Moh Jaiswal 500916860 Professor Laura Turner

Final Assignment

Question 1

Comparing STAR data set regression with the MASS data set, both the coef along with the std. errors are very different. This is not a good indicator for external validity across states as the combined dataset model1 and the STAR data set had similar results, but different results from the MASS data set.

The external validity question for this exercise is whether our results for combined school data specific to "Cali" (dummy variable = 1) can be generalized to "Mass."

Conclusion:
We see some evidence against this given that the variable which matters most is different in the two regressions (log_avginc), and especially from the fact that average income does not seem to matter to the computers per student.

```
# Question 1
rm(list=ls())

library(haven)
library(jtools)
library(clubSandwich)
library(plm)
library(car)
library(lmtest)
library(ivpack)
MASS <- read_dta("MASS.dta")
View(MASS)
STAR_small <- read_dta("STAR_small.dta")
View(STAR_small)

MASS$log_avginc <- log(MASS$percap)
STAR_small$log_avginc <- log(STAR_small$avginc)

# Computers per student is 1/(students per computer)

MASS$comp_stu = 1.0/MASS$s_p_c
MASS$comp_stu[!is.finite(MASS$comp_stu)] <- 0
MASS$s_p_c <- NULL

MASS_sample = subset(MASS, select = c(comp_stu, log_avginc))
STAR_sample = subset(STAR_small, select = c(comp_stu, log_avginc))

#Combine data sets
STAR_sample$state = "Cali"
MASS_sample$state = "Mass"

schooldata <- rbind(MASS_sample, STAR_sample)

#rm(MASS,STAR_small)

schooldata$dummy = ifelse(schooldata$state=="Cali", 1, 0)

model1 <- lm(comp_stu ~ log_avginc + dummy, data = schooldata)
summ(model1, digits=4, robust = "HC1")
model2 <- lm(comp_stu ~ log_avginc, data = subset(schooldata, state=="Cali"))
summ(model2, digits=4, robust = "HC1")
model3 <- lm(comp_stu ~ log_avginc, data = subset(schooldata, state=="Mass"))
summ(model3, digits=4, robust = "HC1")
```

```
> summ(model1, digits=4, robust = "HC1")
MODEL INFO:
Observations: 640
Dependent Variable: comp_stu
Type: OLS linear regression

MODEL FIT:
F(2,637) = 5.5892, p = 0.0039
R² = 0.0172
Adj. R² = 0.0142

Standard errors: Robust, type = HC1
------------------------------------------------------
                      Est.      S.E.    t val.       p
----------------- -------- -------- -------- --------
(Intercept)         0.0669   0.0230   2.9059   0.0038
log_avginc          0.0232   0.0080   2.9187   0.0036
dummy               0.0075   0.0053   1.4168   0.1570
------------------------------------------------------
```

```
> model2 <- lm(comp_stu ~ log_avginc, data = subset(schooldata, state=="Cali"))
> summ(model2, digits=4, robust = "HC1")
MODEL INFO:
Observations: 420
Dependent Variable: comp_stu
Type: OLS linear regression

MODEL FIT:
F(1,418) = 10.8857, p = 0.0011
R² = 0.0254
Adj. R² = 0.0230

Standard errors: Robust, type = HC1
------------------------------------------------------
                      Est.      S.E.    t val.       p
----------------- -------- -------- -------- --------
(Intercept)         0.0662   0.0236   2.8056   0.0053
log_avginc          0.0264   0.0088   3.0023   0.0028
------------------------------------------------------
> model3 <- lm(comp_stu ~ log_avginc, data = subset(schooldata, state=="Mass"))
> summ(model3, digits=4, robust = "HC1")
MODEL INFO:
Observations: 220
Dependent Variable: comp_stu
Type: OLS linear regression

MODEL FIT:
F(1,218) = 0.5139, p = 0.4742
R² = 0.0024
Adj. R² = -0.0022

Standard errors: Robust, type = HC1
------------------------------------------------------
                      Est.      S.E.    t val.       p
----------------- -------- -------- -------- --------
(Intercept)         0.1032   0.0527   1.9589   0.0514
log_avginc          0.0107   0.0184   0.5790   0.5632
------------------------------------------------------
```

Question 2

Fixed effects regression for the whole dataset

$\#\ of\ Vehicular\ Fatalities = \ -0.1660 - (0.6640 * beertax) + (0.0706 * emppop)$

In the model, the obvious hypothesis is that increase in the emp pop should **increase** the #VF, the more of the population is employed, the higher the traffic density, the greater the fatalities . According to model4, an increase in the emppop by 1% **increased** total #VF by 0.0706, holding beertax constant. 1 dollar increase in beer tax holding the emppop constant (assume it is 0) causes a **decrease** in fatality rate by 0.6640.

FIPS Codes:

1984, Missouri - *29*

Intercepts:
- 1984: - 0.1958
- Missouri: - 1.6872

If we assume state and time were binary variables
When we assume that 1984 and Missouri are all equal to 1 (binary variable). The co-effecients are added to the intercept value as they are fixed for that particular fixed effect. Giving us the new regression equation:

1984, Missouri
$\#VF\ for\ 1984,\ Missouri = (3.1443 - 0.1958 - 1.6872) - 0.6640 * Beertax + 0.0706 * emppop$

```
# Question 2

rm(list=ls())

library(haven)
library(jtools)
library(clubSandwich)
library(plm)
library(car)
library(lmtest)
library(ivpack)

ROADS <- read_dta("fatality_data.dta")

View(ROADS)

ROADS$mrall = ROADS$mrall*10000

model4 <- plm(mrall ~ beertax + emppop + factor(state) + factor(year), model = "pooling", data=ROADS, index = c("state","year"))

coef_test(model4, vcov = "CR1", cluster = ROADS$state)
```

```
> model4 <- plm(mrall ~ beertax + emppop + factor(state) + factor(year), model = "pooling", data=ROADS, index = c("state","year"))
>
> coef_test(model4, vcov = "CR1", cluster = ROADS$state)
              Coef. Estimate     SE  t-stat  d.f. p-val (Satt) Sig.
1       (Intercept)  -0.1660 0.7364 -0.2254 20.12      0.82391     
2           beertax  -0.6640 0.3133 -2.1193  8.55      0.06471    .
3            emppop   0.0706 0.0124  5.6804 17.99      < 0.001  ***
4    factor(state)4  -0.8944 0.4265 -2.0970  9.23      0.06466    .
5    factor(state)5  -0.7043 0.3253 -2.1651  8.60      0.05996    .
6    factor(state)6  -1.9912 0.5025 -3.9623  9.57      0.00291   **
7    factor(state)8  -2.3044 0.4982 -4.6251 11.45      < 0.001  ***
8    factor(state)9  -2.6177 0.4812 -5.4400 11.18      < 0.001  ***
9   factor(state)10  -1.9319 0.4942 -3.9090 10.27      0.00278   **
10  factor(state)12  -0.3957 0.1652 -2.3959  9.25      0.03948    *
11  factor(state)13   0.0235 0.2526  0.0929 10.53      0.92774     
12  factor(state)16  -1.1169 0.4186 -2.6679  9.80      0.02396    *
13  factor(state)17  -2.3211 0.4725 -4.9126  9.23      < 0.001  ***
14  factor(state)18  -1.8529 0.4409 -4.2026  9.44      0.00207   **
15  factor(state)19  -2.0765 0.4166 -4.9848 10.31      < 0.001  ***
16  factor(state)20  -1.9090 0.4145 -4.6056 11.48      < 0.001  ***
17  factor(state)21  -1.3126 0.4514 -2.9080  8.63      0.01813    *
18  factor(state)22  -0.8038 0.2635 -3.0498  8.51      0.01473    *
19  factor(state)23  -1.4697 0.2898 -5.0716 10.23      < 0.001  ***
20  factor(state)24  -2.4257 0.4829 -5.0231 10.92      < 0.001  ***
21  factor(state)25  -2.7828 0.4670 -5.9595 10.77      < 0.001  ***
22  factor(state)26  -1.6823 0.3639 -4.6236  8.93      0.00127   **
23  factor(state)27  -2.7491 0.4629 -5.9388 12.18      < 0.001  ***
24  factor(state)28   0.0536 0.1781  0.3007  8.55      0.77083     
25  factor(state)29  -1.6872 0.4305 -3.9195  9.48      0.00318   **
26  factor(state)30  -0.8807 0.4343 -2.0278 10.10      0.06978    .
27  factor(state)31  -2.2459 0.4244 -5.2916 11.66      < 0.001  ***
28  factor(state)32  -1.4819 0.5010 -2.9575 11.86      0.01211    *
29  factor(state)33  -2.2059 0.3778 -5.8395 15.42      < 0.001  ***
30  factor(state)34  -2.5795 0.5070 -5.0880  9.53      < 0.001  ***
31  factor(state)35   0.2414 0.3966  0.6086  8.85      0.55808     
32  factor(state)36  -2.3256 0.4741 -4.9050  8.69      < 0.001  ***
33  factor(state)37  -0.8583 0.1602 -5.3580 20.24      < 0.001  ***
34  factor(state)38  -2.2736 0.4263 -5.3337 11.04      < 0.001  ***
35  factor(state)39  -1.9208 0.3989 -4.8155  9.02      < 0.001  ***
36  factor(state)40  -0.9115 0.2470 -3.6909 10.90      0.00361   **
37  factor(state)41  -1.5854 0.4658 -3.4037  9.46      0.00729   **
38  factor(state)42  -1.8324 0.4304 -4.2578  8.60      0.00235   **
39  factor(state)44  -2.8550 0.4914 -5.8096 10.11      < 0.001  ***
40  factor(state)45   0.2142 0.0829  2.5831 17.84      0.01884    *
41  factor(state)46  -1.7077 0.3527 -4.8415 12.84      < 0.001  ***
42  factor(state)47  -1.0615 0.4213 -2.5195  8.82      0.03329    *
43  factor(state)48  -1.5260 0.4084 -3.7365 10.92      0.00333   **
44  factor(state)49  -1.7989 0.3224 -5.5792 12.72      < 0.001  ***
45  factor(state)50  -1.7812 0.3661 -4.8649 13.92      < 0.001  ***
46  factor(state)51  -1.9543 0.3392 -5.7615 12.67      < 0.001  ***
47  factor(state)53  -1.9928 0.4568 -4.3629  9.18      0.00173   **
48  factor(state)54  -0.2255 0.3687 -0.6116 10.18      0.55422     
49  factor(state)55  -2.3576 0.4923 -4.7884 10.15      < 0.001  ***
50  factor(state)56  -0.9790 0.5363 -1.8255 10.65      0.09607    .
51 factor(year)1983  -0.0849 0.0330 -2.5753 46.97      0.01323    *
52 factor(year)1984  -0.1958 0.0519 -3.7717 41.93      < 0.001  ***
53 factor(year)1985  -0.2832 0.0579 -4.8933 37.11      < 0.001  ***
54 factor(year)1986  -0.2394 0.0696 -3.4374 28.88      0.00180   **
55 factor(year)1987  -0.2989 0.0819 -3.6496 23.21      0.00132   **
56 factor(year)1988  -0.3510 0.0858 -4.0927 21.72      < 0.001  ***
>
```

Question 3

Generalized linear regression

Understanding the coefficients for a logit model is not as straightforward as linear regression. Here we are dealing with the probability of something happening rather than a direct impact of IV on DV.

a) logit positive value = logistic > 1 = increase in the probability of the event when you have a positive change in the IV

b) logit negative value = logistic < 1 = decrease in the probability of the event when you have a positive change in the IV

The larger the magnitude of our coefficients the greater the significance the variable has on predicting the dependent variable outcome. If a logit coefficient has a positive value, we assume the logistic value is greater than 1. This simply means that when there is a positive change in the independent variable, the probability of the event happening is higher, and vice versa for negative values. Except if a logit coefficient value is negative, then the logistic value will be less than 1.

Coefficient Analysis:

piratio: 5.9100 (The probability of being denied increases as the PI ratio increases, significantly)
s33: -0.0027 (The probability of being denied decreases as the price of the house increases, however, it is a minute change)
dummy: 0.4762 (The probability of being denied increases if the applicant is self-employed, albeit marginally)

Conclusion:

The results point toward a high focus on piratio being the most important predictor for the loan, which is a sensible metric to underwrite credit approvals.

Prediction:

The logdif value is negative, indicating that the chances of getting denied decrease by 0.2967% if the price of the home increases from $90,000 to $100,000. Such a prediction would not be significant and thus, unreliable.

```
# Question 3

rm(list=ls())

library(haven)
library(jtools)
library(clubSandwich)
library(plm)
library(car)
library(lmtest)
library(ivpack)

options(warn=-1)

options(scipen = 20)

HMDA <- read_dta("hmda_sw.dta")

View(HMDA)

HMDA$piratio <- HMDA$s46/100
HMDA$deny <- ifelse(HMDA$s7==3,1,0)
HMDA$dummy <- ifelse(HMDA$s27a=="1", 1, 0)
HMDA <- subset(HMDA, s33 < 99999)
HMDA <- subset(HMDA, s17 < 99999)
```

```
> model5 <- glm(deny ~ piratio + s33 + dummy, family = binomial(link = "logit"), data=HMDA)
> summ(model5, digits=4, robust="HC1")
MODEL INFO:
Observations: 2343
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(3) = 98.0546, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.0795
Pseudo-R² (McFadden) = 0.0577
AIC = 1608.6858, BIC = 1631.7226

Standard errors: Robust, type = HC1
------------------------------------------------------
                      Est.     S.E.    z val.        p
----------------- --------- -------- --------- --------
(Intercept)        -3.6352   0.3679   -9.8810   0.0000
piratio             5.9100   1.0159    5.8176   0.0000
s33                -0.0027   0.0008   -3.4551   0.0006
dummy               0.4762   0.2068    2.3021   0.0213
------------------------------------------------------
>
> prediction10 <- predict(model5,
+                         newdata = data.frame("s33" = 90,
+                                              "piratio" = median(HMDA$piratio),
+                                              "dummy" = median(HMDA$dummy)),
+                         type = "response")
>
>
> prediction11 <- predict(model5,
+                         newdata = data.frame("s33" = 100,
+                                              "piratio" = median(HMDA$piratio),
+                                              "dummy" = median(HMDA$dummy)),
+                         type = "response")
>
> logdif <- prediction11 - prediction10
> logdif
        1
-0.002967
```

```
model5 <- glm(deny ~ piratio + s33 + dummy, family = binomial(link = "logit"), data=HMDA)
summ(model5, digits=4, robust="HC1")

prediction10 <- predict(model5,
                        newdata = data.frame("s33" = 90,
                                             "piratio" = median(HMDA$piratio),
                                             "dummy" = median(HMDA$dummy)),
                        type = "response")


prediction11 <- predict(model5,
                        newdata = data.frame("s33" = 100,
                                             "piratio" = median(HMDA$piratio),
                                             "dummy" = median(HMDA$dummy)),
                        type = "response")

logdif <- prediction11 - prediction10
logdif
```

Question 4

Coefficient Analysis:
piratio: 6.2950 (The probability of being denied increases as the PI ratio increases, significantly)
s33: -0.0037 (The probability of being denied decreases as the price of the house increases, however, it is a minute change)
s17: 0.0035 (The probability of being denied increases with a positive change in the income of the applicant)
dummy: 0.4079 (The probability of being denied increases if the applicant is self-employed, albeit marginally)

Conclusion:
The inclusion of the variable income plays a **directly inverse** relationship with Piratio. This will impact our error values, causing them to inflate and making our model unreliable. There is also the probability where the greater the income, the higher the price of the home you purchase will be. Making the loan riskier and thus increasing the probability of being denied. This might be due to the low probability of a steady flow of high income over a very long period of time.

Omitted variable bias:
Using median income to predict the actual income will vary over time but not over states. Giving us entity-specific intercepts. Another way to saying is that our instrument will be autocorrelated, and the error term would also be autocorrelated and not serially correlated. Making it a more precise instrument to measure the dependent variable.

```
# Question 4

model6 <- glm(deny ~ piratio + s33 + s17 + dummy, family = binomial(link = "logit"), data=HMDA)

summ(model6, digits=4, robust="HC1")
```

```
> summ(model6, digits=4, robust="HC1")
MODEL INFO:
Observations: 2343
Dependent Variable: deny
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(4) = 104.4947, p = 0.0000
Pseudo-R² (Cragg-Uhler) = 0.0846
Pseudo-R² (McFadden) = 0.0615
AIC = 1604.2457, BIC = 1633.0417

Standard errors: Robust, type = HC1
--------------------------------------------------------
                       Est.     S.E.     z val.        p
----------------- --------- -------- ---------- --------
(Intercept)         -3.8296   0.3974    -9.6360   0.0000
piratio              6.2950   1.0589     5.9448   0.0000
s33                 -0.0037   0.0009    -3.9774   0.0001
s17                  0.0035   0.0018     1.9088   0.0563
dummy                0.4079   0.2137     1.9090   0.0563
--------------------------------------------------------
```