Ford Final Project

Libby Ford

2023-04-08

Part I Please answer the following questions based on the attached article: Does Attack Advertising Demobilize the Electorate?

1. Is this a causal study? In other words, is the aim of the study to estimate the causal effect of a treatment on an outcome? (2.5 points)

Yes, the aim of this study was to estimate the causal effect of negative/ attacking advertising on voter turnout in elections.

2. Is this a randomized experiment or an observational study? Explain your reasoning. (2.5 points)

This paper includes both a randomized experiment and an observational study to back their claims. The randomized experiment is the opt-in, paid study that randomly shows participants either a non-political advertisement, a positive ad, or a negative or "attacking" ad inside of a 15 minute news-program. They then asked many survey questions to figure out what other variables needed to be controlled for, then they asked if they were registered and planning to vote. This is randomized because the ad shown was randomly selected, and they manipulated the groups to have a treatment and control, making it an experimental study. The observational study in this paper is the real-world replication of their experiment, where they measured valence of political campaigns (classifying them into three categories) and then turnout in individual states by dividing the votes cast by the total voting-age population of the state. Because this study was not manipulated by the researchers and the data was not collected in a controlled, randomized environment, this is an observational study.

3. What is the treatment the study is interested in estimating the effects of? (Technically, the study is interested in the effect of two different treatments. For this whole problem set, just focus on one of them.) (5 points)

The treatment this study mainly focuses on is the negative or attacking political advertisements.

4. What is the outcome variable? (5 points)

The outcome variable of this study is the voter turnout or likelihood to vote.

5. What was the unit of observation? In other words, what does each observation represent? (2.5 points)

Each observation represents an individual voter and their exposure to negative, positive, or neutral ad campaigns, and it also includes control questions and their likelihood to vote in the coming elections.

6. How many people participated in this study? (Hint: you may need to look into the table containing the results of the analysis) (2.5 points)

1,655 people participated in this study.

7. What was the estimated average causal effect of the treatment on the outcome? In other words, what were the findings of the study? (Make sure to include the assumption, why the assumption is reasonable, the treatment, the outcome, as well as the direction, size, and unit of measurement of the average treatment effect) (10 points)

The estimated average effect of negative advertisements on voter turnout was -5%, meaning voters were 5% less likely to vote on average after being shown this advertisement as opposed to a positive ad. The assumption is that these advertisements alone are what is impacting an individual's likelihood to vote and that there is not another explanation for the relationship between X and Y. This is reasonable because voters were surveyed directly after exposure to these ads and because everything else in the ad and broadcast were held constant, from visuals, to audio, to overall theme in terms of the political message, the only difference was a few words that changed the valence of the ad. In the logistic regression model, voter turnout seemed to change by size 1.614 when the predictor has a one unit increase, meaning it had this impact both in the positive direction and the negative direction. After this logit coefficient was converted into a linear probability, the study found that negative ads made voters 2.5% less likely to vote, whereas positive ads made voters 2.5% more likely to vote, meaning the difference between the two was 5% on average.

8. How strong is the internal validity of this study? In other words, have the researchers accurately measured the causal effect on the sample of individuals who were part of the study? Please explain your reasoning. (5 points)

The internal validity is strong because they conducted detailed surveys to get individual's characteristics, voter history, and other important components that they eventually controlled for in their regression model. This means that the coefficient in the regression is likely not due to another relationship between variables that was not controlled for or due to random chance. The standard error on these coefficients also reflects the precision of their estimates, and the controls all demonstrate the accuracy.

9. How strong is the external validity of this study? In other words, can we generalize the results to a population outside of the sample? Please explain your reasoning and be specific about what population you think the findings can or cannot be generalized to. (5 points)

The external validity is relatively strong, because, although this sample was not randomized and required individuals to self-select in order to participate, the sample's distribution of different characteristics was indicative of the whole population. As described in the article, across all experiments, 56% of participants were male, 53% were white, 26% were black, 12% were Hispanic, and 10% were Asian, as well as 49% democratic participants, 24% republican, 21% independent, and 44% were college graduates. This means that there was a wide variety of people participating in this study, making it more indicative of the wide LA population, even though the sample was non-random.

Part II In this exercise, we will analyze student performance from a mathematics class. The objective is to model the relationship between midterm and final exam scores in mathematics so that we can later predict mathematics final exam scores based on midterm scores. The dataset we will use is in the Maths.csv file. Table: Maths below shows the names and descriptions of the variables in this dataset.

```
 Please answer the following questions based on the Maths.csv dataset.
```

```
maths <- read.csv("Maths.csv")
```

10. In this dataset, what does each observation represent? (2.5 points)

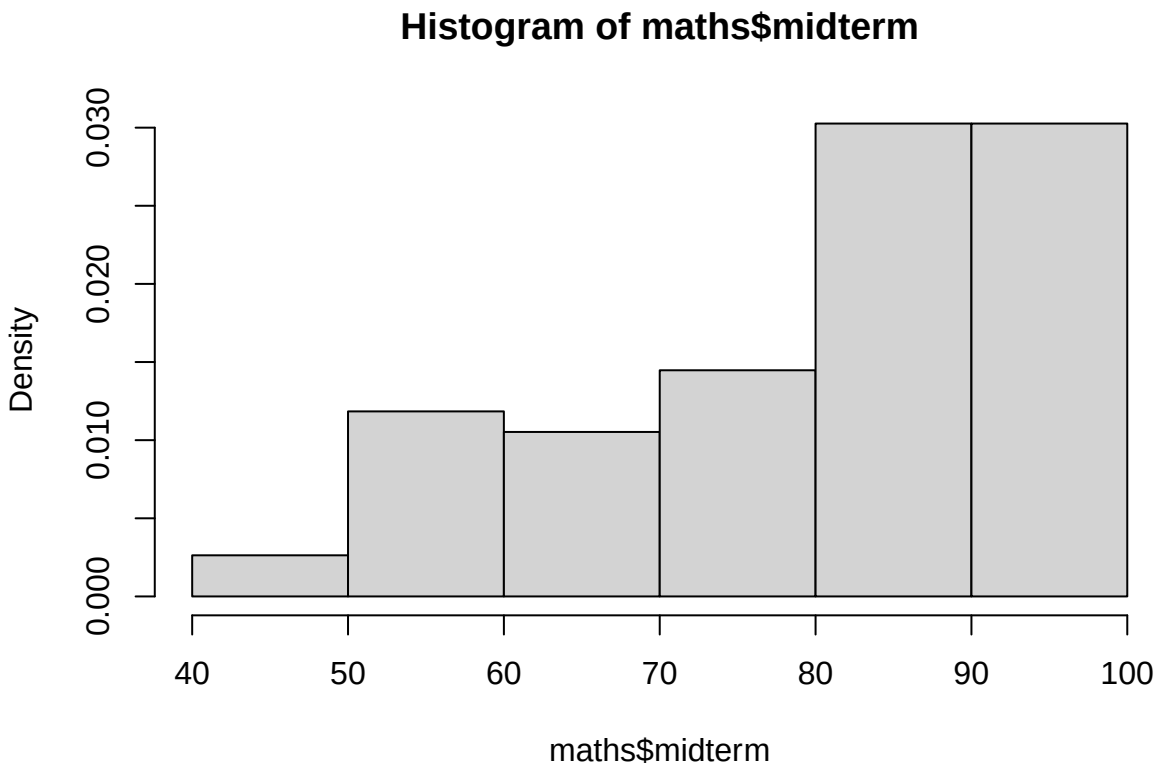
```
head(maths)
```

```
##   midterm final overall gradeA
## 1   79.25 47.00    69.2      0
## 2   96.25 87.75    94.3      1
## 3   58.25 37.75    62.0      0
## 4   54.50 62.00    72.4      0
## 5   83.00 39.75    72.4      0
## 6   41.75 49.50    59.5      0
```

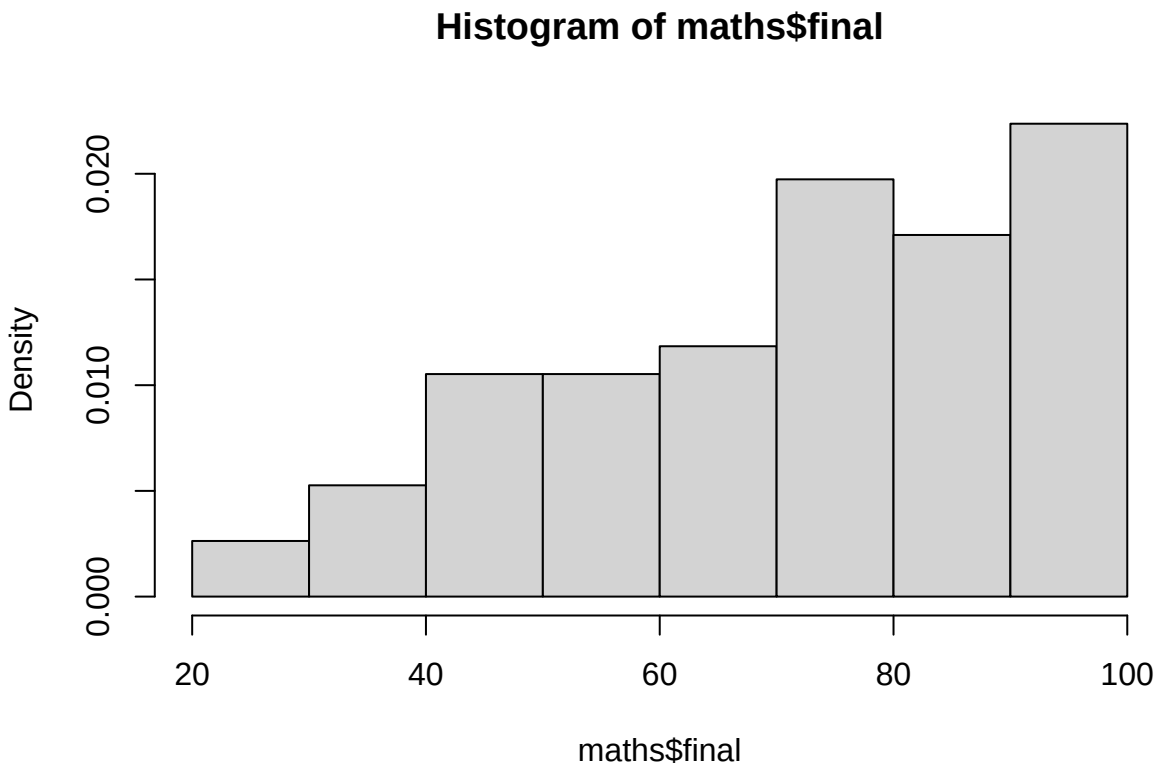
In this dataset, each observation represents an individual student and includes their midterm, final, and overall grades as well as a dummy variable for whether or not this student received an A/A- in their math course.

11. Create a visualization of the distribution of each of the variable in the dataset. (5 points)

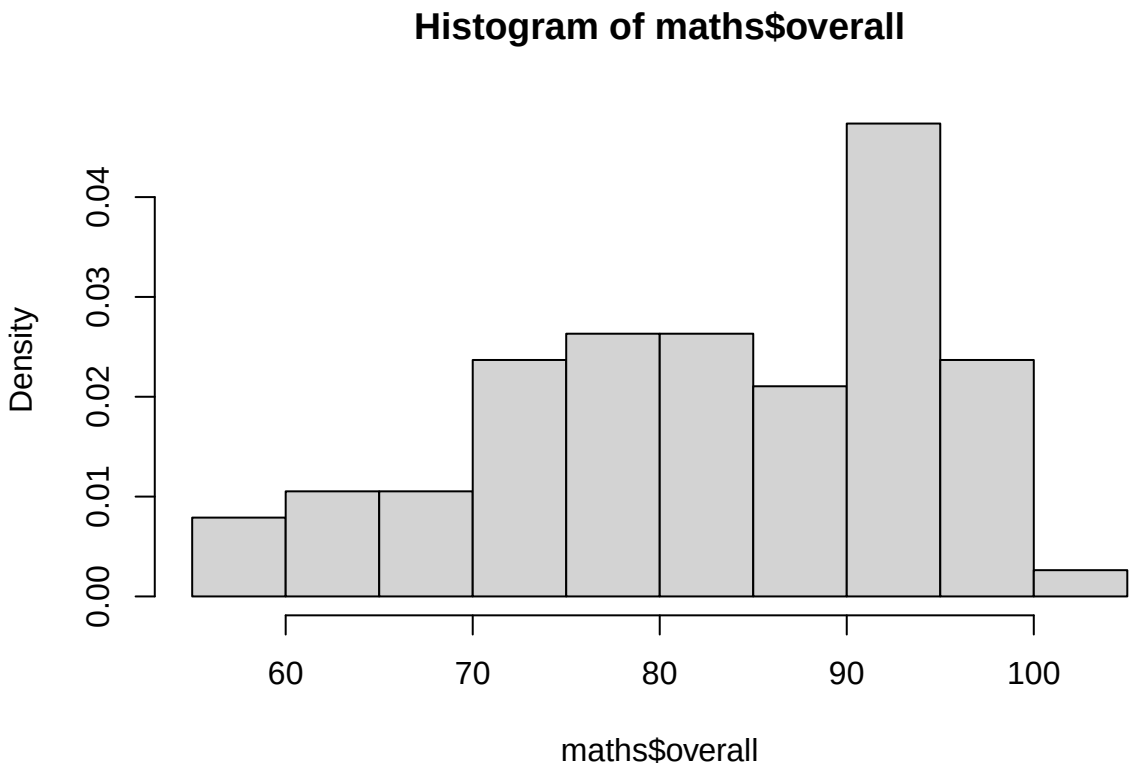
```
hist(maths$midterm, freq = F)
```



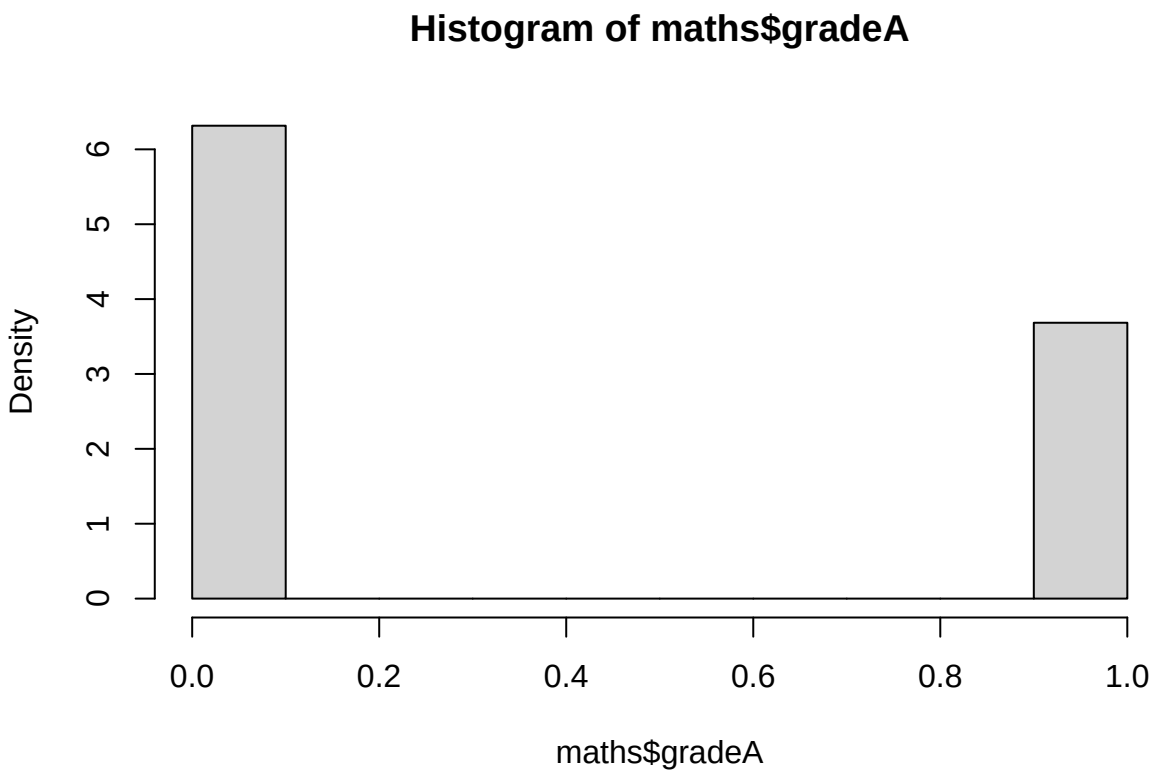

```
hist(maths$final, freq = F)
```



```
hist(maths$overall, freq = F)
```




```
hist(maths$gradeA, freq = F)
```



12. What should be our X variable? In other words, which variable are we going to use as the predictor?

Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

Our X variable or predictor is the midterm score because the objective of this study is to be able to predict final exam scores from the midterm scores.

13. What should be our Y variable? In other words, which variable are we going to use as the outcome variable? Please provide the name of the variable and identify whether it is binary or non-binary. (2.5 points)

Our Y variable or outcome variable is the final exam scores because we are trying to predict these scores based on a given student's midterm scores.

14. Compute the correlation coefficient between X and Y. Is the relationship between X and Y moderately or strongly linear? (2.5 points)

```
cor(maths$midterm, maths$final)
```

```
## [1] 0.7160323
```

The relationship between X and Y is strongly linear because a correlation coefficient of 0.716 demonstrates that these two variables are strongly correlated. The range for a correlation coefficient is between -1 and 1, so having the correlation between midterm and final grades be 0.716 (slightly higher than the moderately linear threshold), which is very close to one, means they are very strongly correlated and linear.

15. Second, let's fit the linear model that we will use to make predictions. Fit a linear model to summarize the relationship between X and Y and store the output in an object called *fit*. Then, ask R to provide the contents of *fit* by running its name. (5 points)

```
fit <- lm(final ~ midterm, data = maths)
fit
```

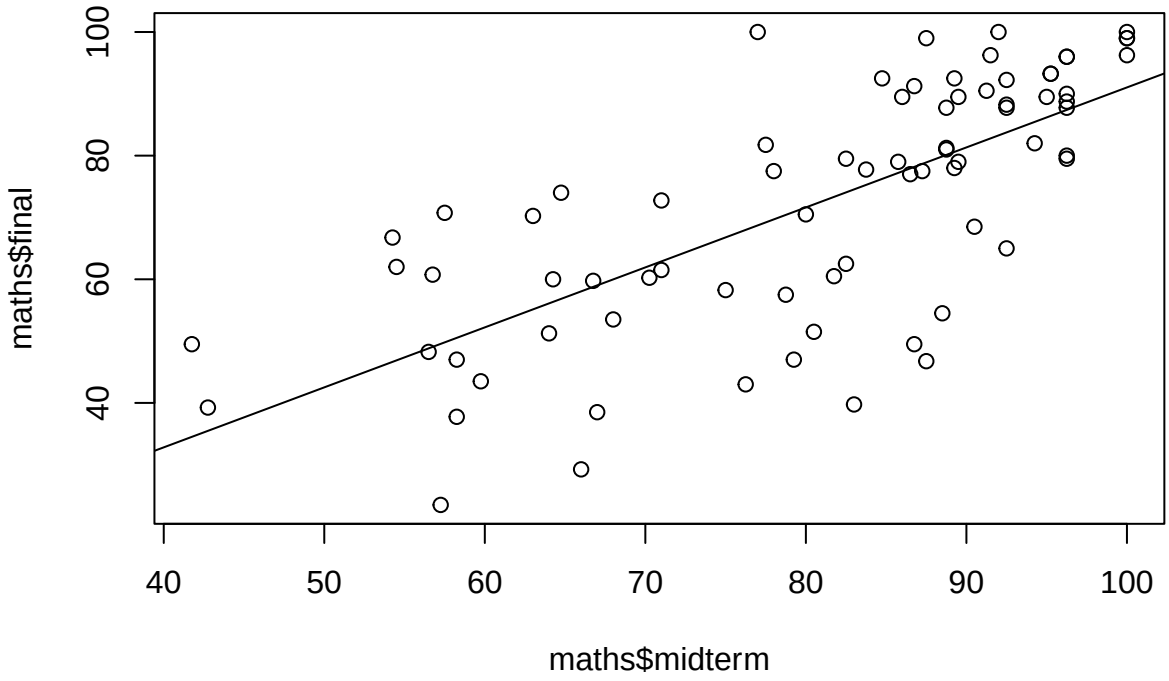
```
##
## Call:
## lm(formula = final ~ midterm, data = maths)
##
## Coefficients:
## (Intercept)      midterm
##     -6.0059       0.9704
```

16. What is the fitted line? In other words, provide the formula Y = a + b where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute a for the estimated value of the intercept coefficient, substitute b for the estimated value of the slope coefficient, and substitute X for the name of the predictor.) (5 points) (had to change alpha and beta so my file would knit!)

Final Exam Score = -6.0059 + Midterm Score(0.9704)

17. Create a visualization of the relationship between X and Y and add the fitted line to the graph using the function abline(). (5 points)

```
plot(maths$midterm, maths$final)
abline(fit)
```



18. Computing Y Hat based on X: Suppose that you earn 80 points in the midterm. What would be your best guess of your predicted final exam score based on your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

```
sum(-6.0059 + 80*(0.9704))
```

```
## [1] 71.6261
```

If I received an 80 on my midterm, I predict that I would receive around 71.626, so an 71/72 on my final exam because I can insert my midterm score where I had previously written "midterm score" in my equation for the fitted line in question 16. When I do this, I keep the other pieces of the line the same, including the intercept point and the slope, and then I calculate my final exam score using the equation above!

19. Computing X based on Y Hat: Now, suppose the best guess of your predicted final exam score is 90 points. What would be your performance in the midterm? Please show your calculations and then answer the question with a full sentence (including units of measurement). (5 points)

```
sum((90 + 6.0059)/0.9704)
```

```
## [1] 98.93436
```

If I received an 90 on my final exam, I predict that I would have receives around an 98.934, so an 98/99 on my midterm exam because I can still utilize my equation for the fitted line in question 16, but change the variable I am solving for. When I do this, I add the intercept to both sides of my equation (because it is negative) and divide by the slope, and this gives me my predicted midterm score based on my final exam score!

20. What is the R2 of the fitted model? And, how would you interpret it? (5 points)

```
summaryfit <- summary(fit)
summaryfit$r.squared
```

```
## [1] 0.5127023
```

The R^2 of this fitted model is 0.512703, and this coefficient measures the goodness of fit of the model. This value for this fitted model means that around 50% of the variability in the outcome of the data cannot be

explained by this model, meaning there are other models that likely fit this data better than our simple regression model.

21. We focus on the following subset of the variables: 'midterm', 'final', and 'overall'. Using the 'scale()' function, scale these variables so that each variable has a mean of zero and a standard deviation of one. Fit the k-means clustering algorithm with four clusters.(10 points)

```
maths.sub <- subset(maths, select = c("midterm", "final", "overall"))
maths.sub.z <- scale(maths.sub)
maths.sub.four.out <- kmeans(maths.sub.z, centers = 4, nstart = 5)
```

22. Create a visualization of the result of the k-means clustering algorithm above and indicate the centroids.(5 points)

```
plot(maths.sub.z, col = maths.sub.four.out$cluster + 1)
points(maths.sub.four.out$centers, pch = 8, cex = 2)
```