Simple Analysis

Grant Esparza

September 6, 2018

Introduction

For this assignment I will be looking at records of political lobbyist activities in the city of Chicago. I obtained this data from the website Kaggle, which was orignally found on the City of Chicago's organization page. There were several datasets to choose from, however I decided to focus on the relationship between a lobbyist's compensation and how that affected the amount of money donated to political organizations. This information was found in two seperate files, `lobbyist_data-compensation.csv` and `lobbyist-data-contributions.csv`.

Data Preparation

```
## load libraries
library(ggplot2)
library(dplyr)
library(pander)
library(boot)

## Read data
comp_dat <- read.csv("lobbyist-data-compensation.csv")
contribute_dat <- read.csv("lobbyist-data-contributions.csv")

## Select variables of interest
comp_dat <- comp_dat %>% select(LOBBYIST_ID, COMPENSATION_AMOUNT)
contribute_dat <- contribute_dat %>% select(LOBBYIST_ID, AMOUNT)

## Join datasets
lobby_dat  <- inner_join(comp_dat, contribute_dat,
                         on = c("LOBBYIST_ID" = "LOBBYIST_ID"))

## Make lobbyist id a factor for grouping
lobby_dat[, "LOBBYIST_ID"] <- as.factor(lobby_dat[, "LOBBYIST_ID"])

## Clean variable names
colnames(lobby_dat) <- c("lobbyist.id", "comp.amount", "contrib.amount")
pander(head(lobby_dat, 4))
```

lobbyist.id	comp.amount	contrib.amount
8081	52500	1500
8081	52500	250
8081	52500	250
6039	2000	500

Fine tuning

After some manipulation with `dplyr` I now have a managable dataset with only the variables I care about. However, notice that there are multiple records for each lobbyist. To take care of that I'll use `group_by` and `summarise` to calculate the total values for each lobbyist.

```
# Group by lobbyist, calculate sum for comp and contriubtion
lobby_summary <- lobby_dat %>%
  group_by(lobbyist.id) %>%
  summarise(comp_sum = sum(comp.amount), contrib_sum = sum(contrib.amount))

## View new tibble
pander(head(lobby_summary))
```

lobbyist.id	comp_sum	contrib_sum
5505	169500	2500
5536	7500	1300
5684	1260000	42750
5703	439000	4000
5728	2800	600
5762	548500	28000

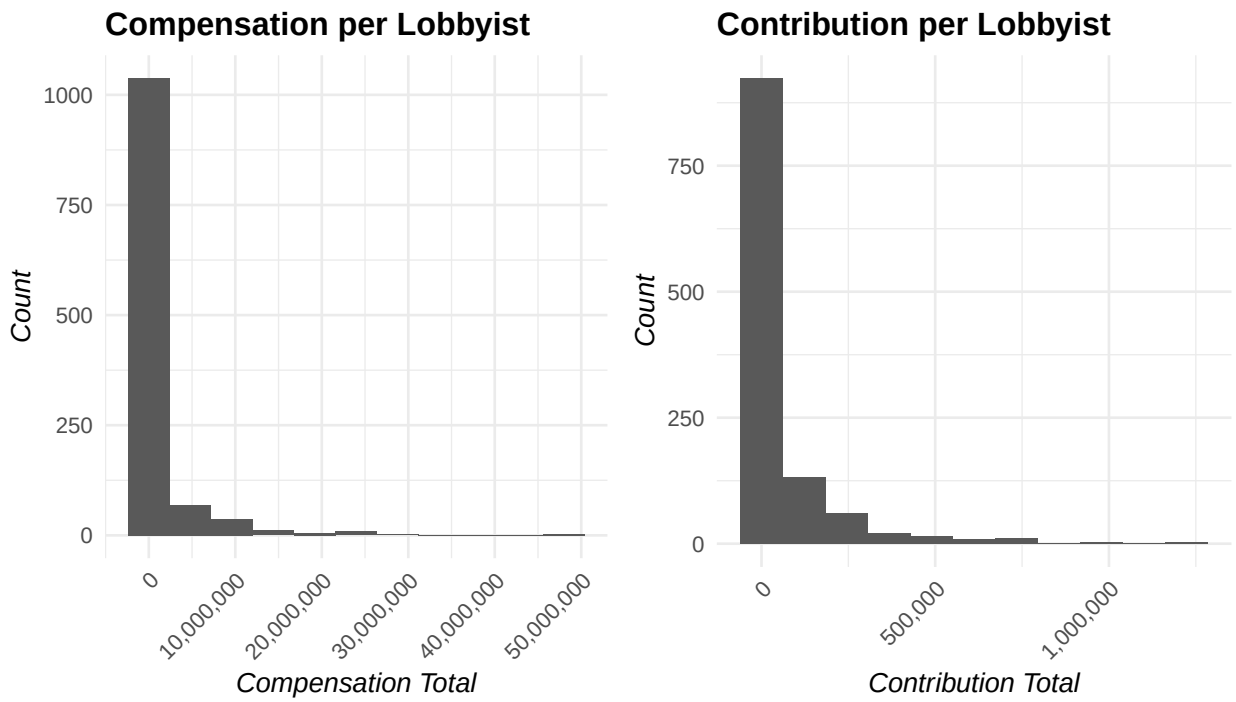
Visualization

```
require(gridExtra)

plot1 <- ggplot(lobby_summary, aes(comp_sum)) +
  geom_histogram(bins = 11) +
  ggtitle("Compensation per Lobbyist") +
  xlab("Compensation Total") + ylab("Count") +
  scale_x_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))

plot2 <- ggplot(lobby_summary, aes(contrib_sum)) +
  geom_histogram(bins = 11) +
  ggtitle("Contribution per Lobbyist") +
  xlab("Contribution Total") + ylab("Count") +
  scale_x_continuous(labels = scales::comma) +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))
```

```
grid.arrange(plot1, plot2, ncol=2)
```



As shown in the plots above, there is a heavy right skew on both variables. Interestingly, there seems to be a few incredibly well compensated lobbyists. The contributions do not reach such high numbers as it seems to cap around $1,500,000. The big takeaway from these plots is that there are a few incredibly expensive endeavours that companies deem worth the money.

Summary Statistics

```
## Summary statistics for lobbyist compensation
pander(summarise(lobby_summary, Mean = mean(comp_sum), Median = median(comp_sum),
          Std_dev = sd(comp_sum), IQR = IQR(comp_sum)), caption = "Compensation per lobbyist")
```

Table 3: Compensation per lobbyist

Mean	Median	Std_dev	IQR
1437068	110586	4397683	517688

Here we can see that the mean for lobbyist compensation is 1437068. The value is so large do to the heavy right skew distribution. Looking at the median 110586 is a much more modest number when compared to the mean. There is a very high standard deviation of 4397683 which indicates that there is a high variation among lobbyist comensation values. The interquartile range is the difference between Q3 and Q1 meaning

that the middle fifty percent of the data has a spread of 517688.

```
pander(summarise(lobby_summary, Mean = mean(contrib_sum),
                 Median = median(contrib_sum), Std_dev = sd(contrib_sum),
                 IQR = IQR(contrib_sum)), caption = "Contribution per lobbyist")
```

Table 4: Contribution per lobbyist

Mean	Median	Std_dev	IQR
61119	9050	138014	44500

The mean for lobbyist contribution is 61119. Similar to the compensation, The value is so large do to the heavy right skew. The median is 9050 which is not as extreme as the mean. The standard deviation of 138014 which indicates that there is a high variation among lobbyist contribution values. The interquartile range is the difference between Q3 and Q1 meaning that the middle fifty percent of the data has a spread of 44500.

Simple Linear Regression

```
lobby_mod <- lm(data=lobby_summary, contrib_sum ~ comp_sum)
pander(summary(lobby_mod))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27016	2772	9.746	1.235e-21
comp_sum	0.02373	0.0005994	39.59	1.899e-218

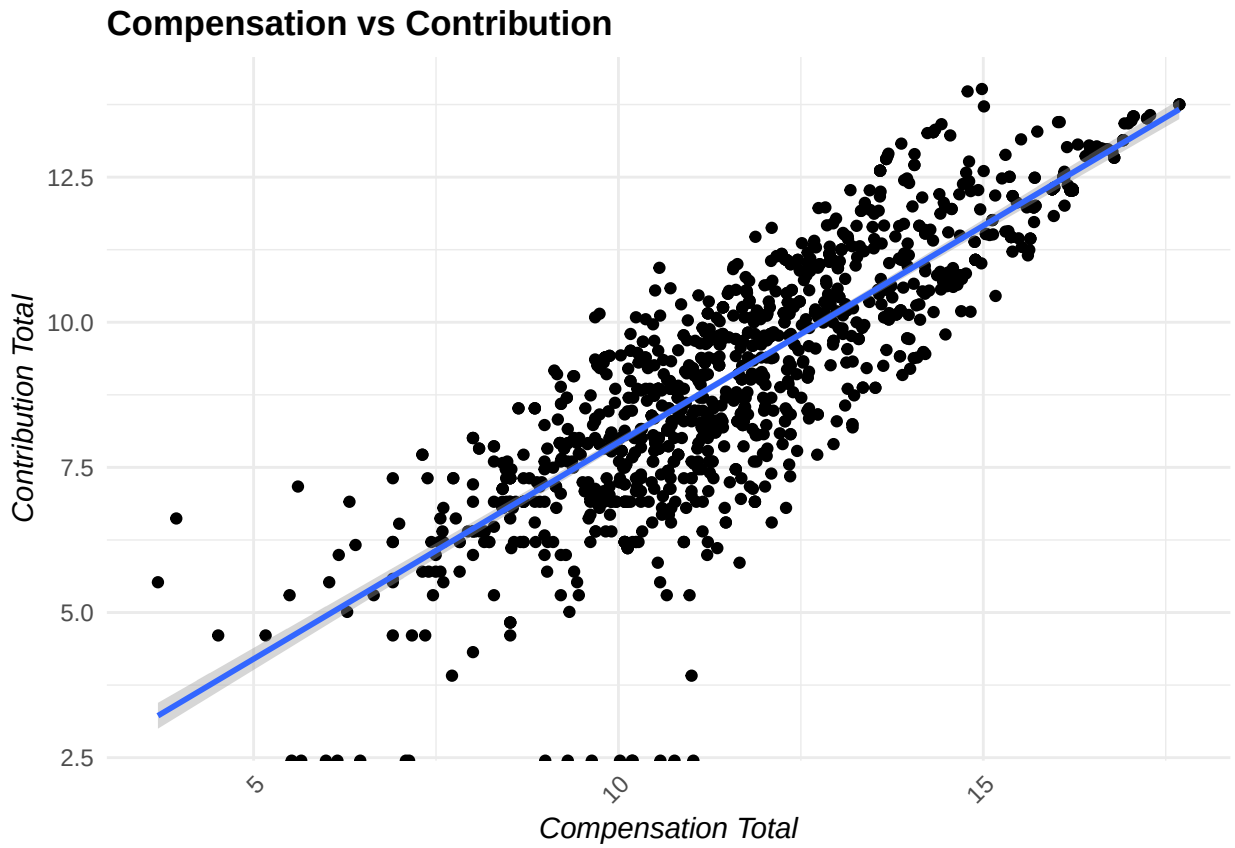
Table 6: Fitting linear model: contrib_sum ~ comp_sum

Observations	Residual Std. Error	R^2	Adjusted R^2
1176	90353	0.5718	0.5714

The model here shows that the total compensation paid to a lobbyist has is significant in explaining the contribution amount put towards a political organization. For every one dollar increase in total compensation, the total contribution amount increases by two percent. When the total compensation amount is zero, we can expect there to be a total political contribution of $27,016. With and R^2 of .57, we can say that 57% of the variation in the total contribution can be explained by the total compensation per lobbyist.

Bivariate Plot

```
## Plot linear regression
ggplot(lobby_summary, aes(log(comp_sum), log(contrib_sum))) +
  geom_point() +
  stat_smooth(method="lm") +
  ggtitle("Compensation vs Contribution") +
  xlab("Compensation Total") + ylab("Contribution Total") +
  theme_minimal() +
  theme(axis.title.x = element_text(face = "italic"),
        axis.title.y = element_text(face = "italic"),
        axis.text.x = element_text(angle=45, hjust=1),
        title = element_text(face = "bold"))
```



Bootstrap Confidence Interval

```
## CI for Mean
mn <- function(d, i) {
  d[i,] %>%
    summarise(mn = mean(comp_sum)) %>%
    .[['mn']]
}

m <- lobby_summary %>%
  na.omit %>%
  boot(., statistic = mn, R = 3000, parallel="multicore", ncpus=4)
bCI <- boot.ci(m, conf = 0.98, type = "bca")
bCI
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = m, conf = 0.98, type = "bca")
##
## Intervals :
## Level       BCa
## 98%   (1201771, 1807252 )
## Calculations and Intervals on Original Scale
```

We can be 98% confident that the mean total compensation amount per lobbyist is between $1,170,669 and $1,783,100.

```
## CI for Median
med <- function(d, i) {
  d[i,] %>%
    summarise(med = median(comp_sum)) %>%
    .[['med']]
}

m <- lobby_summary %>%
  na.omit %>%
  boot(., statistic = med, R = 3000, parallel="multicore", ncpus=4)
bCI <- boot.ci(m, conf = 0.98, type = "bca")
bCI
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = m, conf = 0.98, type = "bca")
##
## Intervals :
## Level       BCa
## 98%   ( 87600, 127599 )
## Calculations and Intervals on Original Scale
```

We can be 98% confident that the median total compensation amount per lobbyist is between $87,771 and $125,993.

```
## CI for Standard Deviation
std_dev <- function(d, i) {
  d[i,] %>%
    summarise(std_dev = sd(comp_sum)) %>%
    .[['std_dev']]
}

m <- lobby_summary %>%
  na.omit %>%
  boot(., statistic = std_dev, R = 3000, parallel="multicore", ncpus=4)
bCI <- boot.ci(m, conf = 0.98, type = "bca")
bCI

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = m, conf = 0.98, type = "bca")
##
## Intervals :
## Level       BCa
## 98%   (3588887, 5586466 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

We can be 98% confident that the standard deviation for total compensation amount per lobbyist is between $3,590,507 and $5,603,371.

```
iqr <- function(d, i) {
  d[i,] %>%
    summarise(iqr = IQR(comp_sum)) %>%
    .[['iqr']]
}

m <- lobby_summary %>%
  na.omit %>%
  boot(., statistic = iqr, R = 3000, parallel="multicore", ncpus=4)
bCI <- boot.ci(m, conf = 0.98, type = "bca")
bCI

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 3000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = m, conf = 0.98, type = "bca")
##
## Intervals :
## Level       BCa
## 98%   (399341, 754195 )
## Calculations and Intervals on Original Scale
```

We can be 98% confident that the IQR for total compensation amount per lobbyist is between $399,675 and $753,745.