Final Exam

Daniel Shapiro

12/19/2022

Question 1 Background:

The case Chavez v. Illinois State Police (1999) was a class-action lawsuit against the Illinois State Police for unconstitutional racial profiling. Among other evidence, the plaintiffs claimed that the percentage of drivers stopped who were Black was not representative of the proportion of the Illinois population that is Black.

The 1990 census found that the racial breakdown of Illinois is roughly 75% White and 15% Black (you can treat these statistics as fixed). The racial breakdown of stopped drivers was 68% White and 25% Black. The standard deviations for the estimated share of stopped drivers was 0.22 (White) and 0.17 (Black), in a sample of 50,000 stopped drivers.

1a) Imagine you are planning to statistically evaluate the plaintiffs' claims that Black drivers were much more likely to get pulled over relative to their population. Clearly specify a null and alternative hypothesis (5 points)

Null hypothesis (μ_0: Black drivers are not more likely to get pulled over relative to their population.

Alternative hypothesis (μ_a): Black drivers are more likely to get pulled over relative to their population.

1b) Calculate a test statistic that allows you to assess the amount of evidence against the claim that the stopped drivers represent a random sample from the Illinois population. Provide an approximate p-value. Explain, in words, what the test statistic and p-value mean in terms of substantive and statistical significance. (10 points)

The t statistic to test hypothesis μ_a against null hypothesis μ_0 can be calculated as: $\frac{\mu_a - \mu_0}{SE(\hat{\mu_a})}$

Here, we see that this is $\frac{.25 - .15}{.17/\sqrt(50000)} = \frac{.1}{.0008} = 131.533$.

The t statistic, at first glance, looks abnormally large; after all, we are supposed to find out whether or not the t statistic fits between -1.96 and 1.96, and this number is nowhere close. But given the data, it makes sense: we have a massive sample size, so the level of precision of this sample will be quite high. The percentage of Black people stopped is **way** higher than would be expected if this sample were random, making it seem that there is almost no chance that this is a random sample of the Illinois population as a whole. A high t statistic indicates large differences between the sample data and the null hypothesis, so this number is logical.

The p-value backs this assessment up. Using the pt() function, we see that the p-value is essentially 0. This number is most certainly below the significance level of $\alpha = 0.05$, so we can reject the null hypothesis at the 5% level of significance and say that the sample of stopped drivers does not represent a random sample from the Illinois population.

```r
t <- -abs(131.533)
pt(q = t, df = 49999)
```

```
## [1] 0
```

```r
# I only did one side (instead of multiplying it by 2) because we are only looking to
# see whether or not Black drivers are more likely to get pulled over, not less.
```

1c) Calculate a confidence interval for your estimate of the difference. Interpret, in words, what this confidence interval means. Does this align with your findings in part b? (5 points)

The equation to find the confidence interval is: $\hat{\mu} \pm 1.96\hat{SE}$. Below, we can set this up.

```r
muhat <- .25
se <- .17/sqrt(50000)

c((muhat - 1.96 * se), (muhat + 1.96 * se))
```

```
## [1] 0.2485099 0.2514901
```

This confidence interval shows us that based on the sample, we can expect that the real "percent Black" of drivers stopped in Illinois is between 24.85% and 25.15%, with 95% certainty. This does fit with our results for part b, as the two bounds are *nowhere close* to 15%, which would be the expected "percent Black" of the sample if Black people were indeed not more likely to get pulled over relative to their population, as stipulated by the null hypothesis.

Question 2 Background:

Imagine that the true population model between our variables of interest (Y and X_1) is causal only when controlling for X_2:

$$Y = -4 + 1.4X_1 - .7X_2 + \epsilon$$

Where:

$$\begin{aligned} X_1 &= N(5, 2) \\ X_2 &= .5X_1^2 - U(0, 4) \\ \epsilon &= N(0, 2) \end{aligned}$$

Unfortunately, a researcher has modeled the relationship between Y and X_1 using the following model:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

2) Draw a random sample from the population distribution of Y, X_1, and X_2 ($n = 1000$). How much does the researcher's model (once you estimate $\hat{\beta}_1$ and $\hat{\beta}_0$) depart from the true population model? Why? Can this be fixed?

First, I set up all of the parameters.

```
dataframe <- data.frame(matrix(ncol = 2, nrow = 1000)) %>%
  mutate(X1 = rnorm(1000, mean = 5, sd = 2)) %>%

# First I make this for X2, then I do a for loop later to address the rest. Seemed easier.

  mutate(X2 = runif(1000, min = 0, max = 4)) %>%
  mutate(epsilon = rnorm(1000, mean = 0, sd = 2))

for(i in 1:1000){
  dataframe$X2[i] <- .5 * (dataframe$X1[i]^2) - dataframe$X2[i]
  dataframe$Y[i] <- -4 + 1.4*dataframe$X1[i] - .7*dataframe$X2[i] + dataframe$epsilon[i]
}
```

Now, I run a regression using the model that the researcher used to see the discrepancy in coefficients.

```
model2 <- lm(Y ~ X1, data = dataframe)
summary(model2)
```

```
##
## Call:
## lm(formula = Y ~ X1, data = dataframe)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1301  -1.5844   0.3578   1.9495   9.6450
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.97096    0.24338   16.32   <2e-16 ***
## X1          -1.97315    0.04545  -43.41   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.941 on 998 degrees of freedom
## Multiple R-squared:  0.6538, Adjusted R-squared:  0.6535
## F-statistic:  1885 on 1 and 998 DF,  p-value: < 2.2e-16
```

As we can see, the model that the researcher used is way wrong and deviates quite a lot. The formula for the true model is $Y = -4 + 1.4X_1 - .7X_2 + \epsilon$, whereas this one ends up as about $Y = 4.85 - 2.13X_1 + \epsilon$, where $\beta_0 = 4.85$ and $\beta_1 = -2.13$. The effect of X_1 on Y is not only different than in the real model – it has the wrong sign as well. The researcher's model shows a negative effect, but the real relationship is positive when the X_2 control gets put in.

Can this be "fixed?" Well, it depends on what you mean by fixed. I don't really see how it can be fixed given the model that the researcher uses, because it's just wrong – the effect is completely misrepresented. If we add in the X_2 variable to the regression, though, we can see the correct effect:

```r
editedmodel2 <- lm(Y ~ X1 + X2, data = dataframe)
summary(editedmodel2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = dataframe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7439 -1.2236  0.0256  1.3009  7.8046
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.27958    0.29527  -14.49   <2e-16 ***
## X1           1.48151    0.10680   13.87   <2e-16 ***
## X2          -0.71448    0.02113  -33.81   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.008 on 997 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8384
## F-statistic:  2592 on 2 and 997 DF,  p-value: < 2.2e-16
```

Using (Y ~ X1 + X2) makes our coefficients look much better. β_0 of -3.95, β_1 of 1.37 and β_2 of -0.69 look very similar to the values given in the equation – -4, 1.4 and -0.7. There is some small differentiation present due to variation in sampling, but the general trends look correct.

Question 3 Background:

To gauge the effect of intrinsic versus extrinsic motives for voting, Gerber, Green, and Larimer conducted a field experiment in Michigan prior to the August 2006 primary election. Voters were randomly assigned to either the control group (no mailer) or one of four treatment groups. Treatment and randomization was at the household level.

1. All four treatments carry the message “DO YOUR CIVIC DUTY - VOTE!” The first type of mailing (*Civic Duty*) provides a baseline for comparison with the other treatments.
2. Households receiving the *Hawthorne* mailing were told “YOU ARE BEING STUDIED!” and informed that researchers would examine their voting behavior by means of public records.
3. The *Self* mailing exerts more social pressure by informing recipients that who votes is public information and listing the recent voting record of each registered voter in the household.
4. The fourth mailing, *Neighbors*, lists not only the household’s voting records but also the voting records of those living nearby. By threatening to “publicize who does and does not vote,” this treatment is designed to apply maximal social pressure.

I have provided you with the original data of Gerber et al. The data is available on the course website (`gerber.dta`). Below is a list of the important variable definitions in your dataset.

- *treatment*: which treatment condition respondents were assigned to. Note the leading spaces on the factor levels.

- *voted* = ‘Yes’ if Respondent voted in the 2006 Primary Election (‘No’ otherwise)
- *yob* - year of birth
- *sex* - ‘male’ or ‘female’
- *hh_id* - a unique household identifier
- *g2002* = ‘yes’ if Respondent voted in the 2002 General Election (‘no’ otherwise)

3a) Use OLS to estimate the average effects of the four treatments on *voting*, not adjusting for any of the other variables. Report the results in a nicely-formatted table. Do you have a lot of confidence in these estimates? Why or why not? Discuss the plausibility of each of the regression assumptions. (15 points)

```
gerber <- read.dta("gerber.dta")
```

First, this is going to require some serious data reworking, because the categories that we are looking at are not exactly formatted well for data analysis. First, I’m going to a) take out the columns that we don’t work with, and b) change some columns around so that they can better work for regression.

```
# Taking out extraneous variables just for ease of looking.

gerber <- gerber %>%
  select(c(sex, yob, g2002, treatment, voted, hh_id))

# Now, changing some columns. I should be able to use recode().

gerber$voted <- recode(gerber$voted, No = 0, Yes = 1)
```

Now, let’s see what happens when we run a regression.

```
model3 <- lm(voted ~ treatment, data = gerber)
stargazer(model3, type = "text")
```

```
##
## ================================================
##                           Dependent variable:
##                       ---------------------------
##                                  voted
## ------------------------------------------------
## treatment Hawthorne           0.026***
##                                (0.003)
##
## treatment Civic Duty          0.018***
##                                (0.003)
##
## treatment Neighbors           0.081***
##                                (0.003)
##
## treatment Self                0.049***
##                                (0.003)
```

```
##
## Constant                          0.297***
##                                    (0.001)
##
## ------------------------------------------------
## Observations                      344,084
## R2                                  0.003
## Adjusted R2                         0.003
## Residual Std. Error      0.464 (df = 344079)
## F Statistic          292.976*** (df = 4; 344079)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Essentially, what this table shows is that being part of the control group implies about a 29.7% chance that the person voted in the August 2006 primary election. With the "Civic Duty" baseline treatment, that percentage jumps to about 31.5% chance of voting (29.7 + 1.8). With the "Hawthorne" treatment, it rises to about 32.3%, with the "Self" treatment, 34.6%, and with the final "Neighbors" treatment, 37.8%.

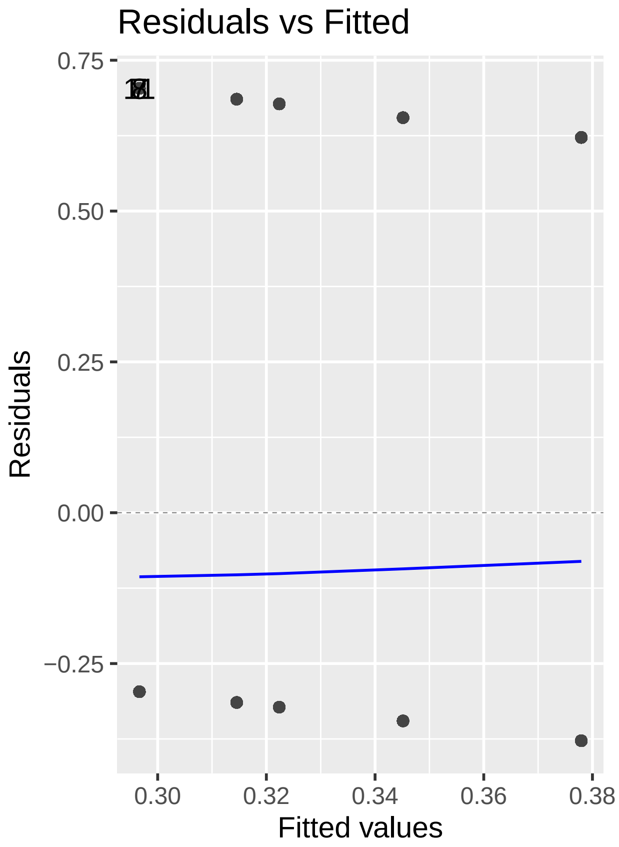
Do I feel confident about these results? Setting aside the question of whether or not using OLS is appropriate for a regression with a binary dependent variable, there's the problem of **cluster dependence** here. We can already tell that regression assumptions will be violated, as sampling and treatment is applied at the household level, but the regression assumes that people are sampled at the *individual* level. We also learned in class that this issue leads to errors and outcomes of individuals being correlated, so by assuming independence, we "overstate the amount of information." We will also have some problems with homoskedasticity.

Taking these issues in mind, let's dive into the assumptions. We will see that linear regression is likely **not** appropriate here, as multiple assumptions fail.

1) Linearity in Parameters:

A good test for linearity in parameters is the residuals vs. fitted values test. Below, we can run this:

```
autoplot(model3, 1)
```



This looks fairly linear; sure, the line is a bit lower than 0, but it is straight at least, and it isn't far from 0. This is a weird example, honestly, because while we do have a linear model here, the variables look a bit different than what we usually see. Usually, we see that at least one side or the other is continuous; here, both are factors. So we can't say "a one-point increase in X leads to a ___ increase in Y;" interpretation is a bit more difficult. The equation is technically linear and the the residuals vs. fitted values test shows linearity, but we must be very careful with interpretation.

2) Random Sampling

This sample seems random, at least among households. The research design is specifically supposed to be random, and the fact that there are 344,084 data points speaks to this factor. I don't see any evidence that the sample is not random, at least at the household level.

3) Variation in X

There is definitely variation in X; there are different values of x.

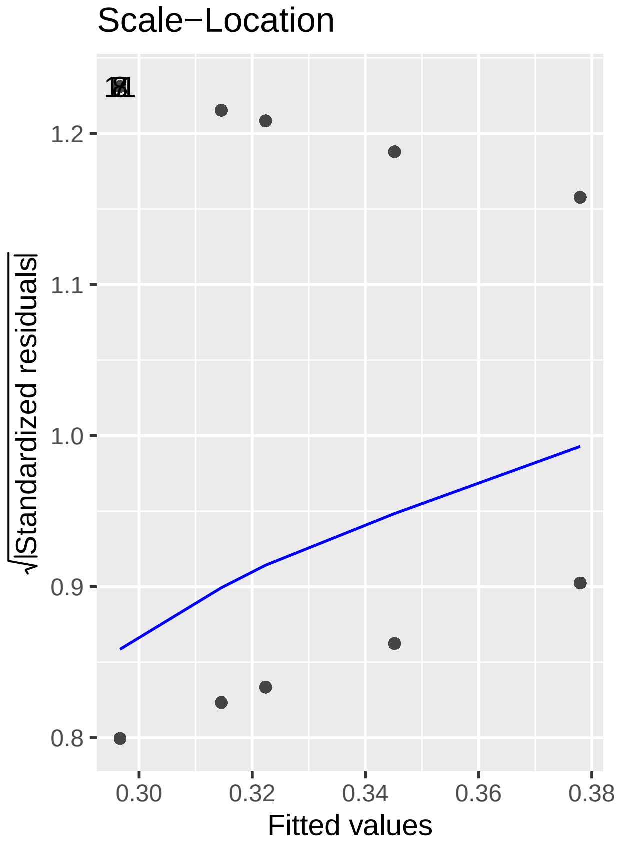
4) Zero conditional mean

In order to say that we have "zero conditional mean" it means that we have to have exogeneity. Here, I think that we can probably say that there is exogeneity; there is no way that voting in the August 2006 primaries influenced receiving the treatment which literally happened before the vote. This parameter is satisfied as well as it can be.

5) Homoskedasticity

This should be a problem and should show us why we don't use OLS for binary dependent variables.

```
autoplot(model3, 3)
```

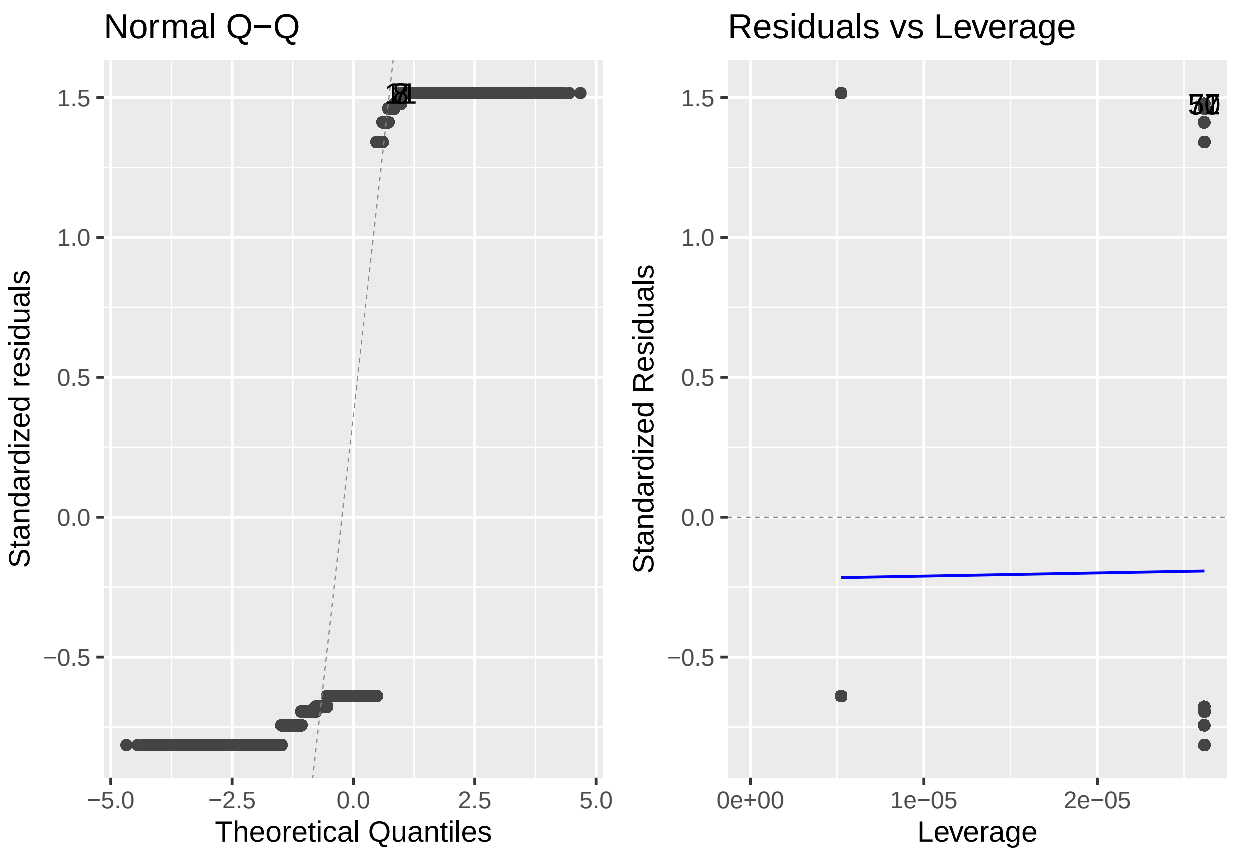


Indeed, we see that this regression does **not** seem homoskedastic – the line should be horizontal.

6) Normality

The errors are nowhere close to normally distributed. We can check by looking at these two plots.

```
autoplot(model3, c(2, 5))
```



3b) Repeat part a) with robust standard errors (avoid using lm_robust(), since this will likely crash your R session). How do your findings change? Why? Which standard errors do you prefer? (6 points)

```r
robusthc2 <- coeftest(model3, vcov = vcovHC(model3, type = "HC2"))
robusthc3 <- coeftest(model3, vcov = vcovHC(model3, type = "HC3"))

robusthc2
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)            0.2966383  0.0010445 283.9984 < 2.2e-16 ***
## treatment Hawthorne    0.0257363  0.0026094   9.8628 < 2.2e-16 ***
## treatment Civic Duty 0.0178993  0.0025947   6.8984 5.269e-12 ***
## treatment Neighbors    0.0813099  0.0026918  30.2071 < 2.2e-16 ***
## treatment Self         0.0485132  0.0026467  18.3295 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
robusthc3
```

```
##
```

```
## t test of coefficients:
##
##                          Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)             0.2966383  0.0010445 283.9977 < 2.2e-16 ***
## treatment Hawthorne     0.0257363  0.0026095   9.8627 < 2.2e-16 ***
## treatment Civic Duty    0.0178993  0.0025947   6.8983 5.272e-12 ***
## treatment Neighbors     0.0813099  0.0026918  30.2067 < 2.2e-16 ***
## treatment Self          0.0485132  0.0026468  18.3293 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Frankly, the results barely change at all. I tried both HC2 and HC3 errors, but very little was different. The Std. Error and t value columns are a bit different, but not by much at all. I do prefer using robust standard errors here, because it's generally good practice, but it doesn't actually make that much of a difference. Our n is so big that it doesn't really do all that much.

3c) Now repeat part a) with robust standard errors clustered at the household level. How do your findings change? Why? Which standard errors do you prefer? (6 points)

Not being able to use lm_robust() does make this harder, but I think that I can do it this way as well. I used HC2 errors, which are considered more robust and are the default in lm_robust().

```
coeftest(model3, vcovCL(model3, cluster = gerber$hh_id, type = "HC2"))
```

```
##
## t test of coefficients:
##
##                          Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)             0.2966383  0.0013096 226.5178 < 2.2e-16 ***
## treatment Hawthorne     0.0257363  0.0032579   7.8995 2.807e-15 ***
## treatment Civic Duty    0.0178993  0.0032367   5.5302 3.202e-08 ***
## treatment Neighbors     0.0813099  0.0033696  24.1303 < 2.2e-16 ***
## treatment Self          0.0485132  0.0033001  14.7006 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Without clustering the standard errors, we are assuming that there is no correlation in observations from the same household. That seems very unlikely, given that voting rates are likely similar within households. So it's good that we cluster our robust errors here. That being said, implementing cluster-robust standard errors doesn't really change the significance of our treatment, which makes sense – given that there are just so many observations, even cutting them max by 3 or so (maybe 3 voters per household, max?), we still will have about 100,000 observations, which is still a very high n and should give us pretty similar levels of error. If we were going from like 1,000 observations to 50, we might lose significance, but here, we're just raising the standard error a bit without actually changing the significance level because our n remains high.

3d) Estimate 95% confidence intervals for the effect of the four treatments, relative to the control condition, based on your results in part c). Describe the substantive and statistical significance of the four treatments. (10 points)

Taking the numbers from above, we can simply get the lower bounds of the effects by taking the Estimate value - 1.96 * the Std. Error value. The upper bound is simply the Estimate value + 1.96 * the Std. Error value. Thus, we end up with:

Hawthorne: c((0.0257363 - 1.96 * 0.0032579), (0.0257363 + 1.96 * 0.0032579)) = **0.019350816, 0.032121784.** Thus, we can say with 95% certainty that the true effect of the Hawthorne treatment on someone voting in the August 2006 midterm elections in Michigan (clustered by household) is between a 1.935% and a 3.212% increase.

Civic Duty: c((0.0178993 - 1.96 * 0.0032367), (0.0178993 + 1.96 * 0.0032367)) = **0.011555368, 0.024243232.** Thus, we can say with 95% certainty that the true effect of the Civic Duty treatment on someone voting in the August 2006 midterm elections in Michigan (clustered by household) is between a 1.156% and a 2.424% increase.

Neighbors: c((0.0813099 - 1.96 * 0.0033696), (0.0813099 + 1.96 * 0.0033696)) = **0.074705484, 0.087914316.** Thus, we can say with 95% certainty that the true effect of the Neighbors treatment on someone voting in the August 2006 midterm elections in Michigan (clustered by household) is between a 7.471% and an 8.791% increase.

Self: c((0.0485132 - 1.96 * 0.0033001), (0.0485132 + 1.96 * 0.0033001)) = **0.042045004, 0.054981396.** Thus, we can say with 95% certainty that the true effect of the Self treatment on someone voting in the August 2006 midterm elections in Michigan (clustered by household) is between a 4.205% and an 5.498% increase.

3e) Estimate whether the effects of the Hawthorne and Neighbors treatments are significantly different from one another (NOTE the leading space in the treatment variable levels). Explain the substantive and statistical significance of these findings. (8 points)

I think that I can just do this in sort of a creative way. I just take out the treatments that I don't want and compare them to one another.

```
edata <- gerber %>%
  filter(treatment != " Control") %>%
  filter(treatment != " Self") %>%
  filter(treatment != " Civic Duty")

model4 <- lm(voted ~ treatment, data = edata)
summary(model4)
```

```
##
## Call:
## lm(formula = voted ~ treatment, data = edata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3780 -0.3780 -0.3224  0.6220  0.6776
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.322375   0.002436  132.32   <2e-16 ***
## treatment Neighbors 0.055574   0.003446   16.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4762 on 76403 degrees of freedom
## Multiple R-squared:  0.003393,   Adjusted R-squared:  0.00338
## F-statistic: 260.1 on 1 and 76403 DF,  p-value: < 2.2e-16
```

Yes, they are significantly different from one another. Essentially, substantively, this means that those households affected by the full-on onslaught of the "Neighbors" treatment have about a 5.55% higher chance of voting in the 2006 midterms than those who were only exposed to the softer "Hawthorne" treatment. The effect is significant at the 0.001 level, so it really does seem to be different.

3f) Using your model from part c), control for g2002. Report what has changed by controlling for previous voting behaviors. Did you find meaningful changes? Why or why not? (5 points)

```
gerber$g2002 <- recode(gerber$g2002, no = 0, yes = 1)
```

```
model4 <- lm(voted ~ treatment + g2002, data = gerber)
coeftest(model4, vcovCL(model4, cluster = gerber$hh_id, type = "HC2"))
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error t value  Pr(>|t|)
## (Intercept)           0.1624257  0.0018806 86.3692 < 2.2e-16 ***
## treatment Hawthorne   0.0253959  0.0032332  7.8548 4.015e-15 ***
## treatment Civic Duty  0.0178638  0.0032106  5.5639 2.640e-08 ***
## treatment Neighbors   0.0812363  0.0033393 24.3273 < 2.2e-16 ***
## treatment Self        0.0484170  0.0032706 14.8037 < 2.2e-16 ***
## g2002                 0.1655117  0.0018905 87.5470 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There are meaningful changes in some areas, but not in others. Overall, the coefficients for the four treatments do not change significantly, but the control group does. I believe that this is because the intercept (the control) now represents those that weren't treated **and** did not vote in 2002, whereas the treatment effects stay unaffected. It looks like those in the "Control" group who did vote in 2002 were almost twice as likely to vote in 2006 as those who did not.

Question 4 Background:

We will continue to work with the Gerber et al data. This time, drop all observations except those in the Control group and Neighbors treatment.

```
q4data <- gerber %>%
  filter(treatment != " Self") %>%
  filter(treatment != " Civic Duty") %>%
  filter(treatment != " Hawthorne")
```

4a) Use a student's t-test to compare the difference in turnout between the Control and Neighbors conditions. Interpret your results. How do they compare to your regression results from Problem 3? (7 points)

```r
ttestprep1 <- q4data %>%
  filter(treatment == " Control") %>%
  select(voted)

ttestprep2 <- q4data %>%
  filter(treatment == " Neighbors") %>%
  select(voted)

t.test(x = ttestprep1, y = ttestprep2)
```

```
##
##  Welch Two Sample t-test
##
## data:  ttestprep1 and ttestprep2
## t = -30.207, df = 52613, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08658577 -0.07603405
## sample estimates:
## mean of x mean of y
## 0.2966383 0.3779482
```

Our data from this t test match the data from Problem 3. We see that turnout for the control population was around 29.66%, whereas the turnout for the "Neighbors" treated population was around 37.79%. This is what we see in Problem 3 as well.

Interpreting our results further, we can see with 95% certainty that the true difference in means lies between 0.0866 and 0.0760, or an 8.66% and a 7.60% difference. "t" stands for the t test statistic, df for degrees of freedom, and p-value is, well, the p-value.

4b) Run a regression of voting on age, without accounting for treatment status (you will have to create your own age variable). Interpret the substantive significance of the coefficient on age. (5 points)

```r
q4bdata <- q4data %>%
  mutate(age = 0)

for(i in 1:nrow(q4bdata)){
  q4bdata$age[i] <- 2006 - q4bdata$yob[i]
}
```

```r
model5 <- lm(voted ~ age, data = q4bdata)
summary(model5)
```

```
##
## Call:
## lm(formula = voted ~ age, data = q4bdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5409 -0.3232 -0.2740  0.6316  0.8123
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.056e-01  3.436e-03   30.73   <2e-16 ***
## age         4.107e-03  6.623e-05   62.01   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4587 on 229442 degrees of freedom
## Multiple R-squared:  0.01648,    Adjusted R-squared:  0.01648
## F-statistic:  3845 on 1 and 229442 DF,  p-value: < 2.2e-16
```

The age coefficient is very small and positive, meaning that for each additional year of age, turnout goes up about 0.4%. This really only starts to matter after a certain period of time; the Estimate value for (Intercept) presumably is for when age hypothetically is equal to 0. Since people can only start voting at 18, the intercept doesn't mean all that much substantively.

4c) Run a regression of voting on *treatment*, *sex*, and an interaction between *treatment* and *sex*. What is the average effect of treatment for men? What is the average treatment effect for women? (Just point estimates, no standard errors necessary) (8 points)

```
model6 <- lm(voted ~ treatment + sex + treatment*sex, data = q4data)
summary(model6)
```

```
##
## Call:
## lm(formula = voted ~ treatment + sex + treatment * sex, data = q4data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3845 -0.3028 -0.2905  0.6286  0.7095
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.3027947  0.0014910 203.087  < 2e-16 ***
## treatment Neighbors           0.0817482  0.0036574  22.351  < 2e-16 ***
## sexfemale                    -0.0123389  0.0021108  -5.846 5.05e-09 ***
## treatment Neighbors:sexfemale -0.0008487  0.0051730  -0.164     0.87
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.4615 on 229440 degrees of freedom
## Multiple R-squared:  0.00447,    Adjusted R-squared:  0.004457
## F-statistic: 343.4 on 3 and 229440 DF,  p-value: < 2.2e-16
```

This is really interesting data. Here we have figures for the Intercept (Control, Male), "treatment Neighbors" (Treatment, Male), "sexfemale" (Control, Female), and "treatment Neighbors:sexfemale" (Treatment, Female). The average treatment effect on men is about +8.17% – the first coefficient. The average treatment effect for women is a bit different – it's the difference between the last two coefficients, which is about +1.15%. So it seems at least superficially that the treatment has a greater effect on men than on women, at least according to this regression.

4d) A friend who hasn't taken PS6800 tells you that Gerber, Green, and Larimer should have controlled for household fixed effects. Is this a good idea? Why or why not? (5 points)

We learned in lecture that whether or not we use fixed effects is in part a question of the type of variation we want to explain. Here, using household fixed effects, we could be controlling for all of the characteristics common to each household. So this would account for all of the variables at the household level that could affect individuals' voting habits. For a lot of the households, there are multiple observations (just looking at the hh_id column in the dataset), so it is possible to use fixed effects here.

I would say that it wouldn't be a bad idea to try it; we may be able to address some of the omitted variable bias within the regression. That being said, I don't feel like I know enough about the results to say whether or not it is a good idea. Is there really such wide variation between households that it is useful to use fixed effects? Do we really care about within-household variation? I think that these are questions that the authors would have likely thought of themselves, so I'm skeptical to say that they should automatically include it just because they can. But I do think that it doesn't hurt to try.

4e) In 2023, it comes out that an error in the 2006 voter files meant that anyone born on the 30th of the month (any month) was reported as not having voted, even if they did in fact vote. Should Gerber, Green, and Larimer be concerned about the validity of their results? Why or why not? (5 points)

This omission will statistically affect somewhere around 11/365 (February doesn't have a 30th day) of the data, so about 3% of the data. This doesn't sound like much, but it's actually over 10,000 observations given the size of the dataset, so it definitely could mean something.

That being said, I would say that they should be only slightly concerned. The thing is, taking out 11 random days in the year is, well, random, so presumably the trends wouldn't change all that much. Unless we find some data that tells us that people born on the 30th are somehow way more or less likely to be receptive to a certain type of treatment, I would think that the omission would end up getting spread out relatively evenly, instead of affecting one area or another way stronger than another. So I would say that perhaps the numbers are a bit higher (30% instead of 29% for control, 33% instead of 32% for Hawthorne, etc.), the trends themselves should stay the same.