrity statement that you have to "sign".

Go to Courseworks Quizzes, start the exam there so that Proctorio starts for you. Keep that tab open during the exam at all times. Then go to Gradescope on a separate tab and start the exam.

Please write your answers neatly by hand on paper, or type them into a Word document. If you choose to copy/paste some of your Stata output to Word best font is `Bold Courier New Size 9` for this purpose. When you are done, upload your Word document or scan and upload your handwritten solutions as a **single** PDF by merging all your answers. **Please also upload your document to Gradescope when you are finished (just like you did with every problem set).** An extra 20 minutes has been provided for these tasks.

Make sure to specify the page number of each question when submitting to Gradescope. Once you upload your solutions to Gradescope, go back to the tab where Quizzes (and Proctorio) is and check the box that you submitted the exam and click submit. This will end your exam (and stop Proctorio).

Some questions ask you to draw a real-world judgment in a problem of practical importance. The quality of that judgment counts. For example, consider the question: "It is 10°F outside. In your judgment, why are so many people wearing heavy coats?" The answer, "To stay warm" would receive more points than the answer, "Because they are fashion-conscious."

Exam Support Options:

- **If you have any questions during the exam. You can ask questions through Piazza** or you may contact your professor and/or TA using the private chat function in your **Zoom** session. If you have internet connection problems, Proctorio will eject you. If it does, once you are reconnected, please contact Proctorio support via **Live Chat** and they will ensure you do not lose any exam time.
- **If you have any issues with Proctorio during the exam:** Proctorio has a 24/7 Live Chat support.
  - At any time before or during the exam, you can initiate a live chat with a support experts by clicking on the **Secure Exam** Proctorio shield icon in the Google Chrome address bar, and selecting the **Live Chat** button; OR, by clicking on the **Live Chat** icon in the **Quiz Tools** utility on your screen.
  - In addition, Proctorio has a 24/7 toll-free number you can call: **866.948.9248** (prior to or after the exam only - unless cellphones are permitted during the exam).

Question 1 (33 points):

The dataset sleep75.dta contains data from Biddle and Hammermesh (1990) that they used to study the tradeoff between time spent sleeping per week and the time spent in paid work. You can find and download the dataset in courseworks under files/Fall2020/. Let's start by studying the regression model

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 educ_i + \beta_3 age_i + u_i$$

Where $sleep_i$ and $totwrk_i$ (total work) are measured in minutes per week, and $educ_i$ and age_i are measured in years.

(a) (11p) Run this regression. If someone works five more hours per week, by how many minutes is sleep predicted to fall?

(b) (11p) Would you say that totwrk, educ, and age explain much of the variation in sleep? What other factors might affect the time spent sleeping? Are these likely to be correlated with totwrk?

(c) (11p) Are either educ or age significant at the 5% level? Are they jointly significant? Does including them change the effect of work on sleep much?

Question 2 (35 points):

These data are taken from the *HighSchool and Beyond* survey conducted by the Department of Education in 1980, with a follow-up in 1986. The survey included students from approximately 1100 high schools. The data in **CollegeDistance_for Midterm.dta** has the following variables:

Name	Description
ed	Years of Education Completed (See below)
ln_ed	Log of education
female	1 = Female/0 = Male
black	1 = Black/0 = Not-Black
dadcoll	1 = Father is a College Graduate/ 0 = Father is not a College Graduate
momcoll	1 = Mother is a College Graduate/ 0 = Mother is not a College Graduate
incomehi	1 = Family Income > $25,000 per year/ 0 = Income ≤ $25,000 per year.
blackxincomehi	black times incomehi
urban	1 = School in Urban Area / = School not in Urban Area
dist	Distance from 4yr College in 10's of miles
blackxdist	black times distance
tuition	Avg. State 4yr College Tuition in $1000's

Years of Education: Rouse computed years of education by assigning 12 years to all members of the senior class. Each additional year of secondary education counted as a one year. Students with vocational degrees were assigned 13 years, AA degrees were assigned 14 years, BA degrees were assigned 16 years, those with some graduate education were assigned 17 years, and those with a graduate degree were assigned 18 years.

You can find and download the dataset in courseworks under files/Fall2020/.

(a) (7p) Run the following regression:

$$ln\_ed_i = \beta_0 + \beta_1 black_i + \beta_2 dist_i + \beta_3 blackxdist_i + \beta_4 momcoll_i + \beta_2 incomehi_i + u_i$$

Why would one want to include the interaction variable, $blackxdist_i$? How do you interpret the coefficient of this variable?

(b) (7p) The interaction variable is not significant, test if $black_i, dist_i, blackxdist_i$ are jointly significant? Explain the reason for your conclusion.

(c) (7p) Run the following regression:

$$ln\_ed_i = \beta_0 + \beta_1 black_i + \beta_2 dist_i + \beta_3 momcoll_i + \beta_4 incomehi_i + \beta_5 blackxincomehi_i + u_i$$

Why would one want to include the interaction variable, $blackxincomehi_i$?

(d) (7p) Write the null hypothesis to check whether there is a significance difference in average education of high income blacks and high income non-blacks? Run the test. What is your conclusion?

(e) (7p) List interval validity threats (reasons for $\hat{\beta}_1$ to be biased) and give an example for two of those threats within this setting.

Question 3 (32 points):

Please answer following questions. First two parts, parts (a) and (b) will consider studying the relationship between the percentage of students at a school that pass a 10th grade math test and the size of the school's budget per student. Parts (c) and (d) are not related to the first two parts.

(a) (8p) We regress the pass rate on expenditure and find the following relationship:

$$\widehat{math10}_s = 13.36 + 2.456 expend_s$$
$$(2.93) \quad (0.660)$$

Where $math10$ is the percentage of students at the school who passed the math test and $expend_s$ is expenditure per pupil, in thousands of dollars. Suppose that we believe $expend_s$ is measured with classical measurement error:

$$expend_s = truespend_s + v_s$$

What can we say about the bias in the coefficient on $expend_s$ in this case? In this data the variance of $expend_s$ is 0.6. Suppose we knew (somehow) that the variance of v_s was 0.15. Is it possible to write the coefficient of $truespend_s$ in the regression below as a function of the $\hat{\beta}_{expend}$, σ^2_{expend} and σ^2_v? If so, calculate the coefficient of $truespend_s$ using the information provided in the question.

(b) (8p) We might think the relationship between expenditure and test scores is non-linear. We add a quadratic term to the regression model:

$$\widehat{math10}_s = \underset{(15.4)}{37.08} - \underset{(6.40)}{7.52}expend_s + \underset{(0.64)}{1.01}expend_s^2$$

The adjusted R-squared of this regression is 0.034 while it is 0.031 in the regression without the quadratic term. Do you think we need to add the quadratic term to capture the nonlinearity of the relationship?

(c) (8p) Assume that you have a regression $Y_i = \beta_0 + \beta_1 X_i + u_i$ with heteroskedastic errors. Can the OLS estimator be unbiased?

(d) (8p) In general, why would you want to include an interaction term between two continuous variables in the following regression? $Rent_i$ is rent amount of an apartment in NYC (in $), $Size_i$ is the size of the apartment in square footage and $Year_i$ is the year that the apartment building was built and $SizexYear_i$ is the interaction variable generated by multiplying $Size_i$ by $Year_i$.

$$Rent_i = \beta_0 + \beta_1 Size_i + \beta_2 Year_i + \beta_3 SizexYear_i + u_i$$