Assignment 3

Question 1 (*Appendix: Figures 1.0 and onwards*)

Cross-sectional regressions for the years 1985 (model2) and 1987 (model1)

$$\# \, of \, Vehicular \, Fatalities \, in \, 1987 = 1.82153 + 0.48304 * Beertax$$
$$\# \, of \, Vehicular \, Fatalities \, in \, 1985 = 1.77116 + 0.39176 * Beertax$$

1.82153 and 1.77116 are intercept values, they are arbitrary.

0.48304 and 0.39176 are the increase in the *# of vehicular fatalities per 10000* people on average with a one-dollar increase on the tax for a case of beer. The co-efficient values are very similar for both the values, indicating that *beertax* could possibly be a time-invariant variable. Due to the p-values of both models being very small, we can assume that the inclusion of beertax into the model is significant (1 and 5%). However, due to the fact that this is a relatively small data set, the intercept value has a higher weightage in predicting the dependent variable.

Changes regression (model3) for the years 1985 and 1987

As we saw in week 4 slides, the effect of subtracting between two time periods nullifies the state-fixed effects as they are time-invariant.

The xdif co-efficient indicates that if the difference in beertax between two time periods increases by 1 dollar per case of the base, then the change in the fatality rate per 10000 residents will decrease by -0.2168. In layman's terms, the effect of time has a positive impact on beertax on predicting the fatality rate. We must understand that the change of -0.2168 relative to an increase in the xdif by 1 dollar is a very small change. However, the standard deviation value is almost 4 times larger than the coefficient of xdif. Making it a volatile variable.

My logical assumption was that the relation between xdif and ydif would be a positive relationship. However, it must be noted that the sample size is small and the difference in the time period is also very small. This was based on the slides from week 4 notes, where the model predicted that the higher the beer tax, the greater the fatalities.

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


ROADS_ASN3 <- read_dta("fatality_data.dta")
View(ROADS_ASN3)

#Question 1

ROADS_ASN3$mrall = ROADS_ASN3$mrall*10000
ROADS_ASN3$mraidall = ROADS_ASN3$mraidall*10000

summary(ROADS_ASN3$mrall)
```

```
summary(ROADS_ASN3$mrall)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.821   1.624   1.956   2.040   2.418   4.218
```

```r
# Next, we run the two cross-sectional regressions from the notes for 1985 and 1987.

model1 <- lm(mrall ~ beertax, data = subset(ROADS_ASN3, year==1987))

model2 <- lm(mrall ~ beertax, data = subset(ROADS_ASN3, year==1985))

summ(model1, digits=5, robust="HC1")

summ(model2, digits=5, robust="HC1")

y2 = ROADS_ASN3$mrall[ROADS_ASN3$year==1987]
y1 = ROADS_ASN3$mrall[ROADS_ASN3$year==1985]
ydif = y2-y1

x2 = ROADS_ASN3$beertax[ROADS_ASN3$year==1987]
x1 = ROADS_ASN3$beertax[ROADS_ASN3$year==1985]
xdif = x2-x1

difdata <- data.frame(ydif,xdif)

model3 <- lm(ydif ~ xdif, data=difdata)

summ(model3, digits = 4, robust="HC1")
```

```
> summm(model1, digits=5, robust="HC1")
MODEL INFO:
Observations: 48
Dependent Variable: mrall
Type: OLS linear regression

MODEL FIT:
F(1,46) = 8.10501, p = 0.00658
R² = 0.14980
Adj. R² = 0.13132

Standard errors: Robust, type = HC1
------------------------------------------------------------
                      Est.       S.E.      t val.          p
---------------- --------- --------- ---------- ---------
(Intercept)        1.82153   0.11270   16.16235   0.00000
beertax            0.48304   0.13287    3.63557   0.00070
------------------------------------------------------------
```

```
> summ(model2, digits=5, robust="HC1")
MODEL INFO:
Observations: 48
Dependent Variable: mrall
Type: OLS linear regression

MODEL FIT:
F(1,46) = 6.41530, p = 0.01479
R² = 0.12239
Adj. R² = 0.10332

Standard errors: Robust, type = HC1
------------------------------------------------------------
                      Est.       S.E.      t val.          p
---------------- --------- --------- ---------- ---------
(Intercept)        1.77116   0.11601   15.26765   0.00000
beertax            0.39176   0.12848    3.04926   0.00380
```

```
> summ(model3, digits = 4, robust="HC1")
MODEL INFO:
Observations: 48
Dependent Variable: ydif
Type: OLS linear regression

MODEL FIT:
F(1,46) = 0.0816, p = 0.7764
R² = 0.0018
Adj. R² = -0.0199

Standard errors: Robust, type = HC1
------------------------------------------------------------
                      Est.      S.E.     t val.          p
---------------- --------- -------- --------- --------
(Intercept)         0.0823    0.0334     2.4606     0.0177
xdif               -0.2168    0.8417    -0.2576     0.7979
```

Question 2 (*Appendix: Figures 2.0 and onwards*)

The β_1 variable is the same in model4 as we as our model3 because the differencing by fixed time effects is mimiced by the data subset function in model4 of both years. This differencing in model4 is done by using the twoways effects function. Which essentially means that it is the "changes" regression but with fixed state effects as well. However, there are certain factors that make our model4 more precise, which is discussed later on.

The xdif coef (model3) and the beertax coef (model4) are the same, -0.2168. This is because when using the plm package, using the twoways effect for fixed effects of time and entity are the same as computing the Δs between time 1 and time 2 (as seen in the "changes" model due to the time being $T = 2$, fixed effects on time can be adjusted by entering time as a **binary variable**.) Here the state is the only fixed effect, this is needed because observations for the same state are not independent even if the states are drawn through random sampling.

What is notable is to look at the errors for model3 (0.8417) and model4 (0.8330) as they are slightly different. We can see the error values are different for both models, this is because the standard errors for the plm model are **clustered** and are **adjusted** for **heteroskedasticity of error**. This makes the model from question 2 more reliable than the model3 in question 1 just like we discussed before.

```
> #Question 2
>
> ROADS_q2 <- subset(ROADS_ASN3, year==1987 | year==1985)
> ROADS_q2
# A tibble: 96 x 43
      state  year spircons unrate perinc emppop beertax sobapt mormon  mlda     dry yngdrv vmiles breath jaild comserd allmort mralll allnite    mralln allsvn a1517 mra1517
  <dbl+lb> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>  <dbl> <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <dbl>   <dbl>   <dbl> <dbl>   <dbl>     <dbl>  <dbl> <dbl>   <dbl>
 1   1 [AL]  1985     1.28   8.90 11333.   55.3  1.65     30.3  0.376  19.7 23.6     0.211  8727.      0     0       0     882  2.19     146 0.0000363     98    66 3.38e-4
 2   1 [AL]  1987     1.18   7.80 11944    57.5  1.56     30.2  0.411  21   23.8     0.216  9166.      0     0       0    1110  2.72     181 0.0000443    114    94 4.59e-4
 3   4 [AZ]  1985     1.86   6.5  13727.   58.6  0.381     3.76 4.66   21    0       0.188  6771.      0     1       1     893  2.80     150 0.0000471     75    48 3.43e-4
 4   4 [AZ]  1987     1.72   6.20 14241    60.2  0.360     3.63 4.49   21    0       0.169  9371.      0     1       1     937  2.77     172 0.0000508     87    50 3.36e-4
 5   5 [AR]  1985     1.12   8.70 11149.   55.0  0.577    23.1  0.376  21   35.9     0.189  7254.      0     0       0     534  2.26      73 0.0000309     43    45 3.98e-4
 6   5 [AR]  1987     1.01   8.10 11537    56.3  0.545    23.1  0.411  21   39.3     0.164  7666.      0     0       0     639  2.68     110 0.0000461     61    58 5.00e-4
 7   6 [CA]  1985     1.97   7.20 16985.   61.3  0.0953    1.76 1.65   21    0       0.168  7874.      0     0       0    4960  1.88     887 0.0000336    499   302 2.73e-4
 8   6 [CA]  1987     1.78   5.80 17846    63.1  0.0900    1.78 1.63   21    0       0.161  8181.      0     0       0    5504  1.99     944 0.0000341    531   302 2.60e-4
 9   8 [CO]  1985     2.05   5.90 15570.   67.9  0.191     2.30 1.86   21    0.0749  0.192  8092.      1     0       1     579  1.79      96 0.0000297     63    35 2.55e-4
10   8 [CO]  1987     1.78   7.70 15605    64.2  0.180     2.30 1.88   21    0.0581  0.193  8182.      1     0       1     591  1.79      90 0.0000273     64    42 2.92e-4
# … with 86 more rows, and 20 more variables: a1517n <dbl>, mra1517n <dbl>, a1820 <dbl>, a1820n <dbl>, mra1820 <dbl>, mra1820n <dbl>, a2124 <dbl>, mra2124 <dbl>,
#   a2124n <dbl>, mra2124n <dbl>, aidall <dbl>, mraidall <dbl>, pop <dbl>, pop1517 <dbl>, pop1820 <dbl>, pop2124 <dbl>, miles <dbl>, unus <dbl>, epopus <dbl>, gspch <dbl>
>
> model4 <- plm(mrall ~ beertax, model="within",  data = ROADS_q2, effect="twoways", index = c("state","year"))
>
> coef_test(model4, vcov = "CR1", cluster = ROADS_q2$state)
    Coef. Estimate    SE t-stat d.f. p-val (Satt) Sig.
1 beertax   -0.217 0.833  -0.26 7.92        0.801
```

Question 3 (*Appendix: Figures 3.0 and onwards*)

Fixed effects regression for the whole dataset

$$\#\, of\, Vehicular\, Fatalities\ = 3.1443\ -\ 0.6422\ *\ Beertax\ +\ 0.0190\ *\ legal\, drinking\, age$$

In the model, the obvious hypothesis is that increase in the legal drinking age should **decrease** the #VF. However, according to model5, an increase in the legal drinking age by 1 **increased** the #VF, holding beertax constant. 1 dollar increase in beer tax holding the minimum drinking age constant (assume it is 0) causes a decrease in fatality rate by 0.6422.

FIPS Codes:

1985, Ohio - *39*
1987, Missouri - *29*

Intercepts:
- 1985: - 0.1303
- 1987: - 0.0679
- Ohio: - 1.6773
- Missouri: - 1.2981

When we assume fixed effects for state and time, we assume that 1985, Ohio, and 1987, Missouri are all equal to 1 (binary variables). Therefore the co-effecients are added to the intercept value as they are fixed for that particular fixed effect. Giving us the new regression equations:

1985, Ohio
$$\#VF\, for\, 1985,\, Ohio\ = (3.1443\ -\ 0.1303\ -\ 1.6773)\ -\ 0.6422\ *\ Beertax\ +\ 0.0190\ *\ legal\, drinking\, a$$
1987, Missouri
$$\#VF\, for\, 1987,\, Missouri\ = (3.1443\ -\ 0.0679\ -\ 1.2981)\ -\ 0.6422\ *\ Beertax\ +\ 0.0190\ *\ legal\, drinki$$

These equations can be used to compare two states at two different time periods to understand the difference between the #VF and plausible explanations for either i) why these numbers are similar to each other or ii) why they are different.

```
#$Question 3
model5 <- plm(mrall ~ beertax + mlda + factor(state) + factor(year), model = "pooling", data=ROADS_ASN3, index = c("state","year"))
coef_test(model5, vcov = "CR1", cluster = ROADS_ASN3$state)
```

```
> coef_test(model5, vcov = "CR1", cluster = ROADS_ASN3$state)
              Coef. Estimate     SE t-stat  d.f. p-val (Satt) Sig.
1       (Intercept)   3.1443 0.8861  3.549 25.11      0.00156   **
2           beertax  -0.6422 0.3533 -1.818  8.50      0.10447
3              mlda   0.0190 0.0310  0.612 26.19      0.54604
4    factor(state)4  -0.5533 0.4632 -1.195  8.45      0.26467
5    factor(state)5  -0.6607 0.3627 -1.822  8.30      0.10466
6    factor(state)6  -1.5084 0.5368 -2.810  8.33      0.02196    *
7    factor(state)8  -1.4845 0.5029 -2.952  8.32      0.01759    *
8    factor(state)9  -1.8470 0.4912 -3.760  8.45      0.00502   **
9   factor(state)10  -1.3019 0.5166 -2.520  8.36      0.03463    *
10  factor(state)12  -0.2621 0.1794 -1.461  8.47      0.18015
11  factor(state)13   0.5166 0.2880  1.794  8.43      0.10870
12  factor(state)16  -0.6431 0.4479 -1.436  8.65      0.18626
13  factor(state)17  -1.9615 0.5112 -3.837  8.33      0.00461   **
14  factor(state)18  -1.4629 0.4730 -3.093  8.32      0.01414    *
15  factor(state)19  -1.5224 0.4377 -3.478  8.57      0.00750   **
16  factor(state)20  -1.2268 0.4178 -2.936  8.31      0.01805    *
17  factor(state)21  -1.2178 0.5025 -2.423  8.32      0.04049    *
18  factor(state)22  -0.8133 0.3055 -2.662  9.30      0.02526    *
19  factor(state)23  -1.1065 0.3031 -3.651  8.34      0.00605   **
20  factor(state)24  -1.7070 0.4946 -3.451  8.32      0.00817   **
21  factor(state)25  -2.1014 0.4813 -4.366  8.38      0.00214   **
22  factor(state)26  -1.4888 0.3983 -3.738  8.30      0.00535   **
23  factor(state)27  -1.8739 0.4617 -4.059  8.58      0.00314   **
24  factor(state)28  -0.0410 0.2026 -0.203  8.36      0.84430
25  factor(state)29  -1.2981 0.4612 -2.815  8.32      0.02183    *
26  factor(state)30  -0.3335 0.4603 -0.725  8.65      0.48781
27  factor(state)31  -1.5173 0.4264 -3.558  8.36      0.00691   **
28  factor(state)32  -0.6011 0.5001 -1.202  8.32      0.26241
29  factor(state)33  -1.2487 0.3439 -3.631  8.38      0.00617   **
30  factor(state)34  -2.0990 0.5431 -3.865  8.36      0.00439   **
31  factor(state)35   0.4235 0.4360  0.971  8.31      0.35876
32  factor(state)36  -2.1647 0.5291 -4.092  8.52      0.00305   **
33  factor(state)37  -0.3057 0.1205 -2.537  8.79      0.03244    *
34  factor(state)38  -1.6263 0.4350 -3.739  8.31      0.00534   **
35  factor(state)39  -1.6773 0.4350 -3.856  8.31      0.00450   **
36  factor(state)40  -0.5551 0.2527 -2.196  8.31      0.05812    .
37  factor(state)41  -1.1684 0.4996 -2.339  8.32      0.04632    *
38  factor(state)42  -1.7686 0.4802 -3.683  8.32      0.00579   **
39  factor(state)44  -2.2580 0.5164 -4.372  8.37      0.00213   **
40  factor(state)45   0.5342 0.0896  5.961 13.70      < 0.001  ***
41  factor(state)46  -1.0102 0.3427 -2.948  8.30      0.01776    *
42  factor(state)47  -0.8630 0.4656 -1.853  8.43      0.09906    .
43  factor(state)48  -0.8957 0.4217 -2.124  8.59      0.06405    .
44  factor(state)49  -1.1715 0.3142 -3.729  8.30      0.00543   **
45  factor(state)50  -0.9358 0.3525 -2.655  8.94      0.02640    *
46  factor(state)51  -1.2971 0.3310 -3.919  8.30      0.00412   **
47  factor(state)53  -1.6601 0.4951 -3.353  8.32      0.00948   **
48  factor(state)54  -0.8708 0.4248 -2.050  8.67      0.07180    .
49  factor(state)55  -1.7269 0.5216 -3.311  8.68      0.00953   **
50  factor(state)56  -0.1902 0.5605 -0.339  8.75      0.74244
51 factor(year)1983  -0.0812 0.0355 -2.290 46.97      0.02658    *
52 factor(year)1984  -0.0753 0.0449 -1.678 46.83      0.10008
53 factor(year)1985  -0.1303 0.0485 -2.685 46.16      0.01005    *
54 factor(year)1986  -0.0492 0.0618 -0.796 44.34      0.43036
55 factor(year)1987  -0.0679 0.0727 -0.935 41.25      0.35547
56 factor(year)1988  -0.0697 0.0737 -0.946 39.48      0.34979
```

Question 4 (*Appendix: Figures 4.0 and onwards*)

Legend
Y = Spirits consumption (spircons)
X = Tax on a case of beer (beertax)
Breath = Binary variable for preliminary alcohol testing law
Mraidall = Alcohol involved vehicle fatality rate
Z = Effect of drinking test law being enforced on the number of fatalities under the influence of alcohol per 10000 citizens (Etlof = breath * mraidall)
Unrate = rate of unemployment within the US

Logic
I wanted to find the predicted variables for beer tax based on the variable effect of preliminary driving test law on fatality rates across different states. The **breath** law variable (1 for yes and 0 for no) weighs its impact on fatalities caused by driving under the influence. Using this data, I wanted to predict the beer tax of that state (This was an attempt to tackle reverse causality between spircons and beertax). This regression was centered around the hypothesis that the stronger endogeneity of the z variable on predicting the X variable, the stronger the instruments significance in predicting the Y variable. After this predicted variable is procured, I wanted to regress it onto **spircons** and see how **beertax** impacts our **endogenous** variable. We have included a control variable **unrate** to make our model coherent by aiding in making a correlation between our variables of interest. This tries to help with our reverse causality problem.

Results

IV Regression

$$Consumption\ of\ spirits\ =\ -\ 13.087\ +\ 32.473\ *\ Beertax\ -\ 0.272\ *\ Unrate$$

- With an increase in beertax by 1 dollar, the per capita consumption of spirits increases by 32.473
- If unemployment increased by 1%, spirits consumption decreased by 0.272.

Given the test statistic for the Durbin-Wu-Hausman test is significant at the 1 and 5% confidence level. The null hypothesis is that the coefficient on the residuals from the first stage regression is zero in the structural model, or that the variable in question is not endogenous. Therefore, with a significant test statistic, we are able to regress this null, and conclude that there is evidence that the variable being instrumented for is, in fact, correlated with the error term and is endogenous.

The High R squared value is negative, this might be because RSS>TSS, which is an acceptable condition for the IV model.

The Sargon-Hansen (J) test is not applicable to the instrument.

Z is considered an exogenous variable, X variable being the endogenous variable.

Moh Jaiswal 500916860 Professor Laura Turner

1) Determining if an instrument is weak
   a) An instrument is valid if it is
      i) Relevant: $corr(Z, X) \neq 0$, we know that beertax and etlof are correlated somewhat, however uncorrelated with spircons
      ii) Exogenous: $corr(Z, U) = 0$
         (1) Because Z_i is uncorrelated with μ_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with μ_i. This **is** the case here.

The instrument variable 'Z' **does not** have have a direct impact on the dependent variable Y, and therefore is considered as a valid instrument.

Omitted variable bias

The etlof value is a weighted dummy variable as it is the effect of only the (alcohol) fatalities in states that have a driving law, I included this specifically to find the relation between beertax and the different etlof state fixed effect. However, the values for the variable mraidall is very small.

Another approach:

To understand the effect of this variable on the consumption of spiritcons, we can also just include **mraidall** as a control variable and keep the exogenous **breath** variable as the instrument. The instrument will be uncorrelated with the error term in this scenario. The relevance of this variable to beertax is up to open interpretation. There would also not a problem of **reverse causality**.

```
> summary(ivmodel, vcov = sandwich, df = 'Inf', diagnostics = TRUE)

Call:
ivreg(formula = spircons ~ beertax + unrate | etlof + unrate,
    data = ROADS_q4, x = TRUE)

Residuals:
   Min     1Q Median     3Q    Max
-61.65  -4.54   4.57  10.14  15.35

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -13.087    186.108   -0.07     0.94
beertax       32.473    376.274    0.09     0.93
unrate        -0.272      1.524   -0.18     0.86

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1  45      0.01   0.930
Wu-Hausman         1  44      3.18   0.081 .
Sargan             0  NA        NA      NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16 on Inf degrees of freedom
Multiple R-Squared: -521,      Adjusted R-squared: -544
Wald test: 0.0488 on 2 DF,  p-value: 0.976

> anderson.rubin.ci(ivmodel, conflevel = 0.95)
$confidence.interval
[1] "(-Infinity, -0.833600299451049 ] union [ 0.472045551765736 ,Infinity)"
```

```
> #Question 4
>
>
> ROADS_q4 <- subset(ROADS_ASN3, year==1985)
> ROADS_q4$etlof <- ROADS_q4$breath*ROADS_q4$mraidall
> ivmodel <- ivreg(spircons ~ beertax + unrate, ~ etlof + unrate, x = TRUE, data = ROADS_q4)
> robust.se(ivmodel)
```

```
> robust.se(ivmodel)
[1] "Robust Standard Errors"

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.087    186.108   -0.07     0.94
beertax       32.473    376.274    0.09     0.93
unrate        -0.272      1.524   -0.18     0.86

> summary(ivmodel, vcov = sandwich, df = 'Inf', diagnostics = TRUE)
```