ECON3140HW7

Nick Gembs

5/1/2022

```
library(haven)

data <- read_dta("C:/Users/Nick/Downloads/hmda-project-1.dta")
```

1. This paper poses the question of whether there is racial discrimination bias in the disperal of mortgage loans. Data collected includes individuals wealth, financial information, race, and whether or not they were approved for a loan. Previous studies were done on this topic by Harold Black et al., (1978), Thomas A. King (1980), and Robert Schafer and Helen F. Ladd (1981). These studies found that being a minority did have a significant decrease on an individuals odds of getting a loan, however, this data is over 10 years old and leaves out significant information that could be highly correlated with race, such as credit history, employment history, and different types of lenders. In the results, OLS estimators show that although economic factors due have some impact on the dispraportionate denial rates of minorities, race is significant influence beyond the 1% significance bound.

```
#2

len = length(data$s7)

for (i in 1:len){
  if (data$s7[i] == 3){
    data$s7[i] = 1
  } else {
    data$s7[i] = 0
  }
}


black = rep(0,len)
cbind(data,black)

for (i in 1:len){
  if (data$s13[i] == 3){
    data$black[i] = 1
  } else if (data$s13[i] == 5) {
    data$black[i] = 0
  } else {
    data$black[i] = NA
  }
}
```

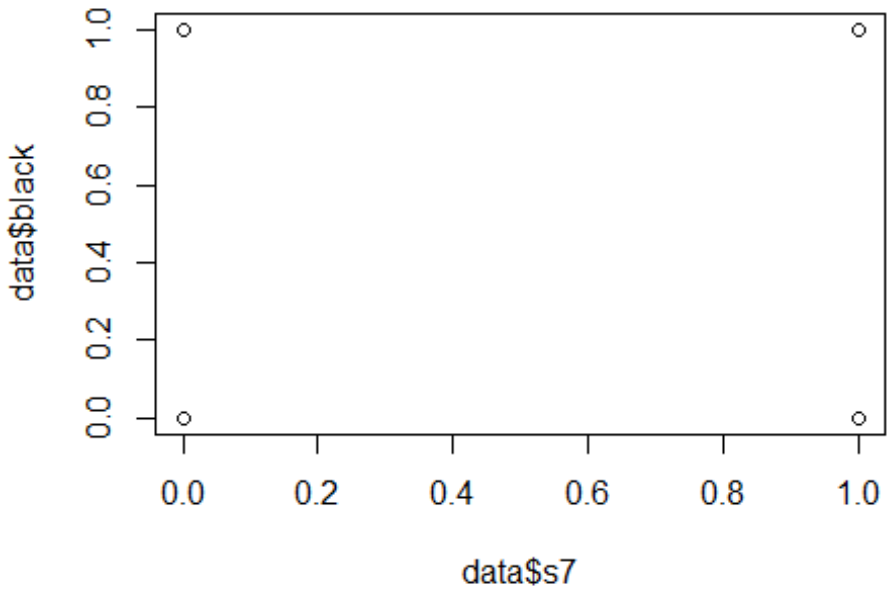
```
## Warning: Unknown or uninitialised column: `black`.
```

```
#3

plot(data$s7,data$black)

bfraction = sum(data$black)/length(data$black)
dfraction = sum(data$s7)/length(data$s7)

library(data.table)
library(survival)
```




```
library(gmodels)
table <- data.frame(data$black, data$s7)
CrossTable(data$black,data$s7)

##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
```

```
##
## Total Observations in Table:  2380
##
##
##               | data$s7
##   data$black |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |      1852 |       189 |      2041 |
##              |     1.709 |    12.560 |           |
##              |     0.907 |     0.093 |     0.858 |
##              |     0.884 |     0.663 |           |
##              |     0.778 |     0.079 |           |
## -------------|-----------|-----------|-----------|
##            1 |       243 |        96 |       339 |
##              |    10.287 |    75.620 |           |
##              |     0.717 |     0.283 |     0.142 |
##              |     0.116 |     0.337 |           |
##              |     0.102 |     0.040 |           |
## -------------|-----------|-----------|-----------|
## Column Total |      2095 |       285 |      2380 |
##              |     0.880 |     0.120 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
# The crosstabulation shows that the B0 should be .093 as that is the percent
of denials when the individual is not black. B1 should be .283-.093 as that
is the difference in percent denials when the individual race is changed to
black.
lm.fit = lm(data$s7~data$black)
lm.fit
```

```
##
## Call:
## lm(formula = data$s7 ~ data$black)
##
## Coefficients:
## (Intercept)   data$black
##      0.0926       0.1906
```

```
#4

lm.fit2 = lm(data$s7~data$s46)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = data$s7 ~ data$s46)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.73070 -0.13736 -0.11322 -0.07097  1.05577
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0799096  0.0211578  -3.777 0.000163 ***
## data$s46     0.0060353  0.0006084   9.920  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3183 on 2378 degrees of freedom
## Multiple R-squared:  0.03974,    Adjusted R-squared:  0.03933
## F-statistic: 98.41 on 1 and 2378 DF,  p-value: < 2.2e-16
```

 # This data is statistically significant to a high degree, showing that being denied for a loan is highly correlated (positively) with a given individual's payment/income ratio. This makes sense as cash outflows are harmful to a person's financials while inflows are a benefit. Banks prefer savers over spenders.

```
# Load libraries
library("lmtest")
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

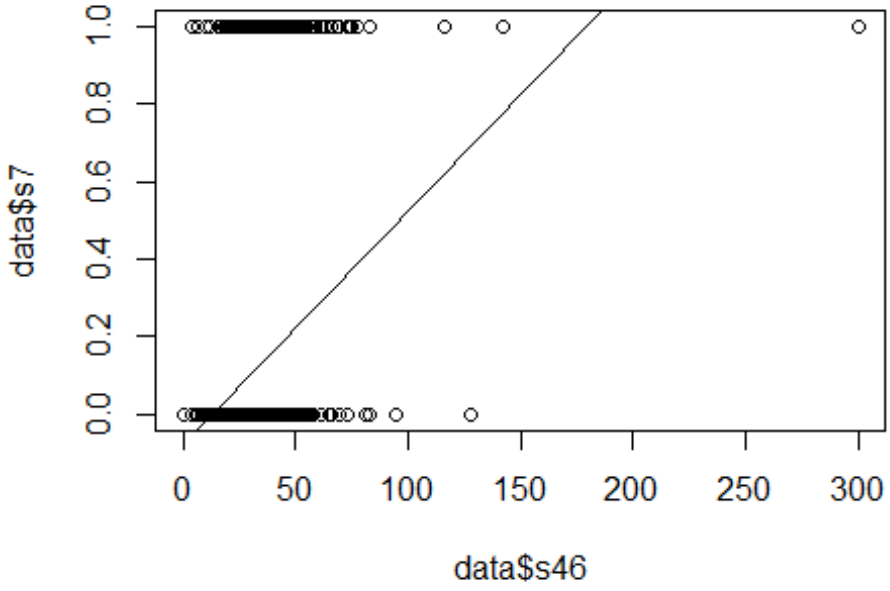
```
library("sandwich")

# Robust t test
robust = coeftest(lm.fit2, vcov = vcovHC(lm.fit2, type = "HC0"))
robust
```

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value Pr(>|t|)
## (Intercept) -0.07990965  0.03195317 -2.5008  0.01246 *
## data$s46     0.00603535  0.00098441  6.1309 1.02e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

 # This test likely suffers from heteroskedasticity, so robust standard errors are preferred.

```
plot(data$s46, data$s7)
abline(robust)
```




```
#5

lm.fit2$coefficients[1] + .2*lm.fit2$coefficients[2]

## (Intercept)
## -0.07870258

lm.fit2$coefficients[1] + .1*lm.fit2$coefficients[2]

## (Intercept)
## -0.07930611

# The value impact on denial is negative for a PI of 20%. This could be
interpereted as a 20% PI decreasing the odds of denial as it is a financially
healthy value. A PI of 10% decreases the value further, but only by a small
amount. To make this amount more accurate. we can use a log model that will
account for percent changes rather than unit changes.

#6

#prevent undefined log error
data$s46[350] = .0001

lm.fit3 = lm(data$s7~log(data$s46))
summary(lm.fit3)

##
## Call:
```

```
## lm(formula = data$s7 ~ log(data$s46))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25820 -0.13262 -0.12096 -0.09652  1.13390
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.22213    0.05880  -3.778 0.000162 ***
## log(data$s46)  0.09899    0.01692   5.851 5.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3225 on 2378 degrees of freedom
## Multiple R-squared:  0.01419,    Adjusted R-squared:  0.01378
## F-statistic: 34.24 on 1 and 2378 DF,  p-value: 5.55e-09
```

```
lm.fit3$coefficients[1] + .2*lm.fit3$coefficients[2]
```

```
## (Intercept)
##  -0.2023305
```

```
lm.fit3$coefficients[1] + .1*lm.fit3$coefficients[2]
```

```
## (Intercept)
##    -0.21223
```

```
#7

lm.fit4 = lm( data$s7 ~ log(data$s46) + data$black)
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = data$s7 ~ log(data$s46) + data$black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35941 -0.10780 -0.09669 -0.07661  1.08624
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.20686    0.05766  -3.587 0.000341 ***
## log(data$s46)  0.08700    0.01663   5.232 1.83e-07 ***
## data$black     0.18348    0.01859   9.869  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3161 on 2377 degrees of freedom
## Multiple R-squared:  0.053,  Adjusted R-squared:  0.0522
## F-statistic: 66.51 on 2 and 2377 DF,  p-value: < 2.2e-16
```

```
# The indicator variable for black is both large (.18348 increase in percent
for black) and statistically significant (p=2x10^-16)
```

8. The problem with Beta2 in example 7 is that it likely suffers from endogeneity. There are many factors that are correlated with minority status that likely have a significant impact on mortgage loan denial that are not included in this model. This will make race absorb some of the causal effect that other variables would have taken on.

```
#9

# s7 = denial likelihood
# black = black indicator
# self = self-employment indicator
# HI = housing expense/income ratio
# LV = loan/value
# CCS = consumer credit score
# MCS = mortgage credit score
# noMI mortgage insurance denied



lm.fittotal = lm(data$s7 ~ log(data$s46) + data$black + data$LV + (data$LV)^2
+ data$HI + data$self + data$CCS + data$MCS + data$NoMI)

summary(lm.fittotal)

##
## Call:
## lm(formula = data$s7 ~ log(data$s46) + data$black + data$LV +
##     (data$LV)^2 + data$HI + data$self + data$CCS + data$MCS +
##     data$NoMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00593 -0.12151 -0.05596 -0.01826  1.08813
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.249079   0.058211  -4.279 1.95e-05 ***
## log(data$s46)  0.026001   0.017135   1.517   0.1293
## data$black     0.110415   0.017614   6.269 4.32e-10 ***
## data$LV        0.036734   0.030778   1.194   0.2328
## data$HI        0.337943   0.069085   4.892 1.07e-06 ***
## data$self      0.072936   0.018422   3.959 7.74e-05 ***
## data$CCS       0.041007   0.003663  11.194  < 2e-16 ***
## data$MCS       0.022184   0.011302   1.963   0.0498 *
## data$NoMI      0.750078   0.042228  17.763  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2871 on 2370 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2213, Adjusted R-squared:  0.2187
## F-statistic: 84.19 on 8 and 2370 DF,  p-value: < 2.2e-16
```

```
lm.fittotalnolog = lm(data$s7 ~ (data$s46) + data$black + data$LV +
(data$LV)^2 + data$HI + data$self + data$CCS + data$MCS + data$NoMI)
```

```
summary(lm.fittotalnolog)
```

```
##
## Call:
## lm(formula = data$s7 ~ (data$s46) + data$black + data$LV + (data$LV)^2 +
##     data$HI + data$self + data$CCS + data$MCS + data$NoMI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00541 -0.12407 -0.05676 -0.01236  1.09002
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2274449  0.0326807  -6.960 4.40e-12 ***
## data$s46     0.0051995  0.0008852   5.874 4.85e-09 ***
## data$black   0.1065980  0.0175067   6.089 1.32e-09 ***
## data$LV      0.0322488  0.0305007   1.057 0.290476
## data$HI     -0.0654563  0.0981261  -0.667 0.504797
## data$self    0.0640292  0.0183518   3.489 0.000494 ***
## data$CCS     0.0393960  0.0036464  10.804  < 2e-16 ***
## data$MCS     0.0266999  0.0112540   2.372 0.017749 *
## data$NoMI    0.7425719  0.0419598  17.697  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2852 on 2370 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2291
## F-statistic: 89.35 on 8 and 2370 DF,  p-value: < 2.2e-16
```

```
# Except for LV, all of the variables are positively significant with regards
to loan denial in at least one of the two regressions. This makes economical
sense because each of the variables (besides black) are either an indicator
or measurement of something that would make an individual a more fiscally
risky investment.Taking the log of PI switches its significance with HI, this
is likely because there is a high degree of correltaion/collinearity between
the two variables.
```

```
#10
```

```
lm.fittotal$coefficients[3]
```

```
## data$black
##  0.1104153

lm.fittotalnolog$coefficients[3]

## data$black
##   0.106598
```

The predicted effect of race on loan denial in the OLS specification is between 10-11% depending on the model used. Economically, this mean that being an individual who is black will increase your odds of being denied a mortgage loan by 10-11% as opposed to being white, all other financial factors being equal.

As stated in question 10, both specifications produce similar values of the coefficient on the indicator variable of black. This is because both the percent and unit increase of PI have a similar overall effect on the model and account for a similar set of variability.

```
#12

(sum(data$LV)/length(data$LV))

## [1] NA

sum(data$LV)

## [1] NA

length(data$LV)

## [1] 2380

which(is.na(data$LV))

## [1] 2246

data$LV[2246] = .75

whiteavg = lm.fittotal$coefficients[1] + lm.fittotal$coefficients[2] *
(sum(log(data$s46))/length(data$s46)) + lm.fittotal$coefficients[3] * 0 +
lm.fittotal$coefficients[4] * (sum(data$LV)/length(data$LV)) +
lm.fittotal$coefficients[4] * (sum((data$LV)^2)/length((data$LV)^2)) +
lm.fittotal$coefficients[5] * (sum(data$HI)/length(data$HI)) +
lm.fittotal$coefficients[6] * (sum(data$self)/length(data$self)) +
lm.fittotal$coefficients[7] * (sum(data$CCS)/length(data$CCS)) +
lm.fittotal$coefficients[8] * (sum(data$MCS)/length(data$MCS)) +
lm.fittotal$coefficients[9] * (sum(data$NoMI)/length(data$NoMI))

whiteavg

## (Intercept)
##   0.1275468
```

```
blackavg = lm.fittotal$coefficients[1] + lm.fittotal$coefficients[2] *
(sum(log(data$s46))/length(data$s46)) + lm.fittotal$coefficients[3] * 1 +
lm.fittotal$coefficients[4] * (sum(data$LV)/length(data$LV)) +
lm.fittotal$coefficients[4] * (sum((data$LV)^2)/length((data$LV)^2)) +
lm.fittotal$coefficients[5] * (sum(data$HI)/length(data$HI)) +
lm.fittotal$coefficients[6] * (sum(data$self)/length(data$self)) +
lm.fittotal$coefficients[7] * (sum(data$CCS)/length(data$CCS)) +
lm.fittotal$coefficients[8] * (sum(data$MCS)/length(data$MCS)) +
lm.fittotal$coefficients[9] * (sum(data$NoMI)/length(data$NoMI))

blackavg

## (Intercept)
##    0.237962

whiteavgnolog = lm.fittotalnolog$coefficients[1] +
lm.fittotalnolog$coefficients[2] * (sum((data$s46))/length(data$s46)) +
lm.fittotalnolog$coefficients[3] * 0 + lm.fittotalnolog$coefficients[4] *
(sum(data$LV)/length(data$LV)) + lm.fittotalnolog$coefficients[4] *
(sum((data$LV)^2)/length((data$LV)^2)) + lm.fittotalnolog$coefficients[5] *
(sum(data$HI)/length(data$HI)) + lm.fittotalnolog$coefficients[6] *
(sum(data$self)/length(data$self)) + lm.fittotalnolog$coefficients[7] *
(sum(data$CCS)/length(data$CCS)) + lm.fittotalnolog$coefficients[8] *
(sum(data$MCS)/length(data$MCS)) + lm.fittotalnolog$coefficients[9] *
(sum(data$NoMI)/length(data$NoMI))

whiteavgnolog

## (Intercept)
##   0.1252158

blackavgnolog = lm.fittotalnolog$coefficients[1] +
lm.fittotalnolog$coefficients[2] * (sum((data$s46))/length(data$s46)) +
lm.fittotalnolog$coefficients[3] * 1 + lm.fittotalnolog$coefficients[4] *
(sum(data$LV)/length(data$LV)) + lm.fittotalnolog$coefficients[4] *
(sum((data$LV)^2)/length((data$LV)^2)) + lm.fittotalnolog$coefficients[5] *
(sum(data$HI)/length(data$HI)) + lm.fittotalnolog$coefficients[6] *
(sum(data$self)/length(data$self)) + lm.fittotalnolog$coefficients[7] *
(sum(data$CCS)/length(data$CCS)) + lm.fittotalnolog$coefficients[8] *
(sum(data$MCS)/length(data$MCS)) + lm.fittotalnolog$coefficients[9] *
(sum(data$NoMI)/length(data$NoMI))

blackavgnolog

## (Intercept)
##   0.2318137

blackavg-whiteavg

## (Intercept)
##   0.1104153
```

```
lm.fittotal$coefficients[3]

## data$black
##  0.1104153

blackavgnolog-whiteavgnolog

## (Intercept)
##    0.106598

lm.fittotalnolog$coefficients[3]

## data$black
##   0.106598
```

As seen from the equation in #12, the effect of indicator black on the average individual as well as the average predicted effect of indicator black on an individual are equivalent to the respective coefficients of the beta value for black in the OLS and altered OLS regressions.