Delaney Gomen
STA4211: *Design of Experiments*
April 17, 2019

2018 Midterm Election Campaigns: Donation Analysis

Motivation and Background

The 2018 midterm elections witnessed a degree of voter participation not seen in decades. Not surprisingly, a documented $5.7 billion dollars was spent on this election, making it the most expensive midterm race in history. The upcoming 2020 presidential election is also projected to break the status-quo as more Americans become involved in politics. Just weeks into her presidential run, Senator Elizabeth Warren (D - MA) has already announced that she will forgo traditional campaign methods such as "fancy receptions or big money fund-raisers only with people who can write the big checks." Large donors are problematic because of fear of "buying" the candidates' legislative action. Voters fear that candidates that accept large donations will work in the best interest of them instead of the majortiy of the population. Keeping politicians financially accountable is now possible with open access to datasets that the Federal Election Commission (FEC) releases which will be analyzed in this report.

Aims

1. Analyze the relationship between reported profession of individual donors and political party of the candidate they donated to
2. Investigate for any difference in average donation sizes of 2020 presidential hopefulls that have advertised their commitment to "grassroots" movements.

Data

The Federal Election Commission hosts a website with campaign finances. For each election cycle, there are datasets for candidate profiles, committee donations, individual donations, and committee profiles. We will focus on donations. Data from election cycles from 2016 to 2020 (and a link to archives from 1980 - 2012) can be found on https://classic.fec.gov/finance/disclosure/ftpdet.shtml#a2017_2018.

Data Collection

Each dataset is in a convenient zip file. They are all .csv files that are cleaned and ready for analysis. There are 5 datasets that are either connected by candidate ID or committee ID. Individual donations are identified by the committee they donated to and each candidate has one (or more) official committees associated with them. By joining the individual donations with the table of candidates on committee ID, I simultaneously selected all donations to official Senate or House campaigns and dropped other donations to PAC's or similar from the dataset. I used SQL to obtain this table, labeled `donations`, which is all 3921383 entries of recorded donation data the FEC has reported for the 2018 midterms. I am taking random samples for each of our aims to compensate for any discrepancies in the data collection process, on my side or the FEC's. Missing records, human error, or incorrect reporting could lead to this data not being representative of the population of donations. To avoid incorrect assumptions, I am doing statistical analysis on random samples from these data to prevent making claims on population parameters and make my conclusions more conservative. Additionally, our factor levels and factor level combinations of interest have varying n for entries in the dataset. Therefore, smaller, uniform samples will allow statistical procedures like Scheffé to be carried out more easily when it comes to dealing with groups.

Data Analysis Plan

This data is observational. For aim (1), analysis of factor effects is most appropriate. Our R output will give us the significance of factor level effects and interactions. For additional analysis, we will perform a Scheffé procedure for multiple contrasts. For aim (2), a one-way ANOVA will suffice to compare average donation sizes, contigent on if significant differences are found. If so, we can perform Tukey analysis to see which means differ.

Aim 1

Overview. We have two factors comparing how they relate to donation size. Factor α represents profession. We picked three professions: professor, lawyer, physician. Factor β represents party. We picked the Republican and Democratic party. Our corresponding columns are represented in a table called `law_prof_md` as:

- `cand_party`: candidate party (`DEM` : Democrat ; `REP`: Republican)
- `occupation`: occupation (`professor`, `lawyer`, `physician`)
- `tran_amount`: donation amount in dollars (integer)

Approach. After loading in all the data, I used filtering techniques to obtain all the donations from all 6 factor level combinations. To avoid discrepancy in the input profession, I used `str_detect` to find the donors who put one of these three professions as their *only* profession. The regular expression for each profession follows: `'^PROFESSOR$'`, `'^LAWYER$|^ATTORNEY$'`, `'^DOCTOR$|^PHYSICIAN$|^MD$'`. After that, I randomly sampled from each factor level combination 1000 entries for a total of 6000 entries to keep factor level combination n_i's equal. I chose 1000 because it is a very large sample size while still being viable to use computationally. There were 409277 observations in total for lawyers, professors, and physicians that I sampled from.

Factor Effects Hypotheses. Define our parameters and state our hypotheses.

- $H_{1,0} : \alpha_1 = \alpha_2 = \alpha_3 = 0; H_{1,a}$: not all α_i equal 0
- $H_{2,o} : \beta_1 = \beta_2 = 0; H_{2,a}$: not all β_i equal 0
- $H_{3,o}$: all $(\alpha\beta)_{ij} = 0; H_{3,a}$: not all $(\alpha\beta)_{ij}$ equal 0 for $i = 1, 2, ..6$.

Where:

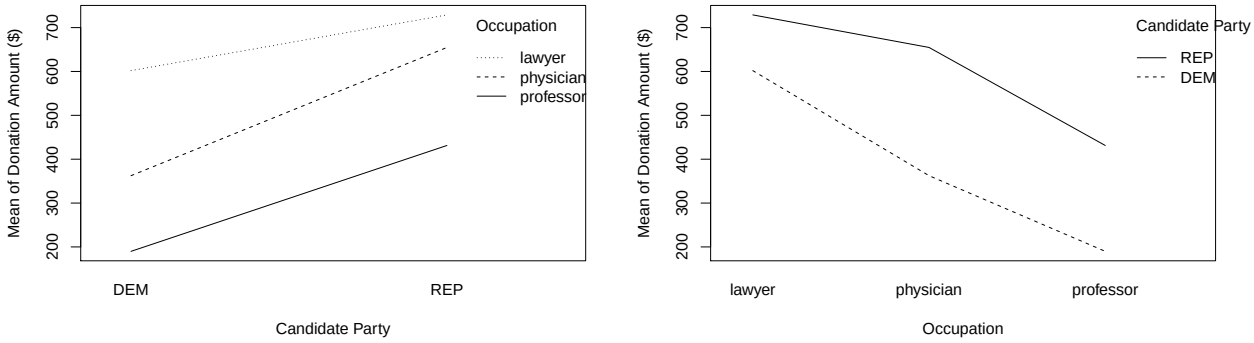
- α_i is the main effect for profession at the ith level ($i = 1, 2, 3$; 1 = lawyer, 2 = professor, 3 = physician).
- β_j is the main effect for political party at the jth level ($i = 1, 2$; 1 = Democrat, 2 = Republican)
- $(\alpha\beta)_{ij}$ is the interaction effect when profession is at the ith level and political party is at the jth level.

Analysis. Set random seed to 42.

```
law_prof_md[1:5,c(3,20,21)]
```

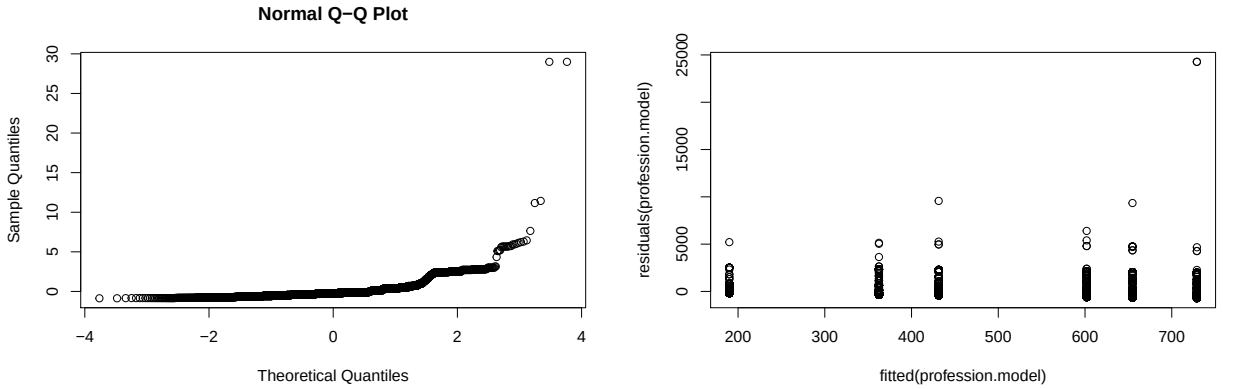
```
##   cand_party occupation tran_amount
## 2        IND     lawyer          50
## 3        IND  professor          25
## 5        IND  physician         100
## 6        IND     lawyer          33
## 7        IND  professor          27
```

Factor Effects Plot

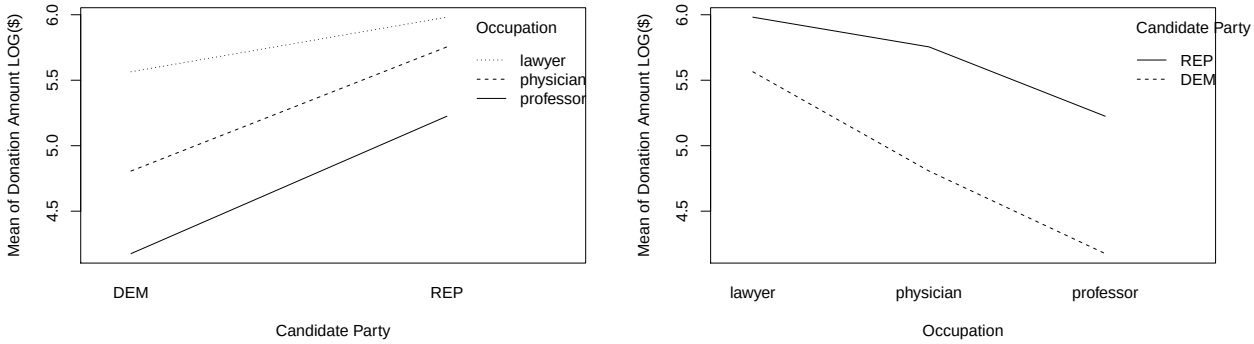



```
profession.model <- aov(tran_amount ~ cand_party*occupation, data = sample_occu)
```

```
qqnorm(rstandard(profession.model))
plot(fitted(profession.model), residuals(profession.model))
```



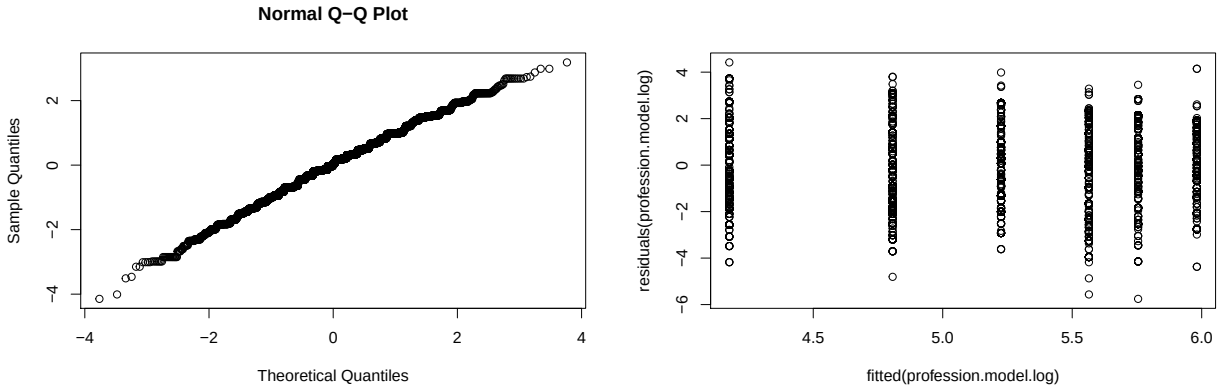
The interaction plot implies that there are main factor effects and interaction effects. The model does not follow our assumptions of normality and equal variance based on the QQNorm plot and residual plot. I observe the interaction effects as important because the slopes are clearly not the same, therefore, I will try a simple log transformation of the data and see if the interaction effects are still important.



Interaction effects still look important, and therefore are not multiplicative. So we will fit the model using a log transformation of base e and use the factor level combination means to examine factor effects jointly. Our model now becomes a log transformation of each paramater stated above for log base e.

Now we check to see if the model meets our assumptions of normality and equal variance.

```
profession.model.log <- aov(log(tran_amount) ~ cand_party*occupation, data = sample_occu)

qqnorm(rstandard(profession.model.log))
plot(fitted(profession.model.log), residuals(profession.model.log))
```



The model does, so we can continue with analysis.

```
anova(profession.model.log)
```

```
## Analysis of Variance Table
##
## Response: log(tran_amount)
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## cand_party                1   971.3  971.31 503.847 < 2.2e-16 ***
## occupation                2  1152.5  576.27 298.930 < 2.2e-16 ***
## cand_party:occupation     2   115.3   57.67  29.917 1.179e-13 ***
## Residuals              5994 11555.1    1.93
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe from the ANOVA summary that three relevant hypotheses tests are tested. They are testing for factor profession (α) and factor candidate party (β) main effects and testing for interactions ($\alpha\beta$).

From our ANOVA output, we can see that the null hypothesis was rejected for all of these hypotheses ($H_{1,0}; H_{2,o}; H_{3,o}$) with 95% confidence. We find that profession, candidate party, and the combination of the two affect donation size. Now to compare factor level effects jointly, I will use a function called `PostHocTest` from the `DescTools` package that will perform a Scheffé contrasts test on the contrasts between factor levels and factor level combinations of the form:

$\sum c_{ij}\mu_{ij}$ with constraint $\sum c_{ij} = 0$

```
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.5.2
```

```r
PostHocTest(profession.model.log, method = 'scheffe', ordered = TRUE, conf.level = 0.99)
```

```
##
##   Posthoc multiple comparisons of means : Scheffe Test
##     99% family-wise confidence level
##
## $cand_party
##              diff    lwr.ci    upr.ci   pval
## REP-DEM 0.8046978 0.6653842 0.9440113 <2e-16 ***
##
## $occupation
##                            diff     lwr.ci     upr.ci   pval
## physician-lawyer    -0.4924284 -0.6630520 -0.3218048 <2e-16 ***
## professor-lawyer    -1.0723774 -1.2430010 -0.9017539 <2e-16 ***
## professor-physician -0.5799490 -0.7505726 -0.4093255 <2e-16 ***
##
## $`cand_party:occupation`
##                                    diff      lwr.ci      upr.ci    pval
## REP:lawyer-DEM:lawyer         0.4169208  0.17562259  0.65821892 1.5e-08 ***
## DEM:physician-DEM:lawyer     -0.7579196 -0.99921777 -0.51662143 < 2e-16 ***
## REP:physician-DEM:lawyer      0.1899836 -0.05131461  0.43128173  0.0957 .
## DEM:professor-DEM:lawyer     -1.3885517 -1.62984991 -1.14725357 < 2e-16 ***
## REP:professor-DEM:lawyer     -0.3392824 -0.58058053 -0.09798419 1.6e-05 ***
## DEM:physician-REP:lawyer     -1.1748404 -1.41613852 -0.93354219 < 2e-16 ***
## REP:physician-REP:lawyer     -0.2269372 -0.46823536  0.01436097  0.0203 *
## DEM:professor-REP:lawyer     -1.8054725 -2.04677066 -1.56417433 < 2e-16 ***
## REP:professor-REP:lawyer     -0.7562031 -0.99750129 -0.51490495 < 2e-16 ***
## REP:physician-DEM:physician   0.9479032  0.70660499  1.18920133 < 2e-16 ***
## DEM:professor-DEM:physician -0.6306321 -0.87193031 -0.38933397 < 2e-16 ***
## REP:professor-DEM:physician   0.4186372  0.17733907  0.65993540 1.3e-08 ***
## DEM:professor-REP:physician -1.5785353 -1.81983347 -1.33723713 < 2e-16 ***
## REP:professor-REP:physician -0.5292659 -0.77056409 -0.28796776 3.5e-14 ***
## REP:professor-DEM:professor  1.0492694  0.80797121  1.29056754 < 2e-16 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the only non-significant difference is that between the log of donation size of physicians who donated to republican candidates and the log of donation size of lawyers who donated to democratic candidates.

In conclusion, we see that there are significant differences between the donation sizes between candidate party and donor profession. We can see in what direction the averages differed by the sign of the "diff" column. Factor level effects and interaction are significant in the log-transformed data.

Aim 2

Overview. We have one factor comparing how the levels relate to donation size which is candidate. There are 4 candidates that we have chosen based off of their recent bids for presidency in the 2020 presidential race. The columns for analysis are represented in a table called `prezhope`:

- `cand_name`: candidate name (`SANDERS, BERNARD`; `GILLIBRAND, KIRSTEN ELIZABETH`; `WARREN, ELIZABETH`; `O'ROURKE, ROBERT (BETO)`

- `tran_amount`: donation amount in dollars (integer)

Approach. After loading in all the data, I used filtering techniques to obtain all the donations from all 4 selected candidates. After that, I randomly sampled from each factor level combination 500 entries for a total of 2000 entries to keep factor level combination n_i's equal. I chose 500 because it is a very large sample size while still being computationally viable in R. There were 467103 observations in total that I sampled from.

Hypotheses Test.

$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4; H_a :$ not all μ_i equal 0

Where:

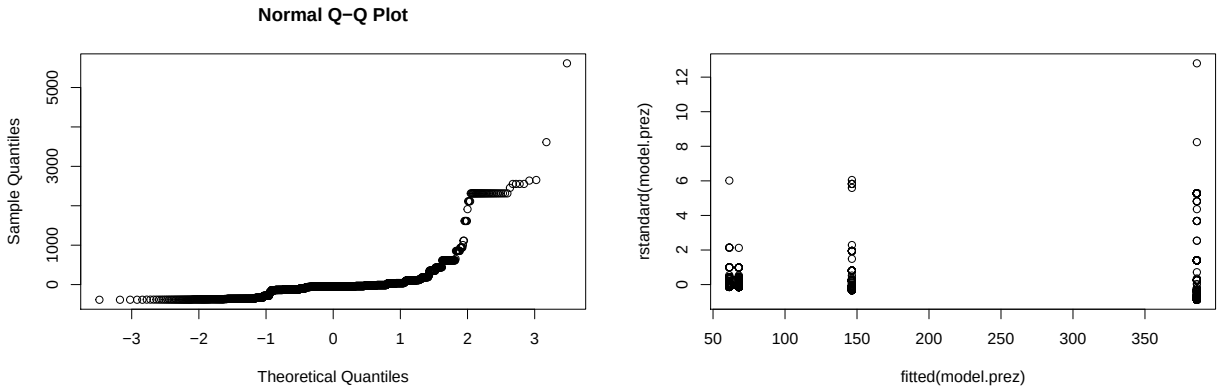
μ_i = donation mean for population of donations to the *ith* candidate ($i = 1, 2, 3, 4$; $1 =$ `SANDERS, BERNARD`; $2 =$ `GILLIBRAND, KIRSTEN ELIZABETH`; $3 =$ `WARREN, ELIZABETH`; $4 =$ `O'ROURKE, ROBERT (BETO)`)

Analysis.

```
prezhope[1:5,c(23,21)]
```

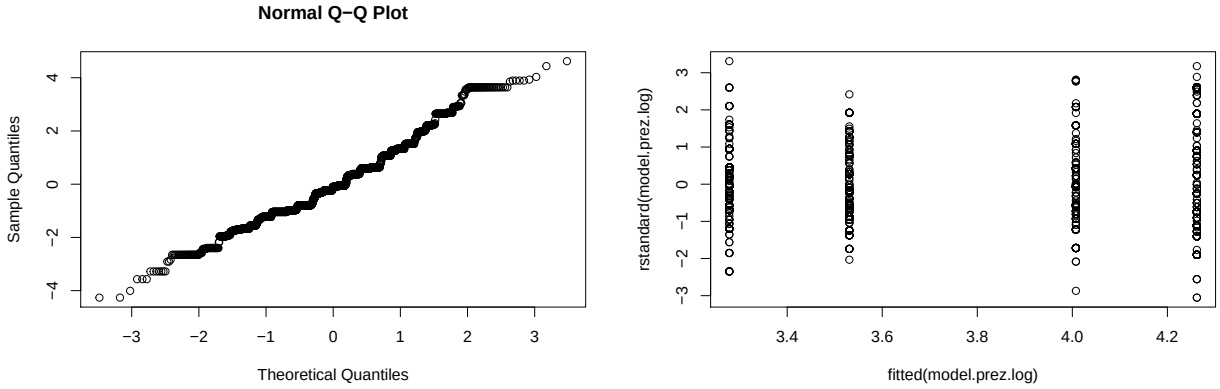
```
##          cand tran_amount
## 265728 WARREN          25
## 101976 WARREN          10
## 47322  WARREN          45
## 326245 WARREN          25
## 90312  WARREN          50
```

```
model.prez <- aov(tran_amount ~ cand, data=prezhope)

qqnorm(residuals(model.prez))
plot(fitted(model.prez), rstandard(model.prez))
```



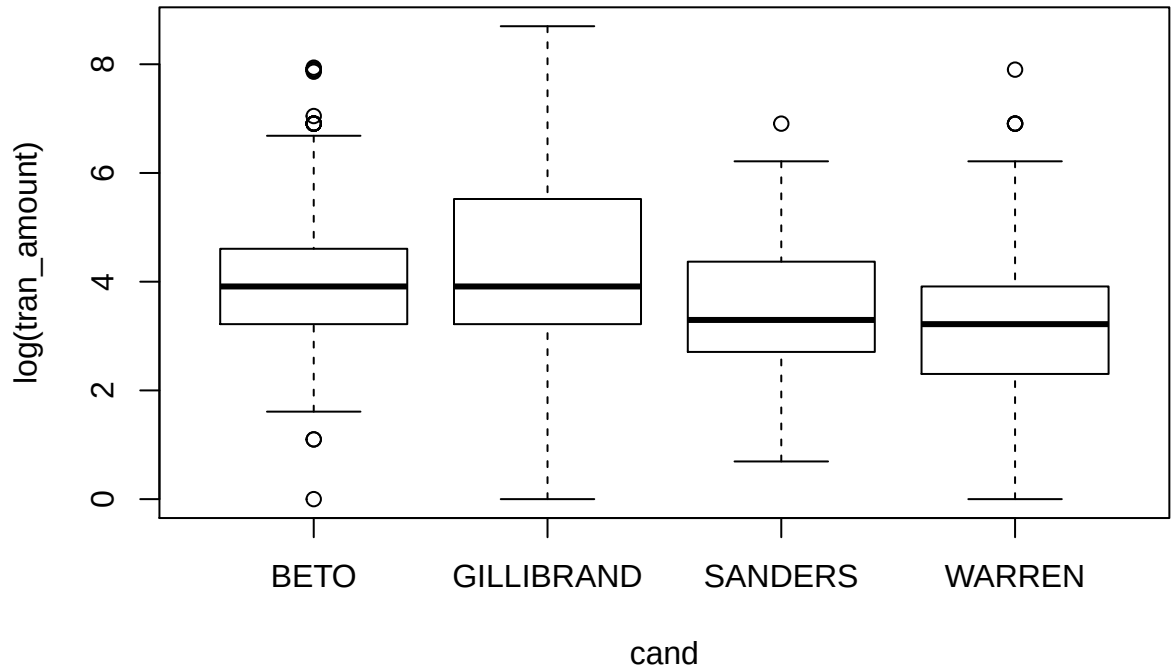
Again we see that the normality assumption does not hold so we perform a log transformation.

```
model.prez.log <- aov(log(tran_amount) ~ cand, data=prezhope)

qqnorm(residuals(model.prez.log))
plot(fitted(model.prez.log), rstandard(model.prez.log))
```



The model now meets our assumptions, so we can continue with analysis.

```
plot(log(tran_amount) ~ cand, data = prezhope)
```




```
summary(model.prez.log)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## cand           3    299   99.59   50.98 <2e-16 ***
## Residuals   1996   3899    1.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis with 95% confidence according to the R output. Now, we will continue analysis with examining the individual mean differences because we rejected the null that at least one pair of means is different. I use the `TukeyHSD` function to test this.

```
TukeyHSD(model.prez.log)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(tran_amount) ~ cand, data = prezhope)
##
## $cand
##                           diff         lwr         upr     p adj
## GILLIBRAND-BETO     0.2545159  0.02723989  0.48179182 0.0209787
## SANDERS-BETO       -0.4766902 -0.70396613 -0.24941420 0.0000005
## WARREN-BETO        -0.7292739 -0.95654985 -0.50199792 0.0000000
## SANDERS-GILLIBRAND -0.7312060 -0.95848198 -0.50393005 0.0000000
## WARREN-GILLIBRAND  -0.9837897 -1.21106570 -0.75651378 0.0000000
## WARREN-SANDERS     -0.2525837 -0.47985969 -0.02530776 0.0223884
```

In conclusion, there are significant differences between the log-transformed means of these four candidates. We can tell what candidates had a lower or higher average donation by the sign of the "diff" column.

Conclusions

In aim (1), I found that the factor levels of candidate party and profession both have significant factor level effects and that the interaction effects are also significant. In aim (2), I found that out of four chosen candidates, at least one pair of mean differences is statistically significant and the sign of the differences between them. These outcomes are important because:

- For those organizing or leading campaigns, knowing the challenges your party faces for gathering donations can have important effects on how one advertises or approaches fundraising.
- For voters, knowing where your candidate is getting the bulk of their donations can help make decisions. "Grassroots" movements pride themselves into serving the average American which can be reflected in average donation size. Candidates who receive larger donations may be supporting legislation not in the common voter's interest.

Future analyses would include finding estimates for the magnitude of differences between mean donations in non-logarithmic numbers. In this paper we established merely the evidence of differences among these factors. Non-transformed difference estimates would be more helpful for campaigners or voters to make more detailed decisions on. The non-transformed sample means will be shown for personal interpretation:

```
tapply(sample_occu$tran_amount,list(sample_occu$occupation,sample_occu$cand_party),mean)
```

```
##               DEM     REP
## lawyer    601.903 728.965
## physician 362.433 654.736
## professor 189.932 431.280
```

```r
tapply(prezhope$tran_amount, prezhope$cand_name, mean)
```

```
## GILLIBRAND, KIRSTEN ELIZABETH       O'ROURKE, ROBERT (BETO)
##                        386.06                        146.45
##              SANDERS, BERNARD             WARREN, ELIZABETH
##                         67.91                         61.31
```

References

Schouten, F. (2019, February 07). A record $5.7 billion was spent on the 2018 elections for Congress. Retrieved from https://www.cnn.com/2019/02/07/politics/midterm-election-costs-topped-5-7-billion/index.html

Herndon, A. W. (2019, February 25). Elizabeth Warren to Forgo Receptions and Fund-Raisers With Big Donors. Retrieved from https://www.nytimes.com/2019/02/25/us/politics/elizabeth-warren-donors-fundraising.html