ECON 710A - Problem Set 2

Alex von Hafften*

2/9/2020

1. Suppose $(Y, X, Z)'$ is a vector of random variables such that $Y = \beta_0 + X\beta_1 + U$, $E[U|Z] = 2$ where $Cov(Z, X) \neq 0$ and $E[Y^2 + X^2 + Z^2] < \infty$. Additionally, let $\{(Y_i, X_i, Z_i)'\}_{i=1}^n$ be a random sample from the model with $\hat{Cov}(Z, X) \neq 0$. Recall, the definition from lecture 3: $\hat{\beta}_1^{IV} = \frac{\hat{Cov}(Z,Y)}{\hat{Cov}(Z,X)} = \frac{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)(X_i - \bar{X}_n)}$ and $\hat{\beta}_0^{IV} = \bar{Y} - \bar{X}\hat{\beta}_1^{IV}$.

(i) Does $\hat{\beta}_1^{IV} \to_p \beta_1$?

Yes.

By the LLN and the CMT, $\hat{Cov}(Z, Y) \to_p Cov(Z, Y)$ and $\hat{Cov}(Z, X) \to_p Cov(Z, X)$. Notice that

$$\begin{aligned}
Cov(Z, U) &= E[(E[Z] - Z)(E[U] - U)] \\
&= E[E[U]E[Z] - E[U]Z - UE[Z] + UZ] \\
&= E[E[U|Z]]E[Z] - E[E[E[U|Z]]Z] - E[UE[Z]] + E[UZ] \\
&= 2E[Z] - E[2Z] - E[E[U|Z]E[Z]] + E[E[U|Z]Z] \\
&= 2E[Z] - 2E[Z] - E[2E[Z]] + E[2Z] \\
&= 0
\end{aligned}$$

By the CMT,

$$\begin{aligned}
\hat{\beta}_1^{IV} &\to_p \frac{Cov(Z, Y)}{Cov(Z, X)} \\
&= \frac{Cov(Z, \beta_0 + X\beta_1 + U)}{Cov(Z, X)} \\
&= \frac{Cov(Z, \beta_0) + Cov(Z, X\beta_1) + Cov(Z, U)}{Cov(Z, X)} \\
&= \frac{\beta_1 Cov(Z, X) + Cov(Z, U)}{Cov(Z, X)} \\
&= \beta_1 + \frac{Cov(Z, U)}{Cov(Z, X)} \\
&= \beta_1 + \frac{0}{Cov(Z, X)} \\
&= \beta_1
\end{aligned}$$

*I worked on this problem set with a study group of Michael Nattinger, Andrew Smith, and Ryan Mather. I also discussed problems with Sarah Bass, Emily Case, Danny Edgel, and Katherine Kwok.

(ii) Does $\hat{\beta}_0^{IV} \to_p \beta_0$?

No.

By (i) and the LLN,

$$\begin{aligned}
\hat{\beta}_0^{IV} &= \bar{Y} - \bar{X}\hat{\beta}_1^{IV} \\
&\to_p E[Y] - E[X]\beta_1 \\
&= E[\beta_0 + X\beta_1 + U] - E[X]\beta_1 \\
&= \beta_0 + E[X]\beta_1 + E[U] - E[X]\beta_1 \\
&= \beta_0 + E[E[U|Z]] \\
&= \beta_0 + 2
\end{aligned}$$

2. Consider the simulation model $Y = \beta_0 + X\beta_1 + U$ and $X = \pi_0 + Z\pi_1 + V$ with $E[U|Z] = E[V|Z] = 0$ where $E[Y^2 + X^2 + Z^2] < \infty$. Additionally, let $\{(Y_i, X_i, Z_i)'\}_{i=1}^n$ be a random sample from this model with $\frac{1}{n}\sum_{i=1}^n (Z_i - \bar{Z}_n)^2 > 0$.

(i) Under what conditions (on $\beta_0, \beta_1, \pi_0, \pi_1$) is Z a valid instrument for X?

For Z to be a valid instrument for X, it needs to be exogenous and relevant. For exogeneoity of the instrument, $E[U|Z] = 0$, which we get by assumption. For relevance of the instrument, we need that $Cov(Z, X) \neq 0$:

$$Cov(Z, X) \neq 0 \iff \frac{Cov(Z, X)}{Var(Z)} = \pi_1 \neq 0$$

(ii) Show that $Y = \gamma_0 + Z\gamma_1 + \varepsilon$ with $E[\varepsilon|Z] = 0$ where γ_0, γ_1, and ε are some functions of $\beta_0, \beta_1, \pi_0, \pi_1$, U, and V. In particular show that $\gamma_1 = \pi_1\beta_1$.

$$\begin{aligned}
Y &= \beta_0 + (\pi_0 + Z\pi_1 + V)\beta_1 + U \\
&= \beta_0 + \pi_0\beta_1 + Z\pi_1\beta_1 + V\beta_1 + U \\
&= \gamma_0 + Z\gamma_1 + \varepsilon
\end{aligned}$$

where $\gamma_0 = \beta_0 + \pi_0\beta_1$, $\gamma_1 = \pi_1\beta_1$, and $\varepsilon = V\beta_1 + U$. Notice that $E[\varepsilon|Z] = E[V\beta_1 + U|Z] = E[V|Z]\beta_1 + E[U|Z] = 0$.

(iii) Let $\hat{\gamma}_1$ and $\hat{\pi}_1$ denote the OLS estimators of γ_1 and π_1, respectively. The ratio $\hat{\gamma}_1/\hat{\pi}_1$ is called the "indirect least squares" estimator of β_1. How does it compare to the IV estimator of β_1 that uses Z as an instrument for X?

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{Cov}(Y,Z)}{\hat{Cov}(X,Z)} \\
\hat{\gamma}_1 &= \frac{\hat{Cov}(Y,Z)}{\hat{Var}(Z)} \\
\hat{\pi}_1 &= \frac{\hat{Cov}(X,Z)}{\hat{Var}(Z)} \\
\frac{\hat{\gamma}_1}{\hat{\pi}_1} &= \frac{\frac{\hat{Cov}(Y,Z)}{\hat{Var}(Z)}}{\frac{\hat{Cov}(X,Z)}{\hat{Var}(Z)}} \\
&= \frac{\hat{Cov}(Y,Z)}{\hat{Cov}(X,Z)} \\
&= \hat{\beta}_1
\end{aligned}$$

(iv) Show that $Y = \delta_0 + X\delta_1 + V\delta_2 + \xi$, $Cov(X,\xi) = Cov(V,\xi) = 0$ where $\delta_0, \delta_1, \delta_2$, and ξ are some functions of $\beta_0, \beta_1, Cov(U,V), Var(V), U, V$. In particular, show that $\delta_1 = \beta_1$.

Let $\delta_2 = \frac{Cov(U,V)}{Var(V)}$ and $\xi = U - V\delta_2$ be the linear project of U onto V: $U = V\delta_2 + \xi$.

$$\begin{aligned}
Cov(X,\xi) &= Cov(\pi_0 + Z\pi_1 + V, U - V\delta_2) \\
&= Cov(Z\pi_1, U) + Cov(Z\pi_1, V\delta_2) + Cov(V,U) + Cov(V, -V\delta_2) \\
&= \pi_1 Cov(Z,U) - \pi_1\delta_2 Cov(Z,V) + Cov(V,U) - \delta_2 Var(V) \\
&= 0 + Cov(V,U) - \frac{Cov(U,V)}{Var(V)} Var(V) \\
&= 0
\end{aligned}$$

because $E[U|Z] = E[V|Z] \implies Cov(Z,U) = Cov(Z,V) = 0$. Furthermore,

$$Cov(V,\xi) = Cov(V, U - V\delta_2) = Cov(V,U) - \delta_2 Var(V) = 0$$

Thus,

$$Y = \beta_0 + X\beta_1 + (V\delta_2 + \xi) = \delta_0 + X\delta_1 + V\delta_2 + \xi$$

where $\delta_0 = \beta_0$ and $\delta_1 = \beta_1$.

(v) Let $\hat{V}_i = X_i - \hat{\pi}_0 - Z_i\hat{\pi}_1$ where $\hat{\pi}_0$ is the OLS estimator of π_0. Furthermore, let $\hat{\delta}_1$ be the OLS estimator from a regression of Y_i on $(1, X_i, \hat{V}_i)$; this estimator is called the "control variable" estimator. How does it compare to the IV estimator of β_1 that uses Z as an instrument for X?

Consider an OLS regression of X on $\hat{V}$: $X = \alpha_0 + \alpha_1\hat{V} + e$:

$$\begin{aligned}
\hat{\alpha}_1 &= \frac{\hat{Cov}(\hat{V}, X)}{\hat{Var}(\hat{V})} \\
&= \frac{\hat{Cov}(\hat{V}, \hat{V} + \tilde{X})}{\hat{Var}(\hat{V})} \\
&= \frac{\hat{Var}(\hat{V})}{\hat{Var}(\hat{V})} \\
&= 1 \\
\hat{\alpha}_0 &= \bar{X} - \hat{\alpha}_1\bar{\hat{V}} \\
&= \bar{X} \\
\hat{e} &= X - \hat{V}\hat{\alpha}_1 - \hat{\alpha}_0 \\
&= X - \hat{V} - \bar{X} \\
&= \tilde{X} - \bar{X}
\end{aligned}$$

where $\tilde{X} = X - \hat{V} = \hat{\pi}_0 + \hat{\pi}_1 Z$. Thus, applying the Frisch-Waugh Lovell Theorem, we can derive $\hat{\delta}_1$ by regressing Y on $\tilde{X}$ and a constant term:

$$\begin{aligned}
\hat{\delta}_1 &= \frac{\hat{Cov}(\tilde{X}, Y)}{\hat{Var}(\tilde{X})} \\
&= \frac{\hat{Cov}(\hat{\pi}_0 + \hat{\pi}_1 Z, Y)}{\hat{Var}(\hat{\pi}_0 + \hat{\pi}_1 Z)} \\
&= \frac{\hat{Cov}(Z, Y)}{\hat{\pi}_1\hat{Var}(Z)} \\
&= \frac{\hat{Var}(Z)\hat{Cov}(Z, Y)}{\hat{Cov}(Z, X)\hat{Var}(Z)} \\
&= \frac{\hat{Cov}(Z, Y)}{\hat{Cov}(Z, X)} \\
&= \hat{\beta}_1^{IV}
\end{aligned}$$

3. The paper "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" by J. Angrist and W. Evans (AE98) considers labor supply responses to the number of children in the household. They consider models of the form $Y = \beta_0 + X_1\beta_1 + X_2'\beta_2 + U$ where Y is some measure of the parents' labor supply, X_1 is a binary variable indicating "more than 2 children in the household", and X_2 is a vector of (assumed) exogenous variables that control for race, age, and whether any of the children is a boy. For the next two questions we will focus on the case where Y is a binary variable indicating whether the mother worked during the year.

(i) Provide a causal interpretation of β_1.

All else equal, β_1 is the change in the probability that the mother worked during the year if the household has more than two children versus two or fewer children.

(ii) Discuss why or why not you think that X_1 could be endogenous. If you think it is, discuss the direction of the (conditional) bias in OLS relative to the causal parameter.

Angrist and Evans (1998) focuses on heteronormative households with one man - a husband - and one woman - a wife. In the U.S., there's a long history of gendered divisions of labor where husbands are more likely to work in the formal labor market outside the home and wives are more likely to work within the home including child-rearing. I think X_1 could be endogenous because parents' labor supply decisions affect the amount of time and money parents have. Children cost time and money, so parents' labor supply decisions could affect their decision to have more than 2 children in the household. Time and money represent two different directions of conditional bias. If higher labor supply results in more labor income, then parents have more financial resources for raising children and thus could raise more children. On the other hand, if higher labor supply results in less time for parenting, then parents may have less children. Although both effects may be present, my prior is that the latter effect dominates for mother's labor supply decisions, so $X_1 = f(Y)$ where f is a decreasing function in $Y \implies \beta_1$ is biased down.

(iii) Repeat the previous two questions when Y is a binary variable indicating whether the husband worked during the year.

All else equal, β_1 is the change in the probability that the husband worked during the year if the household has more than two children versus two or fewer children.

Taking this setup as given, I think β_1 is less likely to be biased if Y just measures the husband's labor decision. Since both time and money resources are needed to raise children, the "default" outcome in this setup if for the husband to work in the formal labor market regardless of the number of children in the household. Whereas, the wife works in formal labor market with no or few children and focus on child-rearing when there are more children.

(iv) Discuss why or why not you think that the binary variable Z_1 which indicates whether the two first children are of the same sex is a valid instrument for X_1.

The rationale for the Z_1 as a valid instrument for X_1 is that parents prefer having children of different sexes. So parents with two male children or two female children are more likely to have a third child than parents with one male child and one female child despite the time and money costs of having children which are likely similar per child regardless of sex.

For an instrument to be valid, it needs to be exogeneous (i.e., $E[U|Z_1] = 0$) and relevant (i.e., $E[ZX']$ is invertible where $Z = (Z_1, X_2')'$ and $X = (X_1, X_2')$). I think it's plausible that Z_1 is exogeneous; I can't think of an argument for why parents' labor supply decisions would change if they have two children of the same sex as opposed to two children of different sex. Furthermore, since a child's sex is more-or-less a 50-50 bernoulli random variable, I think it's plausible that parent's labor supply decision don't affect the sex of their children.

I'm less convinced by relavence; it seems like the time and money cost of an additional child far outweighs impact of a preference for having children with different sexes. Ancedotal evidence suggest that, in some large families (4+ children), the parents kept having children "because they wanted a boy/girl". I'm not convinved that these families would be smaller if their older children where different sexes. It seems like the parents had a preference for a lot of children.

(v) Estimate the reduced form regression of X_1 on Z_1 and X_2, do the results suggest that Z_1 is relevant?

In the regression results from R above, we can see that the coefficient on `samesex` is statistically signficiant at the 0.1 percent level. It remains statistically signficiant given a wide variety of specifications for X_2. These results suggests that Z_1 is a relevant instrument.

Table 1:

	Dependent variable:
	morekids
samesex	0.064^{***} (0.002)
blackm	0.027^{**} (0.013)
hispm	0.080^{***} (0.010)
othracem	0.025^{***} (0.007)
blackd	0.044^{***} (0.013)
hispd	0.091^{***} (0.010)
othraced	0.061^{***} (0.007)
agefstm	-0.027^{***} (0.001)
agefstd	-0.019^{***} (0.001)
agem1	0.012^{***} (0.001)
aged1	0.019^{***} (0.001)
boy1st	-0.010^{***} (0.002)
boy2nd	-0.010^{***} (0.002)
Constant	0.382^{***} (0.008)
Observations	333,707
R^2	0.084
Adjusted R^2	0.084
Residual Std. Error	0.469 (df = 333693)
F Statistic	$2{,}364.383^{***}$ (df = 13; 333693)
Note:	*p<0.1; **p<0.05; ***p<0.01

(vi) (Attempt to) replicate the first three rows of Table 7 columns 1, 2, 5, 7, and 8 in AE98. Interpret the empirical results in relation to your discussion of the previous questions.

sample	dependent_variable	method	coefficient	standard_error	p_value
All Women	Worked for pay	OLS	-0.176	0.002	0.000
All Women	Worked for pay	2SLS	-0.117	0.025	0.000
All Women	Weeks worked	OLS	-8.978	0.072	0.000
All Women	Weeks worked	2SLS	-5.559	1.136	0.000
All Women	Hours/week	OLS	-6.647	0.062	0.000
All Women	Hours/week	2SLS	-4.547	0.966	0.000
Married Women	Worked for pay	OLS	-0.167	0.002	0.000
Married Women	Worked for pay	2SLS	-0.117	0.028	0.000
Married Women	Weeks worked	OLS	-8.025	0.089	0.000
Married Women	Weeks worked	2SLS	-5.272	1.235	0.000
Married Women	Hours/week	OLS	-5.970	0.075	0.000
Married Women	Hours/week	2SLS	-4.784	1.035	0.000
Husbands of Married Women	Worked for pay	OLS	-0.008	0.001	0.000
Husbands of Married Women	Worked for pay	2SLS	0.004	0.009	0.643
Husbands of Married Women	Weeks worked	OLS	-0.901	0.044	0.000
Husbands of Married Women	Weeks worked	2SLS	0.579	0.600	0.335
Husbands of Married Women	Hours/week	OLS	0.151	0.052	0.003
Husbands of Married Women	Hours/week	2SLS	0.509	0.703	0.468

Across all measures of women's labor supply, the 2SLS coefficient is smaller in magnitude than the OLS coefficient. This suggests that there is some endogeneity between labor supply choices and having more than 2 children.

The story for married is largely as discussed in (iii). For married men, the OLS coefficients for worked and weeks worked are significant and negative. The 2SLS coefficients are not significant meaning that after controlling for the effect of children on labor supply decisions having more children does not affect labor supply decisions.

Appendix

R code for 3(v)

```
df <- read_dta("AE80.dta")

regression <- lm(morekids ~
     samesex +
     blackm + hispm + othracem + blackd + hispd + othraced +
     agefstm + agefstd + agem1 + aged1 +
     boy1st+ boy2nd,
   data = df)
```

R code for 3(vi)

```
df2 <- df  %>%
  mutate(age1m = agem1,
         age1d = aged1,
         weeks1m = weeksm1,
         weeks1d = weeksd1)

results <- NULL

for (which_sample in c("All Women", "Married Women", "Husbands of Married Women")) {
  for (which_dep_var in c("worked", "weeks1", "hoursw")) {

    # independent variable
    if (which_sample == "All Women") sample <- df2
    else sample <- df2 %>% filter(msample == 1)

    x_1 <- sample %>% pull(morekids)
    z_1 <- sample %>% pull(samesex)

    # control variables
    if (which_sample == "Husbands of Married Women") var_suffix <- "d"
    else var_suffix <- "m"

    x_2 <- sample %>%
      select(paste0(c("black", "hisp", "othrace", "agefst", "age1"), var_suffix),
             "boy1st", "boy2nd", "const") %>%
      as.matrix()

    x <- cbind(x_1, x_2)
    z <- cbind(z_1, x_2)

    k <- ncol(x)
    n <- nrow(x)

    # dependent variables
    dep_var <- paste0(which_dep_var, var_suffix)
    y <- sample %>% pull(dep_var) %>% as.numeric()
```

```
    # OLS
    coefficients <- solve(t(x) %*% x) %*% (t(x) %*% y)
    residuals <- as.numeric(y - x %*% coefficients)
    varcov <- solve(t(x) %*% x) * sum(residuals^2) / n

    results <- bind_rows(results,
                         tibble(sample=which_sample,
                                dependent_variable = which_dep_var,
                                method = "OLS",
                                coefficient = round(coefficients[1], 3),
                                standard_error = round(sqrt(varcov[1, 1]), 3),
                                p_value = round(2*(1- pt(q = abs(coefficients[1])/sqrt(varcov[1, 1]),
                                                         df = n-k-1)), 3)))

    # 2SLS
    coefficients_1 <- solve(t(z) %*% z) %*% (t(z) %*% x_1)
    x_1_hat <- as.numeric(z %*% coefficients_1)
    x_hat <- cbind(x_1_hat, x_2)
    coefficients_2 <- solve(t(x_hat) %*% x_hat) %*% (t(x_hat) %*% y)
    residuals_2 <- as.numeric(y - x_hat %*% coefficients_2)
    varcov <- solve(t(x_hat) %*% x_hat) * sum(residuals_2^2) / n

    results <- bind_rows(results,
                         tibble(sample=which_sample,
                                dependent_variable = which_dep_var,
                                method = "2SLS",
                                coefficient = round(coefficients_2[1], 3),
                                standard_error = round(sqrt(varcov[1, 1]), 3),
                                p_value = round(2*(1- pt(q = abs(coefficients_2[1])/sqrt(varcov[1, 1]),
                                                         df = n-k-1)), 3)))

  }
}

results %>%
  mutate(dependent_variable =
           case_when(dependent_variable == "worked" ~ "Worked for pay",
                     dependent_variable == "weeks1" ~ "Weeks worked",
                     dependent_variable == "hoursw" ~ "Hours/week")) %>%
  kable()
```