

2021 中级计量经济学作业 2

1. (关于分块回归) 已知 $y = \beta_0 + \beta_1 x_1 + \dots + \beta_K x_K + \varepsilon$. 证明 OLS 回归系数估计 b_K 满足

$$b_K \rightarrow \beta_K + \frac{\text{Cov}(\tilde{x}_K, \varepsilon)}{\text{Var}(\tilde{x}_K)},$$

其中 $\tilde{x}_K = x_K - L(x_K | 1, x_1, \dots, x_{K-1})$, $L(x_K | 1, x_1, \dots, x_{K-1})$ 是 x_K 在 $1, x_1, \dots, x_{K-1}$ 上的 Linear Projection. (假设所需的秩条件和大数定律满足)

证明: 首先, 定义 $\mathbf{M}_{(K)}$ 为除 x_K 之外的自变量 $1, x_1, \dots, x_{K-1}$ 的消灭矩阵.

$$\begin{aligned} \tilde{\mathbf{y}} &= \mathbf{M}_{(K)} \mathbf{y} \\ &= \mathbf{M}_{(K)} (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \mathbf{M}_{(K)} \begin{bmatrix} \mathbf{X}_{(K)} & \mathbf{X}_K \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{(K)} \\ \beta_K \end{bmatrix} + \mathbf{M}_{(K)} \boldsymbol{\varepsilon} \\ &= \mathbf{M}_{(K)} (\mathbf{X}_{(K)} \boldsymbol{\beta}_{(K)} + \mathbf{X}_K \beta_K) + \mathbf{M}_{(K)} \boldsymbol{\varepsilon} \\ &= \tilde{\mathbf{X}}_K \beta_K + \mathbf{M}_{(K)} \boldsymbol{\varepsilon} \end{aligned}$$

其次, 由分块回归知

$$\begin{aligned} b_K &= (\tilde{\mathbf{X}}_K' \tilde{\mathbf{X}}_K)^{-1} (\tilde{\mathbf{X}}_K' \tilde{\mathbf{y}}) \\ &= (\tilde{\mathbf{X}}_K' \tilde{\mathbf{X}}_K)^{-1} (\tilde{\mathbf{X}}_K' (\tilde{\mathbf{X}}_K \beta_K + \mathbf{M}_{(K)} \boldsymbol{\varepsilon})) \\ &= \beta_K + (\tilde{\mathbf{X}}_K' \tilde{\mathbf{X}}_K)^{-1} (\tilde{\mathbf{X}}_K' \mathbf{M}_{(K)} \boldsymbol{\varepsilon}) \\ &\rightarrow \beta_K + \text{E}[\tilde{\mathbf{x}}_K \tilde{\mathbf{x}}_K']^{-1} \text{E}[\tilde{\mathbf{x}}_K \boldsymbol{\varepsilon}]. \end{aligned}$$

注意这里 \tilde{x}_K 是标量, 因此 $b_K \rightarrow \beta_K + \text{E}[\tilde{x}_K^2]^{-1} \text{E}[\tilde{x}_K \boldsymbol{\varepsilon}]$. 同时, 因为 $\mathbf{X}_{(K)}$ 中包含常数项, $\text{E}[\tilde{x}_K] = 0$, 因此 $b_K \rightarrow \beta_K + \text{Var}(\tilde{x}_K)^{-1} \text{Cov}(\tilde{x}_K, \boldsymbol{\varepsilon})$.

2. 在同方差假设下, 比较 OLS 回归和 2SLS 回归的渐进方差, 说明哪一个更加有效 (more efficient).

OLS 和 2SLS 的渐进方差分别为:

$$\text{Avar}(\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})) = \sigma^2 \text{E}[\mathbf{x}\mathbf{x}']^{-1}$$

$$\text{Avar}(\sqrt{N}(\hat{\boldsymbol{\beta}}_{\text{2SLS}} - \boldsymbol{\beta})) = \sigma^2 \text{E}[\hat{\mathbf{x}}\hat{\mathbf{x}}']^{-1},$$

其中 $\hat{\mathbf{x}}$ 是第一阶段用工具变量对 \mathbf{x} 回归得到的 Linear projection, 因此 $\mathbf{x} = \hat{\mathbf{x}} + \mathbf{r}$, $\text{E}(\hat{\mathbf{x}}\mathbf{r}') = \mathbf{0}$. 因此

$$\begin{aligned} \text{E}(\mathbf{x}\mathbf{x}') &= \text{E}(\hat{\mathbf{x}}\hat{\mathbf{x}}' + \hat{\mathbf{x}}\mathbf{r}' + \mathbf{r}\hat{\mathbf{x}}' + \mathbf{r}\mathbf{r}') \\ &= \text{E}(\hat{\mathbf{x}}\hat{\mathbf{x}}' + \mathbf{r}\mathbf{r}') \end{aligned}$$

因此,

$$\sigma^2 \text{E}[\hat{\mathbf{x}}\hat{\mathbf{x}}']^{-1} - \sigma^2 \text{E}[\mathbf{x}\mathbf{x}']^{-1}$$

是半正定的.

3. 考虑如下关于吸烟是否影响新生儿体重的模型：

$$\log(bwght) = \beta_0 + \beta_1 male + \beta_2 parity + \beta_3 \log(faminc) + \beta_4 packs + \varepsilon,$$

其中 $\log(bwght)$ 是新生儿体重的对数, $male$ 是新生儿是否为男孩的虚拟变量, $parity$ 是新生儿出生的顺序, $faminc$ 是家庭收入, $packs$ 是孕妇在怀孕期每天吸烟的平均数.

1. $packs$ 会否和 ε 相关? 为什么?

未观察到的其他影响健康的 (从而影响婴儿体重) 的因素可能和 $packs$ 相关. 比如, 抽烟多的人可能也酗酒, 或者饮食不健康, 等等. 这些遗漏变量会包括在 ε 中.

2. 假设你取得了样本中妇女所在省的香烟平均价格 $cigprice$ 作为 $packs$ 的工具变量. 讨论 $cigprice$ 能否满足工具所需的两个条件.

相关性: 需求理论会认为 $packs$ 和 $cigprice$ 负相关, 但在实际数据中 $cigprice$ 和 $packs$ 相关性未必会大, 因为仍需控制影响 $packs$ 的其他变量, 而这些变量不可得. 例如, 对香烟成瘾的程度决定了 $packs$ 和 $cigprice$ 的需求弹性, 而成瘾程度数据不可得. 同时, 这里的香烟价格是省的平均价格, 也会丢失了很多地区微观的变化.

外生性: $cigprice$ 可能是外生的, 因为香烟价格直觉上不会直接和婴儿体重相关. 然而, 香烟的价格受税率的影响很大, 税率低的地区可能医疗条件等公共服务的水平也相对低, 从而影响母亲的健康状况 (包含在 ε 当中), 从而也有可能不是外生的.

3. 用 BWGHT.DTA 数据估计模型. 先用 OLS. 再用 2SLS, 用 $cigprice$ 作为 $packs$ 的工具变量. 结果是否有很大的不同? (可直接附上 Stata 回归结果, 无需考虑格式)

```
. reg lbwght male parity lfaminc packs
```

Source	SS	df	MS	Number of obs =	1,388
Model	1.76664363	4	.441660908	F(4, 1383) =	12.55
Residual	48.65369	1,383	.035179819	Prob > F =	0.0000
Total	50.4203336	1,387	.036352079	R-squared =	0.0350
				Adj R-squared =	0.0322
				Root MSE =	.18756

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
male	.0262407	.0100894	2.60	0.009	.0064486 .0460328
parity	.0147292	.0056646	2.60	0.009	.0036171 .0258414
lfaminc	.0180498	.0055837	3.23	0.001	.0070964 .0290032
packs	-.0837281	.0171209	-4.89	0.000	-.1173139 -.0501423
_cons	4.675618	.0218813	213.68	0.000	4.632694 4.718542

图 1

OLS 和 2SLS(这里是 IV) 回归的结果相差很大. OLS 中, 多一包烟可减少新生儿体重 8.4%, 且显著. IV 中, $packs$ 的系数很大, 是正数, 但统计上不显著. 这个结果不符合常识.

4. 作第一阶段回归, $cigprice$ 和 $packs$ 的相关性是否足够强? 这个结果如何影响了 2SLS 回归?

```
. ivreg lbwght male parity lfaminc (packs = cigprice)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of obs	=	1,388
Model	-91.350027	4	-22.8375067	F(4, 1383)	=	2.39
Residual	141.770361	1,383	.102509299	Prob > F	=	0.0490
				R-squared	=	.
				Adj R-squared	=	.
Total	50.4203336	1,387	.036352079	Root MSE	=	.32017

lbwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
packs	.7971063	1.086275	0.73	0.463	-1.333819 2.928031
male	.0298205	.017779	1.68	0.094	-.0050562 .0646972
parity	-.0012391	.0219322	-0.06	0.955	-.044263 .0417848
lfaminc	.063646	.0570128	1.12	0.264	-.0481949 .1754869
_cons	4.467861	.2588289	17.26	0.000	3.960122 4.975601

Instrumented: packs
Instruments: male parity lfaminc cigprice

图 2

```
. reg packs male parity lfaminc cigprice
```

Source	SS	df	MS	Number of obs	=	1,388
Model	3.76705108	4	.94176277	F(4, 1383)	=	10.86
Residual	119.929078	1,383	.086716615	Prob > F	=	0.0000
				R-squared	=	0.0305
				Adj R-squared	=	0.0276
Total	123.696129	1,387	.089182501	Root MSE	=	.29448

packs	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
male	-.0047261	.0158539	-0.30	0.766	-.0358264 .0263742
parity	.0181491	.0088802	2.04	0.041	.0007291 .0355692
lfaminc	-.0526374	.0086991	-6.05	0.000	-.0697023 -.0355724
cigprice	.000777	.0007763	1.00	0.317	-.0007459 .0022999
_cons	.1374075	.1040005	1.32	0.187	-.0666084 .3414234

图 3

可以看到，*cigprice* 对 *packs* 的影响不显著。而且，符号也是正的，和我们预想的负号相反。因此，在这里 *cigprice* 是很弱的工具变量。同时，*cigprice* 也可能不是外生的。这造成了 2SLS 回归的不一致以及方差很大。

总结：这道题说明，虽然工具变量从理论上来说很好，但实际操作中，工具变量的选择是比较困难的。做实证研究涉及到工具变量选择时，需要谨慎讨论其合理性。