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, problem sets, answer keys, homeworks, and lecture notes, as well as any and all posts to Campuswire. You are **not** permitted to use ChatGPT for this midterm. You are **not** permitted to review recordings during the midterm.

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
costt4_a	Average cost of attendance (tuition-grants)
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]**

> I assume that richer students are able to pay for college without borrowing money, meaning that lower student debt is associated with a student body that is of a higher socioeconomic status. I further assume that higher socioeconomic status is associated with higher future earnings for a long **long** list of reasons (i.e., higher wealth social networks, greater confidence in bargaining for salary). As such, I theorize that students with less debt come will have higher future earnings. This expectation is explicitly not causal, since I don't think debt **causes** lower future earnings. Rather both debt and future earnings are explained by socioeconomic status.

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

> I hypothesize that the relationship between student debt and future earnings will be negative.

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). **[1 point]**

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
debt <- read_rds('https://github.com/jbisbee1/DS1000_F2023/blob/main/Lectures/2_Intro_to_R/data/sc_debt.Rds?raw=true')
```

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? **[4 points]**

```
summary(debt %>%
          select(grad_debt_mdn,md_earn_wne_p6))
```

```
##  grad_debt_mdn   md_earn_wne_p6
##  Min.   : 2332   Min.   : 10600
##  1st Qu.:13000   1st Qu.: 26100
##  Median :21500   Median : 31500
##  Mean   :19646   Mean   : 33028
##  3rd Qu.:25125   3rd Qu.: 37400
##  Max.   :45881   Max.   :120400
##  NA's   :325     NA's   :240
```

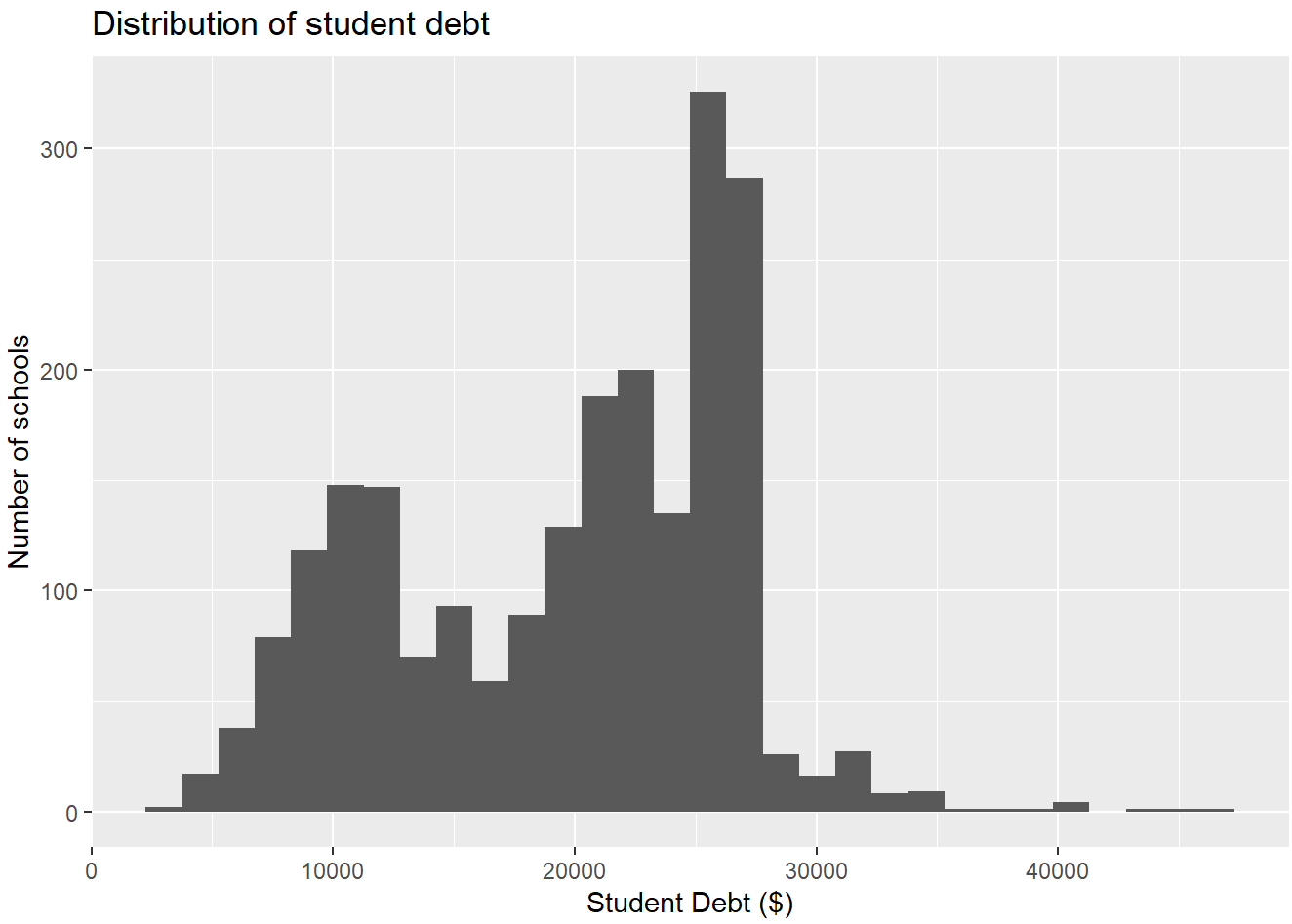
> Both variables appear to be continuous / numeric variables. Both have missing data. There are 325 schools that don't have measures of student debt, and 240 schools that don't have measures of future earnings.

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]**

```
debt %>%
  ggplot(aes(x = grad_debt_mdn)) +
  geom_histogram() +
  labs(x = 'Student Debt ($)',
       y = 'Number of schools',
       title = 'Distribution of student debt')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

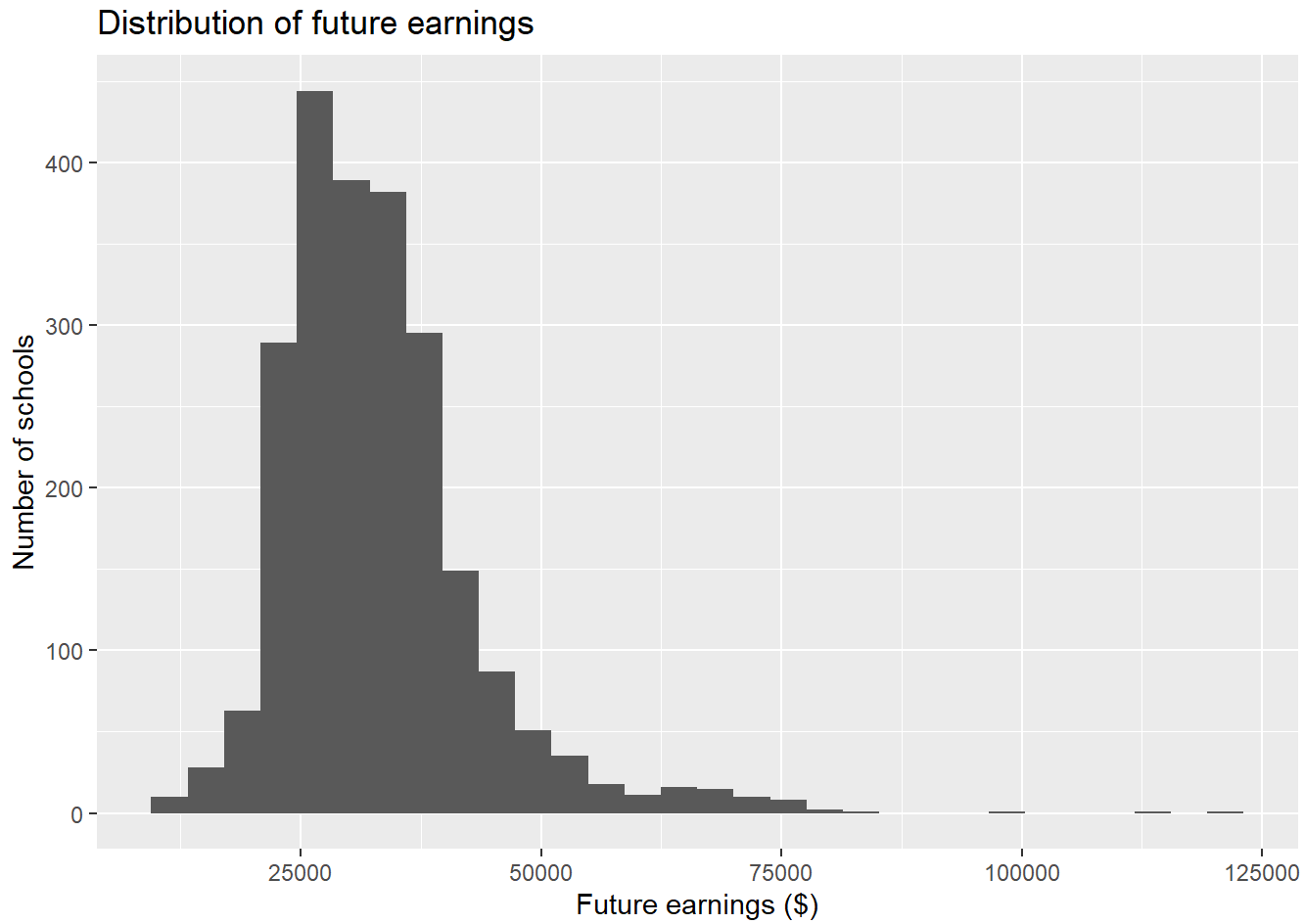
```
## Warning: Removed 325 rows containing non-finite values (`stat_bin()`).
```




```
debt %>%
  ggplot(aes(x = md_earn_wne_p6)) +
  geom_histogram() +
  labs(x = 'Future earnings ($)',
       y = 'Number of schools',
       title = 'Distribution of future earnings')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 240 rows containing non-finite values (`stat_bin()`).
```



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(grad_debt_mdn > 25000,'high','low'))
```

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]**

```
debt %>%
  drop_na(high_debt) %>%
  group_by(high_debt) %>%
  summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T))
```

```
## # A tibble: 2 × 2
##   high_debt avg_earn
##   <chr>        <dbl>
## 1 high        34432.
## 2 low         33467.
```

> This analysis does not support my theory. In fact, it appears to be the opposite. Schools with higher student debt have higher future earnings by roughly $1,000 per year.

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(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- debt %>%
    sample_n(size = nrow(debt),replace = T) %>%
    group_by(high_debt) %>%
    summarise(avg_earn = mean(md_earn_wne_p6,na.rm=T)) %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

bsRes %>%
  drop_na(high_debt) %>%
  spread(high_debt,avg_earn) %>%
  summarise(conf = mean(low > high))
```

```
## # A tibble: 1 × 1
##    conf
##   <dbl>
## 1  0.03
```

> I am only 3% confident that schools with lower student debt produce graduates with higher earnings. In other words, I am 97% confident that schools with higher student debt produce graduates with higher earnings.

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

```
bsRes %>%
  drop_na(high_debt) %>%
  spread(high_debt,avg_earn) %>%
  summarise(diff = mean(high - low))
```

```
## # A tibble: 1 × 1
##    diff
##   <dbl>
## 1  912.
```

> The average difference is approximately $912.

Question 5: 8 points

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

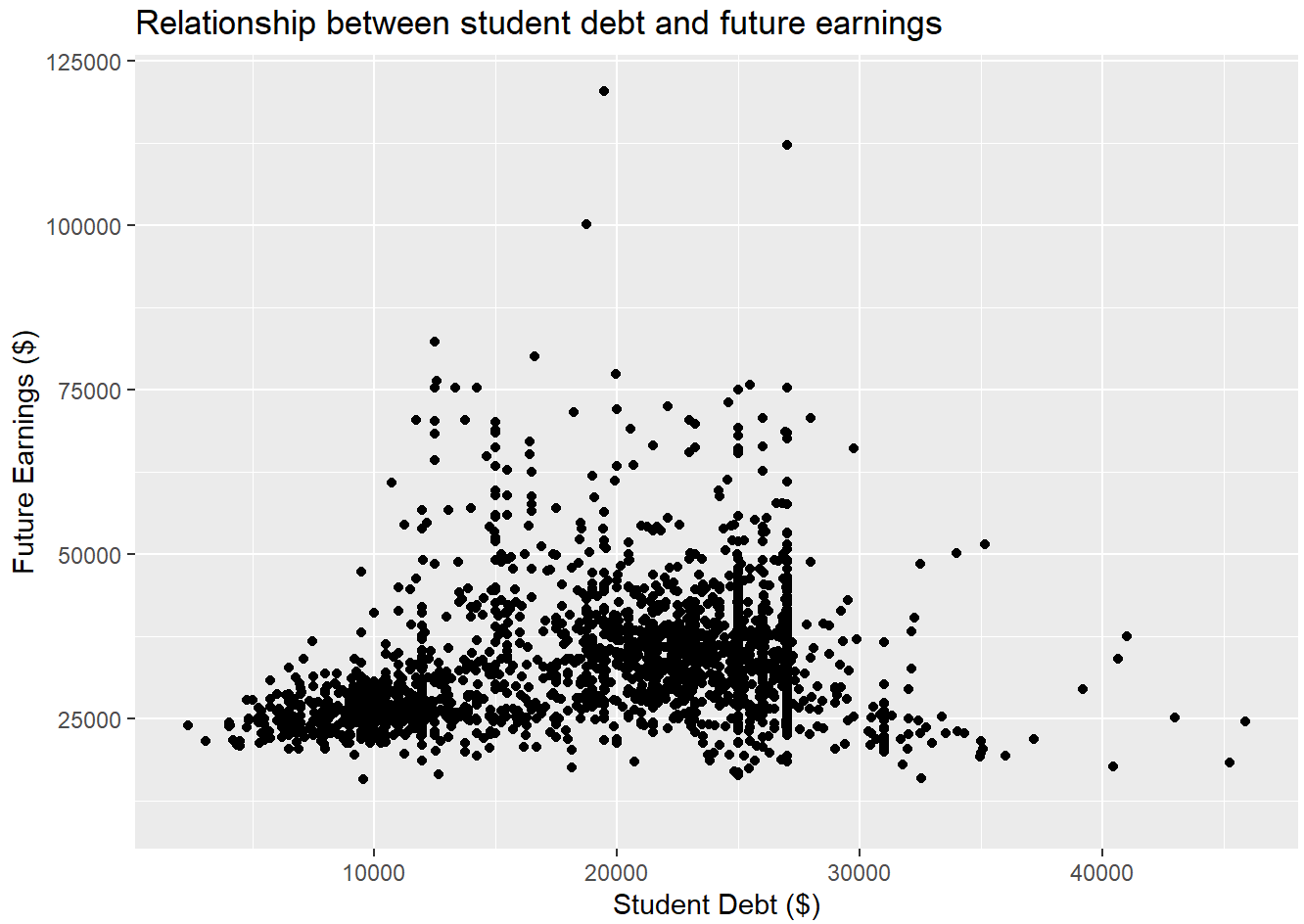
First, based on your theory above and the research question, which variable is the independent / explanatory / predictor variable X? Which variable is the dependent / outcome variable Y? Why? **[2 points]**

> The independent variable is the student debt and the depent variable is future earnings. This is because my intuition is that the student debt measure predicts the future earnings measure.

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]**

```
debt %>%
  ggplot(aes(x = grad_debt_mdn,y = md_earn_wne_p6)) +
  geom_point() +
  labs(x = 'Student Debt ($)',
       y = 'Future Earnings ($)',
       title = 'Relationship between student debt and future earnings')
```

```
## Warning: Removed 378 rows containing missing values (`geom_point()`).
```



The relationship plot suggests that there might be a curvelinear relationship between student debt and future earnings. However, there is clearly a positive association between the two variables across most values of student debt.

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]**

```
model_earn_debt <- lm(md_earn_wne_p6 ~ grad_debt_mdn,data = debt)
```

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]**

```
summary(model_earn_debt)
```

```
##
## Call:
## lm(formula = md_earn_wne_p6 ~ grad_debt_mdn, data = debt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24390  -5419  -1782   3014  86737
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.690e+04  6.219e+02   43.25   <2e-16 ***
## grad_debt_mdn 3.471e-01  2.983e-02   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9707 on 2166 degrees of freedom
##   (378 observations deleted due to missingness)
## Multiple R-squared:  0.05881,    Adjusted R-squared:  0.05838
## F-statistic: 135.3 on 1 and 2166 DF,  p-value: < 2.2e-16
```

According to the regression analysis, schools that have zero student debt have future earnings of approximately $26,900. In addition, the model indicates that, for every additional dollar of student debt, future earnings increases by $0.35 per year. The p-value is tiny, meaning that I am over 99.9% confident that this relationship is not zero.

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.

```
debt_analysis <- debt %>%
  select(grad_debt_mdn,md_earn_wne_p6) %>%
  drop_na()
cvRes <- NULL
for(i in 1:100) {
  inds <- sample(1:nrow(debt_analysis),size = round(nrow(debt_analysis)*.5),replace = F)
  train <- debt_analysis %>% slice(inds)
  test <- debt_analysis %>% slice(-inds)

  m <- lm(md_earn_wne_p6 ~ grad_debt_mdn,train)
  cvRes <- test %>%
    mutate(preds = predict(m,newdata = test)) %>%
    summarise(rmse = sqrt(mean((md_earn_wne_p6 - preds)^2))) %>%
    mutate(cvInd = i) %>%
    bind_rows(cvRes)
}

cvRes %>%
  summarise(mean(rmse))
```

```
## # A tibble: 1 × 1
##   `mean(rmse)`
##          <dbl>
## 1        9760.
```

My model makes mistakes of roughly $9,760 on average.

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?

The data does not support my theory. This might be due to the debt measure capturing more than simple socioeconomic status. Or it might be that my assumptions are wrong. One possible explanation would be that schools with greater debt have more diverse student bodies. A long line of social science and economics literature demonstrates that humans are more productive, more creative, and generally just *better* when we exist in more diverse environments. Exposure to different perspectives makes us literally smarter because it forces us to challenge pre-existing views, think harder, and problem solve better. A way to test this theory would be to predict future earnings as a function of student diversity across a range of dimensions.

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.