Midterm Exam

[YOUR NAME]

2024-03-07

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 March 12th.

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 Spring 2024 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 more expensive schools have fewer undergraduates? First, require `tidyverse` and load the `sc_debt.Rds` (https://github.com/jbisbee1/DS1000_S2023/blob/main/data/sc_debt.Rds?raw=true) dataset from `GitHub` . Save the object as `debt` . **[1 point]**

```
require(tidyverse)
```

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

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.2     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.2     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1     
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
debt <- read_rds('https://github.com/jbisbee1/DS1000_S2024/blob/main/data/sc_debt.Rds?raw=true')
```

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. **[4 points]**

> I think more expensive schools should have fewer undergraduates because fewer people can afford expensive schools. My theory is based on the simple logic of supply and demand from economics, in which higher prices reduce demand.

Write out the hypothesis associated with your theory. What relationship do you expect to see between the number of undergraduates and the cost of attendance? **[1 point]**

> I expect to see that the number of undergraduates is declining as the cost of attendance rises.

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

> Based on my theory, the cost of attendance is the independent variable and the number of undergraduates is the dependent variable. This is because my theory assumes that attendance costs cause undergraduate enrollment.

Question 2: 8 points

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

```
debt %>%
  select(costt4_a,ugds)
```

```
## # A tibble: 2,546 × 2
##    costt4_a  ugds
##       <int> <int>
##  1    23053  5271
##  2    24495 13328
##  3    14800   365
##  4    23917  7785
##  5    21866  3750
##  6    29872 31900
##  7    10493  1201
##  8       NA  2677
##  9    19849  4407
## 10    31590 24209
## # ℹ 2,536 more rows
```

```
summary(debt %>%
          select(costt4_a,ugds))
```

```
##     costt4_a          ugds
##  Min.   : 6525   Min.   :    2
##  1st Qu.:15051   1st Qu.:  829
##  Median :23948   Median : 2021
##  Mean   :29971   Mean   : 4861
##  3rd Qu.:42824   3rd Qu.: 5820
##  Max.   :78555   Max.   :98630
##  NA's   :136     NA's   :1
```

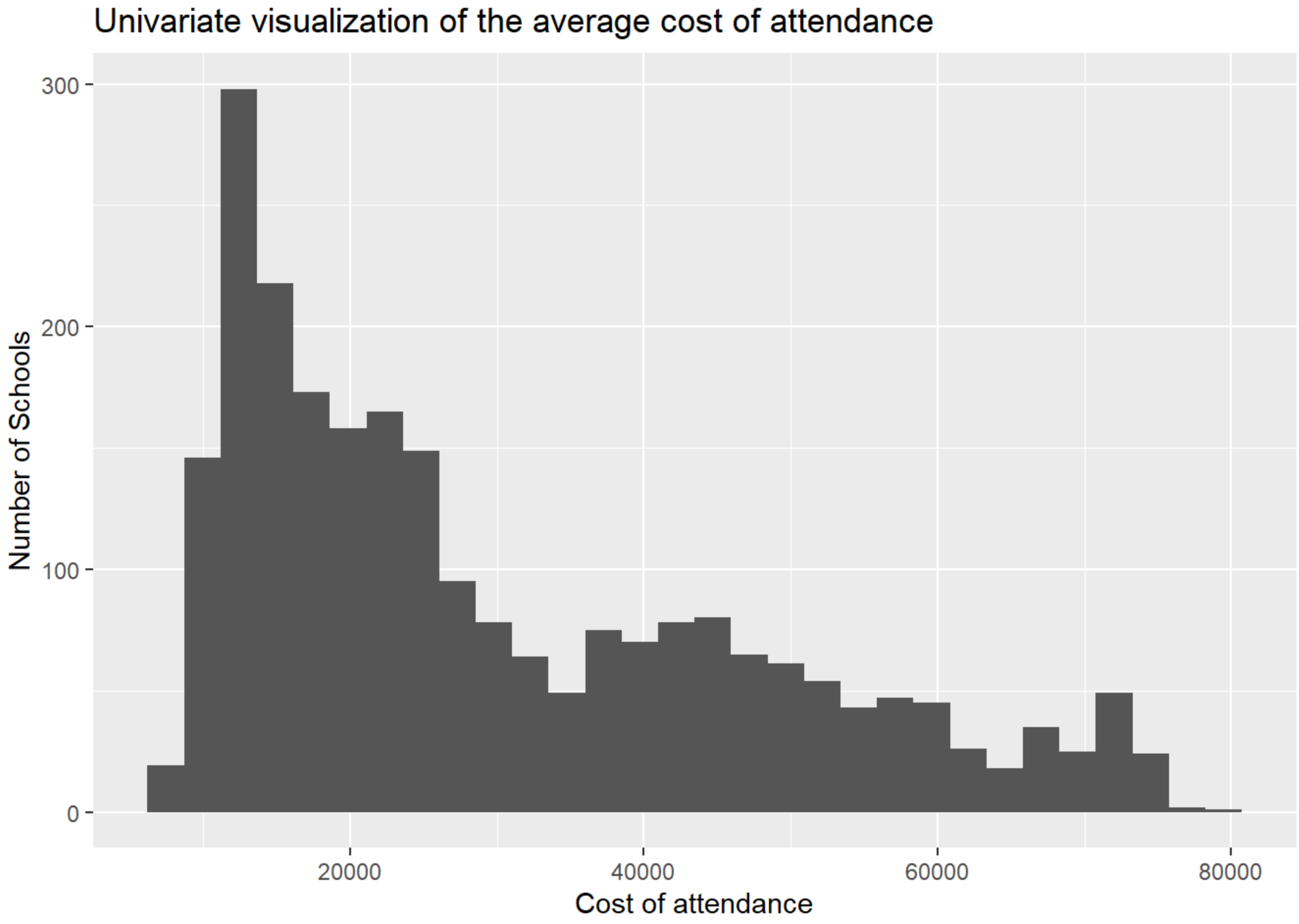
> Both variables appear to be continuous measures. The first several observations are numbers and are stored as integers in the data. There is one missing value in the `ugds` variable, and there are 136 missing values in the `costt4_a` variable. In both cases, these missing values indicate that these schools chose not to share these pieces of information with the researchers who collected the data.

Visualize both variables using *univariate visualizations* with the `ggplot()` function. Make sure to choose the appropriate `geom_...()` function and label your plots! **[4 points]**

```
# Plot 1
debt %>%
  ggplot(aes(x = costt4_a)) + # What goes on the x-axis?
  geom_histogram() + # What is the appropriate geom for this variable?
  labs(x = 'Cost of attendance', # Include good labels!
       y = 'Number of Schools',
       title = 'Univariate visualization of the average cost of attendance')
```

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

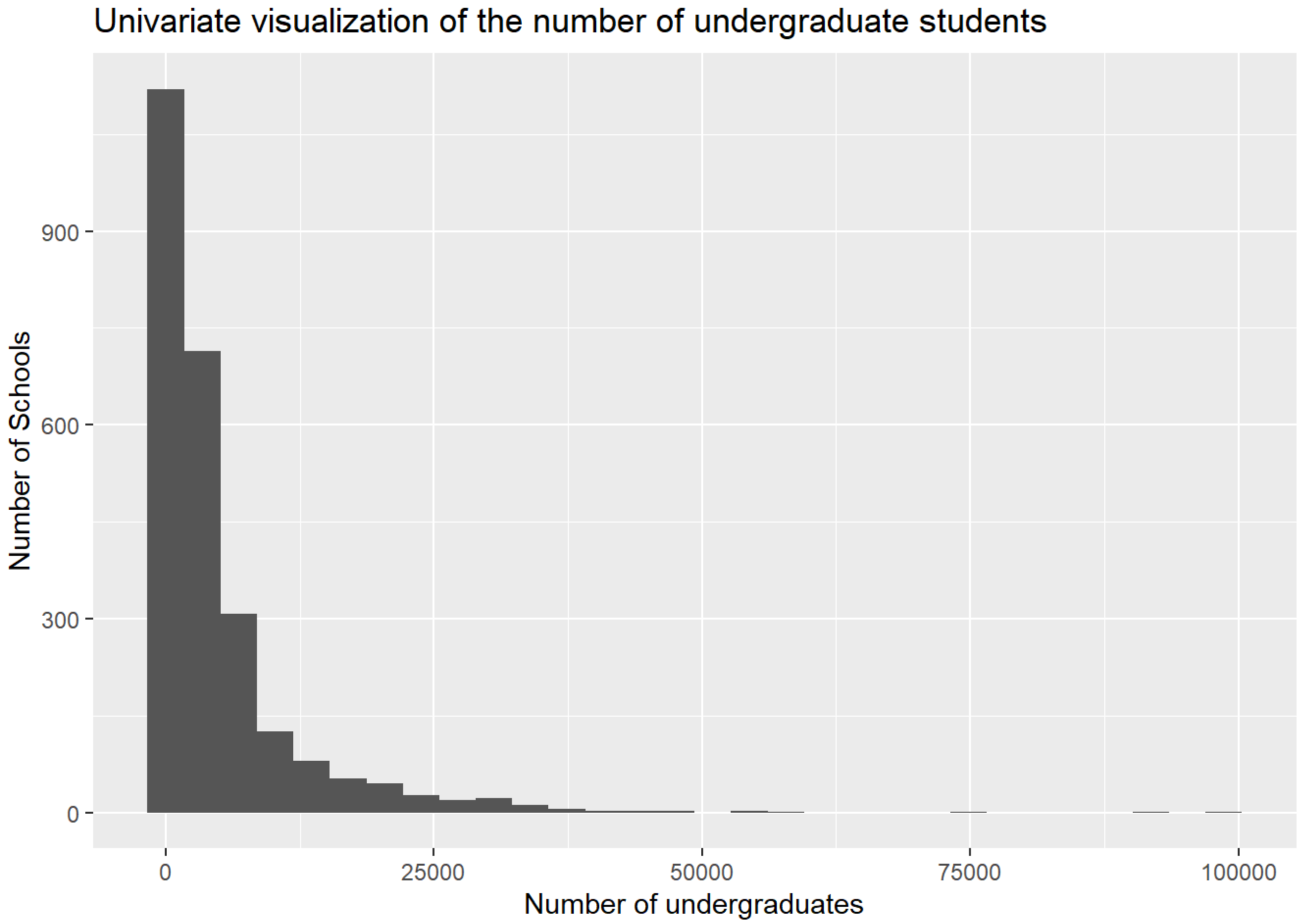
```
## Warning: Removed 136 rows containing non-finite values (`stat_bin()`).
```



```
# Plot 2
debt %>%
  ggplot(aes(x = ugds)) + # What goes on the x-axis?
  geom_histogram() + # What is the appropriate geom for this variable?
  labs(x = 'Number of undergraduates', # Include good labels!
       y = 'Number of Schools',
       title = 'Univariate visualization of the number of undergraduate students')
```

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

```
## Warning: Removed 1 rows containing non-finite values (`stat_bin()`).
```



Question 3: 8 points

Create a new variable called “ `expensive` ” which takes on the value “expensive” if the school’s average cost of attendance is above the median cost of attendance across all schools in the dataset, and “cheap” otherwise (use the `ifelse()` command within a `mutate` function and add this new column to your original dataset using the object assignment operator `<-`). **[4 points]**

```
debt <- debt %>%
  mutate(expensive = ifelse(costt4_a > median(costt4_a,na.rm=T),
                            'expensive','cheap')) # Fill in the ifelse() function
```

Using this new variable, investigate whether expensive schools have more or fewer undergraduates. Answer the question by calculating the average number of students by the new `expensive` variable you created, using the `group_by()` and `summarise()` functions. (Use `drop_na(expensive)` to ignore the `NA` category!) Does this answer support your theory? **[4 points]**

```
debt %>%
  drop_na(expensive) %>%
  group_by(expensive) %>%
  summarise(avg_students = mean(ugds,na.rm=T))
```

```
## # A tibble: 2 × 2
##   expensive avg_students
##   <chr>            <dbl>
## 1 cheap            6093.
## 2 expensive        4058.
```

> Expensive schools have 4,058 students on average, while cheap schools have 6,093 students on average. This supports my initial theory that expensive schools would have fewer students than cheap schools.

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! Restate your answer to the research question, but this time indicate your confidence, along with the overall average difference across bootstrapped realities. **[5 points]**

```
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- debt %>%
    sample_n(size = nrow(debt),replace = T) %>%
    drop_na(expensive) %>%
    group_by(expensive) %>%
    summarise(avg_ugds = mean(ugds,na.rm = T),.groups = 'drop') %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

bsRes %>%
  drop_na(expensive) %>%
  spread(expensive,avg_ugds) %>%
  mutate(diff = expensive - cheap) %>%
  summarise(conf = mean(diff > 0),
            avg = mean(diff))
```

```
## # A tibble: 1 × 2
##    conf    avg
##   <dbl>  <dbl>
## 1     0 -2024.
```

> Expensive schools have, on average, 2,024 fewer undergraduate students than cheap schools. I am more than 99.9% confident in this conclusion.

Re-run the same analysis, except only sample 50 schools at random with replacement, instead of all the schools in the dataset. Does your answer change? **[3 points]**

```
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  bsRes <- debt %>%
    sample_n(size = 50,replace = T) %>%
    drop_na(expensive) %>%
    group_by(expensive) %>%
    summarise(avg_ugds = mean(ugds,na.rm = T),.groups = 'drop') %>%
    mutate(bsInd = i) %>%
    bind_rows(bsRes)
}

bsRes %>%
  drop_na(expensive) %>%
  spread(expensive,avg_ugds) %>%
  mutate(diff = expensive - cheap) %>%
  summarise(conf = mean(diff > 0),
            avg = mean(diff))
```

```
## # A tibble: 1 × 2
##    conf    avg
##   <dbl>  <dbl>
## 1  0.09 -2353.
```

> My answer changes in two ways. First the overall average difference has increased to 2,353 fewer students attending expensive schools compared to cheap schools. Second, I am only 91% confident in this conclusion.

Question 5: 8 points

Now let’s look at the variables in a multivariate way. **NOTE**: we will go back to using the original variable for the cost of attendance, **not** the binary version you created above.

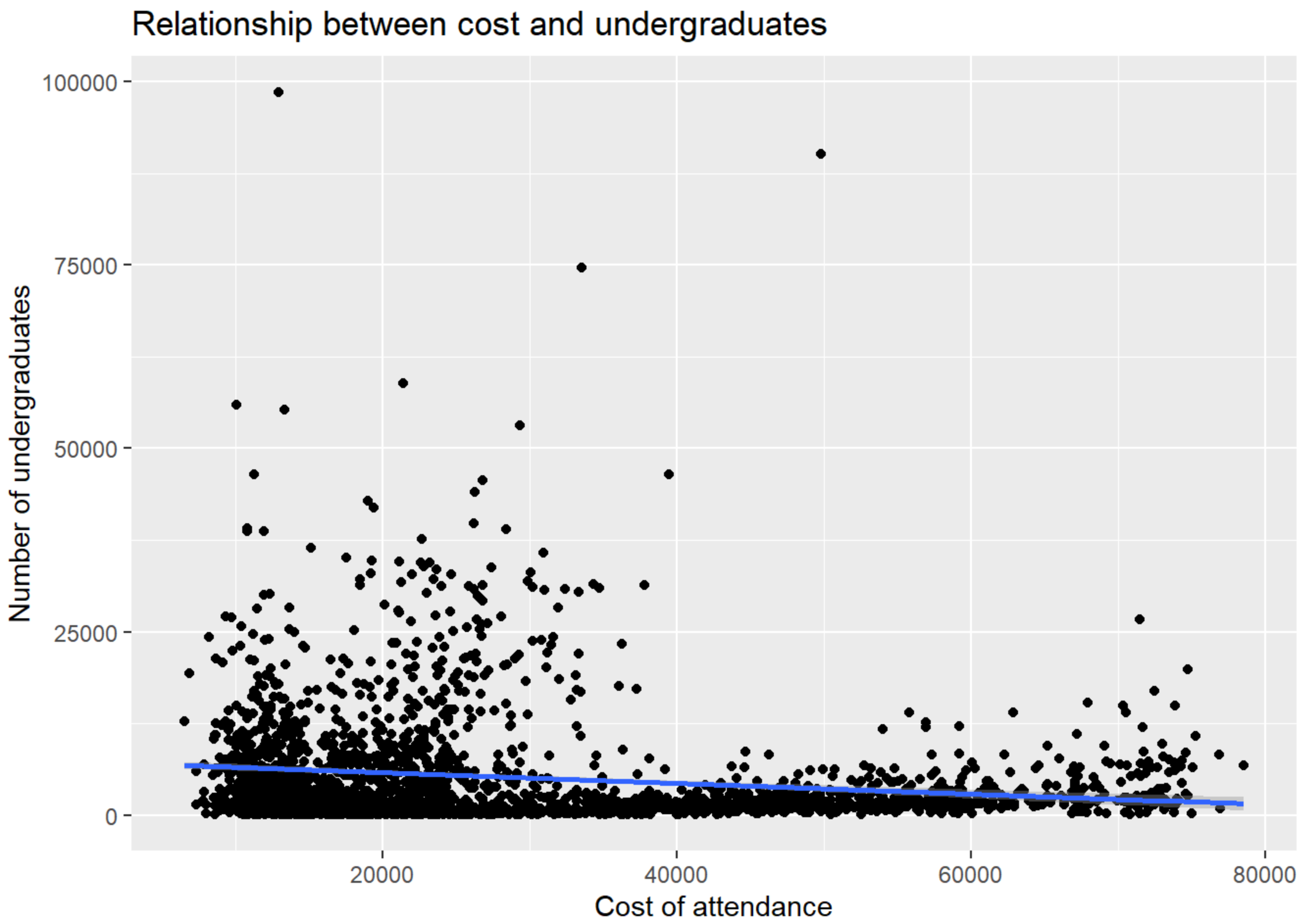
First, visualize the relationship between these two variables using a multivariate visualization. Make sure to choose the appropriate `geom_...()` and place the correct variables on the correct axes, again based on your theory and answers in Question 1. Does the visual inspection change your answer to the research question? **[3 points]**

```
debt %>%
  ggplot(aes(x = costt4_a,y = ugds)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(x = 'Cost of attendance',
       y = 'Number of undergraduates',
       title = 'Relationship between cost and undergraduates')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 136 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 136 rows containing missing values (`geom_point()`).
```



> The multivariate visualization does not change my answer. We still see a negative association between the number of undergraduates and the cost of attendance, with more expensive schools having fewer undergrads.

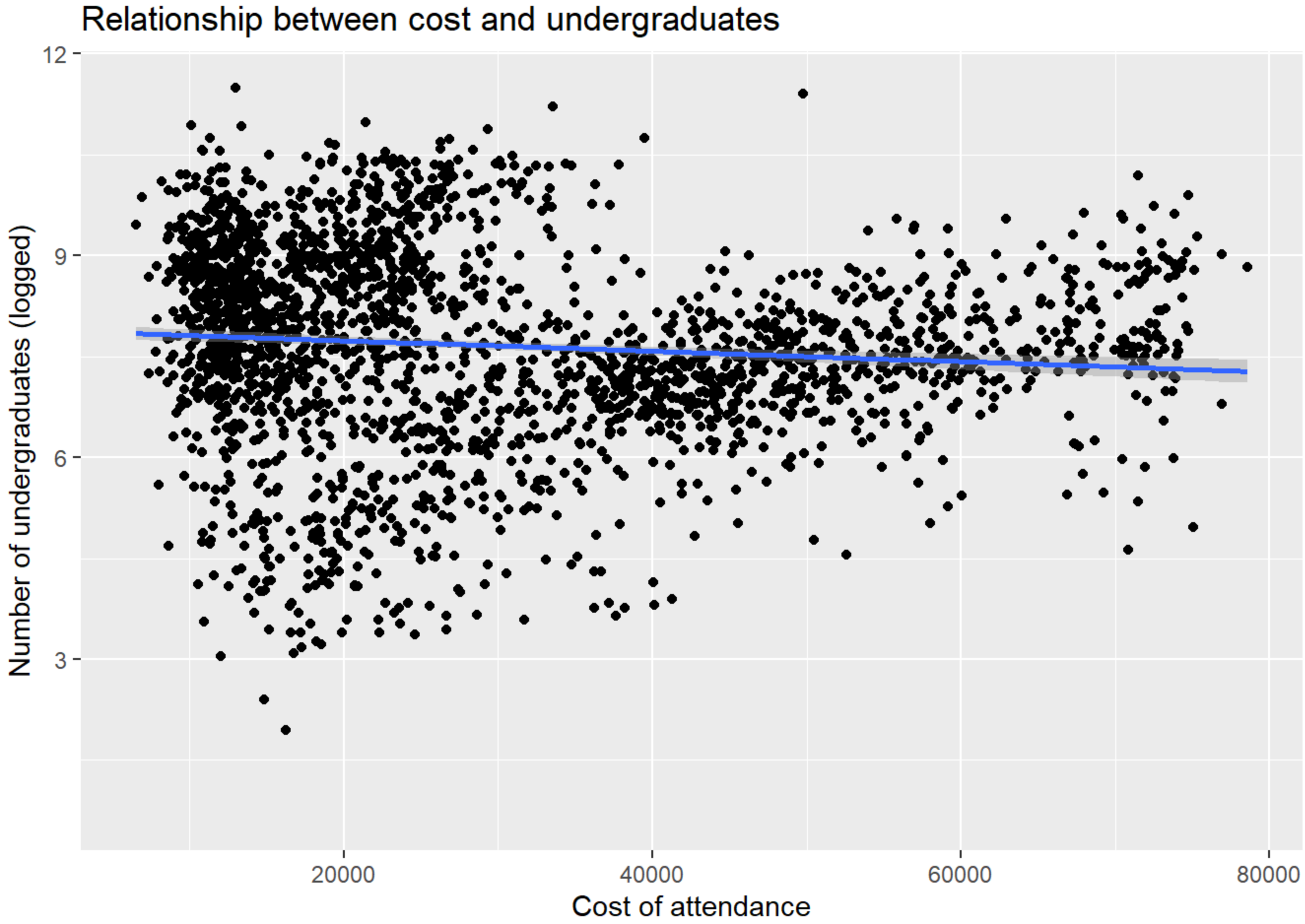
Second, re-create the graph but log the `ugds` variable. Does your answer change? **[2 points]**

```
debt %>%
  mutate(ugds = log(ugds)) %>%
  ggplot(aes(x = costt4_a,y = ugds)) +
  geom_point() +
  geom_smooth(method = 'lm') +
  labs(x = 'Cost of attendance',
       y = 'Number of undergraduates (logged)',
       title = 'Relationship between cost and undergraduates')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 136 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 136 rows containing missing values (`geom_point()`).
```



> The relationship is still negative, although it is very close to zero.

Finally, run the regression of Y on X using the `lm()` function. Save the model to a new object called `model_cost_ugds` using the object assignment operator `<-` and interpret the output of the regression using the `summary(model_cost_ugds)` command. Describe what the intercept (α) and slope (β) mean in substantive terms. Do the results support your theory? How confident are you in this conclusion? (**NOTE**: Do not use the logged version of the `ugds` variable for this analysis.) **[3 points]**

```
# Write code here
summary(model_cost_ugds <- lm(ugds ~ costt4_a,debt))
```

```
##
## Call:
## lm(formula = ugds ~ costt4_a, data = debt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6516  -4122  -2250    901  92320
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.252e+03  3.030e+02  23.933   <2e-16 ***
## costt4_a    -7.263e-02  8.706e-03  -8.342   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7565 on 2408 degrees of freedom
##   (136 observations deleted due to missingness)
## Multiple R-squared:  0.02809,    Adjusted R-squared:  0.02769
## F-statistic:  69.6 on 1 and 2408 DF,  p-value: < 2.2e-16
```

```
-7.263e-02
```

```
## [1] -0.07263
```

```
7.252e+03
```

```
## [1] 7252
```

> The regression model indicates a negative relationship, consistent with the previous results and my hypothesis. The intercept (i.e., the α) is 7,252, meaning that schools with no costs have an average of 7,252 undergraduates. The β_1 coefficient is -7.263e-02 or -0.0726. Substantively, the model suggests that a $1 increase in the cost of attendance reduces the number of undergraduates by 0.07 students. (This is equivalent to concluding that a $100 increase in the cost of attendance reduces the number of undergraduates by approximately 7 students.) I am more than 99.9% confident in this conclusion.

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. Make sure to drop missing data from *both* the X and Y variables before conducting this analysis.

```
set.seed(123)
debt_analysis <- debt %>%
  select(costt4_a,ugds) %>%
  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(ugds ~ costt4_a,train)
  cvRes <- test %>%
    mutate(preds = predict(m,newdata = test)) %>%
    summarise(rmse = sqrt(mean((ugds - preds)^2))) %>%
    mutate(cvInd = i) %>%
    bind_rows(cvRes)
}

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

```
## # A tibble: 1 × 1
##   `mean(rmse)`
##          <dbl>
## 1        7581.
```

> My model makes mistakes of roughly 7,581 students on average.

EC - Conclusion: 4 points

Re-run the regression but subset the data to privately controlled institutions. Does your answer change? Now re-run subsetting the data to publicly controlled institutions? Does your answer change in this subset? Visualize the multivariate relationship again, coloring points by `control` and using the logged version of the `ugds` variable, along with `geom_smooth(method = 'lm')`. Which set of conclusions do you think is correct: the one from above where you ran the regression on the full data, or the one here where you ran two separate regressions? Why?

```
summary(model_cost_ugds_private <- lm(ugds ~ costt4_a,debt %>% filter(control == 'Private')))
```

```
##
## Call:
## lm(formula = ugds ~ costt4_a, data = debt %>% filter(control ==
##     "Private"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3683  -1202   -702    -66  98150
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.189e+02  3.819e+02  -0.573    0.567
## costt4_a     5.391e-02  8.362e-03   6.447 1.65e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4726 on 1207 degrees of freedom
##   (111 observations deleted due to missingness)
## Multiple R-squared:  0.03329,    Adjusted R-squared:  0.03249
## F-statistic: 41.56 on 1 and 1207 DF,  p-value: 1.649e-10
```

```
summary(model_cost_ugds_public <- lm(ugds ~ costt4_a,debt %>% filter(control == 'Public')))
```

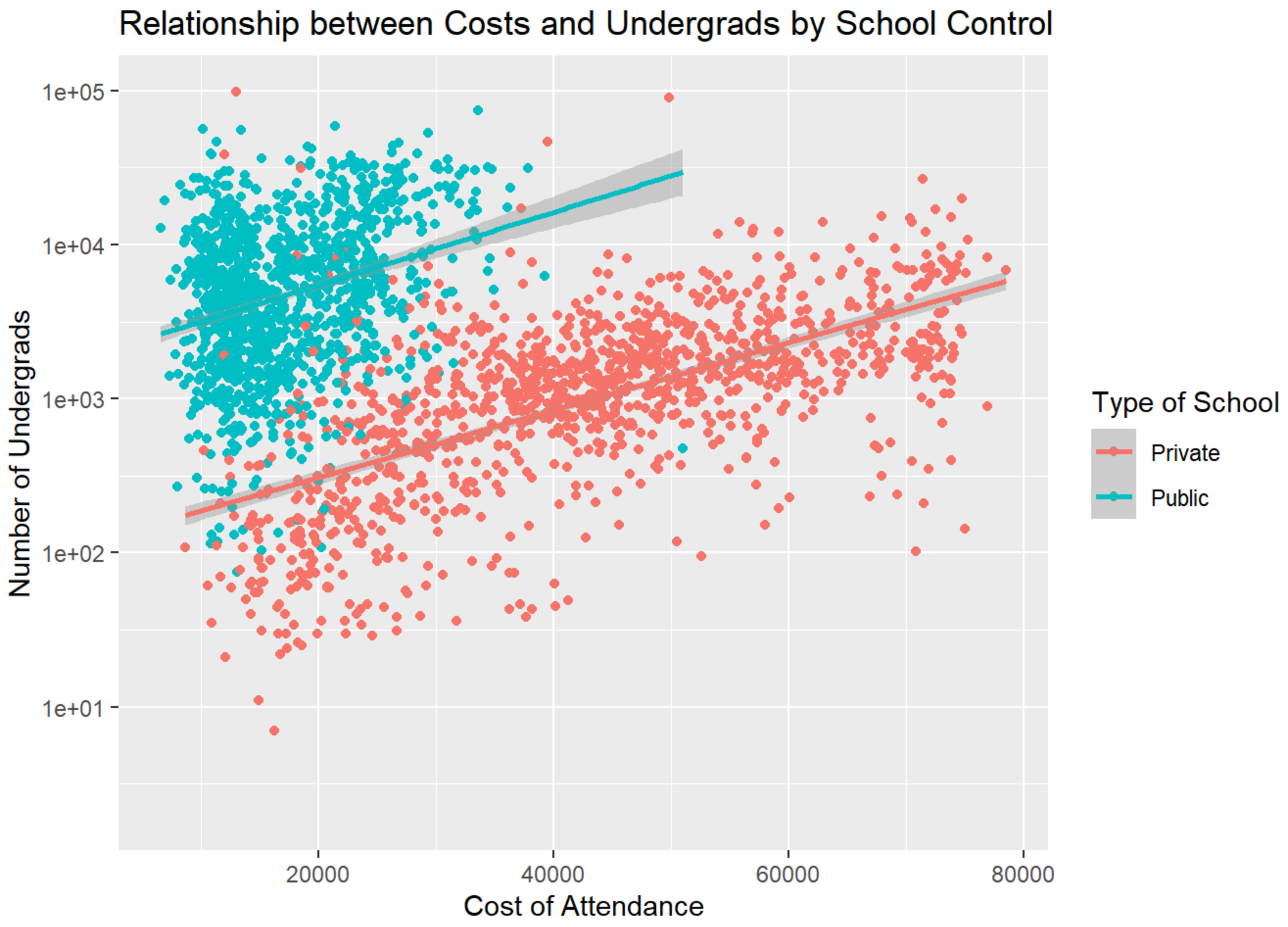
```
##
## Call:
## lm(formula = ugds ~ costt4_a, data = debt %>% filter(control ==
##     "Public"))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23252  -5037  -2417   2686  58955
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 145.25224  723.08503   0.201    0.841
## costt4_a      0.46248    0.03971  11.645   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8310 on 1199 degrees of freedom
##   (25 observations deleted due to missingness)
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.1009
## F-statistic: 135.6 on 1 and 1199 DF,  p-value: < 2.2e-16
```

```
debt %>%
  ggplot(aes(x = costt4_a,y = ugds,color = control)) +
  geom_point() +
  scale_y_log10() +
  geom_smooth(method = 'lm') +
  labs(x = 'Cost of Attendance',
       y = 'Number of Undergrads',
       color = 'Type of School',
       title = 'Relationship between Costs and Undergrads by School Control')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 136 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 136 rows containing missing values (`geom_point()`).
```



> Subsetting the data to either private or public institutions reverses the results of my model. In the private-only regression, a $100 increase in costs corresponds to 5 more students. In the public-only regression, a $100 increase in costs corresponds to 46 more students. I think this set of conclusions is correct, meaning my theory and hypothesis are wrong, because we are controlling for the type of the school. In the full data regression, the results are negative because public schools are both cheaper and bigger than private schools, obscuring the fact that *within* these schools, the relationship between cost and undergraduates is actually positive.

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 Evaluation Completion Code”.

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.