Jeffrey Chang
jlc190004

Question 10.2

The labor supply of married women has been a subject of a great deal of economic research. Consider the following supply equation specification

HOURS = β_1 + β_2*WAGE + β_3*EDUC + β_4*AGE + β_5*KIDSL6 + β_6*KIDS618 + β_7*NWIFEINC + e

where HOURS is the supply of labor, WAGE is hourly wage, EDUC is years of education, KIDSL6 is the number of children in the household who are less than six years old, KIDS618 is the number between 6 and 18 years old, and NWIFEINC is household income from sources other than the wife's employment.

(a) Discuss the signs you expect for each of the coefficients.

- Expecting the WAGE is positive with HOURS, because increase WAGE will cause more quantity of labor supplied.
- The coefficient of EDUC in this supply equation might have two expectation
  (i) more educated workers may have trending to work more.
  (ii) more educated workers may be more efficient and choose to work less.
- AGE can to positive or negative because between 35 to 45 is the age that has enough experience and full of working energy.
- KIDSL6 & KIDSL618 expected to be negative because parents need to spend time to take care the children.
- NWIFEINC expected to be negative, as increased household income reduces the need for the wife's income.

(b) Explain why this supply equation cannot be consistently estimated by least squares regression.

In this supply equation the variables WAGE and HOURS are called endogenous variables because their values are determined within the system that supply and demand jointly determine the hours and wages. An endogenous variable on the right-hand side of an equation makes the least squares estimator inconsistent. We cannot estimate models for demand or supply using least squares estimation.

(c) Suppose we consider the woman's labor market experience EXPER and its square, EXPER^2, to be instruments for WAGE. Explain how these variables satisfy the logic of instrumental variables.

According to the definition of instrumental variables, the instrumental variables must be correlated with the endogenous variable and uncorrelated with the error term. So, there are a correlation between WAGE and EXPER, and WAGE and EXPER^2. EXPER and EXPER^2 are likely exogenous relative to the supply equation, and uncorrelated with the supply equation error term.

(d) Is the supply equation identified? Explain.

The supply equation is identified. One endogenous variable and two instrumental variables. EXPER and EXPER^2 are instrumental variables, and it is satisfying the requirement L > B.

(e) Describe the steps (not computer commands) you would take to obtain 2SLS estimates.

The first stage regression has the endogenous variable on the left-hand side, and all exogenous and instrumental variables on the right-hand side.

- The first stage regression is:

WAGE = γ_1 + γ_2*EDUC + γ_3*AGE + γ_4*KIDSL6 + γ_5*KIDS618 + γ_6*NWIFEINC +θ_1* EXPER + θ_2*EXPER^2 + v_i

The least squares fitted value is: $\widehat{WAGE}$

- The second stage regression is:

$\widehat{HOURS}$= β'_1 + β'_2*$\widehat{WAGE}$ + β'_3*EDUC + β'_4*AGE + β'_5*KIDSL6 + β'_6*KIDS618 + β'_7*NWIFEINC + e*

Compare the coefficient between the original and modify equations. $\beta_2$will be different than β'_2 and find out whether we are overestimated the effect of WAGE. Calculating the difference between standard error and we can see IV/2SLS estimator is whether efficient or not.

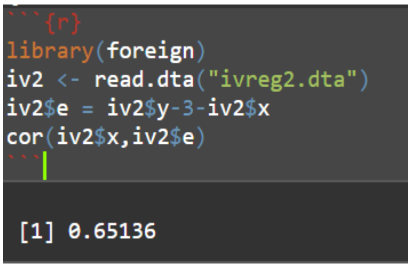
<u>Question 10.6</u>

The 500 values of x, y, z_1, and $z_2$in *ivreg2.dat* were generated artificially. The variable y = β_1 + β_2*x + e = 3 + 1*x + e.

(a) The explanatory variable x follows a normal distribution with mean zero and variance $\sigma^2{}_x$ = 2. The random error e is normally distributed with mean zero and variance $\sigma^2{}_e$ = 1. The covariance between x and e is 0.9. Using the algebraic definition of correlation, determine the correlation between x and e

r_{xe} = cov (x, e)/(var(x)*var(e))^(0.5) = 0.9/(2)^(0.5) = 0.6364

(b) Given the values of y and x, and the values of β_1 = 3 and β_2 = 1, solve for the values of the random disturbances e. Find the sample correlation between x and e and compare it to your answer in (a).

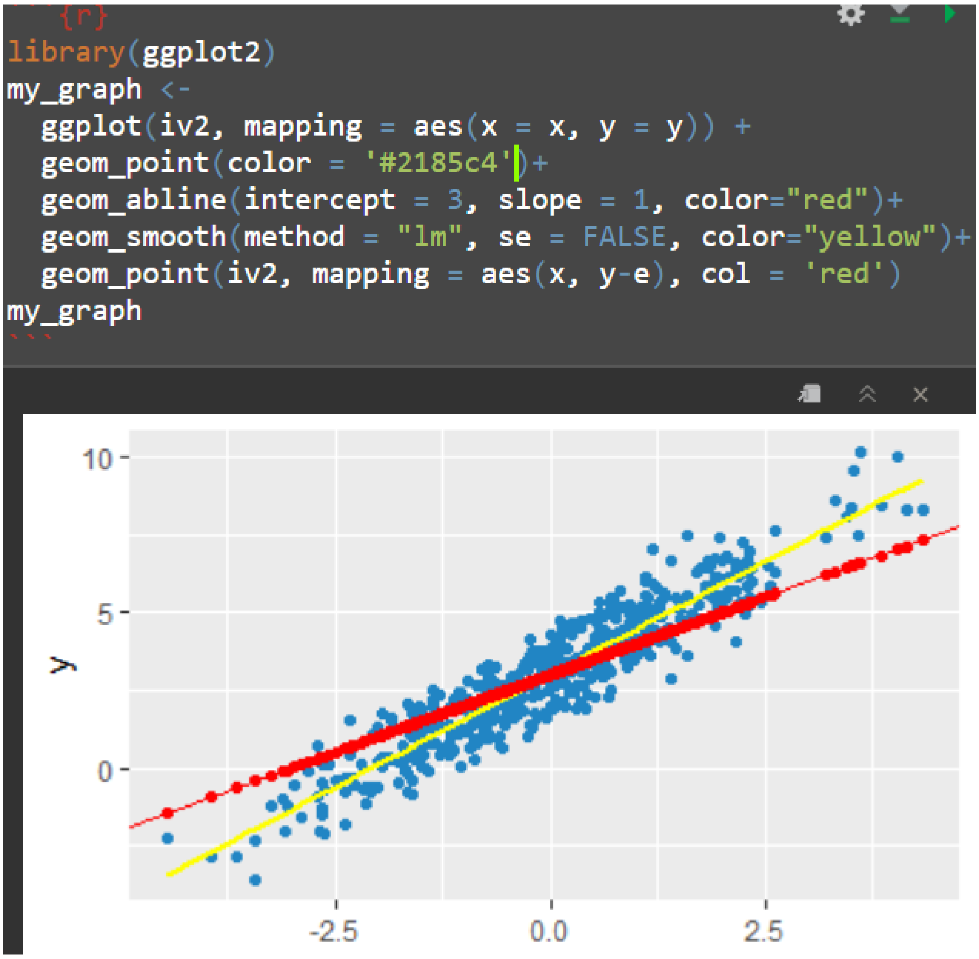


~~~
```{r}
library(foreign)
iv2 <- read.dta("ivreg2.dta")
iv2$e = iv2$y-3-iv2$x
cor(iv2$x,iv2$e)
```

[1] 0.65136
~~~

The sample correlation between x and e is 0.65136, slightly higher than the true value in (a).

(c) In the same graph, plot the value of y against x, and the regression function E(y) = 3 + 1 * x. Note that the data do not fall randomly about the regression function.

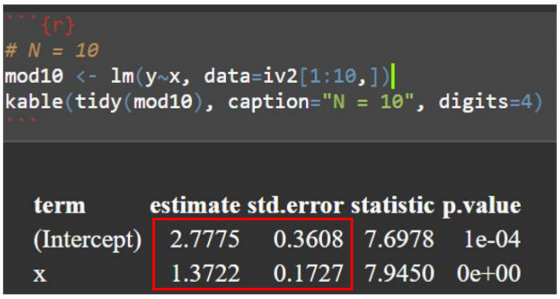


~~~
```{r}
library(ggplot2)
my_graph <-
  ggplot(iv2, mapping = aes(x = x, y = y)) +
  geom_point(color = '#2185c4')+
  geom_abline(intercept = 3, slope = 1, color="red")+
  geom_smooth(method = "lm", se = FALSE, color="yellow")+
  geom_point(iv2, mapping = aes(x, y-e), col = 'red')
my_graph
```
~~~

Red Line: y = 3+1*x, yellow = Regression line of data (***ivreg2.da***).

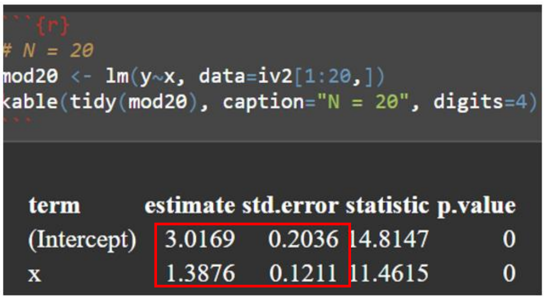
Compare the red and yellow line above, the data tends to fall below the regression line for x < 0 and above the regression line for x > 0

(d) Estimate the regression model y = β_1 + β_2x + e by least squares using a sample consisting of the first N = 10 observations on y and x. Repeat using N = 20, N = 100, and N = 500. What do you observe about the least squares estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not?



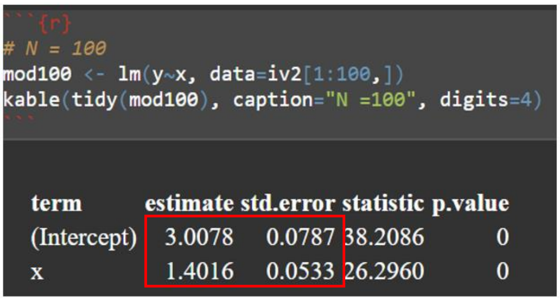
~~~
```{r}
# N = 10
mod10 <- lm(y~x, data=iv2[1:10,])
kable(tidy(mod10), caption="N = 10", digits=4)
```
~~~

```
term          estimate std.error statistic p.value
(Intercept)    2.7775    0.3608   7.6978    1e-04
x              1.3722    0.1727   7.9450   0e+00
```



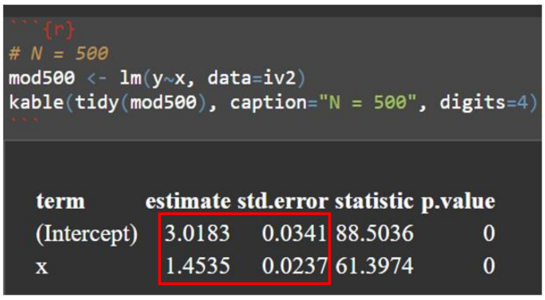
~~~
```{r}
# N = 20
mod20 <- lm(y~x, data=iv2[1:20,])
kable(tidy(mod20), caption="N = 20", digits=4)
```
~~~

```
term          estimate std.error statistic p.value
(Intercept)    3.0169    0.2036  14.8147        0
x              1.3876    0.1211  11.4615        0
```



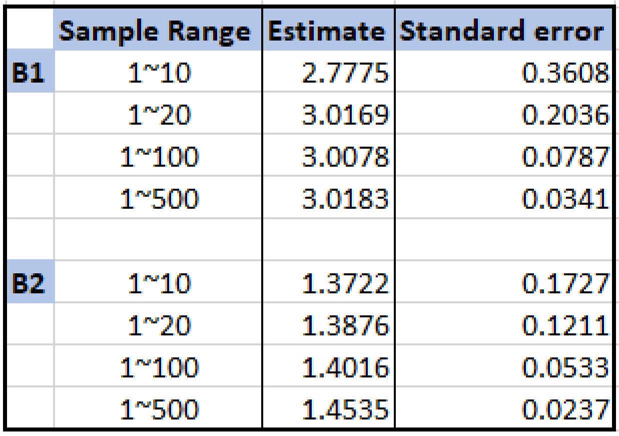
~~~
```{r}
# N = 100
mod100 <- lm(y~x, data=iv2[1:100,])
kable(tidy(mod100), caption="N =100", digits=4)
```
~~~

```
term          estimate std.error statistic p.value
(Intercept)    3.0078    0.0787  38.2086        0
x              1.4016    0.0533  26.2960        0
```



~~~
```{r}
# N = 500
mod500 <- lm(y~x, data=iv2)
kable(tidy(mod500), caption="N = 500", digits=4)
```
~~~

```
term          estimate std.error statistic p.value
(Intercept)    3.0183    0.0341  88.5036        0
x              1.4535    0.0237  61.3974        0
```



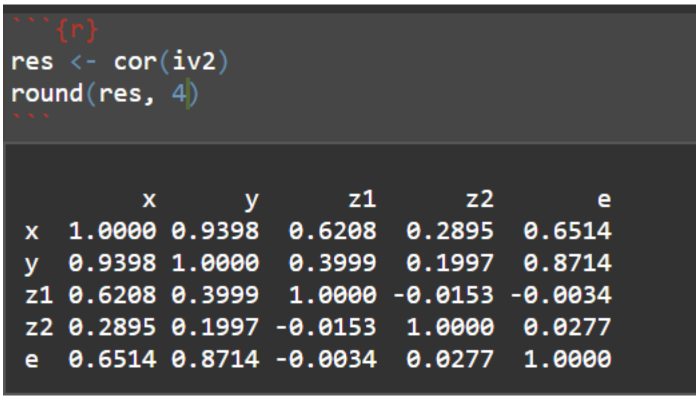
	Sample Range	Estimate	Standard error
B1	1~10	2.7775	0.3608
	1~20	3.0169	0.2036
	1~100	3.0078	0.0787
	1~500	3.0183	0.0341
B2	1~10	1.3722	0.1727
	1~20	1.3876	0.1211
	1~100	1.4016	0.0533
	1~500	1.4535	0.0237

The estimates for β_1: they move closer to the true value as the sample size from 10 to 100. But it does not get closer when the sample increases from 100 to 500.

The estimates for β_2: they move away from the true value as the sample size increases. The estimates do not get closer to the true values as sample size increases because of the inconsistency caused by the correlation between x and e. Furthermore, the inconsistency does not disappear as the sample size increases.

For both $\beta_1$and β_2, the standard errors decrease as the sample size increases.

(e) The variables z_1 and z_2 were constructed to have normal distributions with means zero and variances one, and to be correlated with x but uncorrelated with e. Using the full set of 500 observations, find the sample correlations between z_1, z_2, x and e. Will z_1 and z_2 make good instrumental variables? Why? Is one better than the other? Why?

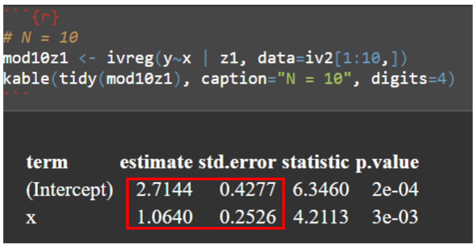


~~~
```{r}
res <- cor(iv2)
round(res, 4)
```
~~~

```
        x      y      z1      z2       e
x  1.0000 0.9398  0.6208  0.2895  0.6514
y  0.9398 1.0000  0.3999  0.1997  0.8714
z1 0.6208 0.3999  1.0000 -0.0153 -0.0034
z2 0.2895 0.1997 -0.0153  1.0000  0.0277
e  0.6514 0.8714 -0.0034  0.0277  1.0000
```

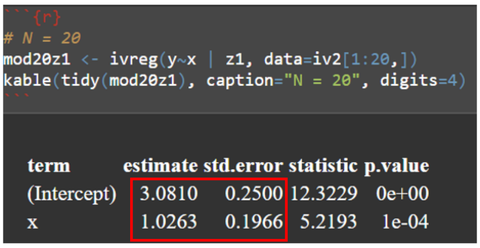
The nonzero correlations between z_1, z_2, x and e:
x and z_1 = (0.6208)
x and z_2 = (0.2894)
e and z_1 = (−0.0034)
e and z_2 = (0.0277)
Both $z_1$and z_2 will be satisfactory instrumental variables.
Because the correlation between x and z_1 (0.6208) is greater than the correlation between x and z_2 (0.2894), z_1 is the better instrumental variable.

(f) Estimate the model y = β_1 + β_2x + e by instrumental variables using a sample consisting of the first N = 10 observations and the instrument z_1. Repeat using N = 20; N = 100, and N = 500. What do you observe about the IV estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not?



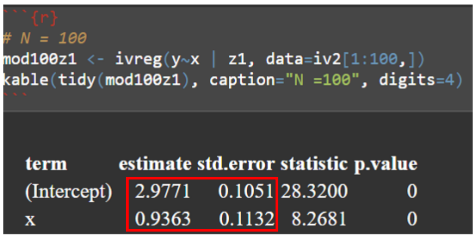
~~~
```{r}
# N = 10
mod10z1 <- ivreg(y~x | z1, data=iv2[1:10,])
kable(tidy(mod10z1), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)    2.7144    0.4277    6.3460   2e-04
x              1.0640    0.2526    4.2113   3e-03
```



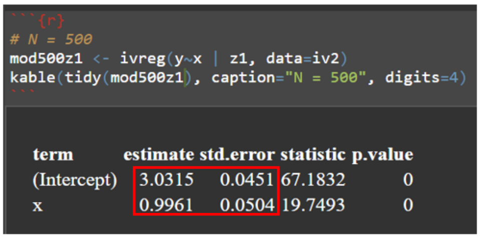
~~~
```{r}
# N = 20
mod20z1 <- ivreg(y~x | z1, data=iv2[1:20,])
kable(tidy(mod20z1), caption="N = 20", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)    3.0810    0.2500   12.3229   0e+00
x              1.0263    0.1966    5.2193   1e-04
```



~~~
```{r}
# N = 100
mod100z1 <- ivreg(y~x | z1, data=iv2[1:100,])
kable(tidy(mod100z1), caption="N =100", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)    2.9771    0.1051   28.3200       0
x              0.9363    0.1132    8.2681       0
```

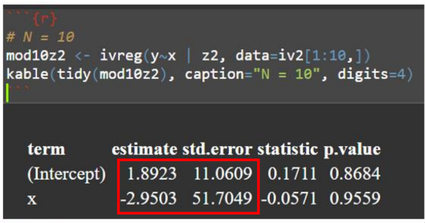


~~~
```{r}
# N = 500
mod500z1 <- ivreg(y~x | z1, data=iv2)
kable(tidy(mod500z1), caption="N = 500", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)    3.0315    0.0451   67.1832       0
x              0.9961    0.0504   19.7493       0
```

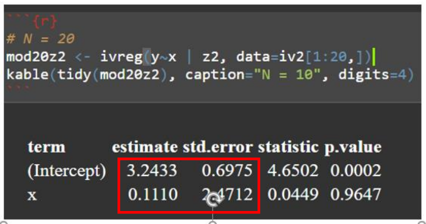
The IV (z_1) estimates for both β_1, β_2 and are getting closer to the true values as the sample size increases, reflecting the consistency of the IV estimator.

(g) Estimate the model y = β_1 + β_2x + e by instrumental variables using a sample consisting of the first N = 10 observations and the instrument z_2. Repeat using N = 20; N = 100, and N = 500. What do you observe about the IV estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not? Comparing the results using z_1 alone to those using z_2 alone, which instrument leads to more precise estimation? Why is this so?



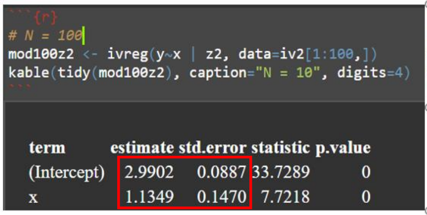
~~~
```{r}
# N = 10
mod10z2 <- ivreg(y~x | z2, data=iv2[1:10,])
kable(tidy(mod10z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   1.8923   11.0609   0.1711  0.8684
x            -2.9503   51.7049  -0.0571  0.9559
```



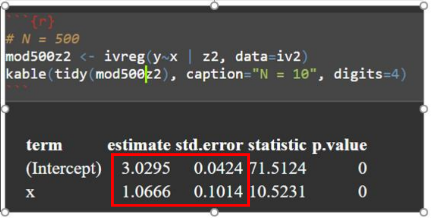
~~~
```{r}
# N = 20
mod20z2 <- ivreg(y~x | z2, data=iv2[1:20,])
kable(tidy(mod20z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   3.2433    0.6975   4.6502  0.0002
x             0.1110    2.4712   0.0449  0.9647
```



~~~
```{r}
# N = 100
mod100z2 <- ivreg(y~x | z2, data=iv2[1:100,])
kable(tidy(mod100z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   2.9902    0.0887  33.7289       0
x             1.1349    0.1470   7.7218       0
```

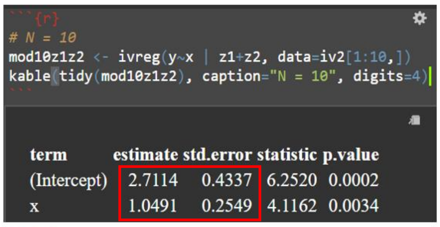


~~~
```{r}
# N = 500
mod500z2 <- ivreg(y~x | z2, data=iv2)
kable(tidy(mod500z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   3.0295    0.0424  71.5124       0
x             1.0666    0.1014  10.5231       0
```

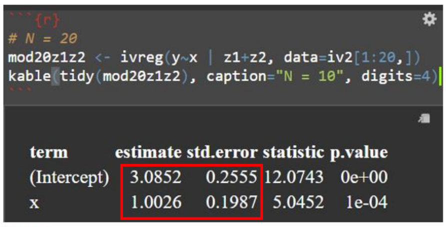
As an instrumental variable (z_2), it gives estimates that are very far away from the true values when the sample sizes are small. (β_1: less than 10 and β_2:less than20). When the sample size is bigger, the estimates move closer to the true values, especially those for β_2. Comparing the results part(f) and part(e), When using only z_1 cause more precise estimation at the smaller sample size. This result happen the correlation between z_1 and x is much higher than the correlation between z_2 and x.
[x and z_1 = (0.6208), x and z_2 = (0.2894)]

(h) Estimate the model y = β_1 + β_2x + e by instrumental variables using a sample consisting of the first N ¼10 observations and the instruments z_1 and z_2. Repeat using N = 20; N = 100, and N = 500. What do you observe about the IV estimates? Are they getting closer to the true values as the sample size increases, or not? If not, why not? Is estimation more precise using two instruments than one, as in parts (f) and (g)?



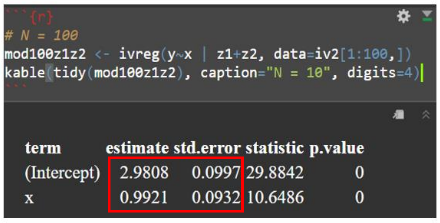
~~~
```{r}
# N = 10
mod10z1z2 <- ivreg(y~x | z1+z2, data=iv2[1:10,])
kable(tidy(mod10z1z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   2.7114    0.4337   6.2520  0.0002
x             1.0491    0.2549   4.1162  0.0034
```



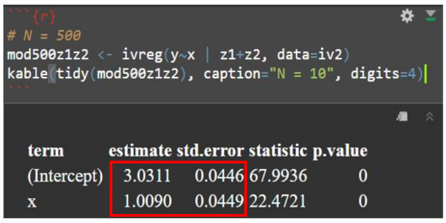
~~~
```{r}
# N = 20
mod20z1z2 <- ivreg(y~x | z1+z2, data=iv2[1:20,])
kable(tidy(mod20z1z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   3.0852    0.2555  12.0743   0e+00
x             1.0026    0.1987   5.0452   1e-04
```



~~~
```{r}
# N = 100
mod100z1z2 <- ivreg(y~x | z1+z2, data=iv2[1:100,])
kable(tidy(mod100z1z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   2.9808    0.0997  29.8842       0
x             0.9921    0.0932  10.6486       0
```



~~~
```{r}
# N = 500
mod500z1z2 <- ivreg(y~x | z1+z2, data=iv2)
kable(tidy(mod500z1z2), caption="N = 10", digits=4)
```
~~~

```
term         estimate std.error statistic p.value
(Intercept)   3.0311    0.0446  67.9936       0
x             1.0090    0.0449  22.4721       0
```

Using instrumental variables $z_1 and\ z_2$, the estimates are getting closer to the true values when the sample size increases. The results, which compare $z_1 + \ z_2$ and only z_1, has been a slight improvement in precision for sample sizes N = 100 and N = 500.