Excel-Stata Block Assignments

Content by Tommy Morgan
Formatting adapted from Natalie Jensen & Emily Leslie
Resources at github.com/tmorg46/ECON378

Last Updated May 2022

Nice to Sheet You: Excel (16 pts)

Answer the following questions in a Word document or Google Doc and turn it in as a PDF. If you turn in your Excel data, you will get a **zero!**

Prep Work™

Open a blank worksheet in Excel or Google Sheets and type "=IF(" into any cell to begin writing an `=IF()` function. When you do, Excel (or Sheets) will bring up a dialogue box that includes the three parameters `logical_expression`, `value_if_true`, and `value_if_false`. (If a little box doesn't appear in Sheets, you might have to click the little question mark button next to the cell you're editing.)

1. (+1) What is a logical expression?
2. (+2) If you write an `=IF()` function that includes `value_if_true` but **does not** include `value_if_false`, what happens differently from writing the same `=IF()` function and including both `value_if_true` and `value_if_false`? (Hint: make up some data and try it!)

Utah State women's gymnastics data exercise

From the Content tab on Learning Suite, download the file "gymnasts_excel.xlsx" and open it in Excel or Google Sheets. This dataset contains meet information and scores from the 2021 and 2022 Utah State University women's gymnastics seasons.

Each column contains a single variable, like vault score or gymnast name, and each row contains a single gymnast's results, so each gymnast will show up multiple times but only once per meet.

3. (+1) Search the internet for an Excel function that can take a column full of duplicate values and return a list of the unique values in that column. Use that function to get the names of each unique gymnast in this dataset. How many unique gymnasts are there?

4. (+1) What is the lowest vault score in the dataset? Which gymnast received that score? (Hint: you can use a `=MIN()` function or filter the dataset)

5. (+2) Use an `=IF()` function to construct a new variable that takes a value of 1 if a performance happened at a meet hosted by BYU and 0 if it happened elsewhere. What percentage of performances in the dataset come from BYU meets?

6. (+2) Use four `=CORREL()` functions to calculate the correlation between a gymnast's being Black and their score in each of the four events. What are those four correlation coefficients?

7. In the NCAA, gymnasts do not need to perform in all four events every time they compete; when they do, they are called an "all-arounder". This dataset does not contain direct information on which gymnasts are all-arounders... or does it?

   (a) (+2) Use four `=MIN()` functions to find the lowest score in each event, then use a `=SUM()` function to total those low scores into one cell. What is that total? In this dataset, would a gymnast that performs in all four events in one meet (an all-arounder) be able to score lower than that total?

   (b) (+3) Use a nested `=IF(SUM())` function to create a new variable that takes a value of 1 if the gymnast in that row was an all-arounder for that meet and 0 otherwise. How many of the observations in the dataset come from all-arounders?

8. (+2) Create a histogram of scores for any **one** of the four events, and change the title to "A Super Cool Histogram". Copy and paste it into your document!

A Blind Date with Big Data: Stata 1 (19 pts)

Answer the following questions in the assignment do file, then turn in that do file on Learning Suite. If you turn in anything else, you will get a **zero!**

Prep Work™

From the Content tab on Learning Suite, download the file "stata1.do" and open it in Stata. This is an example .do file that is designed to get you familiar with best practices in Stata. You'll use it for the rest of this assignment!

1. (+2) In stata1.do, modify the two `globals` at the top of the assignment so that they identify you.

More NCAA women's gymnastics data!

From the Content tab on Learning Suite, download the file "gymnasts_stata.dta". This dataset contains meet information and scores from the regular seasons of every women's gymnastics team that has visited BYU from 2017 to 2022. The rows and columns are identical to those you saw in the Excel assignment, but this dataset has over 4000 observations!

2. (+1) In stata1.do, modify the path in the `cd` command to point to the folder where you downloaded your copy of "gymnasts_stata.dta". Run the `cd` command you just modified and the `use` command just below it. This will load in the dataset!

3. (+1) Add an asterisk to the beginning of the line with the `cd` command to turn the whole line into a comment. (The TA grading these will be really sad if you skip this.)

4. (+2) Run a `browse` command in the command window (not in the do file) to take a look at the data. How many observations are there in the dataset? Fill that number into `global observations` in the do file.

5. (+2) Modify the `destring` command in the do file so that it turns the four event score variables from string to numeric. (Remember, `force` is always an option!)

6. (+3) Now that the scores are numeric, run a `sort` command in the command window that sorts the dataset by floor score. What is the sixth lowest floor score in the dataset? What gymnast received that score, and what team did she play for? Fill your answers into their corresponding `globals`.

7. MyKayla Skinner is an Olympic gymnast who competed for the University of Utah when she was in college. Her teammates are wondering how different her average scores were on each event from the team's overall averages. Coincidentally, the dataset you have is perfect for this question!

   (a) (+2) In the command window, run a `summarize` command with an `if` statement to obtain the University of Utah's average score in each of the four events across the whole dataset. Fill each event's average into the corresponding `global Utah_average_event` in the do file.

   (b) (+2) Now run a `summarize` command with an `if` statement to obtain MyKayla Skinner's average score in each of the four events. Fill each event's average into the corresponding `global MyKayla_average_event` in the do file.

8. Former President Barack Obama is interested in using your data for a project at the Obama Foundation. Of the four events, he only cares about the uneven bars, and he doesn't need to know about the team a gymnast played for or what school hosted each meet. Also, because the interns at the Obama Foundation aren't used to Stata, he's asked that you share whatever you share with him as a .csv file.

   (a) (+1) Modify the first `drop` command to get rid of the the three event score variables and two non-score variables that President Obama told you he doesn't need.

   (b) (+1) Modify the second `drop` command to get rid of all of the observations that are missing a bars score.

   (c) (+2) Modify the `export` command so that it will output a comma-separated values (.csv) file like President Obama asked. (Hint: Running `help export` in the command window might be helpful here!)

   (d) (+0) Run the three commands you just modified. Double-check that the .csv file that the commands create contains all the modifications that President Obama asked you to make, and make sure you save it somewhere you can access later, as you may find it useful very soon...

More Data, No Problems: Stata 2 (20 pts +2)

Answer the following questions in the assignment do file, then turn in that do file on Learning Suite. If you turn in anything else, you will get a **zero!**

Prep Work™

From the Content tab on Learning Suite, download the file "stata2.do" and open it in Stata. This is a nearly empty .do file that is designed to evaluate your ability to use the `help` menus and your knowledge of Stata syntax to write great code. You'll write most of the file in this assignment!

1. (+2) In stata2.do, modify the two `globals` at the top of the assignment so that they identify you.

2. (Bonus +2) Write a comment describing what you're doing above every command that **you** write in this assignment.

A new internship at the Obama Foundation

Congratulations! It's your first day as an intern at the Obama Foundation! You've been brought on to examine whether Black and non-Black gymnasts in NCAA women's gymnastics have different experiences on the uneven bars as seasons progress. To do this, your supervisor has provided you with a .csv file that they got from a student who... happens to share your exact name? Interesting! They also gave you a list of instructions for analyzing this data in Excel, but...

3. (+2) In stata2.do, write a `cd` command and an `import` command that together will bring the data from your .csv file from last assignment into Stata.

4. (+1) Go back and comment out the `cd` command you wrote in question 3.

5. (+2) To see if gymnasts improve in bars as the season goes on, use `correl` commands with `if` statements in the command window to find the correlation of bars scores with meet number for both Black and non-Black gymnasts. Report those correlation coefficients in stata2.do within their corresponding `globals`.

6. Because doing that first project in Stata instead of Excel left you with two extra hours at the end of your first day, you have time to clean some data from the 1930 US Census that the intern you replaced was working on! From the Content tab on Learning Suite, download the file "c1930.dta" and get to work!

(a) (+2) Write a line that changes the directory to the folder where you downloaded idk_lol.dta (even if you downloaded it to the directory you're already in) and a line that opens the file into Stata. After the dataset is open in Stata, go back and comment out the directory changing command like you did in question 4.

(b) (+1) Write code that changes the names of the variables from "var1" and "var2" to "years_of_education" and "household_income".

(c) (+2) Write a command that changes the household_income variable from string to numeric.

(d) Your goal with this data is to regress household income on years of education, but the values of the education variable might be labeled strangely. To see if this is the case, run a `tab` command followed by a `tab, nolabel` command in the command window to figure out the coded values behind all of the "years_of_education" labels.

   i. (+4) Write code that changes the numeric values of the years_of_education variable so that they match with the number of years of education represented by their labels; for example, if the label "Kindergarten" was assigned to the value 12, you would recode it to 1. To start, recode the value labeled "99" to the missing value and the value labeled "None" to 0. (Warning: running this code will make the labels in the browse window or tab commands wrong, but that's okay!)

   ii. (+0) After running that code, you can run `label drop _all` to get rid of the newly incorrect labels on education.

   iii. (+2) In the command window, regress household income on education. Report the constant term in `global beta_not` and the slope coefficient in `global beta_one`.