BT2101 GA1 Group 67 Submission

Lim Zhen Yong A0236495U, Lim Shaun Lii A0236519Y, Jared Lau ZiLek A0236485W, Lo Zhi Hao A0236437B

2022-10-04

1 Introduction to R

Please use the discrim dataset from the Wooldridge package in R to answer this question. You can also read the given .xlsx file if you cannot load the data directly in R. Please carefully read the document of variable description.

This dataset contains zip code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

```
## Setting up the environment for further studies

## install.packages("wooldridge")
## install.packages("dplyr")
## install.packages("MASS")

library(wooldridge)
library(dplyr)
library(MASS)
library(knitr)
library(corrplot)

## documentation for MASS is at https://cran.r-project.org/web/packages/MASS/MASS.pdf

## Downloading the dataset

data('discrim')
## ?discrim
```

```
summary(discrim)
```

```
##      psoda           pfries         pentree          wagest
##  Min.   :0.730   Min.   :0.670   Min.   :0.490   Min.   :4.250
##  1st Qu.:0.980   1st Qu.:0.850   1st Qu.:0.950   1st Qu.:4.250
##  Median :1.060   Median :0.930   Median :1.020   Median :4.500
##  Mean   :1.045   Mean   :0.922   Mean   :1.322   Mean   :4.616
##  3rd Qu.:1.085   3rd Qu.:1.000   3rd Qu.:1.470   3rd Qu.:4.950
##  Max.   :1.490   Max.   :1.270   Max.   :3.950   Max.   :5.750
##  NA's   :8       NA's   :17      NA's   :12      NA's   :20
##      nmgrs           nregs          hrsopen           emp
##  Min.   : 1.00   Min.   :1.000   Min.   : 7.00   Min.   : 3.00
##  1st Qu.: 3.00   1st Qu.:3.000   1st Qu.:12.00   1st Qu.:11.38
##  Median : 3.00   Median :3.000   Median :15.50   Median :16.38
##  Mean   : 3.42   Mean   :3.608   Mean   :14.44   Mean   :17.62
##  3rd Qu.: 4.00   3rd Qu.:4.000   3rd Qu.:16.00   3rd Qu.:21.00
##  Max.   :10.00   Max.   :8.000   Max.   :24.00   Max.   :80.00
##  NA's   :6       NA's   :22                      NA's   :6
##      psoda2         pfries2          pentree2        wagest2
##  Min.   :0.410   Min.   :0.6900   Min.   :0.410   Min.   :4.250
##  1st Qu.:1.000   1st Qu.:0.8400   1st Qu.:0.940   1st Qu.:5.050
##  Median :1.050   Median :0.9400   Median :1.040   Median :5.050
##  Mean   :1.045   Mean   :0.9412   Mean   :1.354   Mean   :4.996
##  3rd Qu.:1.103   3rd Qu.:1.0100   3rd Qu.:2.053   3rd Qu.:5.050
##  Max.   :1.400   Max.   :1.3700   Max.   :2.850   Max.   :6.250
##  NA's   :22      NA's   :28       NA's   :24      NA's   :21
##      nmgrs2          nregs2         hrsopen2          emp2
##  Min.   :0.000   Min.   :1.000   Min.   : 8.00   Min.   : 0.00
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:12.00   1st Qu.:11.50
##  Median :3.000   Median :3.000   Median :15.00   Median :17.00
##  Mean   :3.484   Mean   :3.608   Mean   :14.47   Mean   :17.57
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:16.00   3rd Qu.:22.50
##  Max.   :8.000   Max.   :8.000   Max.   :24.00   Max.   :55.50
##  NA's   :6       NA's   :22      NA's   :11      NA's   :13
##     compown           chain          density          crmrte
##  Min.   :0.0000   Min.   :1.000   Min.   :  163   Min.   :0.00518
##  1st Qu.:0.0000   1st Qu.:1.000   1st Qu.: 1666   1st Qu.:0.02888
##  Median :0.0000   Median :2.000   Median : 2868   Median :0.04312
##  Mean   :0.3439   Mean   :2.117   Mean   : 4562   Mean   :0.05338
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.: 5660   3rd Qu.:0.06219
##  Max.   :1.0000   Max.   :4.000   Max.   :41437   Max.   :0.35971
```

```
##                                NA's   :1       NA's   :1
##      state          prpblck           prppov            prpncar
##  Min.   :1.000   Min.   :0.00000   Min.   :0.004298   Min.   :0.00000
##  1st Qu.:1.000   1st Qu.:0.01165   1st Qu.:0.029710   1st Qu.:0.04353
##  Median :1.000   Median :0.04144   Median :0.044441   Median :0.07389
##  Mean   :1.193   Mean   :0.11349   Mean   :0.071297   Mean   :0.11487
##  3rd Qu.:1.000   3rd Qu.:0.12106   3rd Qu.:0.082159   3rd Qu.:0.12348
##  Max.   :2.000   Max.   :0.98166   Max.   :0.418480   Max.   :0.62724
##                  NA's   :1         NA's   :1          NA's   :1
##      hseval          nstores          income           county
##  Min.   : 33900   Min.   :1.000   Min.   : 15919   Min.   : 1.00
##  1st Qu.:107900   1st Qu.:2.000   1st Qu.: 37883   1st Qu.: 6.00
##  Median :142300   Median :3.000   Median : 46272   Median :14.00
##  Mean   :147399   Mean   :3.139   Mean   : 47054   Mean   :13.66
##  3rd Qu.:176800   3rd Qu.:4.000   3rd Qu.: 54981   3rd Qu.:20.00
##  Max.   :473400   Max.   :8.000   Max.   :136529   Max.   :29.00
##  NA's   :1                        NA's   :1
##      lpsoda            lpfries           lhseval         lincome
##  Min.   :-0.31471   Min.   :-0.40048   Min.   :10.43   Min.   : 9.675
##  1st Qu.:-0.02020   1st Qu.:-0.16252   1st Qu.:11.59   1st Qu.:10.542
##  Median : 0.05827   Median :-0.07257   Median :11.87   Median :10.742
##  Mean   : 0.04032   Mean   :-0.08781   Mean   :11.83   Mean   :10.720
##  3rd Qu.: 0.08155   3rd Qu.: 0.00000   3rd Qu.:12.08   3rd Qu.:10.915
##  Max.   : 0.39878   Max.   : 0.23902   Max.   :13.07   Max.   :11.824
##  NA's   :8          NA's   :17         NA's   :1       NA's   :1
##     ldensity            NJ               BK              KFC
##  Min.   : 5.094   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 7.418   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 7.961   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   : 7.959   Mean   :0.8073   Mean   :0.4171   Mean   :0.1951
##  3rd Qu.: 8.641   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :10.632   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :1
##        RR
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2415
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
head(discrim)
```

```
##   psoda pfries pentree wagest nmgrs nregs hrsopen  emp psoda2 pfries2 pentree2
## 1  1.12   1.06    1.02   4.25     3     5    16.0 27.5   1.11    1.11     1.05
## 2  1.06   0.91    0.95   4.75     3     3    16.5 21.5   1.05    0.89     0.95
## 3  1.06   0.91    0.98   4.25     3     5    18.0 30.0   1.05    0.94     0.98
## 4  1.12   1.02    1.06   5.00     4     5    16.0 27.5   1.15    1.05     1.05
## 5  1.12     NA    0.49   5.00     3     3    16.0  5.0   1.04    1.01     0.58
## 6  1.06   0.95    1.01   4.25     4     4    15.0 17.5   1.05    0.94     1.00
##   wagest2 nmgrs2 nregs2 hrsopen2 emp2 compown chain density    crmrte state
## 1    5.05      5      5     15.0 27.0       1     3    4030 0.0528866     1
## 2    5.05      4      3     17.5 24.5       0     1    4030 0.0528866     1
## 3    5.05      4      5     17.5 25.0       0     1   11400 0.0360003     1
## 4    5.05      4      5     16.0   NA       0     3    8345 0.0484232     1
## 5    5.05      3      3     16.0 12.0       0     1     720 0.0615890     1
## 6    5.05      3      4     15.0 28.0       0     1    4424 0.0334823     1
##     prpblck    prppov   prpncar hseval nstores income county     lpsoda
## 1 0.1711542 0.0365789 0.0788428 148300       3  44534     18 0.11332869
## 2 0.1711542 0.0365789 0.0788428 148300       3  44534     18 0.05826885
## 3 0.0473602 0.0879072 0.2694298 169200       3  41164     12 0.05826885
## 4 0.0528394 0.0591227 0.1366903 171600       3  50366     10 0.11332869
## 5 0.0344800 0.0254145 0.0738020 249100       1  72287     10 0.11332869
## 6 0.0591327 0.0835001 0.1151341 148000       2  44515     18 0.05826885
##       lpfries  lhseval  lincome ldensity NJ BK KFC RR
## 1  0.05826885 11.90699 10.70401 8.301521  1  0   0  1
## 2 -0.09431065 11.90699 10.70401 8.301521  1  1   0  0
## 3 -0.09431065 12.03884 10.62532 9.341369  1  1   0  0
## 4  0.01980261 12.05292 10.82707 9.029418  1  0   0  1
## 5          NA 12.42561 11.18840 6.579251  1  1   0  0
## 6 -0.05129331 11.90497 10.70358 8.394799  1  1   0  0
```

(a) Please find the sample means of `prpblck` and `income` in the data set, along with their sample standard deviations.

```
## finding mean and SD for prpblck

## discrim$prpblck
## missing data is omitted when performing the calculation
meanPRPBLCK <- mean(discrim$prpblck, na.rm = TRUE)
sdPRPBLCK <- sd(discrim$prpblck, na.rm = TRUE)

data.frame(meanPRPBLCK, sdPRPBLCK)
```

```
##   meanPRPBLCK sdPRPBLCK
## 1   0.1134864 0.1824165
```

```
## finding mean and SD for income

## discrim$income
## missing data is omitted when performing the calculation
meanINCOME <- mean(discrim$income, na.rm = TRUE)
sdINCOME <- sd(discrim$income, na.rm = TRUE)

data.frame(meanINCOME, sdINCOME)
```

```
##   meanINCOME sdINCOME
## 1   47053.78 13179.29
```

mean of prpblck = 0.1134864

mean of income = 47053.78

sd of prpblck = 0.1824165

sd of income = 13179.29

(b) Use the rlm command from the MASS package in R to estimate the linear model below. It is okay to use rlm with default settings. Report the results in equation form, including the sample size and R2. Interpret the coefficient of prpblck. Is it meaningful? Is it worth looking into?

$$psoda = \beta 0 + \beta 1 \times prpblck + \mu \quad (1)$$

```
## running the linear model

lm <- lm(psoda ~ prpblck, data = discrim)
summary(lm)
```

```
##
## Call:
## lm(formula = psoda ~ prpblck, data = discrim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30884 -0.05963  0.01135  0.03206  0.44840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.03740    0.00519  199.87  < 2e-16 ***
## prpblck      0.06493    0.02396    2.71  0.00702 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0881 on 399 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.01808,    Adjusted R-squared:  0.01561
## F-statistic: 7.345 on 1 and 399 DF,  p-value: 0.007015
```

```
## Sample size = 401
## Multiple R squared is 0.01808
## Adjusted R squared is 0.01561
```

The relationship is as follows:

$psoda = 1.03740 + 0.06493 \times prpblck$

Sample size = 401

Multiple R squared is 0.01808, and Adjusted R squared is 0.01561.

A 1 unit change in proportion of black residents in the zipcode area is associated with a 0.06493 change in the price of medium soda. This marks an approximately 0.06493 / 1.0374 x 100% = 6.25% increase in the price of soda in areas where the population is all black compared to areas in which there are no black population. This relationship is statistically significant (p-value = 0.00702 < 0.05), which suggests that we are statistically confident that a change in `prpblck` is associated with a change in `psoda`. Although the relationship is small in scale, its statistical significance suggests that this is a meaningful relationship, and thus is worth looking into.

(c) Can you use the model above to make causal claims? Why or Why not? Do you see any potential threats to internal validity?

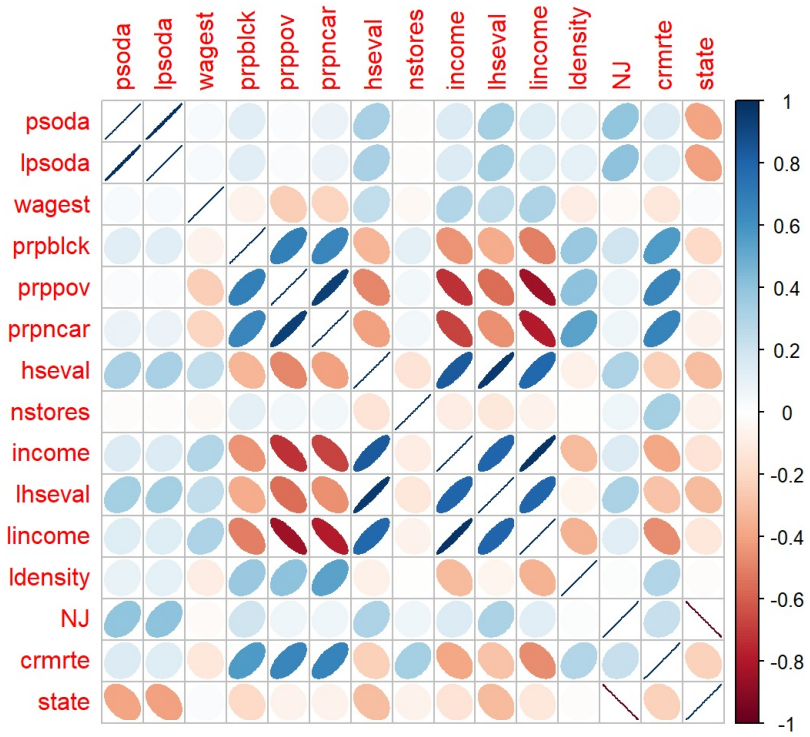
```
## creating a correlation matrix to investigate and have a rough understanding of the relationship between differ
ent variables

cor.model <- cor(discrim[, c('psoda', 'lpsoda', 'wagest', 'prpblck', 'prppov', 'prpncar', 'hseval', 'nstores', 'i
ncome', 'lhseval', 'lincome', 'ldensity', 'NJ', 'crmrte', 'state')], use = "complete.obs")
round(cor.model, 2)
```

```
##          psoda lpsoda wagest prpblck prppov prpncar hseval nstores income
## psoda     1.00   1.00   0.03    0.12   0.02    0.09   0.32   -0.01   0.14
## lpsoda    1.00   1.00   0.03    0.12   0.02    0.09   0.32   -0.01   0.15
## wagest    0.03   0.03   1.00   -0.07  -0.25   -0.22   0.24   -0.03   0.30
## prpblck   0.12   0.12  -0.07    1.00   0.68    0.65  -0.34    0.12  -0.44
## prppov    0.02   0.02  -0.25    0.68   1.00    0.93  -0.49    0.06  -0.73
## prpncar   0.09   0.09  -0.22    0.65   0.93    1.00  -0.41    0.06  -0.68
## hseval    0.32   0.32   0.24   -0.34  -0.49   -0.41   1.00   -0.14   0.84
## nstores  -0.01  -0.01  -0.03    0.12   0.06    0.06  -0.14    1.00  -0.10
## income    0.14   0.15   0.30   -0.44  -0.73   -0.68   0.84   -0.10   1.00
## lhseval   0.33   0.33   0.25   -0.36  -0.56   -0.46   0.96   -0.12   0.80
## lincome   0.13   0.14   0.30   -0.51  -0.84   -0.78   0.79   -0.07   0.97
## ldensity  0.10   0.10  -0.10    0.37   0.40    0.53  -0.08    0.00  -0.32
## NJ        0.39   0.40  -0.02    0.20   0.06    0.07   0.30    0.06   0.14
## crmrte    0.14   0.14  -0.12    0.56   0.65    0.67  -0.24    0.34  -0.39
## state    -0.39  -0.40   0.02   -0.20  -0.06   -0.07  -0.30   -0.06  -0.14
##          lhseval lincome ldensity    NJ crmrte state
## psoda       0.33    0.13     0.10  0.39   0.14 -0.39
## lpsoda      0.33    0.14     0.10  0.40   0.14 -0.40
## wagest      0.25    0.30    -0.10 -0.02  -0.12  0.02
## prpblck    -0.36   -0.51     0.37  0.20   0.56 -0.20
## prppov     -0.56   -0.84     0.40  0.06   0.65 -0.06
## prpncar    -0.46   -0.78     0.53  0.07   0.67 -0.07
## hseval      0.96    0.79    -0.08  0.30  -0.24 -0.30
## nstores    -0.12   -0.07     0.00  0.06   0.34 -0.06
## income      0.80    0.97    -0.32  0.14  -0.39 -0.14
## lhseval     1.00    0.80    -0.06  0.32  -0.29 -0.32
## lincome     0.80    1.00    -0.34  0.13  -0.46 -0.13
## ldensity   -0.06   -0.34     1.00  0.01   0.29 -0.01
## NJ          0.32    0.13     0.01  1.00   0.22 -1.00
## crmrte     -0.29   -0.46     0.29  0.22   1.00 -0.22
## state      -0.32   -0.13    -0.01 -1.00  -0.22  1.00
```

```
## Creating a correlation matrix plot to better visualise the relationship

corrplot(cor.model, method = 'ellipse')
```



No. This is due to the omitted variable bias that might be present in the model. For example, as `income` and `hseval` has a relationship with price of medium soda as well as are correlated to the proportion of black residents in the area, `income` and `hseval` might be some underlying confounding variables that affects the linear model.

Other potential confounding variables includes the crime rate `crmrte` of the zipcode area. From the correlation matrix, we can notice that the crime rate is positively correlated with the proportion of black residents in the area. As the crime rate of an area increases, there might be less customers that are willing to visit the fast food store. Thus, the store might need to reduce the price to attract more customers.

Thus, as some of the potential confounding variables are omitted, it suggests that there are omitted variables bias, which violates the internal validity and OLS assumptions. Thus, the model cannot give causal inference between prpblck and psoda.

(d) What could be potential confounding variables affecting causal inference in the model above? Try to be comprehensive and clearly lay out the logic behind each variable you list out.

Some of the confounding variables might include different types of fast food chains, the purchasing power of residents living in the area, and the demographic of the township where the data is recorded

For types of fast food chains (`chain`), different types of fast food chains have different pricings, and the black population might have a preference for certain fast food chains which leads to uneven distribution of fast food chains. Without taking into account these fast food chains and fixating on only one of them, we are unable to conclude that there is a relationship between price of soda and proportion of black residents.

For the purchasing power of residents living in the area, if the purchasing power of the residents in the town is higher, it suggests that the area overall has a higher living standard and thus might lead to a higher price for soda. Similarly, as shown by the correlation between housing value (`hseval`) and proportion of black residents (`prpblck`) and the correlation between income (`income`) and proportion of black residents (`prpblck`), we can notice that in general the purchasing power of the residents in the area and the proportion of black residents is negatively associated. Thus, it might be a confounding variable as it has a relationship with both price of soda and proportion of black residents.

For the demographic of the township where the data is recorded, different townships have different population and race constituents, which suggests a possible difference in purchasing habits and consumer preference. As demand increases for a certain product (such as soda), the price will also increase accordingly. Meanwhile, the demographic of the township is also associated with the proportion of black residents, as proportion of black residents make up a part of what makes the township what it is. Different races might have different purchasing and living habits, and thus the proportion of black residents in a district might be associated with the demographic of the town. Thus, it might be a confounding variable as it has a relationship with both price of soda and proportion of black residents.

(e) Which of the potential confounding variables above are measurable and which of them are not measurable? Give reasons for your classification. Which of them do you have access to in the data set? Can you think of proxies for those that are not measurable?

For types of fast food chains (`chain`), it is a measurable variable that can be measured by taking into account the `chain` variable in the model above.

On the other hand, purchasing power of residents living in the area is an immeasurable variable that cannot be measured directly from the data set. As we are unable to quantitatively measure purchasing power in the population, we decided to use `income` and `hseval` as proxies to represent the purchasing power.

Meanwhile, the demographic of the township where the data is recorded is also another immeasurable variable that cannot be measured directly from the data set. As we are unable to quantitatively measure the demographic of the township, we have decided to use `NJ` as a proxy to represent the demographic of the township.

(f) Include the ones that are measurable and available in your regression equation and rerun the regression. How does the coefficient of prpblck change? How do the Adjusted R2 and Multiple R2 change? What is your new interpretation of the estimated relationship between prpblack and psoda?

```
## running the linear model

linear.model <- lm(psoda ~ prpblck + income + hseval + as.factor(chain) + as.factor(NJ), data = discrim)
summary(linear.model)
```

```
##
## Call:
## lm(formula = psoda ~ prpblck + income + hseval + as.factor(chain) +
##     as.factor(NJ), data = discrim)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20037 -0.03358 -0.00373  0.03553  0.42855
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.971e-01  1.594e-02  62.547  < 2e-16 ***
## prpblck            5.783e-02  2.106e-02   2.745  0.00632 **
## income            -2.146e-06  4.910e-07  -4.371 1.59e-05 ***
## hseval             7.023e-07  1.157e-07   6.072 2.99e-09 ***
## as.factor(chain)2 -4.809e-02  9.200e-03  -5.227 2.80e-07 ***
## as.factor(chain)3  5.146e-02  8.586e-03   5.994 4.64e-09 ***
## as.factor(chain)4 -7.741e-02  9.992e-03  -7.748 8.06e-14 ***
## as.factor(NJ)1     5.855e-02  9.449e-03   6.197 1.46e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06597 on 393 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.4577, Adjusted R-squared:  0.448
## F-statistic: 47.38 on 7 and 393 DF,  p-value: < 2.2e-16
```

```
## coefficient for prpblck changes by 0.05783 - 0.06493 = - 0.0071
## Multiple R squared increased from 0.01808 to 0.4577
## Adjusted R squared increased from 0.01561 to 0.448
```

The relationship is as follows:

`psoda` = 0.9971 + 0.05783 x `prpblck` - 2.146e-06 x `income` + 7.023e-07 x `hseval` - 0.04809 x `KFC` + 0.05146 x `Roy Rogers` - 0.07741 x `Wendy's` + 0.05855 x `NJ`

Coefficient for prpblck changes by 0.05783 - 0.06493 = - 0.0071

Multiple R squared increased from 0.01808 to 0.4577

Adjusted R squared increased from 0.01561 to 0.448

Holding other factors constant, a 1 unit change in proportion of black residents in the zipcode area is associated with a 0.05783 change in the price of medium soda. This marks an approximately 0.05783 / 0.9971 x 100% = 5.8% increase in the price of soda in areas where the population is all black compared to areas in which there are no black population. This relationship is statistically significant (p-value = 0.00632 < 0.05), which suggests that we are statistically confident that a change in `prpblck` is associated with a change in `psoda`. Although the relationship is small in scale, its statistical significance suggests that this is a meaningful relationship, and thus is worth looking into.