ASSIGNMENT 1

In this assignment, you will:

1. Load up, summarize, and clean data into STATA.
2. Run and interpret a linear regression model.
3. Think through important considerations for the empirical design of a quantitative study.

There are two datasets that will be generated and used—one with observations at the country level and one with observations at the individual firm level.

Our datasets contain information on the countries that the firms are in, the strength of the intellectual property protection in the countries (measured using a Likert scale with possible values from **one to seven**; higher values indicate stronger IP protection), firm innovation (measured as the number of patents generated by the firm), a unique identifier for each firm, whether or not the firm is in a service industry (**zero is not; one is yes**), and the number of employees working for the firm. Suppose we are interested in studying the effects of IP protection in these countries on the innovativeness of a sample of firms in those countries, controlling for whether or not the firm is in a service industry and the number of employees.

1: Load up and combine the two datasets, merging the country-level data to the firm-level data using the variable containing country numbers (this implies the firm-level data is your “master” file, and the country data your “using” file). This will add the country-level data to the observations for all the firms.

What kind of merge did you need to do?

|  |
| --- |
|  |

2: We should always first assess whether our merge was successful: how many firms in your dataset were not matched to the country data?

|  |
| --- |
|  |

3: Next, we should always inspect our data. Summarize your variables, and inspect all the information. Here, focus on the existence of impossible values (see the introduction text above to see a description of the variables). Do you need to engage in any data cleaning? How many observations are affected by your cleaning (note: if you did not need to clean please fill in ‘0’ here)?

|  |
| --- |
|  |

4: If you answered ‘yes’ for the previous question, which variables did you clean and how did you clean these? If no, why not?

|  |
| --- |
|  |

5: After cleaning, you prepare summary statistics for your variables. What is the average score for the ‘number of employees’ variable (rounded to three digits)?

|  |
| --- |
|  |

6: You also check the bivariate correlations between your different variables in order to create a correlation matrix. What is the correlation between the service industry indicator and firm innovation (rounded to three digits)?

|  |
| --- |
|  |

7: To investigate our research question, we need to run an OLS regression model where we take innovativeness as the dependent variable. We first start with a basic model including only control variables, being the indicator of whether or not the firm is in a service industry and the number of employees.

Run this model in Stata, **making sure to have robust standard errors**.

What is the effect of the number of employees on firm innovativeness, according to this model? (Note: Base this interpretation on the coefficient and *p*-value of this variable!)

|  |
| --- |
|  |

8: We now run the same model, but now include our main variable of interest—the strength of the IP protection in the country of the firm.

Suppose that our hypothesis is that the strength of the IP protection in the country of the firm has a positive effect on firm innovativeness. Do we find support for our hypothesis? Why (not)?

|  |
| --- |
|  |

9: Compare the number of observations between this model and the first model. Are they different? If so, what caused these differences? If not, is this what you would have expected—given the nature of the data?

|  |
| --- |
|  |

10: Suppose that our estimate for the effect of the number of employees was less precise: this would lead to a higher standard error. How would the t-statistic and p-value change for the ‘employees’ variable in the regression?

|  |
| --- |
|  |