Midterm Exam

[YOUR NAME]

2023-10-25

Overview

This is your midterm exam. It consists of five questions plus two additional extra credit questions.

Survey EC

In addition, there is an additional extra credit opportunity if you respond to a short survey about this course. The survey is not part of Vanderbilt's official teaching evaluations. I use it to help me improve the course in the second half of the semester, and respond to your specific needs. To receive the extra credit, take the survey and then submit the secret completion code to Brightspace (quiz name: "Midterm Survey") receive an additional four points. The survey is anonymous, meaning that the completion code is the same for everyone (so please don't share it!).

Grading

Each of the five questions is worth 8 points, while the two extra credit questions and the survey are worth four points each. Note that the survey can be taken any time outside of class up until October 31st.

When you have finished, please upload a PDF of your midterm to Brightspace under the "Midterm Exam" assignment.

Resources

You are permitted to rely on course resources from the first part of the Fall 2023 semester. These include all lecture slides, recordings, problem sets, answer keys, homeworks, and lecture notes, as well as any and all posts to Campuswire.

Codebook

The midterm uses the `sc_debt.Rds` dataset, the codebook for which is reproduced below:

Name	Description
unitid	Unit ID
instnm	Institution Name
stabbr	State Abbreviation
grad_debt_mdn	Median Debt of Graduates
control	Control Public or Private
region	Census Region

Name	Description
preddeg	Predominant Degree Offered: Associates or Bachelors
openadmp	Open Admissions Policy: 1= Yes, 2=No,3=No 1st time students
adm_rate	Admissions Rate: proportion of applications accepted
ccbasic	Type of institution– see here (https://data.ed.gov/dataset/9dc70e6b-8426-4d71-b9d5-70ce6094a3f4/resource/658b5b83-ac9f-4e41-913e-9ba9411d7967/download/collegescorecarddatadictionary_01192021.xlsx)
selective	Institution admits fewer than 10 % of applicants, 1=Yes, 0=No
research_u	Institution is a research university 1=Yes, 0=No
sat_avg	Average SAT Scores
md_earn_wne_p6	Average Earnings of Recent Graduates
ugds	Number of undergraduates

Question 1: 8 points

Our overarching research question is: do schools with higher student debt produce graduates who make more money in their future earnings?

Propose a theory that answers this question. There are no wrong answers to this question, but the best answers are those that clearly describe the assumptions on which the theory rests. **[6 points]**

> Write answer here. (3 - 6 sentences)

Write out the hypothesis associated with your theory. What relationship do you expect to see between student debt and future earnings? **[2 points]**

> Write answer here. (1 sentence)

Question 2: 8 points

Require `tidyverse` and load the `sc_debt.Rds` (https://github.com/jbisbee1/DS1000_F2023/blob/main/Lectures/2_Intro_to_R/data/sc_debt.Rds?raw=true) dataset from `GitHub` (https://github.com/jbisbee1/DS1000_F2023/blob/main/Lectures/2_Intro_to_R/data/sc_debt.Rds). Save the object as `debt`. **[1 point]**

```
# WRITE CODE HERE
```

Now let's look at the data. What type of variable is student debt? What type of variable is future earnings? Do either of them have missing values? HINT: use the `summary()` function to make this super easy. **[4 points]**

```
# WRITE CODE HERE
```

> Write answer here. (1 - 3 sentences)

Finally, visualize both variables using *univariate visualizations* with the `ggplot()` function. Make sure to choose the appropriate `geom_...()` function and label your plots! **[3 points]**

```
# Plot 1
debt %>%
  ggplot(aes(x = )) + # What goes on the x-axis?
  geom_...() + # What is the appropriate geom for this variable?
  labs(x = '', # Include good labels!
       y = '',
       title = '')
```

```
## Error in debt %>% ggplot(aes(x = )): could not find function "%>%"
```

```
# Plot 2
debt %>%
  ggplot(aes(x = )) + # What goes on the x-axis?
  geom_...() + # What is the appropriate geom for this variable?
  labs(x = '', # Include good labels!
       y = '',
       title = '')
```

```
## Error in debt %>% ggplot(aes(x = )): could not find function "%>%"
```

Question 3: 8 points

Create a new variable called " `high_debt` " which takes on the value "high" if the school's average student debt is above $25,000, and "low" otherwise (use the `ifelse()` command within a `mutate` function and add this new column to your original dataset using the object assignment operator `<-`). **[3 points]**

```
debt <- debt %>%
  mutate(high_debt = ifelse()) # Fill in the ifelse() function
```

```
## Error in debt %>% mutate(high_debt = ifelse()): could not find function "%>%"
```

Using this new variable, investigate whether schools with "high" student debt have higher or lower future earnings. Answer the question by calculating the average median future earnings by `high_debt` using the `group_by()` and `summarise()` functions. (Use `drop_na(high_debt)` to ignore the `NA` category!) Does this answer support your theory? **[5 points]**

```
# Write code here
```

> Write answer here. (1 - 3 sentences)

Question 4: 8 points

How confident are you in the conclusion drawn in question 3? Use 100 bootstrapped simulations using a `for()` loop and `sample_n()` with `size` set to the number of rows in the data and `replace` set to `TRUE` to express your confidence. Make sure to instantiate an empty object `bsRes` to store your bootstrapped analyses, and to `set.seed(123)` at the beginning of your code! **[5 points]**

```
# Set seed here
# Instantiate the empty bsRes object here
# Start the for() loop here
#
#
#

# Analyze the results of the for() loop here
```

Write answer here. (1 - 2 sentences)

How large is the average difference across these bootstraps? **[3 points]**

```
# Write code here
```

Write answer here. (1 sentence)

Question 5: 8 points

Now let's look at the original variable for student debt in a multivariate way.

First, based on your theory above and the research question, which variable is the independent / explanatory / predictor variable X? Which variable is the dependent / outcome variable Y? Why? **[2 points]**

Write answer here. (2 sentences)

Second, visualize the relationship between these two variables using a multivariate visualization. Make sure to choose the appropriate `geom_...()` ! Does the visual inspection change your answer to the research question? **[2 points]**

```
# Write code here
```

Write answer here. (1 - 2 sentences)

Third, run the regression of Y on X using the `lm()` function. Save the model to a new object called `model_earn_debt` using the object assignment operator `<-` . **[1 point]**

```
# Write code here
```

Finally, interpret the output of the regression using the `summary(model_earn_debt)` command. Describe what the intercept (α) and slope (β) mean in substantive terms. Do the results support your theory? How confidence are you in this conclusion? **[3 points]**

```
# Write code here (super easy...use summary() on your model object)
```

> Write answer here (2 - 4 sentences).

EC - Question 6: 4 points

How good is your regression model from question 5? Calculate the RMSE using 100-fold cross validation with a 50-50 split between the train and test sets.

```
# Write code here
```

> Write answer here (1 sentence).

EC - Conclusion: 4 points

In conclusion, what would you say about your analysis? Did it support your theory? Why or why not? What additional ideas do you have for future analyses?

> Write answer here (3 - 5 sentences).

Survey

Please complete this **anonymous** course evaluation. This does not influence Professor Bisbee's career or position in the university and will only be used to improve the course. You can find the anonymous survey here (https://nyu.qualtrics.com/jfe/form/SV_b7t5vqhhbalgGZ8). Upon completing the survey, you will be given a completion code. To receive the extra credit points, please paste the completion code into the Brightspace quiz titled "Midterm Survey".

NOTE: There is only one completion code to ensure that all responses are anonymized and can't be linked back to the midterm exams. To prevent students from sharing the code with their friends to get the 4 extra credit points without completing the survey, these 4 points are only provided if the number of midterms with the completion code *exactly equals the number of survey responses*. In other words, if there are 150 exams with the completion code, but only 50 completed surveys, **all students will forfeit their extra credit points**. The purpose of this strict rule is to disincentivize the sharing of this code either by those who would fill out the survey and then share the code, or by those who would ask to be given the code without filling out the survey.