HW6

Vivian Bui

10/3/2021

```r
justice_df <- read.csv('https://raw.githubusercontent.com/dpuelz/Policy-Research-Laboratory/main/data/ju
momwage_df <- read.csv('https://raw.githubusercontent.com/dpuelz/Policy-Research-Laboratory/main/data/yu
```

PROBLEM 1:

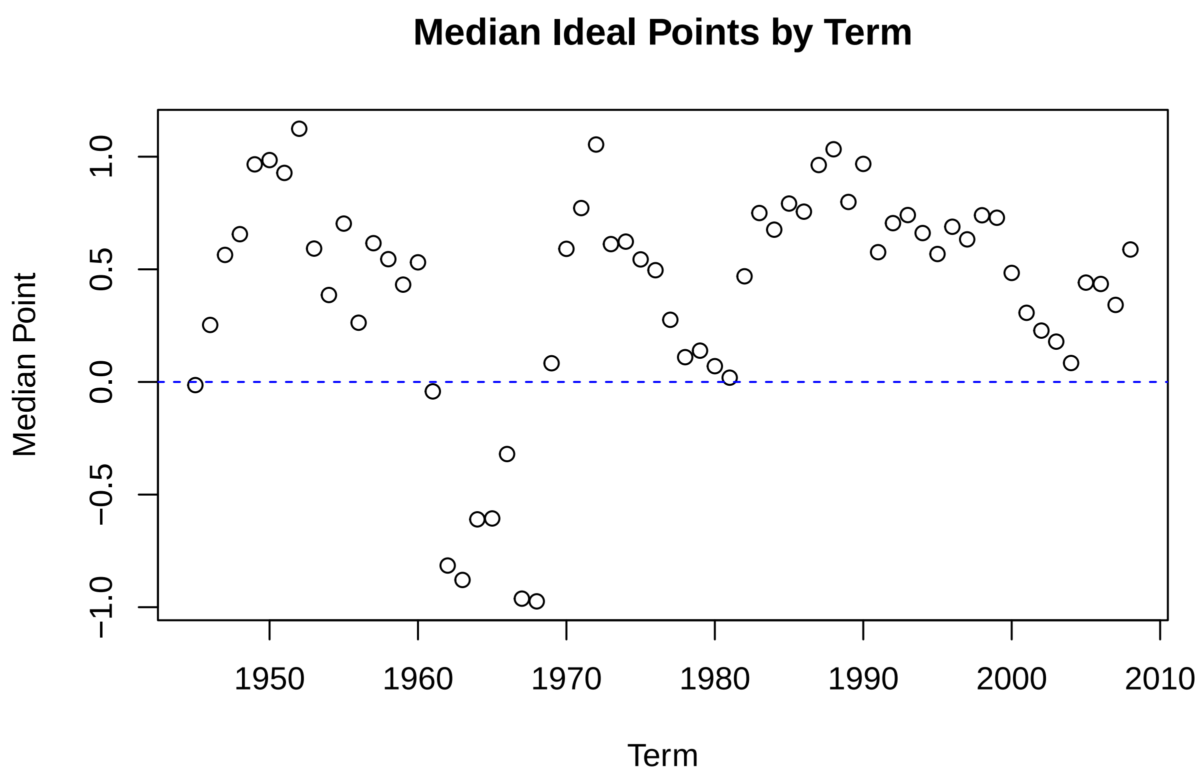
a. We wish to know the median ideal point for the Court during each term included in the dataset. First, calculate the median ideal point for each term of the Court. Next, generate a plot with term on the horizontal axis and ideal point on the vertical axis. Include a dashed horizontal line at zero to indicate a "neutral" ideal point. Be sure to include informative axis labels and a plot title. Ans:

```r
#Find list of terms
terms <- unique(justice_df$term)

#Initiate
median_point <- c()
loop=0

#Loop to find median ideal pts by term
for (term in terms)
  {
  loop=loop+1
  ideal_pt = justice_df$idealpt[justice_df$term == term]
  median_point[loop] <- median(ideal_pt)
  }

#Plot
plot(x=terms,y=median_point, main='Median Ideal Points by Term', xlab='Term', ylab='Median Point')
abline(h=0, col='blue', lty=2)
```



b. Next, we wish to identify the name of the justice with the median ideal point for each term. Which justice had the median ideal point in the most (potentially nonconsecutive) terms? How long did this justice serve on the Court overall? What was this justice's average ideal point over his/her entire tenure on the Court? Ans:

See below.

```
#Initiate
name <- c()
loop2 = 0

#Loop to find list of justice names
for (point in median_point)
  {
  loop2=loop2+1
  year = terms[loop2]
  name[loop2] <- justice_df$justice[justice_df$idealpt==point & justice_df$term==year]
  }

#Justice had the median ideal point in the most terms
name_freq <- as.data.frame(table(name))
name_max <- as.character(name_freq$name[name_freq$Freq == max(name_freq$Freq)])
print(name_max)
```

```
## [1] "White"
```

```r
#Serving time
serving_terms <- justice_df$term[justice_df$justice==name_max]
serving_year = length(serving_terms)
print(serving_year)
```

```
## [1] 32
```

```r
#Initiate
sum=0

#Find sum of ideal point of justice White throughout his terms
for (year in serving_terms)
  {
  justice_idealpt = subset(justice_df, justice_df$term == year & justice_df$justice == name_max, select
  sum=sum+justice_idealpt
  }

#Average ideal point of justice White throughout his terms
avg = sum/serving_year
print(avg)
```

```
##       idealpt
## 152 0.4401563
```

c. We now turn to the relationship between Supreme Court ideology and the president. Specifically, we want to see how the ideology of the Supreme Court changes over the course of each presi- dent's time in office. Begin by creating two empty "container" vectors: one to hold Democratic presidents, and another for Republican presidents. Label each vector with the presidents' names. Ans:

```r
#Create empty vectors
dem_president <- c()
rep_president <- c()

#Label dem vector
dem_president = as.factor(unique(justice_df$pres[justice_df$pparty=='D']))
dem_president =  factor(1:5, labels =  unique(justice_df$pres[justice_df$pparty=='D']))

#Label rep vector
rep_president = as.factor(unique(justice_df$pres[justice_df$pparty=='R']))
rep_president = factor(1:6, labels = unique(justice_df$pres[justice_df$pparty=='R']))
```

d. Next, for each Democratic president, calculate the shift in Supreme Court ideology by subtracting the Court's median ideal point in the president's first term from its median ideal point in the president's last term. Use a loop to store these values in your Democratic container vector. Repeat the same process for Republican presidents. Ans:

```r
#Function: calculate the shift point for each president in each party
calc_shift <- function(president_party)
  {
    for (president in president_party)
```

```
    {
    pres_term = unique(justice_df$term[justice_df$pres == president])
    first_term = pres_term[1]
    last_term = pres_term[length(pres_term)]
    medpt_first = median(justice_df$idealpt[justice_df$term == first_term])
    medpt_last = median(justice_df$idealpt[justice_df$term == last_term])
    return (pres_shift = medpt_first - medpt_last)
    }
  }

#Apply the function
dem_df <- as.data.frame(tapply(dem_president, dem_president, calc_shift))
rep_df <- as.data.frame(tapply(rep_president, rep_president, calc_shift))

#Set name for df columns
dem_df <- setNames(cbind(rownames(dem_df), dem_df, row.names = NULL), c("pres", "shift_pt"))
rep_df <- setNames(cbind(rownames(rep_df), rep_df, row.names = NULL), c("pres", "shift_pt"))

#Output
print(dem_df)
```

```
##       pres shift_pt
## 1  Truman   -1.138
## 2 Kennedy    0.837
## 3 Johnson    0.364
## 4  Carter    0.206
## 5 Clinton    0.257
```

```
print(rep_df)
```

```
##         pres    shift_pt
## 1 Eisenhower  0.06099999
## 2      Nixon -0.54000003
## 3       Ford  0.04800004
## 4     Reagan -1.01399999
## 5 BushSenior  0.09400004
## 6 BushJunior -0.28099999
```

e. What was the mean and standard deviation of the Supreme Court ideology shifts you just calculated when looking only at the Democratic presidencies? What about the Republican presidencies? Which Republican president's tenure had the largest conservative (positive) shift on the Court? Which Democratic president's tenure had the largest liberal (negative) shift? Ans:

See below.

```
#Mean and sd for D pres
mean(dem_df$shift_pt)
```

```
## [1] 0.1052
```

```r
sd(dem_df$shift_pt)
```

```
## [1] 0.7384542
```

```r
#Mean and sd for R pres
mean(rep_df$shift_pt)
```

```
## [1] -0.272
```

```r
sd(rep_df$shift_pt)
```

```
## [1] 0.4403894
```

```r
#Republican president with the largest shift
rep_df$pres[rep_df$shift_pt == max(rep_df$shift_pt)]
```

```
## [1] "BushSenior"
```

```r
#Democratic president with the largest shift
dem_df$pres[dem_df$shift_pt == min(dem_df$shift_pt)]
```

```
## [1] "Truman"
```

f. Create a plot that shows the median Supreme Court ideal point over time. Then, add lines for the ideal points of each unique justice to the same plot. The color of each line should be red if the justice was appointed by a Republican and blue if he or she was appointed by a Democrat. (You can assume that when a Justice first appears in the data, they were appointed by the president sitting during that term.) Label each line with the justice's last name. Briefly comment on the resulting plot. Ans:

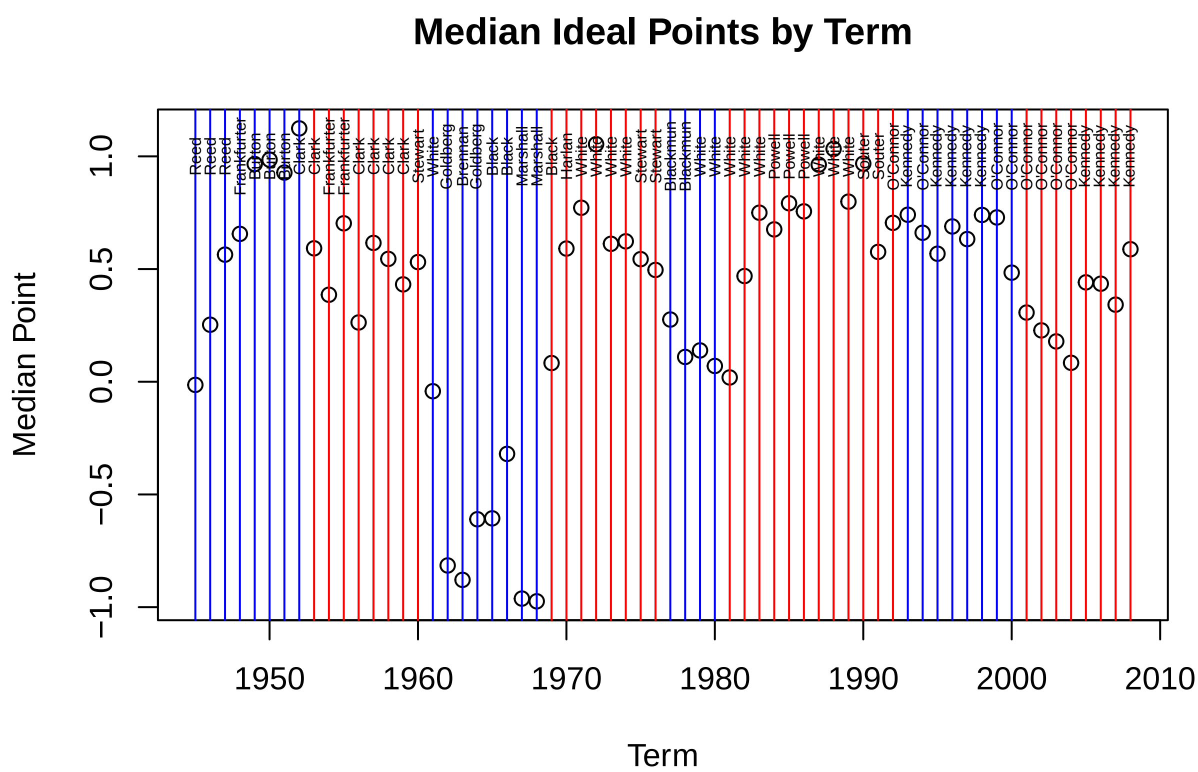
Turns out that the ideal point does not necessarily means a judge's preference towards one specific party. A high, and positive median pt can still affiliate with Democrat rather than what we initially assumed as Republican. Nonetheless, the highest positive median pt instead tied to the Democrats. One thing that we are more confident to claim is that the majority of median pts which are negative will be more likely pointed towards a preference for the blue party. Other than that, the median pts in general is not a substantial and adequate indication of party preference.

```r
#Initiate
loop3=0

#Plot the Supreme Court median ideal point over time
plot(x=terms,y=median_point, main='Median Ideal Points by Term', xlab='Term', ylab='Median Point')

#Draw line for each justice
for (i in 1:length(terms)) {
  loop3=loop3+1
  x = terms[loop3]
  jt_name = name[loop3]
  jt_party = justice_df$pparty[justice_df$term == x & justice_df$justice == jt_name]
  if (jt_party == 'D')
```

```
    {
    abline(v=x, col='Blue')
    text(x=x, y=1, srt=90, cex = 0.5, labels = jt_name)
    }
  else
  { abline(v=x, col='Red')
      text(x=x, y=1, srt=90, cex = 0.5, labels = jt_name)}}
```



PROBLEM 2:

a. How many different women are in the data? How many observations per year? We will refer to each row as a "person-year observation" since the row contains data on a given person in a particular year. In a few sentences, describe one advantage and one disadvantage of using a contemporary cohort of women rather than an older cohort in estimating the predictors of the mother wage gap. Ans:

Advantage: At the start of the contemporary cohort, we can be ascertain that the participants 'start fresh' (i.e. has not expose to any studying variables), thus, we can observe the effects of exposure and improve our analysis pf outcomes based on collected data of contemporary cohort.
Disadvantage: Certain exposures take time to develop observable outcomes, thus, using contemporary cohort rather than the older one will potentially increase the risk of losing valuable data. Additionally, there are occasions where we want to study the aggregated effect of exposures, which can only be observed on a substantially same cohort over time.

```r
#Different women in the data
length(unique(momwage_df$PUBID))
```

```
## [1] 1484
```

```r
#Observations per year
years <- unique(momwage_df$year)
for (year in years) {print(sum(momwage_df$year == year))}
```

```
## [1] 867
## [1] 895
## [1] 860
## [1] 877
## [1] 853
## [1] 274
## [1] 752
## [1] 890
## [1] 436
## [1] 846
## [1] 840
## [1] 837
## [1] 888
## [1] 584
## [1] 873
```

b. numChildren is the variable representing how many children the woman had at the time of an observation. Please provide a table that shows the proportion of observations by the number of children. Provide a brief substantive interpretation of the results. Ans:

More than 60% of female participants do not have children, and 19% to 12% of participants have one to two children. Less than 0.3% of women in the study have 6 children, and 6 is the maximum number of children the women in this study have.
This, however, indicates the possibility for sample selection bias that potentially affects our analysis for the mom wage gap.

```r
prop.table(table(momwage_df$numChildren))
```

```
##
##            0            1            2            3            4            5
## 0.6244382993 0.1972865538 0.1214137573 0.0440718977 0.0114932596 0.0009505703
##            6
## 0.0003456619
```

c. Create a new indicator variable isMother that takes a value of 1 if the woman has at least one child in that year and a value of 0 otherwise. Tabulate the new variable. Briefly comment on the results. Ans:

In the studying sample, only 4346 or less than 38% of the studying women are mothers. As previously mentioned, this may affect our analysis for the mom wage gap.

```r
momwage_df$isMother <- ifelse(momwage_df$numChildren != 0, 1, 0)
tabulate(momwage_df$isMother)
```

```
## [1] 4346
```

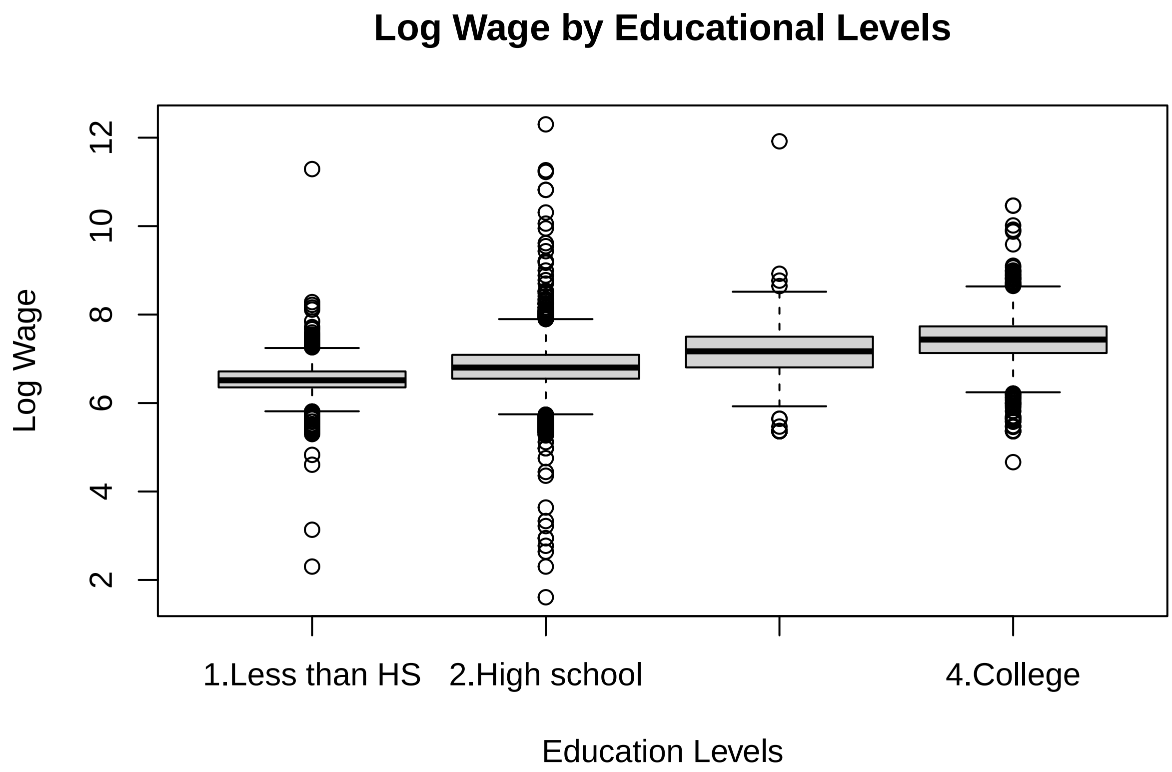
d. Create a new variable called logwage that is the log of wage. Make two boxplots, one for wage and the other for logwage, as a function of educational level (educ). Compare the two boxplots and discuss the purpose of the log transformation. Ans:

The log transformation helps remove skewness that greatly enhance our visibility and analysis on the plots.

```r
#Creat logwage
momwage_df$logwage <- log(momwage_df$wage)

#Boxplots
boxplot(wage~educ, data=momwage_df,
        main = 'Wage by Educational Levels',
        xlab = 'Education Levels',
        ylab = 'Wage')
```


```r
boxplot(logwage~educ, data=momwage_df,
        main = 'Log Wage by Educational Levels',
        xlab = 'Education Levels',
        ylab = 'Log Wage')
```



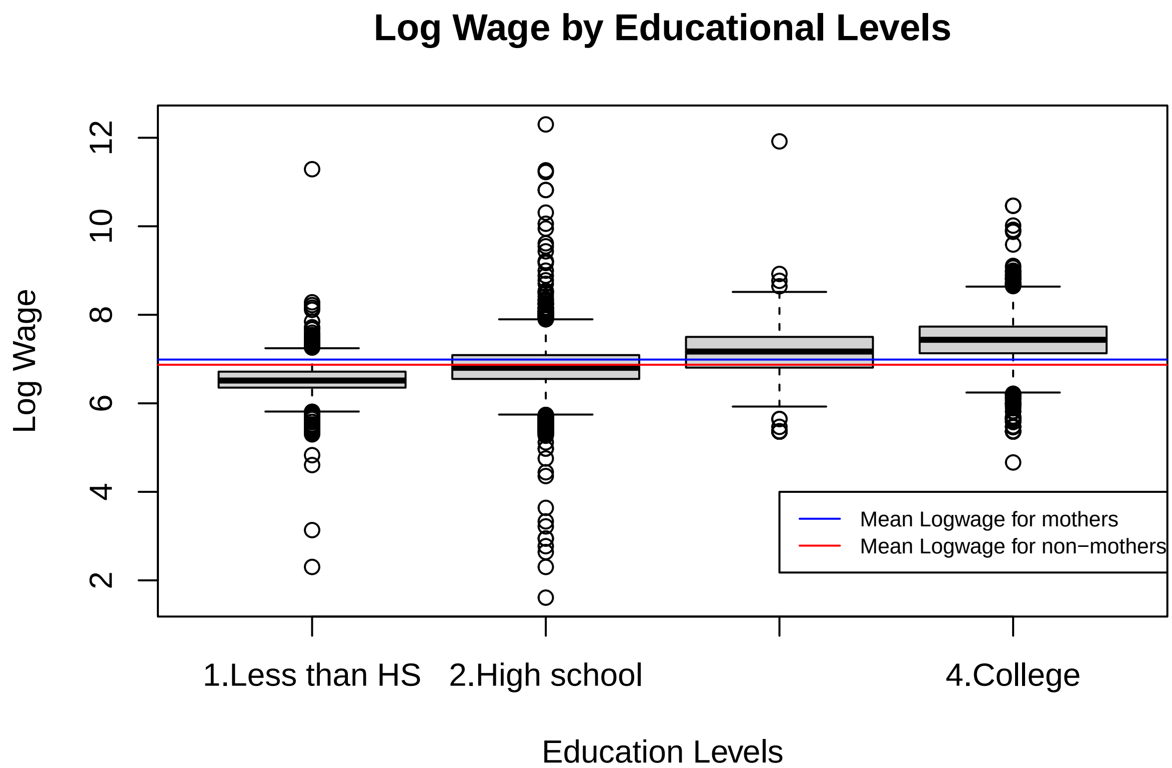
e. In the same graph, plot the mean logwage against year for mothers, then for non-mothers in a different color or line type. Include a legend and a proper title. Make sure the figure and axes are clearly labeled. Give a brief interpretation of the results. Ans:

Overall, regardless of having children or not, women will make more money if having some college or higher. Though not significant, calculated results show that the mean wage for mothers is higher than the mean wage for non-mothers. However, noted that our sample has more than 60% are non-mothers, this difference between the two mean wage may not be applicable in the population.

```
#Mean logwage for mothers
logwage_isMother = momwage_df$logwage[momwage_df$isMother==1]
mean_logwage_isMother = mean(logwage_isMother)

#Mean logwage for non-mothers
logwage_nonMother = momwage_df$logwage[momwage_df$isMother==0]
mean_logwage_nonMother = mean(logwage_nonMother)

#Plot
boxplot(logwage~educ, data=momwage_df,
        main = 'Log Wage by Educational Levels',
        xlab = 'Education Levels',
        ylab = 'Log Wage')
abline(h=mean_logwage_isMother, col='blue')
abline(h=mean_logwage_nonMother, col='red')
legend(3, 4, legend=c("Mean Logwage for mothers", "Mean Logwage for non-mothers"),
       col=c("blue", "red"), lty=1:1, cex=0.7)
```



f. Run a regression using fixed effects for both woman and year. You should be sure to include variables for number of children (numChildren) and job characteristics (fullTime, firmSize, multipleLocations, unionized, industry). Note: that you should not use the isMother variable you created earlier in this model. Report the coefficient of numChildren. Provide a brief substantive interpretation of this coefficient and the coefficients for any two other variables. (Hint: fixed effects means including the relevant factor variables in the regression model – see the bolded statement in the problem introduction). Ans:

The coefficient of numChildren is 303.726. This means that, on average per woman, having an additional child gives a women 303.726 positive cashflow in wage (child support, perhaps?).
Other vars such as full-time status, company's geography, union labor, etc., when being TRUE, will be more likely to have a positive effect on the women's wage. The effect of industries on wage, however, varies depend on the type of industries.

```
install.packages("plm", repos = "http://cran.us.r-project.org")
```

```
##
##   There is a binary version available but the source version is later:
##     binary source needs_compilation
## plm  2.4-2  2.4-3             FALSE
```

```
## installing the source package 'plm'
```

```r
library(plm)

plm(wage ~ numChildren
    + fullTime
    + multipleLocations
    + unionized
    + industry, data=momwage_df,
    index=c("PUBID", "year"),
    model="within")
```

```
##
## Model Formula: wage ~ numChildren + fullTime + multipleLocations + unionized +
##     industry
##
## Coefficients:
##                                                                                         numChildren
##                                                                                             303.726
##                                                                                        fullTimeTRUE
##                                                                                             314.945
##                                                                                   multipleLocations
##                                                                                             108.521
##                                                                                           unionized
##                                                                                             216.710
##                     industryArts, Entertainment, Recreation, Accommodations, and Food Services
##                                                                                            -171.469
##                                                                               industryConstruction
##                                                                                             -82.727
##                                                                 industryEducational and Social Services
##                                                                                             175.870
##                                industryFinance, Insurance, Real Estate, and Rental and Leasing
##                                                                                             294.873
##                                                                                 industryHealth Care
##                                                                                             501.012
##                                                              industryInformation and Communications
##                                                                                             183.219
##                                                               industryManufacturing - Durable Goods
##                                                                                             300.856
##                                                            industryManufacturing - Nondurable Goods
##                                                                                             478.508
##                                                                                      industryMining
##                                                                                             759.238
##                                                                                       industryOther
##                                                                                             616.286
##                                                                              industryOther Services
##                                                                                             742.909
## industryProfessional, Scientific, Management, Administrative, and Waste Management Services
##                                                                                             358.657
##                                                         industryPublic Administration and Armed Forces
##                                                                                             474.544
##                                                                                industryRetail Trade
##                                                                                             -11.876
##                                                              industryTransportation and Warehousing
##                                                                                             219.053
```

```
##                                                     industryUtilities
##                                                               463.133
##                                             industryWholesale Trade
##                                                               190.806
```