

Practical: Instrumental Variables (Solutions)

11 December 2014

A Short Course on Concepts and Methods in Causal Inference
V Edition

Dataset

For this practical, we will use data from a cluster-randomized community intervention trial on the effect of vitamin A supplements on one-year mortality in Indonesian children [1]. Of 450 villages, 229 were assigned to a treatment in which village children were provided two oral doses of vitamin A. Children in the 221 control communities were not provided vitamin A supplements. This design resulted in 12,094 children assigned to supplements and 11,588 to control. We will analyze this data with the 2SLS approach using the randomization indicator as an IV, ignoring clustering for simplicity.

You will find the data for this practical in the file `vitamina.txt`. You will need to copy them into an appropriate folder to be accessed from Stata (use the command `insheet` using ‘‘*path-to-file*\vitamina.txt’’). The dataset contains the following variables:

`id` sequential id number of study subjects;

`assvita` treatment assignment indicator; 1 if assigned to vit. A, 0 otherwise.

`recvita` received treatment indicator; 1 if received vit. A, 0 otherwise.

`death` one-year mortality indicator; 1 if deceased during follow-up, 0 otherwise.

If you are not familiar with Stata, remember: `help command-name` is your friend!

Exercises

1. Perform a conventional as-treated analysis by regressing `death` on `recvita` with the command `regress death recvita, vce(robust)` (we use robust standard to account for the heteroschedasticity of the binary outcome). How do you interpret the estimated coefficient of `recvita`?

We use the `regress` command to fit the model $E(\text{death}|\text{recvita}) = \delta_0 + \delta_1 \text{recvita}$:

```
. regress death recvita, vce(robust)
```

Linear regression

Number of obs = 23682
 F(1, 23680) = 62.08
 Prob > F = 0.0000
 R-squared = 0.0020
 Root MSE = .07093

		Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
death							
recvita		-.0064701	.0008212	-7.88	0.000	-.0080797	-.0048606
_cons		.0077104	.0007391	10.43	0.000	.0062617	.0091591

The estimated coefficient of `recvita` equal to $\hat{\delta}_1 = -.0064701 = -6.47$ perthousands points. Since the model is linear, this is a risk-difference estimate: children who received vitamin A supplements have a one-year risk of death lower than those who did not receive supplements by about 6.47 per-thousands points.

2. Compute a cross-tabulation of `assvita` by `recvita` using the `tab2` command. Do you think the monotonicity assumption could be satisfied?

The output of the `table` command is the following:

```
. tab2 assvita recvita, row
```

-> tabulation of `assvita` by `recvita`

```
+-----+
| Key          |
|-----|
| frequency    |
| row percentage |
+-----+
```

	recvitaA		
assvitaA	0	1	Total
0	11,588	0	11,588
	100.00	0.00	100.00
1	2,419	9,675	12,094
	20.00	80.00	100.00
Total	14,007	9,675	23,682

| 59.15 40.85 | 100.00

No childer assigned to no supplements received any supplements, in accordance with the study design. Thus the monotonicity assumption is satisfied, i.e. there are no defiers. To see this, let Z indicate whether a children is assigned to vitamin A ($Z = 1$) or no vitamin A ($Z = 0$) and let X indicate whether a children actually received vitamin A ($X = 1$) or not ($X = 0$). Remeber that defiers are exactly those subjects with $X^0 = 1$ and $X^1 = 0$, but in our setting all childrens have $X^0 = 0$ by design.

3. Compute the first-stage regression of `recvita` on `assvita` and compute the predicted values `recvitapred` obtained from the model. Is `assvita` a strong IV? Use the commands:

- `regress recvita assvita`
- `predict recvitapred`

The output of the considered commands is the following:

```
. regress recvita assvita
```

Source	SS	df	MS	
Model	3787.23359	1	3787.23359	
Residual	1935.16	23680	.081721284	
Total	5722.39359	23681	.24164493	

Number of obs =	23682
F(1, 23680) =	46343.30
Prob > F =	0.0000
R-squared =	0.6618
Adj R-squared =	0.6618
Root MSE =	.28587

recvita	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
assvita	.7999835	.0037161	215.27	0.000	.7926997 .8072673
_cons	-7.42e-14	.0026556	-0.00	1.000	-.0052052 .0052052

```
. predict recvitapred
(option xb assumed; fitted values)
```

The first-stage model is $E(\text{recvita}|\text{assvita}) = \alpha_0 + \alpha_1 \text{assvita}$, with estimates $\hat{\alpha}_0 = -7.42\text{e-}14$ and $\hat{\alpha}_1 = .7999835$. The almost-null $\hat{\alpha}_0$ is compatible with the fact that, by study design, children assigned to no supplements could not receive any supplements: $E(\text{recvita}|\text{assvita} = 0) = \alpha_0 = 0$.

From the output we see that the IV is very strong, since the F-statistic $F = 46343.30$ is much greater than 10. In fact, 80% of children assigned to vitamin A actually received vitamin A supplements (compare with the results of the previous exercise).

4. Compute the second-stage regression of `death` on `recvitapred` with the command `regress death recvitapred`. How do you interpret the estimated coefficient of `recvita`?

The output of the second-stage regression $E(\text{death}|\text{recvitapred}) = \beta_0 + \beta_1 \text{recvitapred}$ is the following:

```
. regress death recvitapred
```

Source	SS	df	MS	Number of obs	=	23682
Model	.039463858	1	.039463858	F(1, 23680)	=	7.83
Residual	119.352479	23680	.005040223	Prob > F	=	0.0051
Total	119.391943	23681	.005041677	R-squared	=	0.0003
				Adj R-squared	=	0.0003
				Root MSE	=	.07099

death	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
recvitapred	-.003228	.0011536	-2.80	0.005	-.0054892 -.0009669
_cons	.0063859	.0006595	9.68	0.000	.0050932 .0076786

Since the monotonicity assumption is satisfied, the estimated coefficient $\hat{\beta}_1 = -.003228$ is an estimate of the Complier Average Casual Effect. Specifically, results suggest that, among children who would taken or not taken the supplements as assigned, vitamin A supplements reduced the one-year risk of death by about 3.23 perthousands points.

5. Compare the results obtained in the conventional analysis with those obtained by the 2SLS procedure. How do you explain the observed differences?

The effect of vitamin A supplements estimated via 2SLS methods (a risk reduction of 3.23/1000 points) is weaker effect than that assessed in the conventional analysis (a risk reduction of 6.47/1000 points). This suggests that never-takers were at a higher risk of death than compliers.

To see this, observe that by design $X^0 \equiv 0$ for all childred, so there are no defiers and no always-treated. Hence:

$$\begin{aligned}
 E(Y|X = 1) &= E(Y^1|X = 1) = E(Y^1|CO, Z = 1) = E(Y^1|CO) \\
 E(Y|X = 0) &= E(Y^0|X = 0) = E(Y^0|CO, Z = 0)\omega + E(Y^0|NT)(1 - \omega) \\
 &= E(Y^0|CO)\omega + E(Y^0|NT)(1 - \omega)
 \end{aligned}$$

with $\omega = \frac{\pi_{CO}P(Z=0)}{\pi_{CO}P(Z=0) + \pi_{NT}}$.

Therefore, the estimand δ_1 in the conventional analysis is

$$\begin{aligned}
 \delta_1 &= E(Y|X = 1) - E(Y|X = 0) \\
 &= E(Y^1|CO) - E(Y^0|CO)\omega - E(Y^0|NT)(1 - \omega) \\
 &= CACE + (E(Y^0|CO) - E(Y^0|NT))(1 - \omega)
 \end{aligned}$$

while the estimand in the 2SLS analysis is $\beta_1 = CACE$. The bias factor $B = \delta_1 - \beta_1 = (E(Y^0|CO) - E(Y^0|NT))(1 - \omega)$ is negative if $E(Y^0|CO) < E(Y^0|NT)$, i.e. if the never treated have a greater baseline risk of mortality than compliers.

6. Implement the 2SLS procedure directly using the `ivregress` procedure with the following command:

- `ivregress 2sls death (recvita=assvita), vce(robust)`

Compare the results with those obtained previously.

The output of the `ivregress` procedure is the following:

```
. ivregress 2sls death (recvita=assvita), vce(robust)
```

Instrumental variables (2SLS) regression				Number of obs =		23682	
				Wald chi2(1) =		7.76	
				Prob > chi2 =		0.0054	
				R-squared =		0.0015	
				Root MSE =		.07095	

death		Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
recvita		-.003228	.0011592	-2.78	0.005	-.0055 -.0009561
_cons		.0063859	.00074	8.63	0.000	.0049356 .0078362


```
Instrumented:  recvita
Instruments:   assvita
```

As expected, the the point estimate of the CACE is exactly the same as obtained previously, but the corresponding S.E. is slightly larger. In general, the S.E. obtained at the second stage regression understate the true variability of the estimates because they do not take into account the uncertainty in the first-step regression.

In general, the correct large-sample covariance matrix of the vector of 2SLS estimates is given by the following formula

$$\widehat{V}(\widehat{\beta}_{2SLS}) = \left(\sum_{i=1}^N \widehat{x}_i \widehat{x}_i^T \right)^{-1} \left(\sum_{i=1}^N \widehat{e}_i^2 \widehat{x}_i \widehat{x}_i^T \right) \left(\sum_{i=1}^N \widehat{x}_i \widehat{x}_i^T \right)^{-1},$$

where, in our setting, $\widehat{x}_i^T = [1 \text{ recvitapred}_i]$ and $\widehat{e}_i = y_i - \widehat{\beta}_{0,2SLS} - \widehat{\beta}_{1,2SLS} \text{recvita}_i$ for each children $i = 1, \dots, N$ ($N = 23,682$). This covariance matrix is heteroschedasticity-robust [2] and it's what `ivregress` computes when the option `vce(robust)` is specified (see `help robust` and the corresponding Stata manual page).

References

- [1] Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000; 29:722-9
- [2] Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, 2002.