Empirical Project - due December 1 at 11:59pm

In this project, you will apply the data analysis and econometrics skills that you have learned in the class so far. As in the labs, you will be analyzing real data and using it to reach a conclusion on an important economic question. Unlike in the labs, you will not have detailed guidance on how to complete each step of the project: you will have to decide on how to merge, rearrange, and clean the data sets to make them usable. As in the real world, there is no single correct way to do this project. A number of the techniques that you have learned in class could be used to answer this question, and any of these will be fine, as long as you make it clear why your approach is appropriate (but you only need to pick one).

For this project, you will be working as a consultant for a company that has run a pilot project of allowing some of its employees to work from home rather than at a central office. The company was having problems retaining their workforce, where many workers were quitting as a result of the long commutes they faced to get to work. They decided to try out allowing some employees to work from home to address this problem, but were concerned about whether working from home may hurt employee productivity. They were also concerned about how this affects employee satisfaction and retention. They have hired you to analyze data and provide recommendations on whether to expand working from home for their full workforce. They are particularly interested in your findings on the effect of working from home on employee performance and retention.

The company is a leading Chinese travel agency that aggregates information on hotels, flights and tours; makes reservations and obtains tickets for clients; and generates revenue through commissions. Most of their business is done over the phone, and they tested out work from home for their call center employees, who are responsible for answering calls from customers and booking orders or resolving issues. Calls to the company hotline are routed to the next available call center employee, and employees work in shifts. Each employee typically works five shifts a week, and their monthly earnings are the sum of flat wage ($160 per month) plus a bonus based on the number of calls that they answered in that month.

For the pilot program, 249 call center employees were randomly allocated into a treatment group (131 employees) and a control group (118 employees). The treatment group worked from home for four out of their five shifts, and spent the fifth shift in the office. Otherwise their specified work hours (9 am to 5pm) and the system of receiving incoming calls was the same. Employees working from home used the same computer terminals and communications software, so the only differences between the treatment and control groups were the location from which they were working.

The pilot began in December 2010 and lasted for nine months, until August 2011. Employees in the work from home group had to remain in that group for the duration of the pilot, while the control group was not allowed to work from home during the pilot. The company has provided you with a number of datasets from before the pilot began as well as when it was running that you can use in your analysis. You do not need to use all of them, but pick whichever you think are the most useful for your analysis. The data sets are available on the class dropbox folder under “Empirical Project” and the variables are labelled:

- `EmployeeStatus.dta` – contains the employee ID of each employee in the pilot and a variable denoting whether they were assigned to the treatment group or control group (equal to one if assigned to treatment).

- `EmployeeCharacteristics.dta` – contains personal information on each employee assigned to the pilot at the start of the pilot (e.g. marital status, age, if completed high school)
- `Quits.dta` – contains a variable for whether the employee quit during the study period
- `QuitDate.dta` – contains information on the month during which employees quit during the study period (if the employee did quit). This is a panel data set at the monthly level, where each employee has nine observations: one for each month of the experiment. There is a dummy variable for whether the employee quit in that month and whether the employee was still working in that month.
- `Attitudes.dta` – surveys of employees on their satisfaction, with one survey from prior to the start of the pilot and another from after. The firm did five rounds of surveys, with the first conducted prior to the start of the pilot and the next four conducted after the pilot had begun. The firm did not survey all their employees, but did survey a random sample of 107 from treatment and 64 from control. The surveys measured three variables: satisfaction with the job (on a scale from 1 to 7), satisfaction in general (on a scale from 0 to 100) and satisfaction with their life trajectory (on a scale from 0 to 40).
- `Performance_Panel.dta` – monthly performance data on workers for the months before and after the experiment. Data is at the monthly level, where the variable `post` indicates if the month is in the pre-experiment period (January to November 2010) or in the period during the experiment (December 2010 to August 2011). The measures of performance are (1) average calls made per hour worked during the month; (2) total calls made in the month; and (3) a performance evaluation rating of their performance in that month (on a scale from 0 and 100).
- `Performance.dta` – data on average performance of workers over the period before and after the experiment. Each worker has one observation in the pre-experiment period and one in the period during the experiment (denoted by the variable `post).`

In order to use the data, you will likely have to do some combination of the following: (i) merging data sets; (ii) recoding variables; (iii) cleaning variables (e.g. checking for and removing non-nonsensical values); (iv) creating new variables for analysis; and (v) reshaping the data to be in a format you want to use for analysis. For example, there may be places with clear data entry errors that you will need to replace with missing values in order to carry out the analysis. One good way to spot data entry errors is to tab or plot the variables and look for nonsensical values. You may want to refer back to the week 1 slides introducing you to different Stata functionalities, and you are welcome to explore features in Stata that we have not previously used in the class: there are some references in the week 1 slides on using Stata, as well as a wealth of information available on the internet through searching on google (e.g. entries on websites like StackExchange and Statalist).

You will write up a short report describing your analysis of the data. You must use Stata or a comparable software like R (i.e. not Excel) to produce the figures. Your tables should look professional and should not simply be copied and pasted from Stata output (see the labs in which you made professional looking tables). Similarly, your figures should look professional, with labeled axes and reasonable looking visualizations. Tables and figures will not count against the page limit. The report should be written in language so that someone who has taken this class can understand it (i.e., you don't have to simplify the technical parts for the client).

When writing this report, it should be broken into three sections. The first section should describe the data that you use and any data cleaning that you did. This should give sufficient detail that someone else could emulate what you did. If you find data entry errors in any variable in the data sets that you use, you should list them here with the name of the data set, information that could be used to identify this observation (e.g. personid, year, month), name of the variable, the potentially erroneous value of the variable, and reason for your data cleaning decision. You should do this as a table, where the table should mention each of the data sets you use, including if there are no problem values in that data set. It should follow the format below, listing each data set you used alphabetically.

To give an example, suppose you had used the datasets `EmployeeStatus.dta, EmployeeCharacteristics.dta and Performance_Panel.dta` for your report and found two errors in the variable calls_per_hour in `Performance_Panel.dta`, but none in the other two datasets. Then your table should look like the following. *Please make sure to follow this format so you are graded correctly!*

Data set	**Observation**	**Variable**	**Value**	**Change and reason for change**
Employee Characteristics.dta	No problem values			
EmployeeStatus.dta	No problem values			

Performance_Panel.dta	personid = 3906
	year = 2010
	month =3
		calls_per_hour	999999	This is an outlier and does not seem possible. I changed it to a missing value

Performance_Panel.dta	personid = 5993
	year = 2012
	month =1
		calls_per_hour	-3	Negative calls per hour is not possible. I changed it to a missing value

The second section should describe the empirical strategy that you use, why you use it, and what variables you will be analyzing. The third should discuss your findings and include the graphs and tables that you produced.

For this project, you can either work alone or as a pair with another student in the same section of the class. This decision is completely at your discretion. If you choose to work alone, your report should be no more than 4 single spaced pages and must include at least two tables and one figure. If you work in a pair, your report should be no more than 6 single spaced pages and must include a total of at least 5 tables and figures.

In addition to your report, you should create and hand in a single Stata .dofile that performs all of the cleaning and analysis. You must write this yourself or within your pair. **You should not copy it from your classmates outside of your pair** -- that would be considered plagiarism and treated accordingly, where we will check the code files for similarities. Your grade will be based on the following rubric.

	Grading Rubric	
	Data Cleaning (25pts)	
	Merge and rearrange data for analysis using Stata	
	Identify and address data quality issues, such as problem values for the variables of interest. Note any issues in your report	
	Empirical strategy (25pts)	
	Apply correct empirical strategy to successfully estimate the *causal* effect of working from home	
	Selection of appropriate dependent variable(s) to answer this question	
	Tables and Figures (25pts)	
	Creation of tables that are well-explained in the report. These should be in a professional-looking format (e.g. using **esttab** in Stata) with the relevant portions of the table well-labeled.	
	Creation of figures that are well-explained in the report. These should be in a professional-looking format with axes and other relevant portions of the graph labelled.	
	Discussion/Conclusion (25pts)	
	Accurately and concisely summarize the results of your analysis and take-aways in a write-up of no more than 4 pages. This should answer the assignment from the company on the effect of working from home	
	Correctly interpret signs, magnitudes and statistical significance for any of the variables of interest	