Mid Term

Muhammad Bashir, Andrew Issa, Tejaswini Suresh Gaikwad

25 April, 2022

Contents

1 Exercise-I

1.1 Question-1

The model given is

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{u}$$

From M1 and data generating process, we know that $cov(x_1, x_2) \neq 0$ and therefore M2 suffers from endogeneity because error term $\tilde{u}$ contains x_2 and therefore we have omitted variable bias. Therefore, given data generating process, M2 does not satisfy Gauss-Markov Assumption for conditional independence or exogeneity. If we estimate this model using OLS, we would not get causal effect of x_1 on y as $(M2)$ suffers from OVB.

1.2 Question-2

Suppose we estimate M2 using OLS, then we know from Lecture-1

$$plim_{n->\infty}(\hat{\tilde{\beta}}_1) = \beta_1 + \frac{Cov(\tilde{u}, x_1)}{V(x_1)}$$

And we know from GDP that $\tilde{u} = \beta_2 x_2 + u$ and thereore $Cov(\tilde{u}, x_1) = Cov(\beta_2 x_2 + u, x_1) = \beta_2 Cov(x_2, x_1) + Cov(x_1, u)$ and from data generating process we know that $Cov(x_1, u) = 0$ and therefore,

$$plim_{n->\infty}(\hat{\tilde{\beta}}_1) = \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{V(x_1)}$$

Since, from data generating process, we know $Cov(x_1, x_2) \neq 0$ and $\beta_2 \neq 0$, $\tilde{\beta}_1$ does not identify β_1. From M1, we know that true value of $\beta_1 = 1, \beta_2 = 2, Cov(x_1, x_2) = 0.5, V(x_1) = 1$. Hence, the probability limit would be 1+2*0.5=2 instead of true value of 1.

1.3 Preliminaries

Here, I import important libraries.

```
rm(list = ls())
library(MASS)
library(ivreg)
library(latex2exp)
library(haven)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.8
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```
library(readxl)
library(dplyr)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

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

1.4 Defining Variance-Covariance Matrix

In order to define Sigma, I first generate an identity matrix of size 5x5.

```
n=1000
Sigma=diag(5)
Sigma[1,2]=0.5
Sigma[2,1]=0.5
Sigma[1,3]=0.8
Sigma[3,1]=0.8
Sigma[1,4]=0.1
Sigma[4,1]=0.1
Sigma[1,5]=0.4
Sigma[5,1]=0.4
Sigma[2,5]=0.6
Sigma[5,2]=0.6
Sigma
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  1.0  0.5  0.8  0.1  0.4
## [2,]  0.5  1.0  0.0  0.0  0.6
## [3,]  0.8  0.0  1.0  0.0  0.0
## [4,]  0.1  0.0  0.0  1.0  0.0
## [5,]  0.4  0.6  0.0  0.0  1.0
```

1.5 Define Data Generating Process

```
DGP<-function(n,Sigma){
  data=mvrnorm(n, mu=rep(0,5), Sigma)
  data<- as.data.frame(data)
  colnames(data)<-c("x1","x2","z1","z2","z3")
  return(data)
}
```

1.6 Question-3

```
set.seed(12)
data=DGP(n,Sigma)
head(data)
```

```
##            x1          x2          z1          z2          z3
## 1  2.6205749 -0.7844534  2.6228346 -0.6573054 -0.2694350
## 2 -1.6671827 -1.2029625 -1.1709611 -1.3528139 -0.3628528
## 3  0.2004340  2.2933629 -0.4376838 -0.6393087  0.9088283
## 4  0.7675721  0.4460242  0.3539917 -1.5143868  1.3479513
## 5  1.4268094  2.3688818  0.5086201 -1.3231946  1.7675555
## 6  0.9183137 -0.1773647  0.5437102  0.7239065 -0.7810318
```

```r
u=rnorm(n)
data$y=5+data$x1+2*data$x2+u
M2=lm(y~x1,data=data)
M2
```

```
##
## Call:
## lm(formula = y ~ x1, data = data)
##
## Coefficients:
## (Intercept)           x1
##       5.057        1.987
```

```r
m=summary(M2)
m$coefficients
```

```
##             Estimate Std. Error  t value      Pr(>|t|)
## (Intercept) 5.056830 0.06368978 79.39783  0.000000e+00
## x1          1.986908 0.06565561 30.26257 2.877195e-143
```

We observe positive bias in $\hat{\hat{\beta}}_1$ because x_1 and x_2 are positively correlated and β_2 is also positive. It is not close to its population value because we knew that model is not identified correctly. We argued in Q-1 that it should be close to 2 and that is the case.

1.7   Question-4

```r
R=25000
beta_se=matrix(0,R,2)
dim(beta_se)
```
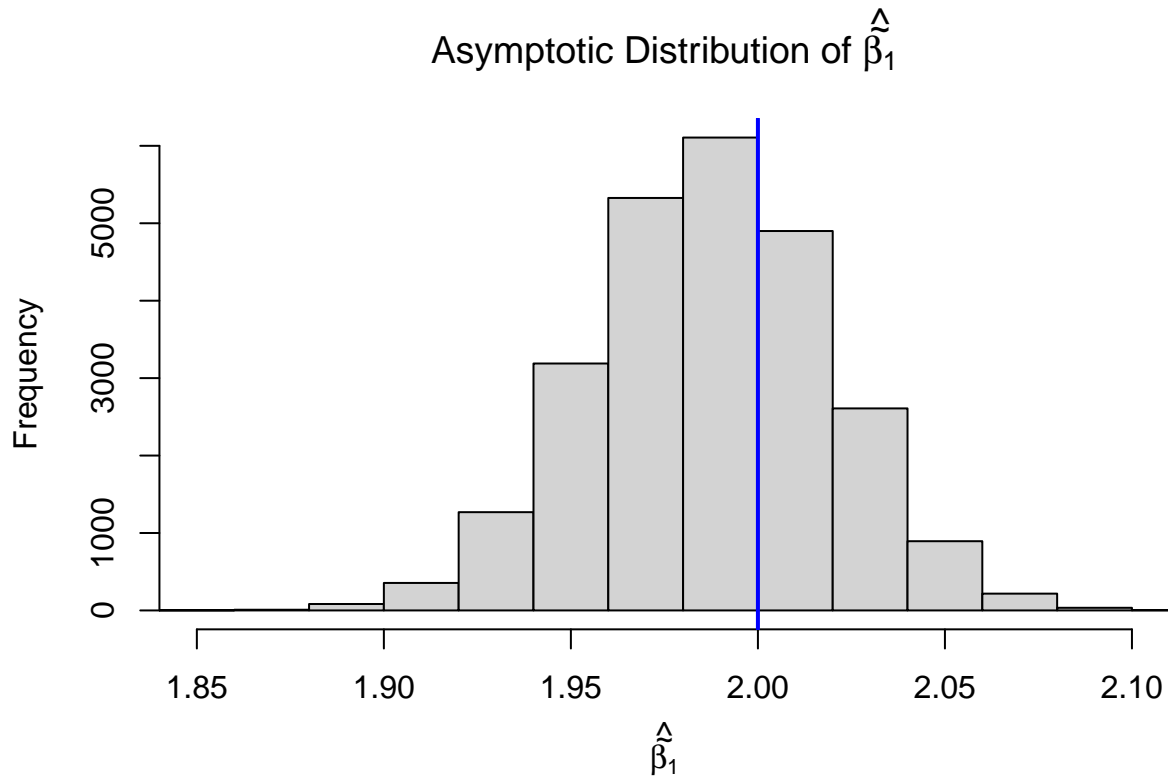
```
## [1] 25000     2
```

```r
set.seed(42)
data=DGP(n,Sigma)
for (i in 1:R){
  set.seed(i)
  u=rnorm(n)
  data$y=5+data$x1+2*data$x2+u
  M2=lm(y~x1,data=data)
  m=summary(M2)
  beta_se[i,1]=m$coefficients[2,1]
  beta_se[i,2]=m$coefficients[2,2]
  }
mean(beta_se[,1])
```

```
## [1] 1.987105
```

```r
mean(beta_se[,2])
```

```
## [1] 0.06329957
```

```
hist(beta_se[,1],main=TeX(r'(Asymptotic Distribution of $\hat{\tilde{\beta_{1}}}$)'),xlab=TeX(r'($\hat{
abline(v=2,col="blue",lwd=2)
```

Asymptotic Distribution of $\hat{\tilde{\beta}}_1$

We see that mean of simulated values is 1.987105 which is very close to 2. We clearly see that this estimator is converging to a value of 2 which is what theory predicts. And yes, it was completely expected. Since model suffers from OVB, this does not converge to true value of β_1 which is 1.

1.8   Question-5

z_1 is best instrument for x_1 in this case because this is highly correlated with x_1 and is not correlated with x_2 and therefore $Cov(\tilde{u}, z_1) = 0$ and hence satisfies the exclusion restriction as well. z_3 is correlated with x_2 and therefore does not satisfy exclusion restriction. z_2 on the other hand, has very weak correlation with the x_1 and therefore is a weak instrument despite the fact that it satisfies exclusion restriction. Hence, z_3 is worst instrument while z_1 is best instrument.

1.9   Question-6

```
set.seed(42)
u=rnorm(n)
data=DGP(n,Sigma)
data$y=5+data$x1+data$x2+u
iv1=ivreg(y~x1|z1,data=data)
iv2=ivreg(y~x1|z2,data=data)
iv3=ivreg(y~x1|z3,data=data)
iv4=ivreg(y~x1|z1+z2+z3,data=data)
stargazer(iv1,iv2,iv3,iv4,type='text',add.lines=list(c("Instrument used","z1","z2","z3","z1,z2 and z3")
```

```
##
## =======================================================================
##                              Dependent variable:
##                          ---------------------------------------
##                                              y
##                          (1)        (2)        (3)        (4)
## ---------------------------------------------------------------------
## x1                     0.963*** 1.148*** 2.576***   1.277***
##                        (0.059)   (0.294)   (0.146)    (0.048)
##
## Constant               4.968*** 4.967*** 4.954***   4.966***
##                        (0.045)   (0.043)   (0.054)    (0.042)
##
## ---------------------------------------------------------------------
## Instrument used               z1        z2        z3   z1,z2 and z3
## Observations               1,000     1,000     1,000      1,000
## R2                         0.484     0.524     0.262      0.542
## Adjusted R2                0.483     0.524     0.262      0.542
## Residual Std. Error (df = 998)  1.417     1.361     1.694      1.335
## =======================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

We see that when we use z_1 as IV, we get 0.963 which is close to true value of 1. z_3 substantially overestimates β_1 because it is positively correlated with x_2. When we use all three together as IVs, it is better than using z_3. When using all three togather, we get minimum standard erros which is expected because this includes z_1 as well. z_1 also has very small SE and is very good IV. z_2 is a relatively weak IV and z_3 does not even identify true β_1.

1.10 Question-7

```r
R=25000
data1=matrix(0,R,2)
data2=matrix(0,R,2)
data3=matrix(0,R,2)
data4=matrix(0,R,2)
set.seed(42)
data=DGP(n,Sigma)
beta_se=matrix(0,R,2)
for (i in 1:R){
  set.seed(i)
  u=rnorm(n)
  data$y=5+data$x1+2*data$x2+u
  iv1=ivreg(y~x1|z1,data=data)
  iv2=ivreg(y~x1|z2,data=data)
  iv3=ivreg(y~x1|z1+z2,data=data)
  data1[i,1]=summary(iv1)$coefficients[2,1]
  data1[i,2]=summary(iv1)$coefficients[2,2]
  data2[i,1]=summary(iv2)$coefficients[2,1]
  data2[i,2]=summary(iv2)$coefficients[2,2]
  data3[i,1]=summary(iv3)$coefficients[2,1]
  data3[i,2]=summary(iv3)$coefficients[2,2]
}
mean(data1[,1])
```

```
## [1] 0.97664
mean(data2[,1])
```
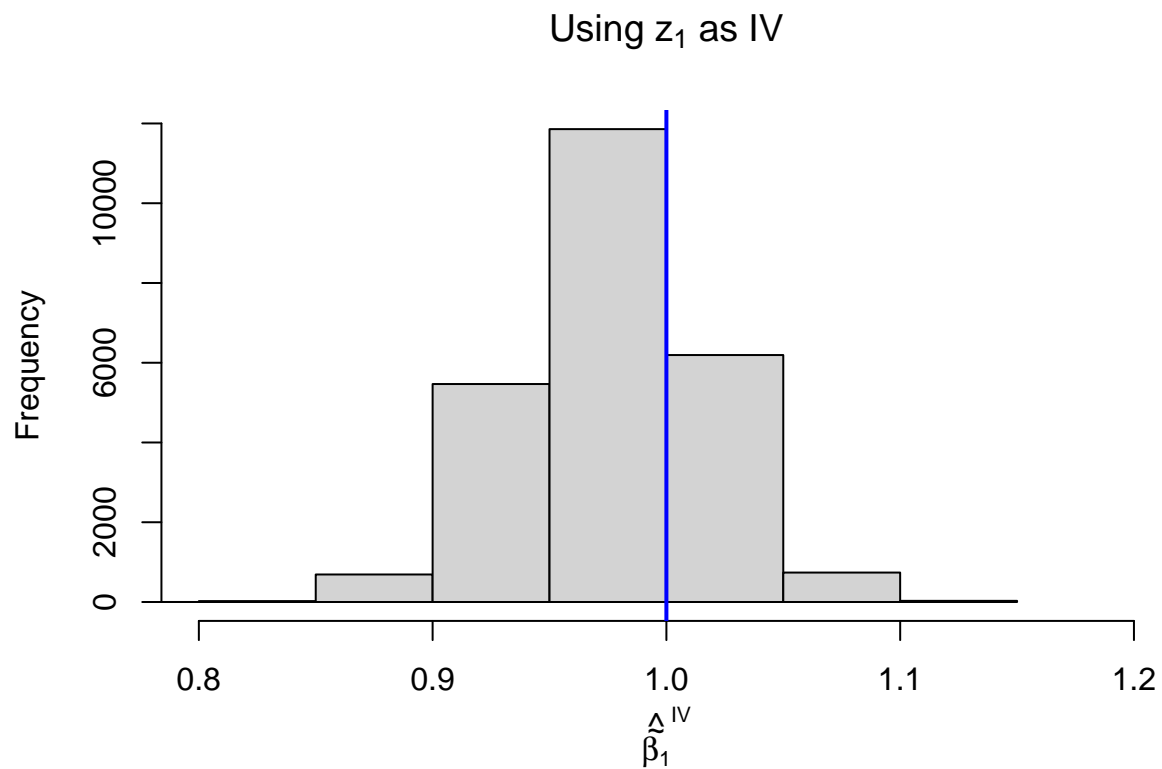
```
## [1] 1.014698
mean(data3[,1])
```

```
## [1] 0.9771157
hist(data1[,1],main=TeX("Using $z_{1}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}^{IV}$)'),breaks=30,
abline(v=1,col="blue",lwd=2)
```

Using z_1 as IV


```
hist(data2[,1],main=TeX("Using $z_{2}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}^{IV}$)'),breaks=30)
abline(v=1,col="blue",lwd=2)
```

Using z_2 as IV

$\hat{\tilde{\beta}}_1^{IV}$
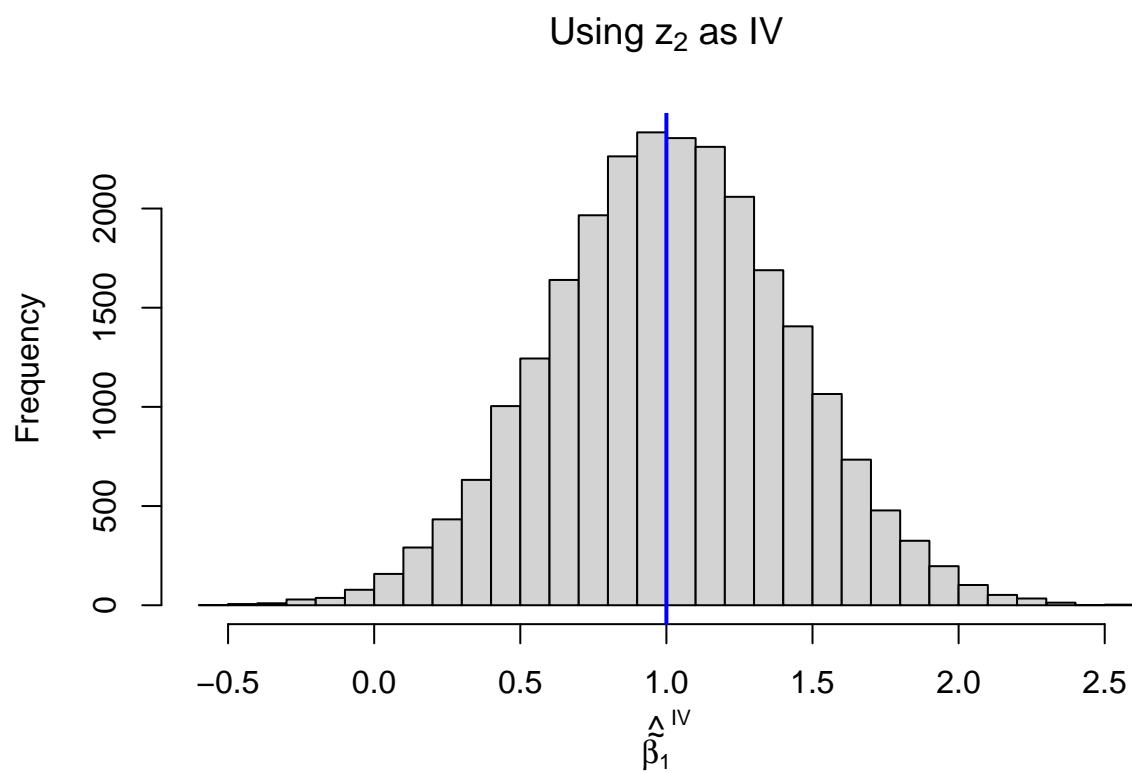
```
hist(data3[,1],main=TeX("Using $z_{1}$ and $z_{2}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}^{IV}$)'
abline(v=1,col="blue",lwd=2)
```

Using z_1 and z_2 as IV

$\hat{\tilde{\beta}}_1^{IV}$
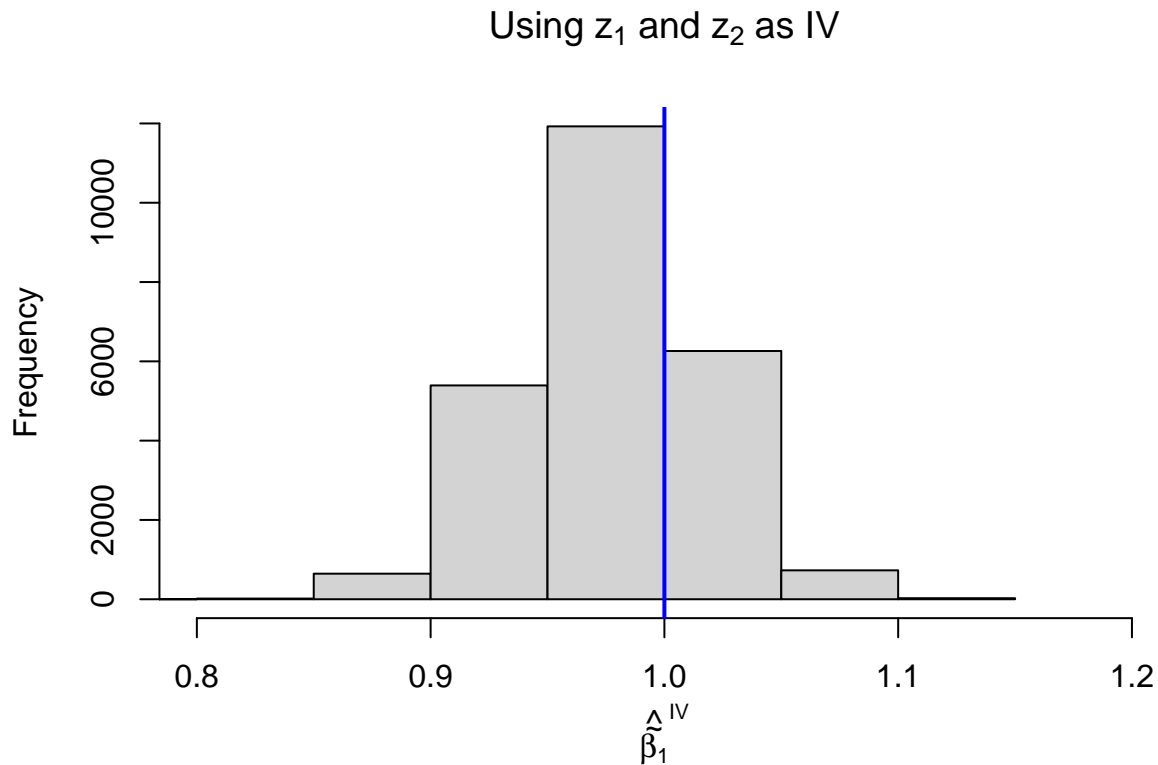
As expected, z_1, z_2 and $z_1 + z_2$ all converge to true value of 1 because they satisfy exclusion restriction. However, z_1 has less variance or dispersion because it is highly correlated with x_1 while z_2 has small correlation which means it has more variance. Using both together as an instrument, reduces variance compared to both seperately. This was expected given Σ matrix.

1.11 Question-8

1.11.1 Question-4'

```
R=25000
n=30
Sigma[1,2]=Sigma[2,1]=0
set.seed(12)
data=DGP(n,Sigma)
beta_se=matrix(0,R,2)
for (i in 1:R){
  set.seed(i)
  u=rnorm(n)
  data$y=5+data$x1+2*data$x2+u
  M2=lm(y~x1,data=data)
  m=summary(M2)
  beta_se[i,1]=m$coefficients[2,1]
  beta_se[i,2]=m$coefficients[2,2]
  }
mean(beta_se[,1])
```

```
## [1] 0.5907062
```
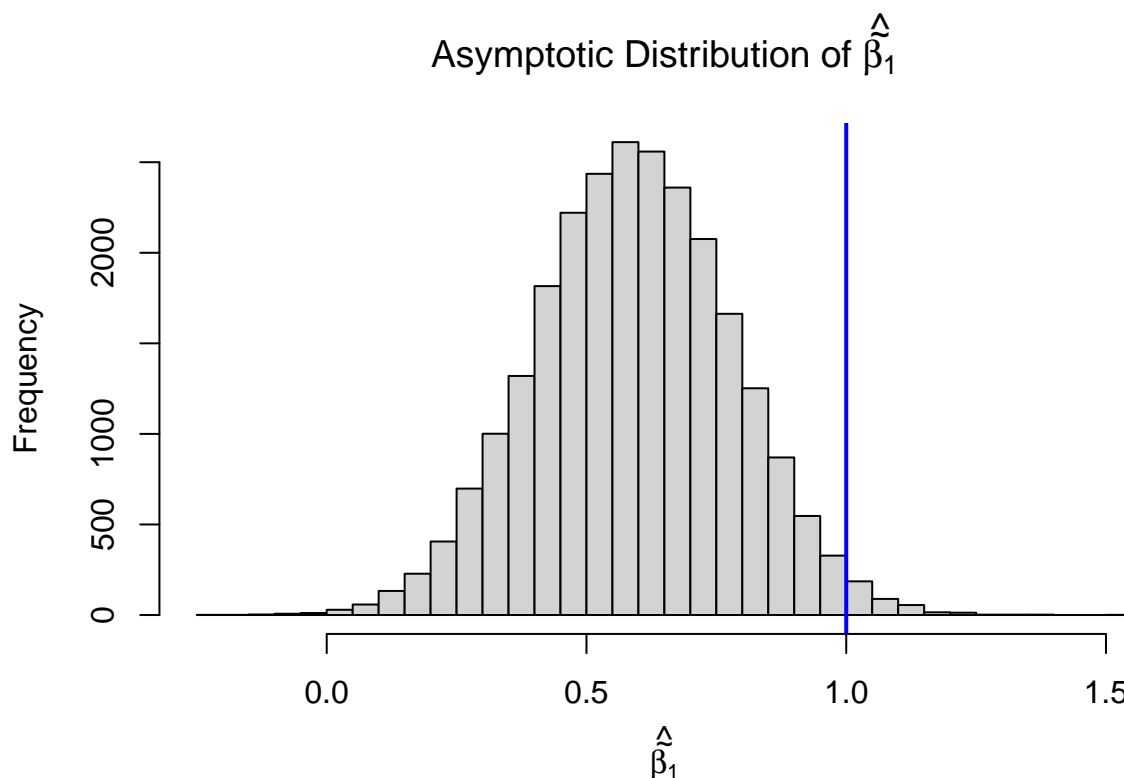
```
mean(beta_se[,2])
```

[1] 0.3253247

```
hist(beta_se[,1],main=TeX(r'(Asymptotic Distribution of $\hat{\tilde{\beta_{1}}}$)'),xlab=TeX(r'($\hat{
abline(v=1,col="blue",lwd=2)
```


Asymptotic Distribution of $\hat{\tilde{\beta}}_1$

Since sample size is very small and x_2 explains major part of y, we do not observe convergence in this small sample to true value of 1. However, in large samples, it does converge. The mean of estimated values is 0.59 and mean estimated SE is 0.32. This was not expected and when I vary sample size, I do observe convergence. This shows that mispecification of form of regression might create problems when sample size is small although Gauss-Markov assumptions hold.

1.12 Question-8-5'

When we set $\sigma_{x_1 x_2} = 0$, then M2 does not suffer from endogeneity because x_1 is not correlated with x_2 and we do not need to use IVs for estimation. Since z_3 is correlated with x_2, it does not satisfy exclusion restriction and therefore cannot act as IV and would not converge to true value. The one with highest correaltion with x_1 would be best IV. Therefore, z_1, z_2 satisfy exclusion while z_1 has highest correlation because covariance is highest and variance in every case is 1 and that is why I just compare covariance instead of correlation. Therefore, z_1 is best IV in this case and while z_3 is worst IV.

1.13 Question-8-6'

```
set.seed(42)
data=DGP(n,Sigma)
u=rnorm(n)
```

```
data$y=5+data$x1+2*data$x2+u
iv1=ivreg(y~x1|z1,data=data)
iv2=ivreg(y~x1|z2,data=data)
iv3=ivreg(y~x1|z3,data=data)
iv4=ivreg(y~x1|z1+z2+z3,data=data)
stargazer(iv1,iv2,iv3,iv4,type='text',column.labels=c("z1","z2","z3","z1,z2 and z3"))
```

```
##
## =============================================================================
## Dependent variable:
## ------------------------------------------
## y
## z1       z2       z3     z1,z2 and z3
## (1)      (2)      (3)        (4)
## -----------------------------------------------------------------------------
## x1                              1.345***  0.274    3.363***   1.960***
##                                (0.456)  (1.579)  (0.869)     (0.403)
##
## Constant                       4.742*** 4.678*** 4.862***    4.779***
##                                (0.430)  (0.518)  (0.583)     (0.437)
##
## -----------------------------------------------------------------------------
## Observations                     30       30       30         30
## R2                              0.379    0.124   -0.138      0.359
## Adjusted R2                     0.356    0.093   -0.179      0.336
## Residual Std. Error (df = 28)   2.352    2.792    3.182      2.388
## =============================================================================
## Note:                                     *p<0.1; **p<0.05; ***p<0.01
```

We see that as expected, z_1 estimates true effect better than other two or three combined, and this was expected because z_1 has highest correlation with x_1 and satisfies exclusion restriction. The IV estimator using z_1 has smallest standard error as well except the case where we use all three as IV, but we see a clear asymptotic bias in that case and little improved SE is simply because we are using more regressors. z_3 has severe asymptotic bias because it does not satisfy exclusion restriction and is correlated with error term.

1.14 Question-8-7'

```
data1=matrix(0,R,2)
data2=matrix(0,R,2)
data3=matrix(0,R,2)
data4=matrix(0,R,2)
set.seed(42)
data=DGP(n,Sigma)
for (i in 1:R){
  set.seed(i)
  u=rnorm(n)
  data$y=5+data$x1+2*data$x2+u
  iv1=ivreg(y~x1|z1,data=data)
  iv2=ivreg(y~x1|z2,data=data)
  iv3=ivreg(y~x1|z1+z2,data=data)
  data1[i,1]=summary(iv1)$coefficients[2,1]
  data1[i,2]=summary(iv1)$coefficients[2,2]
  data2[i,1]=summary(iv2)$coefficients[2,1]
  data2[i,2]=summary(iv2)$coefficients[2,2]
```

```
    data3[i,1]=summary(iv3)$coefficients[2,1]
    data3[i,2]=summary(iv3)$coefficients[2,2]
}
mean(data1[,1])
```

```
## [1] 1.4326
```

```
mean(data2[,1])
```

```
## [1] 0.6844612
```

```
mean(data3[,1])
```

```
## [1] 1.447166
```

```
hist(data1[,1],main=TeX("Using $z_{1}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}$)'),breaks=30)
abline(v=1,col="blue",lwd=2)
```
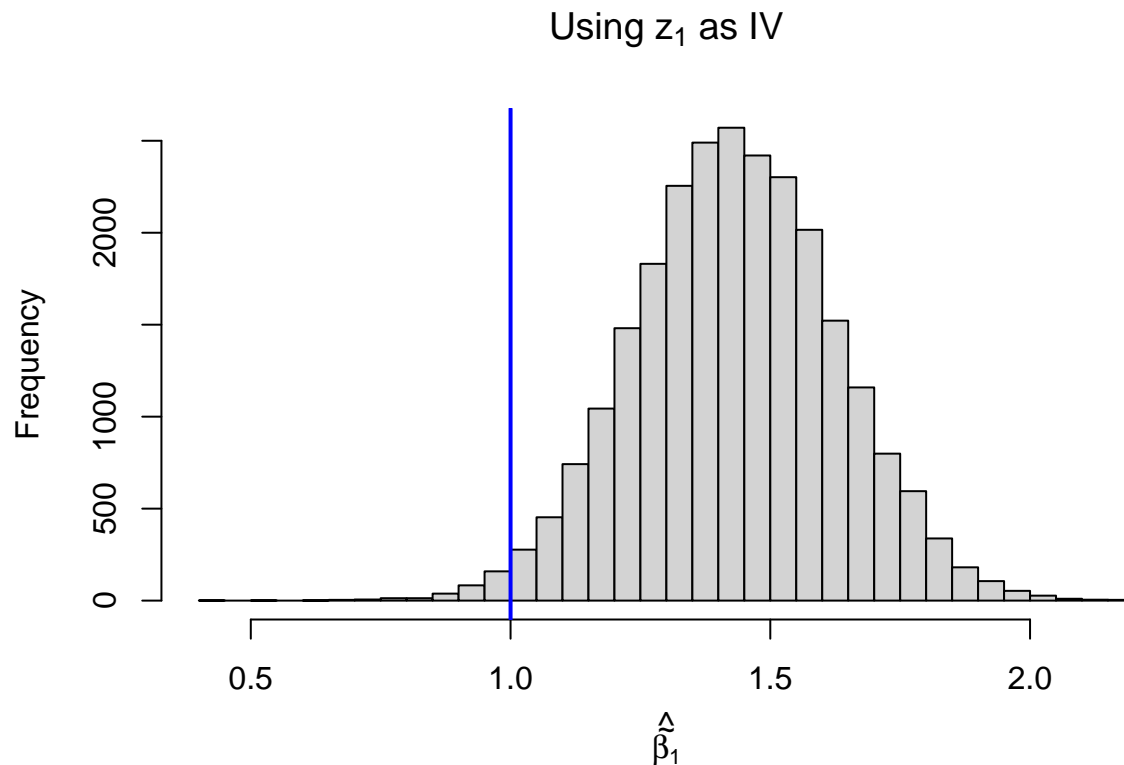
Using z_1 as IV


```
hist(data2[,1],main=TeX("Using $z_{2}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}$)'),breaks=30)
abline(v=1,col="blue",lwd=2)
```

Using z_2 as IV


```r
hist(data3[,1],main=TeX("Using $z_{1}$ and $z_{2}$ as IV"),xlab=TeX(r'($\hat{\tilde{\beta_{1}}}$)'),brea
abline(v=1,col="blue",lwd=2)
```

Using z_1 and z_2 as IV

We observe that z_2 performs relatively better although sample size is small. In general we know that IV estimator is not unbiased in small samples and that would be poten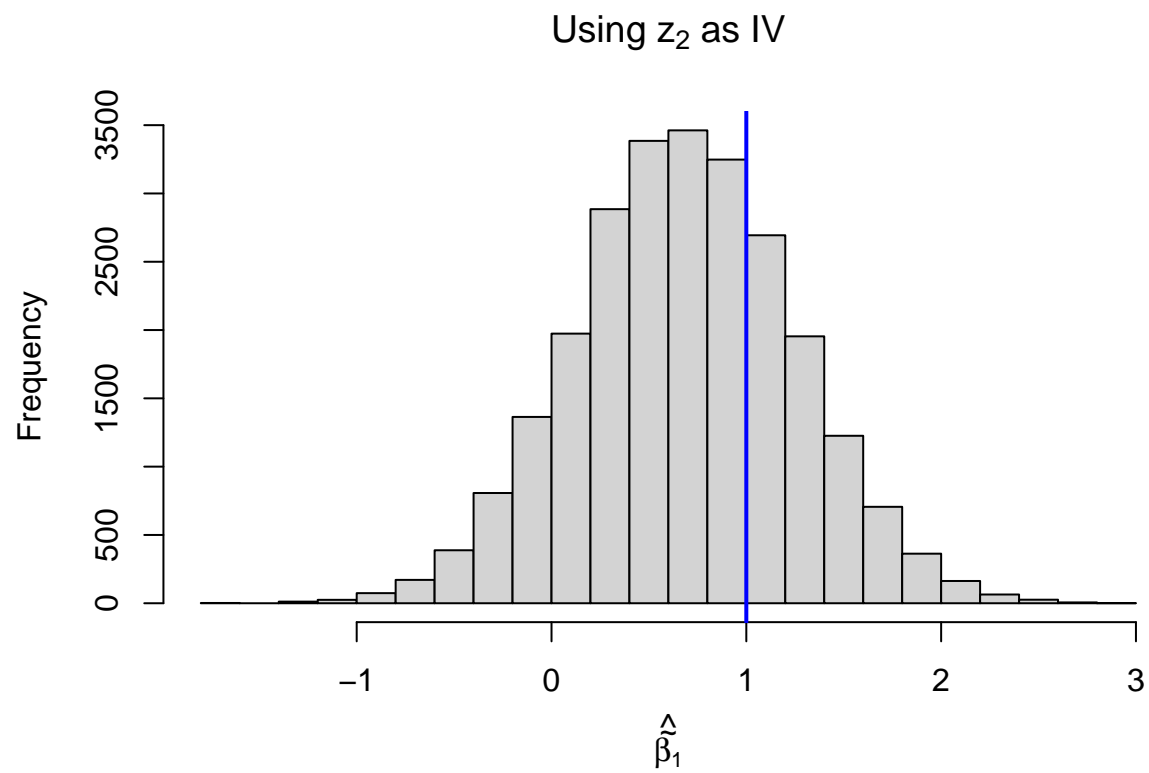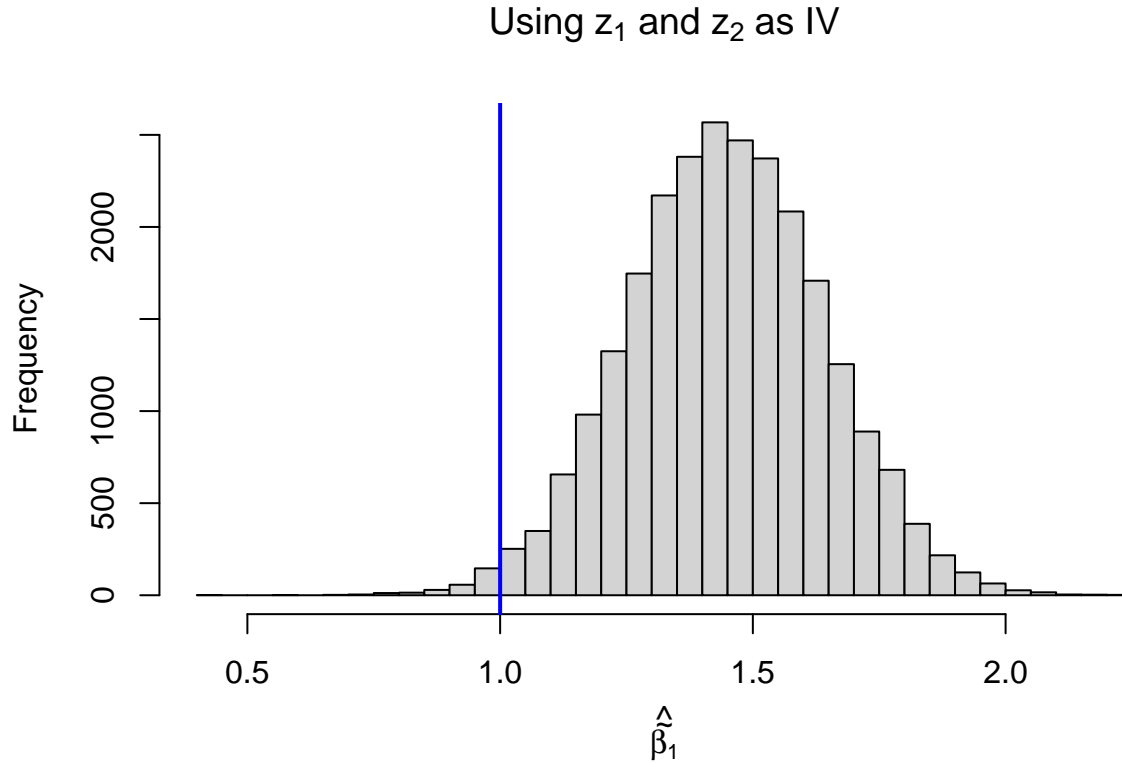tial reason for bias in all three scenarios. However, the smaller bias in z_2 is just because of chance. Since sample size is very small, we should not expect IV estimators to reveal real properties.

2 Exercise-II

2.1 Research Question

- Do counties with intensive social media campaign have less COVID-19 infections compared to control group?

- Do residents of counties with intensive social media campaigns tend to stay more at home compared to control group?

- Do residents of counties with intensive social media campaigns tend to move less compared to control group?

2.2 Question-1

We know that both variables of interest, staying at home due to public campaign by experts and COVID-19 cases are determined by many factors other than influence of experts. Different families and areas differ in their income size, types of jobs, beliefs and past experience, political leanings and other factors that also determine these outcomes of interest i.e public's risk behavior towards COVID, participation in outdoor activities over the vacations and COVID-19 spread. Therefore, in order to find causal impact of campaign, we actually need to run a randomized control trials so that this endogeneity is addressed. Possible research question would be, Does exposure to video messages from experts encouraging for staying at home change

14

mobility decisions and travel around two vacation occasions i.e Thanks Giving and Christmas Holiday? The authors needed to randomize it at zip code or state level given enormous similarities in families in a given area and household. This also helps us reduce movement across treatment arms or spill-overs which would lead to severe contaminatin if we randomized it at individual level.

2.3 Question-2

```r
set.seed(1234)
setwd("C:/bashar/Microeconometrics/mid term")
data1=read_dta("fb_movement_data.dta")
data2=read_excel("randomized_sample_thanksgiving.xlsx")
names(data1)[1] <- "county"
data1$county=as.integer(data1$county)
data2$county=as.integer(data2$county)
data2=data2 %>%group_by(county) %>% mutate(total=n())
data2=data2 %>%group_by(county) %>% mutate(s= sum(urban))
data2$more_urban=data2$s>data2$total/2
data2=subset(data2,select=-c(total,s,urban,treat,zip,state))
data2=data2[!duplicated(data2),]

merge=merge(x=data1,y=data2, by = "county",all.y=TRUE)
rm("data1","data2")
merge$stay_home=100*merge$stay_home
dat=distinct(subset(merge, select =c(county_name,date,movement_ch,stay_home,high_county,more_urban)))
head(dat)
```

```
##   county_name       date movement_ch stay_home high_county more_urban
## 1      Apache 2020-05-05    -0.02260    18.338           0      FALSE
## 2      Apache 2020-05-03    -0.45194    38.484           0      FALSE
## 3      Apache 2020-05-04    -0.03093    17.211           0      FALSE
## 4      Apache 2020-07-16    -0.00131    16.436           0      FALSE
## 5      Apache 2020-03-22    -0.20931    23.837           0      FALSE
## 6      Apache 2020-03-23    -0.22154    22.077           0      FALSE
```

```r
dat=dat[dat$date>"2020-11-22" & dat$date<"2020-11-25",]
dat$left_home=100-dat$stay_home
dat$county_name.f=factor(dat$county_name)
```

We note that in 2020, thanks giving was on 26th November. When we merge, we need to keep entire treatment data. However, we do not have outcome data for 62 counties or 232 zip codes.If this is non-random, our results might be biased. However, authors argue that these areas have very small number of facebook users and therefore would not substainally effect our estimate of average treat effect.

I use the variable more_urban as control because this controls for the fact if a county is overwhelmingly urban or not. This is important because we know COVID-19 infections behave differently between urban and rural areas because there is more interaction in urban areas. Moreover, people in rural and urban counties may be systematically different which comes up when we address treatment effect hetrogeneity. The important variable to use as control in expaiming share of individuals staying home would be county specific fixed effects because poeple in different counties may have different behaviours regarding going out.

2.4 Question-3

```r
model1 =lm(movement_ch~high_county,data=dat)
model2=lm(stay_home~high_county,data=dat)
```

15

```r
model3=lm(stay_home~high_county+more_urban,data=dat)
stargazer(model1,model2,model3,dep.var.labels=c("Movement","Stay Home","Stay Home"),keep=c("high_county
```

```
##
## Treatment effects with traditional standard errors
## ==========================================================================
##                                          Dependent variable:
##                           ------------------------------------------------
##                               Movement              Stay Home
##                                 (1)            (2)              (3)
## ------------------------------------------------------------------------
## High Intensity County          -0.007         0.080            0.065
##                               (0.005)        (0.148)          (0.147)
##
## County is more urban                                          0.630***
##                                                               (0.149)
##
## Constant                      -0.064***      19.412***        19.154***
##                               (0.004)        (0.106)          (0.121)
##
## ------------------------------------------------------------------------
## Observations                   1,382          1,382            1,382
## R2                             0.001          0.0002           0.013
## Adjusted R2                    0.001          -0.001           0.012
## Residual Std. Error    0.096 (df = 1380)   2.751 (df = 1380)    2.734 (df = 1379)
## F Statistic            1.768 (df = 1; 1380) 0.293 (df = 1; 1380) 9.082*** (df = 2; 1379)
## ==========================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Following authors, I just observations between '2020-11-22' to '2020-11-22-25' to calculate impact at thanks giving. This is the time we were targeting with our intervention as to how people owuld behave right before thanks giving vacations.

We observe that on average, counties with high intensity treatment tend to have less movement by 0.007. Moreover, the percentage of people staying home goes up by 0.080 percentage points in high intensity counties and this effect reduces to 0.065 when we introduce control of more_urban. However, in all three model specifications, effect is not statistically significant and we cannot say that it is significantly different from 0. This would mean that we would not even get a significant result from HC or clustered standard errors because that would increase standard erros in general.

2.5 Question-4

```r
model1 =lm(movement_ch~high_county,data=dat)
model2=lm(stay_home~high_county,data=dat)
model3=lm(stay_home~high_county+more_urban,data=dat)
model1=coeftest(model1,vcovHC(model1,type=c("HC1")))
model2=coeftest(model2,vcovHC(model2,type=c("HC1")))
model3=coeftest(model3,vcovHC(model3,type=c("HC1")))
stargazer(model1,model2,model3,column.labels=c("Movement","Stay Home","Stay Home"),type='text',keep=c("
        covariate.labels=c("High Intensity County","County is more urban"))
```

```
##
## Using Hetroskedasticity Robust Standard Errors
## ====================================================
##                             Dependent variable:
```

```
##                              ----------------------------
##
##                      Movement  Stay Home Stay Home
##                        (1)        (2)       (3)
## -------------------------------------------------------
## High Intensity County  -0.007     0.080     0.065
##                        (0.005)   (0.148)   (0.147)
##
## County is more urban                        0.630***
##                                             (0.148)
##
## Constant               -0.064*** 19.412*** 19.154***
##                        (0.004)   (0.103)   (0.120)
##
## ======================================================
## ======================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

After introducing hetroskadasticity robust standard errors, the effect is obviously still insignificant and there is little change standard errors.

2.6   Question-5

```r
model1 =lm(movement_ch~high_county,data=dat)
model2=lm(stay_home~high_county,data=dat)
model3=lm(stay_home~high_county+more_urban,data=dat)
model4=lm(stay_home~high_county+county_name.f,data=dat)
model1=coeftest(model1,vcovCL(model1,cluster = ~county_name))
model2=coeftest(model2,vcovCL(model2,cluster = ~county_name))
model3=coeftest(model3,vcovCL(model3,cluster = ~county_name))
model4=coeftest(model4,vcovCL(model4,cluster = ~county_name))
stargazer(model1,model2,model3,model4,keep=c("high_county","more_urban","Constant"),title="Results using
```

```
##
## Results using Cluster Standard Errors
## ==============================================================
##                          Dependent variable:
##               ----------------------------------------
##
##               Movement  Stay Home Stay Home Stay Home
##                 (1)        (2)       (3)       (4)
## --------------------------------------------------------------
## high_county    -0.007     0.080     0.065    -0.194
##                (0.007)   (0.198)   (0.196)   (0.506)
##
## more_urban                          0.630***
##                                    (0.199)
##
## Constant       -0.064*** 19.412*** 19.154*** 16.009***
##                (0.005)   (0.135)   (0.158)   (0.000)
##
## --------------------------------------------------------------
## County Fixed Effects    No         No        No        Yes
## ==============================================================
```

When we cluster standard errors around county of observation, the standard error goesup and therefore result still remains insignificant. We are clustering at county level because other observed factors for observations from a given county are going to be correlated. When we control for county fixed effects, the treatment effect still remains insignificant.

2.7 Question-6

For treatment effect hetrogeneity across rural and urban areas, we add a new regressor as a interaction between urban dummy and treatment dummy. So, our model becomes

$$y_{it} = \beta_0 + \beta_1 * high\_county_i + \tau * urban_i + \epsilon_{it}$$

where y_{it} denotes outcome variable for county i on date t, β_1 is average treatment effect for rural counties while $\beta_1 + \tau$ is treatment effect for urban counties. Here again, we cluster standard errors with respect to county.

```
model1 =lm(movement_ch~high_county+I(high_county*more_urban),data=dat)
model2=lm(stay_home~high_county+I(high_county*more_urban),data=dat)
model1=coeftest(model1,vcovCL(model1,cluster = ~county_name))
model2=coeftest(model2,vcovCL(model2,cluster = ~county_name))
stargazer(model1,model2,title="Results using Cluster Standard Errors",type='text',add.lines = list(c("N
```

```
##
## Results using Cluster Standard Errors
## ===========================================================
##                            Dependent variable:
##                        ----------------------------
##
##                          Movement        Stay Home
##                            (1)              (2)
## -----------------------------------------------------------
## high_county                0.007           -0.235
##                           (0.008)          (0.234)
##
## I(high_county * more_urban)  -0.033***      0.727***
##                           (0.009)          (0.279)
##
## Constant                   -0.064***       19.412***
##                           (0.005)          (0.135)
##
## -----------------------------------------------------------
## N                          1386             1386
## ===========================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

We now observe that although still insignificant, the symbol of treatment effect for rural counties is reversed compared to average effect we were getting before. This shows that treatment actually had opposite results on rural areas than we expected.

We observe that coefficient on interaction term is significant for movement and therefore there is hetrogeneity in treatment effect. This also shows that the rural areas actually have positive treatment effect while urban areas have net negative treatment effect. However, treatment effect for rural areas is not significant. Similarly, people in high intensity but rural counties tend to stay home less while urban counties with treatment have increase in people staying home. However, these effects are also not statistically significant.