Exam 2 - Data Science for the Social World

Michael G. Findley* Michael Denly†

Instructions

This is a two-hour, open-book, open-internet exam. As soon as you click on the exam on Canvas is when your timer starts. If you submit the exam late, you will be penalized by 1 point for each minute late that you submit. For example, if you submit 5 minutes late, you will lose 5 points on your overall exam grade.

Each question is worth 5 points. Because the test only has 19 questions, it means that everyone will start the test with 5 free points. There is also a fun bonus question at the end. If you answer it correctly, you can earn 5 extra points. If you do not answer it correctly, there is no penalty.

In terms of your submission, Canvas will only allow you to submit one file. That one file should be either a **PDF file or Word document corresponding to the output of your `R` Markdown `.Rmd` file.** Kindly note that we will not accept Google Docs, and the resulting PDF file or Word document must be generated using `R` Markdown. We will be able to easily discern when that is not the case. Submission of an `R` script instead of a working `R` Markdown `.Rmd` file will incur a 20-point penalty.

Using a comment on your Canvas submission of your PDF file or Word document output from `R` Markdown, please also provide a link to a working GitHub repo that stores all of your exam files. Exams that fail to provide a working GitHub repo link will incur a 30-point penalty. Essentially, please do *not* email us anything unless you can't get your GitHub repo working, and email is your only method of submitting everything–i.e., given that Canvas will only allow you to submit one file.

Your GitHub repo should contain your `R` Markdown `.Rmd` file, your `R` Project, and the exam dataset, etc. Basically, the repo needs to include everything you would need such that your repo is communicating perfectly with the files on your computer. You may name the repo anything that you would like, but maybe something like "exam2" (no spaces!) would be appropriate. Since GitHub provides time stamps for everything, we will be able to discern if you modify the files outside your two-hour exam window; in short, please respect the

*Professor, Department of Government, UT Austin, mikefindley@utexas.edu
†PhD Candidate, Department of Government, UT Austin, mdenly@utexas.edu

two-hour window. Kindly also note that you can get your GitHub repo working before the exam, so there is no need to email us anything and take the 30-point penalty.

Please annotate your `R` code chunks in your `R` Markdown `.Rmd` file with comments, and make sure that the text surrounding it sufficiently explains what you are doing. We will remove points when you do not provide clear comments or explanations to tell us exactly what you are doing with your code. We have this policy in place because it is considered good practice in academia and industry to have clear code files and explanations.

Please work independently. You may *not* consult anyone in the class or outside the class for help, and you may not post the exam questions on the Stack Exchange, Google Groups, or any similar website–though you may visit these websites or others. Please also do not discuss the questions or answers over WhatsApp, GroupMe, text message, or any other platform, especially because everyone will be taking the exams at different times. We will be monitoring accordingly, and anyone who violates any one of these policies will receive a zero on the exam.

We have endeavored to make the exam self-explanatory, but feel free to email the instructors if you have questions. At least one of us will be available over email for the entire 4-hour exam. However, please email both of us if you have a question (i.e., do not email only one of us), because we will be taking shifts.

And one final hint: use your time wisely. If you can't answer one question, move on to the next one, and come back to it once you are done with the ones that you can answer more quickly. Good luck!

Data Overview

For all questions, you will be working off the `2021_exam2_data.xlsx` Excel file, which you have downloaded along with the exam. The Excel file contains separate, labeled tabs of the following different datasets:

- `college_scorecard`: An individual-level dataset on earnings and labor market outcomes for past/graduated students who attended numerous universities in the United States.
- `avocados`: A dataset of weekly avocado sales in the state of California for the 2015-2018 period.
- `training`: A dataset detailing labor market outcomes for individuals who were eligible to participate in a job training program. The dataset is a very famous one in the discipline of economics, originating in Lalonde's (1986) study published in *The American Economic Review*.
- `titanic`: An individual-level dataset detailing which passengers survived the infamous Titanic shipwreck–yes, the same one chronicled by James Cameron's 1997 movie, starring Leonardo Di Caprio and Kate Winslet.

Below, there will be subsections dedicated to the analysis of each dataset. Each of these subsections will provide a full overview of the variables in the respective dataset.

College Scorecard Questions

For the following questions, you will be working with the `college_scorecard` data. It contains data on earnings and labor market outcomes for past/graduated students who attended numerous universities in the United States, represented by the following variables:

- `unitid`: ID variables for the college/university
- `inst_name`: Name of the college or university
- `state_abbr`: Two-letter abbreviation for the state where the college/university
- `earnings_med`: Median earnings among students (a) who received federal financial aid, (b) who began as undergraduates at the institution ten years prior, (c) with positive yearly earnings
- `pred_degree_awarded_ipeds`: Predominant degree awarded. 1 = less-than-two-year, 2 = two-year, 3 = four-year+
- `year`: Year in which outcomes are measured
- `earnings_med`: Median earnings among students (a) who received federal financial aid, (b) who began as undergraduates at the institution ten years prior, (c) with positive yearly earnings
- `count_not_working`: Number of students who are (a) not working (not necessarily unemployed), (b) received federal financial aid, and (c) who began as undergraduates at the institution ten years prior
- `count_working`: Number of students who are (a) working, (b) who received federal financial aid, and (c) who began as undergraduates at the institution ten years prior

1. Please clear the environment in `R`.

2. Load the `college_scorecard` dataset in `R`, and call it "college_scorecard".

3. Provide summary statistics for the `college_scorecard` dataset.

4. Create a smaller dataset with just data measured in 2014 and 2015 on former students who graduated from four-year+ colleges and universities located in Texas (`state_abbr`: "TX") and Louisiana (`state_abbr`: "LA"). Call the resulting data frame "small_scorecard".

5. Collapse the "small_scorecard" data frame to get both (a) the average of number people working who graduated from universities in Texas and Lousiana; and (b) the total number of people working who graduated from universities in Texas and Lousiana. Call your resulting data frame "even_smaller_scorecard". (Hint: Your resulting data frame should only have two observations–one for Texas, the other for Lousiana.)

6. Use the "even_smaller_scorecard" data frame to provide a bar graph detailing the *percent* of people working. Make sure to label the axes and provide an appropriate title for the graph. (Hint: you will need to create a new variable to answer this question.)

7. On the basis of your graph, did people who graduated from four-year col-

leges/universities located in Texas or Louisiana have a better chance of being employed? More broadly, do you think that going to college/university in one state gives people a better chance at getting a job? (Hints: (a) you will want to take a look at the summary statistics of the "even_smaller_scorecard" data frame; and (b) you will want to take a look at the universities included in the "smaller_scorecard")

Questions on Avocado Sales

The following questions will be based on the `avocados` dataset. As mentioned above, it contains data on weekly avocado sales in the state of California for the 2015-2018 period. Here is a description of the dataset's variables:

- `date`: Date of observation
- `average_price`: Average (undeflated) price of the avocados
- `total_volume`: Total volume of avocados sold

8. Load the avocados dataset in `R`, and call the data frame "avocados".

9. Create a new variable called "year" that only captures the year in which the avocados were sold.

10. Using the data from the World Bank's World Development Indicators helpfully stored in the `WDI` library, deflate the `average_price` variable. When doing so, capture that deflated price with a new variable called "deflated price_XXXX", and replace "XXXX" with the relevant base year for those prices. Make sure to also save your deflator variable, and call it "deflator". (Hints: (a) Google "World Bank GDP Deflator" to get the indicator code; (b) make sure the get relevant GDP deflator for the US, because we are working in US dollars; and (c) you can tell which value is the base year for the deflator on the basis on which year has a deflator value of 100. You may want to keep the start year at 1960 so that have a long enough time series to tell which year is actually the base year.)

11. Collapse the resulting data frame to obtain the average deflated price of the avocados for each year. Call the resulting data frame "collapsed_avocados", and show your resulting output with the `head()` command.

12. Reshape the deflated `collapsed_avocados` data frame wide, and call the resulting data frame `wide_avocados`. Once you are done, use the `head()` command to show your resulting data frame.

13. Label your variables in the `wide_avocados` data frame.

Training Dataset Questions

The following questions will be based off the `training` dataset. As mentioned above, it contains data on labor market outcomes for people who were eligible to participate in a job training program. The dataset contains the following variables:

- `training_program`: Dummy variable to indicate whether the person participated in the job training program. Participant = 1; Non-participant
- `age`: Age in years
- `educ`: Years of education
- `black`: Black person = 1; Non-black person = 0
- `hisp`: Hispanic person = 1; Non-hispanic person = 0
- `marr`: Married person = 1; Single person = 0
- `re_74`: Real (deflated) earnings in 1974
- `re_75`: Real (deflated) earnings in 1975
- `re_78`: Real (deflated) earnings in 1978

14. Load the `training` dataset in `R`. Call the data frame `training`.

15. Reshape the `training` data frame long so that all of the earnings variables are in a single column. After you are done with the reshape, provide summary statistics for your long data frame. (Note: it is good practice to create an `id` variable to capture each observation with `training$id = 1:nrow(training)` in the data frame before you reshape. This way, you can easily see which observation corresponds to which person).

Titanic Questions

The following questions will be based on the `titanic` dataset. As mentioned above, it is an individual-level dataset detailing which passengers survived the infamous Titanic shipwreck. Here is a description of the dataset's variables:

- `class`: Class of the passenger's ticket
- `age`: Child vs. Adult. Child = 0 and Adult = 1
- `female`: Female passenger = 1 and Male passenger = 0
- `survived`: Surviving passenger = 1 and Deceased passenger = 0

16. Load the `titanic` in `R` and call the resulting data frame `titanic`.

17. Provide summary statistics for the `titanic` data frame.

18. Create a cross-tab to compare average rates of survivorship by gender. Tell us in words what are the results.

19. Use `ifelse` to create a new variable called `first_class`, indicating whether or not the passenger had a first-class ticket. Then, provide a frequency table to show your results.

Bonus. The grammy-wining theme song of the 1997 blockbuster movie "Titanic", starring Leonardo Di Caprio and Kate Winslet, frequently plays on repeat on Professor Findley's iPod. What is the name of the song, and who is the song's artist? (Hint: See the iconic picture below.)

"My Heart Will Go On" by Céline Dion

Housekeeping/GitHub repo link

Save all of the files–i.e., `.Rmd`, `R` Project, `.xlsx`, `.pdf`/Word Doc)–push them to your GitHub repo, and provide us with the link to that repo.

GitHub link here. Professor Findley to verify that all files are pushed to repo.