Assingment 8

Chitresh Kumar

```r
if(!require("pacman")) install.packages("pacman")
pacman::p_load(tidyverse, reshape, gplots, ggmap, RStata,haven,
                data.table,margins,pastecs,MASS,lmtest,broom,car,stargazer,sa
ndwich,knitr,dplyr)
search()
theme_set(theme_classic())
```

```r
df<-read_dta('ivreg2.dta')
head(df)
```

```
## # A tibble: 6 x 4
##        x     y     z1      z2
##    <dbl> <dbl>  <dbl>   <dbl>
## 1 -0.965  1.16  0.438 -1.17
## 2 -2.33   1.53 -2.51  -1.43
## 3  0.472  4.78 -0.449 -0.0394
## 4 -3.43  -3.58 -0.848  0.530
## 5  0.138  2.14  0.729  0.0836
## 6 -1.53   1.03 -0.638 -0.603
```

PART b

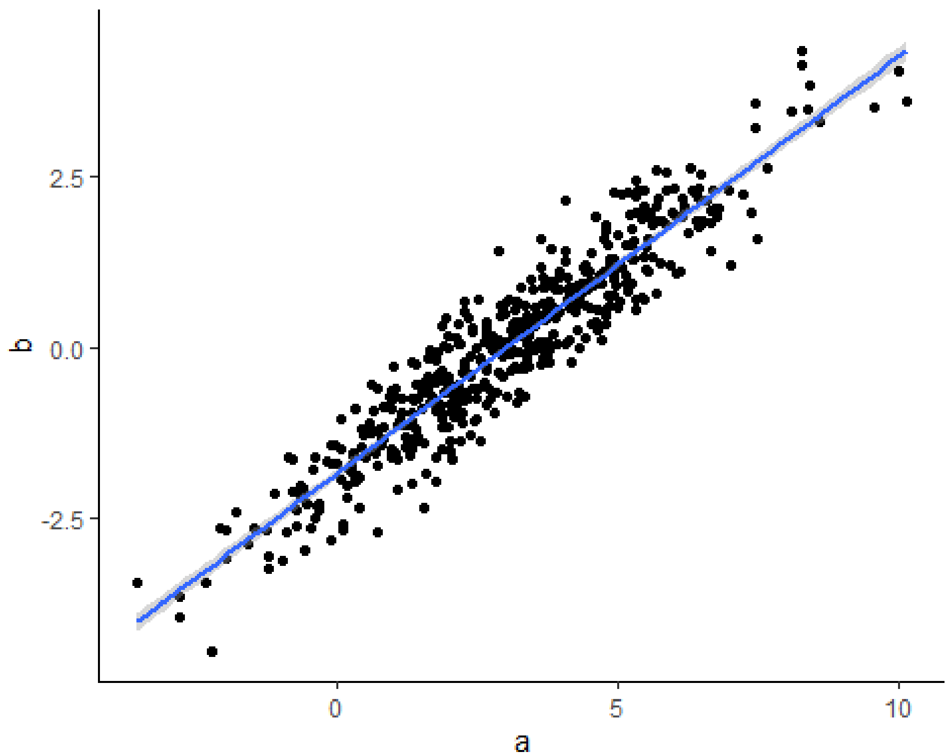
```r
a=df$y
b=df$x


e=a-3-b
```

```r
cor(b,e)
```

```
## [1] 0.65136
```

PART c

```r
e_y=3+b
ggplot(df,aes(x=a,y=b))+
  geom_point()+
  geom_smooth(method="lm")
```



PART D

```
lmdata_1<- df %>% slice(1:10)
lm_1<-lm(y~x,data=lmdata_1)
summary(lm_1)
```

```
##
## Call:
## lm(formula = y ~ x, data = lmdata_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6450 -0.6888 -0.2390  0.4484  1.9556
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7775     0.3608   7.698 5.76e-05 ***
## x             1.3722     0.1727   7.945 4.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 8 degrees of freedom
## Multiple R-squared:  0.8875, Adjusted R-squared:  0.8735
## F-statistic: 63.12 on 1 and 8 DF,  p-value: 4.589e-05
```

```r
lmdata_2<- df %>% slice(1:20)
lm_2<-lm(y~x,data=lmdata_2)
summary(lm_2)

## 
## Call:
## lm(formula = y ~ x, data = lmdata_2)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.83171 -0.52577  0.08304  0.45379  1.75205 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.0169     0.2036   14.81 1.59e-11 ***
## x             1.3876     0.1211   11.46 1.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9056 on 18 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8728 
## F-statistic: 131.4 on 1 and 18 DF,  p-value: 1.053e-09

lmdata_3<- df %>% slice(1:100)
lm_3<-lm(y~x,data=lmdata_3)
summary(lm_3)

## 
## Call:
## lm(formula = y ~ x, data = lmdata_3)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.1199 -0.5289  0.0271  0.5255  1.7940 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3.00783    0.07872   38.21   <2e-16 ***
## x            1.40164    0.05330   26.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.7864 on 98 degrees of freedom
## Multiple R-squared:  0.8759, Adjusted R-squared:  0.8746 
## F-statistic: 691.5 on 1 and 98 DF,  p-value: < 2.2e-16

lmdata_4<- df %>% slice(1:500)
lm_4<-lm(y~x,data=lmdata_4)
summary(lm_4)
```

```
##
## Call:
## lm(formula = y ~ x, data = lmdata_4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20345 -0.51588 -0.01086  0.52412  2.26606
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.01825    0.03410    88.5   <2e-16 ***
## x            1.45352    0.02367    61.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7624 on 498 degrees of freedom
## Multiple R-squared:  0.8833, Adjusted R-squared:  0.8831
## F-statistic:  3770 on 1 and 498 DF,  p-value: < 2.2e-16
```

PART e

```
df$e<-e
cor(df)
```

```
##            x         y           z1          z2            e
## x  1.0000000 0.9398447  0.620821104  0.28948601  0.651359982
## y  0.9398447 1.0000000  0.399870154  0.19965601  0.871374239
## z1 0.6208211 0.3998702  1.000000000 -0.01530765 -0.003447192
## z2 0.2894860 0.1996560 -0.015307651  1.00000000  0.027708992
## e  0.6513600 0.8713742 -0.003447192  0.02770899  1.000000000
```

PART f

```
lmdata_1<- df %>% slice(1:10)
lm_i1<-lm(x~z1,data=lmdata_1)
x_fit<-fitted(lm_i1)
sls_1<-lm(y~x_fit,data=lmdata_1)
summary(sls_1)
```

```
##
## Call:
## lm(formula = y ~ x_fit, data = lmdata_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8950 -1.4894 -0.5804  2.4329  3.0017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7144     0.8705   3.118   0.0143 *
## x_fit         1.0640     0.5142   2.069   0.0723 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.733 on 8 degrees of freedom
## Multiple R-squared:  0.3486, Adjusted R-squared:  0.2672
## F-statistic: 4.282 on 1 and 8 DF,  p-value: 0.07231
```

```
lmdata_2<- df %>% slice(1:20)
lm_i2<-lm(x~z1,data=lmdata_2)
x_fit2<-fitted(lm_i2)
sls_2<-lm(y~x_fit2,data=lmdata_2)
summary(sls_2)
```

```
##
## Call:
## lm(formula = y ~ x_fit2, data = lmdata_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7627 -1.2284  0.2782  1.3325  3.6378
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0810     0.5024   6.132 8.61e-06 ***
## x_fit2        1.0263     0.3952   2.597   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.225 on 18 degrees of freedom
## Multiple R-squared:  0.2726, Adjusted R-squared:  0.2322
## F-statistic: 6.745 on 1 and 18 DF,  p-value: 0.01821
```

```
lmdata_3<- df %>% slice(1:100)
lm_i3<-lm(x~z1,data=lmdata_3)
x_fit3<-fitted(lm_i3)
sls_3<-lm(y~x_fit3,data=lmdata_3)
summary(sls_3)
```

```
##
## Call:
## lm(formula = y ~ x_fit3, data = lmdata_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8057 -1.4028  0.2217  1.5811  4.0127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9771     0.2058  14.463  < 2e-16 ***
## x_fit3        0.9363     0.2217   4.223 5.41e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.053 on 98 degrees of freedom
## Multiple R-squared:  0.1539, Adjusted R-squared:  0.1453
## F-statistic: 17.83 on 1 and 98 DF,  p-value: 5.408e-05
```

```
lmdata_4<- df %>% slice(1:500)
lm_i4<-lm(x~z1,data=lmdata_4)
x_fit4<-fitted(lm_i4)
sls_4<-lm(y~x_fit4,data=lmdata_4)
summary(sls_4)
```

```
##
## Call:
## lm(formula = y ~ x_fit4, data = lmdata_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8886 -1.4938  0.0333  1.3726  7.6623
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.03150    0.09153  33.119   <2e-16 ***
## x_fit4       0.99613    0.10232   9.736   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.046 on 498 degrees of freedom
## Multiple R-squared:  0.1599, Adjusted R-squared:  0.1582
## F-statistic: 94.78 on 1 and 498 DF,  p-value: < 2.2e-16
```

PART g

```
lmdata_g1<- df %>% slice(1:10)
lm_g1<-lm(x~z2,data=lmdata_g1)
x_fitg1<-fitted(lm_g1)
sls_g1<-lm(y~x_fitg1,data=lmdata_g1)
summary(sls_g1)
```

```
##
## Call:
## lm(formula = y ~ x_fitg1, data = lmdata_g1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9100 -1.5470 -0.7599  2.0799  5.8201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.892      3.697   0.512    0.623
## x_fitg1       -2.950     17.282  -0.171    0.869
```

```
##
## Residual standard error: 3.38 on 8 degrees of freedom
## Multiple R-squared:  0.00363,    Adjusted R-squared:  -0.1209
## F-statistic: 0.02915 on 1 and 8 DF,  p-value: 0.8687
```

```
lmdata_g2<- df %>% slice(1:20)
lm_g2<-lm(x~z2,data=lmdata_g2)
x_fitg2<-fitted(lm_g2)
sls_g2<-lm(y~x_fitg2,data=lmdata_g2)
summary(sls_g2)
```

```
##
## Call:
## lm(formula = y ~ x_fitg2, data = lmdata_g2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8580 -1.4576  0.2981  1.5408  5.0120
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2433     0.7499   4.325 0.000408 ***
## x_fitg2       0.1110     2.6571   0.042 0.967140
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.609 on 18 degrees of freedom
## Multiple R-squared:  9.693e-05, Adjusted R-squared:  -0.05545
## F-statistic: 0.001745 on 1 and 18 DF,  p-value: 0.9671
```

```
lmdata_g3<- df %>% slice(1:100)
lm_g3<-lm(x~z2,data=lmdata_g3)
x_fitg3<-fitted(lm_g3)
sls_g3<-lm(y~x_fitg3,data=lmdata_g3)
summary(sls_g3)
```

```
##
## Call:
## lm(formula = y ~ x_fitg3, data = lmdata_g3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8871 -1.3593 -0.0861  1.5005  5.3739
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9902     0.2137  13.996  < 2e-16 ***
## x_fitg3       1.1349     0.3542   3.204  0.00183 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.124 on 98 degrees of freedom
## Multiple R-squared:  0.09483,    Adjusted R-squared:  0.08559
## F-statistic: 10.27 on 1 and 98 DF,  p-value: 0.001828
```

```
lmdata_g4<- df %>% slice(1:500)
lm_g4<-lm(x~z2,data=lmdata_g4)
x_fitg4<-fitted(lm_g4)
sls_g4<-lm(y~x_fitg4,data=lmdata_g4)
summary(sls_g4)
```

```
##
## Call:
## lm(formula = y ~ x_fitg4, data = lmdata_g4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8896 -1.3373 -0.1368  1.4042  7.5446
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.02946    0.09804  30.901  < 2e-16 ***
## x_fitg4      1.06665    0.23458   4.547 6.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.187 on 498 degrees of freedom
## Multiple R-squared:  0.03986,    Adjusted R-squared:  0.03793
## F-statistic: 20.68 on 1 and 498 DF,  p-value: 6.839e-06
```

PART h

```
data_1<- df %>% slice(1:10)
comb_1<-lm(x~z2+z1,data=data_1)
fit_1<-fitted(comb_1)
sls_lm1<-lm(y~fit_1,data=data_1)
summary(sls_lm1)
```

```
##
## Call:
## lm(formula = y ~ fit_1, data = data_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0639 -1.3528 -0.4129  2.4800  2.9570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7114     0.8746   3.100   0.0147 *
## fit_1         1.0491     0.5140   2.041   0.0755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.745 on 8 degrees of freedom
## Multiple R-squared:  0.3425, Adjusted R-squared:  0.2603
## F-statistic: 4.166 on 1 and 8 DF,  p-value: 0.07553

data_2<- df %>% slice(1:20)
comb_2<-lm(x~z2+z1,data=data_2)
fit_2<-fitted(comb_2)
sls_lm2<-lm(y~fit_2,data=data_2)
summary(sls_lm2)

##
## Call:
## lm(formula = y ~ fit_2, data = data_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9387 -0.9457  0.3604  1.4182  3.6502
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0852     0.5045   6.115 8.92e-06 ***
## fit_2         1.0026     0.3924   2.555   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.235 on 18 degrees of freedom
## Multiple R-squared:  0.2661, Adjusted R-squared:  0.2254
## F-statistic: 6.528 on 1 and 18 DF,  p-value: 0.01989

data_3<- df %>% slice(1:100)
comb_3<-lm(x~z2+z1,data=data_3)
fit_3<-fitted(comb_3)
sls_lm3<-lm(y~fit_3,data=data_3)
summary(sls_lm3)

##
## Call:
## lm(formula = y ~ fit_3, data = data_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1040 -1.4575  0.2193  1.5744  3.9015
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.9808     0.1962  15.191  < 2e-16 ***
## fit_3         0.9921     0.1833   5.413 4.41e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.958 on 98 degrees of freedom
## Multiple R-squared:  0.2302, Adjusted R-squared:  0.2223
## F-statistic:  29.3 on 1 and 98 DF,  p-value: 4.407e-07
```

```r
data_4<- df %>% slice(1:500)
comb_4<-lm(x~z2+z1,data=data_4)
fit_4<-fitted(comb_4)
sls_lm4<-lm(y~fit_4,data=data_4)
summary(sls_lm4)
```

```
##
## Call:
## lm(formula = y ~ fit_4, data = data_4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1205 -1.3817 -0.0526  1.2986  7.3349
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.03113    0.08920   33.98   <2e-16 ***
## fit_4        1.00899    0.08984   11.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.994 on 498 degrees of freedom
## Multiple R-squared:  0.2021, Adjusted R-squared:  0.2005
## F-statistic: 126.1 on 1 and 498 DF,  p-value: < 2.2e-16
```

10.2) $Hours = \beta_1 + \beta_2 Wage + \beta_3 Educ + \beta_4 Age +$
$\beta_5 Kids L6 + \beta_6 Kids 618 +$
$\beta_7 NWifeINC + e.$

(a) We expect β_2 to be positive. As the wage increase, the supply of labor should increase.

The education coefficient may be positive or negative. As we know everyone has different intellectual level and the comfors of doing one work if someone is interested in the work & is educated they will work more. But if one is not interested in the work, They might not work rigorously even though they are educated.

The age variable may be positive or negative depending on the demography of the workforce. The younger population can work more with efficiency while the older people can ~~work~~ ~~less~~ complete less work how with efficiency.

The coefficient of KIDSL6 & KIDS618 should be negative. since the married women are expected to focus more on their child than working.

NWIFEINC → The coefficient should be negative. Because extra income from other sources will force not to take the jobs because it is less required.

(6) Wage is the income of the labour. and the Hours is the supply of labour. from the lecture, we discussed the demand & supply of coffee. This applies here on the Hours & the Wage of the Labour. The variables Hours & Wage are endogenous.

We have also seen the variable which was omitted in the lecture was Ability. Ability is correlated with wage & education. More educated person will have more ability in turn will earn more wage. So the least square estimator will be biased as well as inconsistent.

(c) To check for instrumental variables,

(i) Exper & $Exper^2$ should not be correlated with regression error term e.

(ii) Exper & $Exper^2$ should be strongly correlated with wage.

(iii) Exper & $Exper^2$ should not explain ~~and~~ or have direct effect on the supply of labor.

Workers with experiences tend to have more salary.

(d) The number of instrumental variable should more than or equal to endogenous variable.

IV $(Exper, Exper^2) \geq$ endogenous (wage)

(e) We should run first regression for 2SLS

~~$Wage = \gamma_1 + \gamma_2 Exper + \gamma_3 Exper^2 +$~~

$$Wage = \gamma_1 + \gamma_2 Educ + \gamma_3 Age + \gamma_4 Kids L6 + \gamma_5 Kids618 + \gamma_6 NWIFEINC + \theta_1 Exper + \theta_2 Exper^2 + e.$$

We will find the fitted value of wage by regressing the first regression and put the value of $\hat{wage}$ in place of the endogenous variable wage.

And then we will apply second regression

10.6) $y = \beta_1 + \beta_2 x + e = 3 + (1 \times x) + e$

$\beta_1 = 3, \ \beta_2 = 1$

(a) $\sigma_x^2 = 2$ Mean $= 0$

$\sigma_e^2 = 1$ Mean $= 0$

$\text{cov}(x, e) = 0.9$

Correlation b/w x & e

$$r_{xe} = \frac{\text{cov}(x, e)}{\sqrt{\sigma_x^2 \cdot \sigma_e^2}}$$

$$= \frac{0.9}{\sqrt{2 \times 1}} = 0.636$$

(b) $y = 3 + 1 \times x + e$

Sample correlation b/w x & e is equal to 0.65136 which is almost equal to 0.636.

(d) for $N = 10$

$$\hat{y} = 2.775 + 1.3772\,x$$

se $\quad 0.3608 \qquad 0.1727$

for $N = 20$

$$\hat{y} = 3.0169 + 1.3876\,x$$

se $\quad 0.2036 \qquad 0.1211$

for $N = 100$

$$\hat{y} = 3.0078 + 1.4016\,x$$

$0.0787 \qquad 0.0533$

for $N = 500$

$$\hat{y} = 3.01825 + 1.4535\,x$$

$0.034 \qquad 0.0236$

β_1, β_2 are not getting closer to the true value of 3 & 1.

$\therefore$ We might have endogenous variable

(e) Sample correlation b/w $z1$ & x is 0.6208 which is strong (>0.5)
sample correlation b/w z_2 & x is 0.2895 ∴ it is weak (< 0.5)
z_1 is more suitable to be used as instrument variable

(f)

for $N = 10$

$\hat{y} = 2.7144 + 1.0640\, x$
SE 0.8705 0.5142

$N = 20$

$\hat{y} = 3.081 + 1.0263\, x$
SE 0.5024 0.3952

for $N = 100$

$\hat{y} = 2.9971 + 0.9363\, x$
SE 0.2058 0.2217

for $N = 500$

$\hat{y} = 3.03150 + 0.99613\, x$
SE 0.0985 0.1023

The interaction variable coefficient
$\beta_1 = 3.03150 \approx 3$ (true parameter)
$\beta_2 = 0.99613 \approx 1$ (")
when we increase the sample size.

(g) $N = 10$

$\hat{y} = 1.892 - 2.950x$

SE 3.697 17.282

$N = 20$

$\hat{y} = 3.2433 + 0.1110x$

SE 0.7499 2.6871

$N = 100$

$\hat{y} = 2.9902 + 1.1349x$

SE 0.2137 0.3542

$N = 500$

$\hat{y} = 3.0294 + 1.0666x$

SE 0.09804 0.2346

Again as we are increasing the sample size, the β coefficients are approaching the true parameter of the regressor.

As we have seen from correlation, Z_1 & x are more correlated compared to Z_2 & x. Z_1 will perform better than Z_2 as we know we need the interaction variable which is stronger in correlation with the regressor.

(h) $N = 10$

$$\hat{y} = 2.7114 + 1.0491x$$
SE $\quad 0.87463 \quad 0.514$

$N = 20$

$$\hat{y} = 3.0852 + 1.0026x$$
SE $\quad 0.5045 \quad 0.2924$

$N = 100$

$$\hat{y} = 2.9808 + 0.9921x$$
SE $\quad 0.1962 \quad 0.1833$

$N = 500$

$$\hat{y} = 3.03113 + 1.00894x$$
SE $\quad 0.0892 \quad 0.0898$

The small increase in sample size is also making the interaction variable converge to true value of the parameter. The combination will perform better and reach the true parameter more quickly.