STATS 506 HW 3

Anuraag Ramesh

November 30, 2022

Contents

Q1.

1.

Configuration for Git is set on three levels System, Global and Local. The system that I am using is a Mac.
- The global git configuration file is in my Users folder on the Macintosh HD
- The local git configuration file is in the .git folder in the folder hw3 with .R and .Rmd file.

```
writeLines(readLines("/Users/anuraagramesh/.gitconfig", 10))
```

```
[filter "lfs"]
    clean = git-lfs clean -- %f
    smudge = git-lfs smudge -- %f
    process = git-lfs filter-process
    required = true
[user]
    name = Anuraag Ramesh
    email = anuraagr@umich.edu
```

```
writeLines(readLines(".gitignore", 10))
```

```
.Rhistory
.gitignore
.DS_Store
```

2.

```
df_func <- function(n1, n2){
  headers = read.table('Data/2020_Business_Academic_QCQ.txt', sep = ',', nrows = 1,
                       encoding = 'UTF-8')
  df = read.table('Data/2020_Business_Academic_QCQ.txt', sep = ',', skip = n1,
```

```
                  nrows = n2 - n1, fileEncoding="latin1")
  colnames(df) = headers
  df = df %>% select('State', 'County Code',
                'Employee Size (5) - Location',
                'Sales Volume (9) - Location',
                'Census Tract')
  colnames(df) = gsub('-', '', colnames(df))
  colnames(df) = gsub('\\(\\d\\)', '', colnames(df))
  colnames(df) = gsub(' ', '', colnames(df))
  colnames(df) = tolower(colnames(df))
  df <- df[complete.cases(df), ]

  return(df)
}
```

3.

```
df = data.frame()
for (i in seq(1, 300001, by = 20000)){
  if(i == 1){
    n1 = i
  }
  else{
    n2 = i
    df1 = df_func(n1, n2)
    n1 = n2
    df = rbind(df, df1)
  }
}

print(head(df))
```

```
  state countycode employeesizelocation salesvolumelocation censustract
1    AL         73                    3                  98       10706
2    AL         73                    3                 165       12401
3    AL         73                    6                1793         800
4    AL         73                    3                 586        4500
5    AL         73                   15                 738       10804
6    AL         73                    2                 297       14404
```

```
agg = df %>% group_by(countycode, censustract) %>%
  summarise(employeesize = sum(employeesizelocation),
            salesvolume = sum(salesvolumelocation))
```

```
‘summarise()‘ has grouped output by ’countycode’. You can override using the
‘.groups‘ argument.
```

```
df1 = data.frame(agg)

df1$censustract = as.character(df1$censustract)
```

```
print(head(df1))
```

```
  countycode censustract employeesize salesvolume
1          1        2300            2         280
2          1       20100          214       34928
3          1       20200         1494       75970
4          1       20300          877       86037
5          1       20400          585       76889
6          1       20500         3740      404804
```

4.

Table creation in SQL:

CREATE TABLE df1 (countycode VARCHAR(255), censustract VARCHAR(255), employeesize INT, salesvolume INT)

```
host_name = 'localhost'
port_no = 3306
db_name = 'Hw3db'
u_id = 'root'

pwd = Sys.getenv('sqlpwd')

conn = RMySQL::dbConnect(RMySQL::MySQL(), host = host_name, port = port_no, user = u_id,
                         password = pwd, dbname = db_name)

RMySQL::dbWriteTable(conn, name = 'df1', df1, overwrite = TRUE, row.names = FALSE)
```

```
[1] TRUE
```

5.

```
res = dbGetQuery(conn, 'SELECT countycode, censustract, salesvolume FROM df1 ORDER
                BY salesvolume DESC LIMIT 10')
print(res)
```

```
   countycode censustract salesvolume
1          16         200     6860101
2          20        1900     5973563
3         103        5101     4777101
4         131         100     4646421
5          73        4500     4459034
6          20        2502     4351429
7          89         201     3854197
8          20        1000     3273042
9         119        4400     2949919
10        119        4800     2896424
```

6.

git commit -m “Question 5 Completed” git push origin main

git branch new
git checkout new

7.

```
df2 = read.csv('Data/AL.csv')
df2 = df2[c('FIELD19', 'FIELD20', 'FIELD22', 'FIELD45', 'FIELD64', 'FIELD65')]

colnames(df2) = c('householdwealth', 'income', 'home_value', 'state', 'countycode', 'censustract')
df2 = df2[df2$home_value != 0, ]
df2 = df2 %>% group_by(countycode, censustract) %>%
  summarise(wealth = mean(householdwealth),
            income = mean(income),
            homevalue = mean(home_value))
```

```
‘summarise()‘ has grouped output by ’countycode’. You can override using the
‘.groups‘ argument.
```

```
df2$censustract = as.character(df2$censustract)

print(head(df2))
```

```
# A tibble: 6 x 5
# Groups:   countycode [1]
  countycode censustract wealth income homevalue
       <int> <chr>        <dbl>  <dbl>     <dbl>
1          1 20100        1463.   53.6      157.
2          1 20200         990.   30.0      109.
3          1 20300        1383.   46.7      130.
4          1 20400        1560.   57        141.
5          1 20500        1663.   75.3      201.
6          1 20600        1756.   52.6      161.
```

8.

CREATE TABLE df2 (countycode VARCHAR(255), censustract VARCHAR(255), wealth FLOAT, income FLOAT, homevalue FLOAT)

```
RMySQL::dbWriteTable(conn, name = 'df2', df2, overwrite = TRUE, row.names = FALSE)
```

```
[1] TRUE
```

9.

Git Log:

commit d398fbb8fa29c92fc9e878a62b0932129cd3c963 (HEAD -> new, origin/new) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 22:55:19 2022 -0500

```
Question 8 Completed
```

commit 210c0f5c9c2219a076f8e70392aa9eb245843c7e (origin/main, main) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 21:51:26 2022 -0500

```
Q5 Completed
```

commit e9353aadc7e50f2f6c1bf456fce013fd6f9321b4 (origin/master) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 19:23:03 2022 -0500

```
Q5 Completed
```

HEAD means the current branch that is being checked out and points out the last commit.

10.

```
Sys.getenv("CENSUS_API_KEY")
```

```
[1] "c9c803849168d3116b2ba3cc7407cce3b36bae56"
```

```
census <- get_decennial(geography = 'tract', variables = c('H006001', 'H006002',
                                                           'H006003', 'H006004','H006005',
                                                           'H006006'),
                        year = 2010, state = '01')
```

```
Getting data from the 2010 decennial Census

Using Census Summary File 1
```

```
census = census %>% spread(variable, value)

census$whitepercent = census$H006002/census$H006001
census$blackpercent = census$H006003/census$H006001
census
```

```
census$countycode = as.numeric(substr(census$GEOID, 3, 5))
census$countycode = as.character(census$countycode)

census$tract = substr(census$GEOID, 6, 11)
census$tract = as.numeric(census$tract)
census$tract = as.character(census$tract)

head(census)
```

```
RMySQL::dbWriteTable(conn, name = 'census', census, overwrite = TRUE, row.names = FALSE)
```

```
[1] TRUE
```

11.

```
combine = dbGetQuery(conn, 'SELECT d2.censustract, d2.wealth, d2.income,
d2.homevalue, d1.employeesize, d1.salesvolume,
c.whitepercent, c.blackpercent FROM
df2 d2 JOIN df1 d1 ON d2.countycode = d1.countycode AND
d2.censustract = d1.censustract
JOIN census c ON c.countycode = d1.countycode AND
c.tract = d1.censustract')
head(combine)
```

Two variables that I would like to control the effects of are **average home size(sq. ft)** and **quality of construction rating**. Both these factors might affect average home valuations in a tract.

12.

Commands to do the following:

git commit -m "Question 11 Completed"
git log
git checkout main
git merge new

Git Log:

commit a5226df2424f4f91b254264b54fa1ca938c77a17 (HEAD -> new) Author: Anuraag Ramesh anuraagr@umich.edu Date: Sat Nov 26 23:49:33 2022 -0500

```
Question 11 Completed
```

commit d398fbb8fa29c92fc9e878a62b0932129cd3c963 (origin/new) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 22:55:19 2022 -0500

```
Question 8 Completed
```

commit 210c0f5c9c2219a076f8e70392aa9eb245843c7e (origin/main, main) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 21:51:26 2022 -0500

```
Q5 Completed
```

commit e9353aadc7e50f2f6c1bf456fce013fd6f9321b4 (origin/master) Author: Anuraag Ramesh anuraagr@umich.edu Date: Fri Nov 25 19:23:03 2022 -0500

```
Q5 Completed
```

The way to reset back to the old repository is to use: `git reset --hard HEAD^`

13.

```
#Correlation matrix
cor(combine[c('wealth', 'income', 'homevalue', 'employeesize',
              'salesvolume', 'blackpercent', 'whitepercent')], use = 'complete.obs')
```

```
                  wealth      income  homevalue employeesize salesvolume
wealth        1.000000000  0.92933366  0.8385221 -0.006344264  0.02607519
income        0.929333659  1.00000000  0.8669981  0.015612614  0.04740549
homevalue     0.838522110  0.86699808  1.0000000  0.135766394  0.14834484
employeesize -0.006344264  0.01561261  0.1357664  1.000000000  0.60530642
salesvolume   0.026075189  0.04740549  0.1483448  0.605306421  1.00000000
blackpercent -0.555901439 -0.51471906 -0.4237204 -0.003762786 -0.01040004
whitepercent  0.568537675  0.51955336  0.4201699 -0.014707577 -0.01047178
             blackpercent whitepercent
wealth       -0.555901439   0.56853767
income       -0.514719061   0.51955336
homevalue    -0.423720375   0.42016988
employeesize -0.003762786  -0.01470758
salesvolume  -0.010400044  -0.01047178
blackpercent  1.000000000  -0.99428734
whitepercent -0.994287335   1.00000000
```

```
#Model only having the percentage of white people in a tract
model_1 = lm(homevalue ~ whitepercent, data = combine)
summary(model_1)
```

```
Call:
lm(formula = homevalue ~ whitepercent, data = combine)

Residuals:
    Min      1Q  Median      3Q     Max
-132.49  -48.52  -15.32   19.67  809.36

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.710      5.927   9.231   <2e-16 ***
whitepercent  129.476      8.161  15.865   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.35 on 1174 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1765,    Adjusted R-squared:  0.1758
F-statistic: 251.7 on 1 and 1174 DF,  p-value: < 2.2e-16
```

```
#Model only having the percentage of black people in a tract
model_2 = lm(homevalue ~ blackpercent, data = combine)
summary(model_2)
```

```
Call:
lm(formula = homevalue ~ blackpercent, data = combine)

Residuals:
    Min      1Q  Median      3Q     Max
-132.11  -48.56  -14.06   21.04  813.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   179.210      3.395   52.79   <2e-16 ***
blackpercent -128.479      8.016  -16.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 82.2 on 1174 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1795,    Adjusted R-squared:  0.1788
F-statistic: 256.9 on 1 and 1174 DF,  p-value: < 2.2e-16
```

```
#First model
model_3 = lm(homevalue ~ wealth + employeesize + salesvolume + blackpercent,
             data = combine)
summary(model_3)
```

```
Call:
lm(formula = homevalue ~ wealth + employeesize + salesvolume +
    blackpercent, data = combine)

Residuals:
    Min      1Q  Median      3Q     Max
-240.54  -26.08   -5.30   16.04  487.08

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.733e+01  5.639e+00 -13.713  < 2e-16 ***
wealth        1.595e-01  3.354e-03  47.554  < 2e-16 ***
employeesize  2.647e-03  4.930e-04   5.368  9.6e-08 ***
salesvolume   1.568e-05  4.687e-06   3.345 0.000850 ***
blackpercent  1.883e+01  5.556e+00   3.388 0.000727 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.36 on 1171 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7284,    Adjusted R-squared:  0.7274
F-statistic: 784.9 on 4 and 1171 DF,  p-value: < 2.2e-16
```

```
#Variance Inlation Factors
vif(model_3)
```

```
      wealth employeesize  salesvolume blackpercent
    1.449519     1.579839     1.580767     1.447584
```

```r
#Standardizing the data
combine_standardized <-
  combine %>%
  mutate(wealth_s = scale(wealth),
         income_s = scale(income),
         employeesize_s = scale(employeesize),
         salesvolume_s = scale(salesvolume),
         whitepercent_s = scale(whitepercent),
         blackpercent_s = scale(blackpercent),
         homevalue_s = scale(homevalue))

model_4 = lm(homevalue ~ wealth_s + employeesize_s + salesvolume_s +
               blackpercent_s, data = combine_standardized)
summary(model_4)
```

```
Call:
lm(formula = homevalue ~ wealth_s + employeesize_s + salesvolume_s +
    blackpercent_s, data = combine_standardized)

Residuals:
    Min      1Q  Median      3Q     Max
-240.54  -26.08   -5.30   16.04  487.08

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     140.561      1.381 101.775  < 2e-16 ***
wealth_s         79.163      1.665  47.554  < 2e-16 ***
employeesize_s    9.318      1.736   5.368  9.6e-08 ***
salesvolume_s     5.808      1.737   3.345 0.000850 ***
blackpercent_s    5.632      1.662   3.388 0.000727 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 47.36 on 1171 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.7284,    Adjusted R-squared:  0.7274
F-statistic: 784.9 on 4 and 1171 DF,  p-value: < 2.2e-16
```

From the correlation matrix as well as single variable models(model_1 and model_2), we can clearly see that the percentage of white people in a tract is positively correlated with home valuation. On the other hand, the percentage of black people in tract is negatively correlated with home valuation.

We can see that income and wealth is highly correlated so we remove one of them while creating a model(model_3). We calculate the variance inflation factor(VIF) to check if the co-variates are highly correlated.

In the final model, we can see that whitepercent/blackpercent is significant in predicting home valuation. This helps us answer thw initial question if racial bias is inherent in home valuattion.
One confounding factor is that the blackpercent co-variate has a positive coefficient. This is conflicting with the correlation matrix and single variable models, which suggest a negative realtion. This can be justified by the fact that the coefficients for multi-variable regression cannot be taken at face value. The fact that blackpercent and wealth are correlated plays an essential part in changing the coefficients, a high blackpercent could lead to a lower average wealth in a tract.

Q2.

1.

A core is a computational unit that actually runs the code that we want on the HPC cluster. It can actually consist of multiple processors.

A node is the server/computer where the memory as well as the processors are installed.

A compute node is where the computation happens. The user specifies the job/application to be run on the compute node using a fixed set of resources. The compute node consists of processors, memory and GPU.

A login node is a place for users to interact with a cluster. Users can use a login node to view files, results as well as submit jobs.

2.

For an interactive session we use salloc instead of submitting a batch job.

```
salloc --nodes=1 --mem-per-cpu=32GB --cpus-per-task=4 --time=03:00:00
```

3.

Path to my scratch directory: `/scratch/stats506s001f22_class_root/stats506s001f22_class/anuraagr`

Creating a symbolic link from my scratch to home directory:

```
ln -s /scratch/stats506s001f22_class_root/stats506s001f22_class/anuraagr
```

Deleting a symbolic link does not affect the original directory, as it is a soft link.