

**Economics 102: Analysis of Economic Data Spring 2021**  
**Department of Economics, U.C.-Davis**  
**Professor Diana Moreira**  
Stata Take-Home Exam

**Version A**

**Description:**

This is an open book exam. You may use any class materials to help you. You can also use the internet to get help with Stata commands. However, you must work independently. You are required to submit your own work, and are **not** allowed to work together with others or get help from other people.

Please submit your exam as a **single PDF**. Clearly answer the questions below, inserting any relevant Stata output directly into your answers. Additionally, you must turn in your **.do file** and **.log file**. Please include these at the end of the **same PDF** with the .do file first and then the .log file. **You will receive 2 points if you comply with this format.**

For this exam, I have provided cross-sectional data on many variables that you can use. You must use **Data A (see brief description below)**. Your answers will depend on the data that you use and the variables you select for the answers below. You will all have unique answers since you will all choose independently which variable you use. Some questions do not have a single correct answer, but you will be graded on the quality of your answers and explanations.

**Brief description of the dataset:** Data A. Contains health and economic information for individuals in the US state of MA. There are multiple individuals per household - often husband and wife participated in the survey and are each represented in a different observation. Note: if there are variables in log and you prefer to work in levels, just transform the variable using `gen new=exp(old)`.

1. Use the command "describe" and present the Stata output you obtained. How many variables are there? and how many observations?
2. Propose a question about the association between two variables using this data (For example, from our house price example, we asked how house price is associated with size or number of bedrooms).
  - (a) What is the question?
  - (b) Why is this question interesting?
  - (c) Describe how each of these two variables are measured in your data (you may transform variables into new variables as needed, e.g., logs, rates or shares, etc.), what the unit of observation of the data set is, and why this is cross-section data.
3. Using these two variables, specify which will be the dependent (y) variable and which will be the independent (x) variables. Why?
4. Provide summary statistics using Stata for each of these variables. Comment. What summary statistic do you think is most useful to know for each variable? Why?
5. Provide a histogram of your dependent variable.
  - (a) Choose either the number of bins or the bin width. Explain why you made this choice.

- (b) Does the data appear skewed? Explain.
  - (c) Are there outliers? Please explain how can you identify the outliers and how many are there?
6. Run a bivariate regression using your dependent variable and the independent variable. Show the regression output.
- (a) Interpret the coefficients.
  - (b) Are the coefficients statistically significant? Explain.
  - (c) Predict the conditional mean of your dependent variable in Stata at the mean of your independent variable
  - (d) Make a graph that includes (1) the actual data points (you can use a scatter plot of the dependent and independent variable), (2) the prediction of the conditional mean (fitted values) ( Remember `||` can be used to combine graphs with the same y and x variables in Stata.) Make your graph look professional (something you would be proud to create for a client or boss) by adjusting the formatting and including a title, axis labels, and any other desired components. (Type “help graph\_intro” in Stata for an intro to making graphs with links to help with making titles, labels, changing colors, etc. Or do a Google search.)
  - (e) What is the correlation coefficient between your dependent and independent variable? (You can use the “correlate” command.) Explain how you could calculate the correlation coefficient using only the information from your bivariate regression table output.
  - (f) Interpret the correlation coefficient
  - (g) Test the hypothesis that the true value of the slope coefficient is 20 percent higher than the coefficient you estimated. Explain how you arrived at your conclusion and use significance level of 1%. [Hint: you will need to calculate the value for  $\beta^*$  based on the value of your estimated coefficient.]
7. Now I want you to continue using the same dependent variable and propose a new independent variable. The new independent variable must be a variable that you believe is associated with the dependent variable
- (a) Explain what are the 3 variables you considered to be the “new independent variable” and the reason you chose the one you chose.
  - (b) Run a bivariate regression using your dependent variable and the new independent variable. Show the regression output and interpret the slope coefficient.
  - (c) calculate the residual and present the summary command for the residual
  - (d) If you had to compare the model in item 6 and the model in 7.b, which one would you say is better? i) Explain the criteria you used; ii) The interpretation of the criteria; iii) and a limitation of this criteria.
  - (e) Suppose you could observe any variable that you want (not restricted to the variables in the dataset). Can you propose a variable that you believe is associated with your dependent variable and also with your independent variable (in model of item 7.b)?
  - (f) What does the existence of this variable imply for the slope coefficient estimated in 7.b? [hint: think about the 4 assumptions for statistical inference and whether they will all hold]

Reminder: Please submit your exam as a **single PDF**. Insert any relevant Stata output directly into your answers. Additionally, attach your **.do file** and **.log file** at the end of the **same PDF** with the .do file first followed by the .log file.