

界线。从图 20.3 可以看出,在样本内的 20 世纪 70 年代中期与 80 年代早期,联邦基金利率与通胀率都曾达到很高的水平。另一方面,这两个变量在样本外都曾达到很低的水平,反映出 2001 年 9·11 事件后宽松的货币政策以及 2008 年全球金融风暴的影响。与此相反,无论在样本内还是样本外,失业率的波动范围大体相同。以上特点也可通过对比样本内外的统计指标来看出。

```
. sum inflation unrate fedfunds if date <= tq(2002q1)
```

其中,“tq(2002q1)”表示季度数据格式。

Variable	Obs	Mean	Std. Dev.	Min	Max
inflation	169	3.792319	2.494935	.5674297	11.79167
unrate	169	5.927811	1.502554	3.4	10.66667
fedfunds	169	6.497337	3.190249	1.683333	17.78

```
. sum inflation unrate fedfunds if date > tq(2002q1)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
inflation	40	2.23731	1.063853	-.448682	4.558042
unrate	40	6.591667	1.930454	4.433333	9.933333
fedfunds	40	1.929667	1.851749	.0733333	5.256667

为了估计 VAR,首先需要根据信息准则确定 VAR 模型的阶数。

```
. varsoc inflation unrate fedfunds if date <= tq(2002q1), maxlag(13)
```

其中,选择项“maxlag(13)”表示最多滞后 13 阶。

Selection-order criteria						Number of obs	=	156
Sample:	1963:2 - 2002:1							
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	-1000.39				77.5091	12.864	12.8878	12.9227
1	-461.535	1077.7	9	0.000	.086931	6.07096	6.16625	6.30557
2	-414.252	94.566	9	0.000	.053221	5.58015	5.7469	5.99071*
3	-396.258	35.988	9	0.000	.047442	5.46485	5.70306*	6.05136
4	-386.713	19.09	9	0.024	.047141	5.45786	5.76754	6.22033
5	-376.51	20.407	9	0.016	.046466	5.44243	5.82358	6.38085
6	-359.906	33.206	9	0.000	.042213	5.34495	5.79756	6.45933
7	-354.848	10.117	9	0.341	.044491	5.39549	5.91956	6.68581
8	-343.743	22.209	9	0.008	.043423	5.36851	5.96404	6.83478
9	-329.929	27.629	9	0.001	.040964*	5.30678	5.97379	6.94901
10	-322.47	14.919	9	0.093	.041961	5.32654	6.065	7.14472
11	-311.765	21.41	9	0.011	.041269	5.30468*	6.11461	7.29882
12	-302.939	17.652*	9	0.039	.041623	5.30691	6.18831	7.477
13	-299.568	6.7431	9	0.664	.045074	5.37907	6.33193	7.72511

Endogenous: inflation unrate fedfunds
Exogenous: cons

其中,“LL”表示对数似然函数;“LR”表示似然比检验,即对最后一阶系数的联合显著性进行似然比检验,随后的 df 与 p 分别表示此似然比统计量的自由度与 p 值;“FPE”表示 Akaike's Final Prediction Error,度量向前一期预测的均方误差(MSE of one-step ahead forecast)。上表显示,不同信息准则所选择的滞后阶数并不一致(上表中打星号者)。根据最简洁的 SBIC 准则,只要滞后 2 阶就够了。根据 HQIC 准则,需滞后 3 阶。根据 FPE,需滞后 9 阶。根据 AIC 准则,需滞后 11 阶。而根据 LR 检验,则需滞后 12 阶之多。如果根据 SBIC 准则,选择滞后 2 阶,可能过于简洁;反

之,如果根据 AIC 准则,选择滞后 11 阶,共需估计  $3 \times 34 = 102$  个参数,将损失较多样本容量。根据 Lutkepohl(2005),SBIC 与 HQIC 提供了对真实滞后阶数的一致估计,而 FPE 与 AIC 可能高估滞后阶数。作为折中,Stock and Watson(2001)与 Beckett(2013)均选择滞后 4 阶。然而,滞后 4 阶的 VAR 模型依然不能保证扰动项为白噪声,故本例选择滞后 5 阶,刚好能保证扰动项为白噪声(参见下文的检验)。

下面估计 5 阶向量自回归模型:

```
. var inflation unrate fedfunds if date <= tq(2002q1), lags(1/5)
```

Vector autoregression						
			No. of obs	=	164	
			AIC	=	5.39581	
			HQIC	=	5.764131	
			SBIC	=	6.303088	
Equation	Parms	RMSE	R-sq	chi2	P>chi2	
inflation	16	1.03136	0.8446	891.4284	0.0000	
unrate	16	.230874	0.9791	7676.319	0.0000	
fedfunds	16	.886433	0.9292	2153.037	0.0000	
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
inflation						
inflation						
L1.	.5168961	.0764965	6.76	0.000	.3669658	.6668265
L2.	-.1584442	-.0840164	1.89	0.059	-.0062249	.3231133
L3.	.1001178	.084663	1.18	0.237	-.0658185	.2660542
L4.	.28014	.0849989	3.30	0.001	.1135452	.4467348
L5.	-.0425256	-.0806997	-0.53	0.598	-.2006942	.1156429
unrate						
L1.	-.8904664	.386067	-2.31	0.021	-.1.647144	-.133789
L2.	1.350818	.6414896	2.11	0.035	.0935211	2.608114
L3.	-.237796	.6696613	-0.36	0.723	-.1.550308	1.074716
L4.	-1.403182	.6447665	-2.18	0.030	-2.666901	-.1394632
L5.	1.068276	.3606713	2.96	0.003	.3613729	1.775178
fedfunds						
L1.	.1095813	.1023397	1.07	0.284	-.0910007	.3101634
L2.	-.0274367	-.1390409	-0.20	0.844	-.2999518	.2450784
L3.	-.0821163	-.1371351	-0.60	0.549	-.3508962	.1866636
L4.	-.0217318	.1356754	-0.16	0.873	-.2876508	.2441872
L5.	-.0059257	.1012347	-0.06	0.953	-.2043421	.1924907
_cons		.7982082	.3714537	2.15	0.032	.0701724
						1.526244
unrate						
inflation						
L1.	.0088075	.017124	0.51	0.607	-.0247549	.04237
L2.	-.0080949	.0188074	-0.43	0.667	-.0449567	.0287668
L3.	.0154295	.0189521	0.81	0.416	-.021716	.0525749
L4.	-.0134909	.0190273	-0.71	0.478	-.0507837	.023802
L5.	.0095475	.0180649	0.53	0.597	-.0258591	.0449541
unrate						
L1.	1.43892	.0864225	16.65	0.000	1.269535	1.608305
L2.	-.5031447	.1435998	-3.50	0.000	-.784595	-.2216944
L3.	.0987478	.1499061	0.66	0.510	-.1950627	.3925583
L4.	-.2194761	.1443333	-1.52	0.128	-.5023641	.063412
L5.	.1231384	.0807376	1.53	0.127	-.0351044	.2813811
fedfunds						
L1.	-.0071153	.0229091	-0.31	0.756	-.0520163	.0377857
L2.	.0609474	.0311248	1.96	0.050	-.0000561	.1219508
L3.	-.0378636	.0306982	-1.23	0.217	-.098031	.0223037
L4.	.0238004	.0303714	0.78	0.433	-.0357266	.0833273
L5.	-.0085719	.0226618	-0.38	0.705	-.0529881	.0358443
_cons		.1091412	.0831512	1.31	0.189	-.0538322
						2.721146
fedfunds						
inflation						
L1.	.1414521	.065747	2.15	0.031	.0125904	.2703138
L2.	.1786844	.0722101	2.47	0.013	.0371551	.3202137
L3.	-.0471542	.0727659	-0.65	0.517	-.1897727	.0954643
L4.	-.0023731	.0730546	-0.03	0.974	-.1455574	.1408113
L5.	-.1204686	.0693595	-1.74	0.082	-.2564108	.0154736
unrate						
L1.	-1.586394	.3318157	-4.78	0.000	-2.236741	-.936047
L2.	1.783011	.5513455	3.23	0.001	.7023939	2.863629
L3.	-1.057614	.5755584	-1.84	0.066	-2.185687	.0704601
L4.	1.101875	.5541619	1.99	0.047	.0157373	2.188012
L5.	-.3179867	.3099886	-1.03	0.305	-.9255533	.2895798
fedfunds						
L1.	.9448705	.0879586	10.74	0.000	.7724748	1.117266
L2.	-.3834779	.1195024	-3.21	0.001	-.6176984	-.1492575
L3.	.3654699	.1178645	3.10	0.002	.1344598	.59648
L4.	.0063154	.1166099	0.05	0.957	-.2222358	.2348666
L5.	.005741	.0870089	0.07	0.947	-.1647933	.1762753
_cons		.2710914	.3192558	0.85	0.396	-.3546385
						.8968214

即使仅滞后5阶,此VAR模型依然包含了48个参数;这些系数如此之多,以至于无法解释其经济含义。因此,在实证论文中,甚至不汇报VAR的回归系数,而主要汇报脉冲响应函数、预测方差分解与格兰杰因果检验。在上述命令var中,由于样本容量为164,故没有使用选择项“dfk”或“small”进行小样本自由度调整。

下面检验各阶系数的联合显著性:

. varwle

Equation: inflation

lag	chi2	df	Prob > chi2
1	63.62012	3	0.000
2	8.46243	3	0.037
3	1.819772	3	0.611
4	17.74984	3	0.000
5	12.07998	3	0.007

Equation: unrate

lag	chi2	df	Prob > chi2
1	356.3783	3	0.000
2	22.35949	3	0.000
3	2.902869	3	0.407
4	4.008033	3	0.261
5	3.539652	3	0.316

Equation: fedfunds

lag	chi2	df	Prob > chi2
1	255.0479	3	0.000
2	35.38302	3	0.000
3	16.31482	3	0.001
4	4.265025	3	0.234
5	3.71692	3	0.294

Equation: All

lag	chi2	df	Prob > chi2
1	572.2714	9	0.000
2	46.94062	9	0.000
3	20.02311	9	0.018
4	27.19206	9	0.001
5	20.22245	9	0.017

虽然单一方程的某些阶系数不显著,但作为三个方程的整体,各阶系数均高度显著。

下面检验残差是否为白噪声(residual whiteness),即残差是否存在自相关:

. varlmar

Lagrange-multiplier test			
lag	chi2	df	Prob > chi2
1	12.0759	9	0.20906
2	10.3352	9	0.32404

H0: no autocorrelation at lag order

结果显示,可以接受残差“无自相关”的原假设,即认为扰动项为白噪声。

进一步检验此 VAR 系统是否稳定(为平稳过程),结果见下表与图 20.4:

. varstable, graph

Eigenvalue stability condition	
Eigenvalue	Modulus
.9694669 + .04441834i	.970484
.9694669 - .04441834i	.970484
.01442202 + .7336039i	.733746
.01442202 - .7336039i	.733746
.6764154 + .2053908i	.706911
.6764154 - .2053908i	.706911
-.6844224	.684422
-.3249446 + .5288642i	.620714
-.3249446 - .5288642i	.620714
-.0073735 + .5922748i	.592321
-.0073735 - .5922748i	.592321
.4677774 + .154557i	.49265
.4677774 - .154557i	.49265
-.4349533	.434953
.4285352	.428535

All the eigenvalues lie inside the unit circle.  
VAR satisfies stability condition.

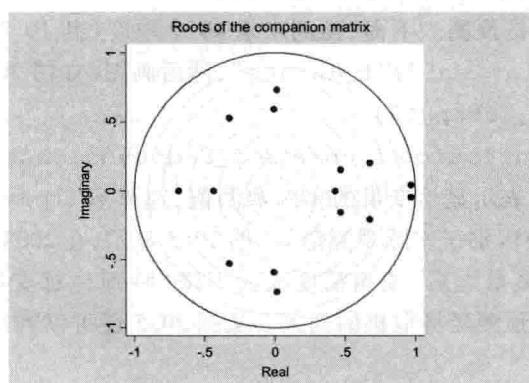


图 20.4 VAR 系统稳定性的判别图

上表与图 20.4 显示,所有特征值均在单位圆之内,故此 VAR 系统是稳定的;但有两个根十分接近单位圆,这意味着有些冲击有较强的持续性(persistence),详见第 21 章。

进一步检验 VAR 模型的残差是否服从正态分布：

. varnorm

#### Jarque-Bera test

Equation	chi2	df	Prob > chi2
inflation	7.020	2	0.02989
unrate	51.470	2	0.00000
fedfunds	217.041	2	0.00000
ALL	275.531	6	0.00000

#### Skewness test

Equation	Skewness	chi2	df	Prob > chi2
inflation	.33178	3.009	1	0.08281
unrate	.79646	17.339	1	0.00003
fedfunds	1.1369	35.332	1	0.00000
ALL	55.680	3		0.00000

#### Kurtosis test

Equation	Kurtosis	chi2	df	Prob > chi2
inflation	3.7662	4.011	1	0.04520
unrate	5.2349	34.131	1	0.00000
fedfunds	8.1567	181.709	1	0.00000
ALL	219.851	3		0.00000

上表显示,绝大多数的检验结果均可在 5% 的显著性水平上拒绝这三个变量的扰动项服从正态分布的原假设。尽管扰动项不服从正态分布对 VAR 模型的影响不大,但残差项的非正态性暗示模型可能偏离了真实的数据生成过程(DGP),并且使得对变量未来值的预测区间(forecast intervals)变得不可信(该预测区间依赖于正态假设)。

VAR 模型的用途之一是预测。下面,预测未来 40 个季度(共 10 年)的变量取值,分别记为“f\_fedfunds”,“f\_inflation”与“f\_unrate”,然后画图(如图 20.5)：

```
. fcast compute f_,step(40)
. fcast graph f_inflation f_unrate f_fedfunds,observed lpattern("_")
```

其中,选择项“observed”表示显示变量的实际观测值,选择项“lpattern(“\_”)”表示以虚线来表示变量的预测值(以便区别于实际观测值)。图 20.5 显示,在 2008 年金融风暴之前,预测效果尚可;但在 2008 年金融风暴之后,预测精度大大下降(特别是对变量 unrate 与 fedfunds 的预测)。显然,VAR 模型无法预测经济危机的到来。从图 20.5 还可以看出,预测的时期越长,则预测的精确度越低。

除了预测之外,我们还想知道对某变量的冲击对该变量及其他变量产生怎样的动态影响,比如提高联邦基金利率一个百分点会对一年后的通胀率有多大影响。这就需要用到正交化的脉冲响应函数(未正交化的脉冲响应函数无法厘清各变量冲击的单独影响,意义不大),但正交化的脉冲响应函数依赖于变量的排序(上文的预测不依赖于变量排序)。为此,分别考察变量之间

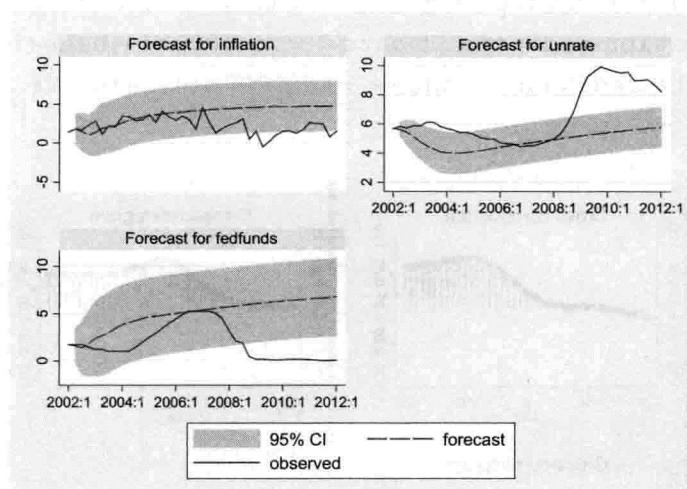


图 20.5 对未来 40 个季度的预测

的格兰杰因果关系与交叉相关图。首先,考察这三个变量之间的格兰杰因果关系。

```
. vargranger
```

Granger causality Wald tests					
Equation	Excluded	chi2	df	Prob > chi2	
inflation	unrate	14.702	5	0.012	
inflation	fedfunds	2.5547	5	0.768	
inflation	ALL	31.159	10	0.001	
unrate	inflation	2.0977	5	0.835	
unrate	fedfunds	17.344	5	0.004	
unrate	ALL	40.496	10	0.000	
fedfunds	inflation	26.041	5	0.000	
fedfunds	unrate	35.561	5	0.000	
fedfunds	ALL	50.517	10	0.000	

上表上部显示,在以 inflation 为被解释变量的方程中,如果检验变量 unrate 系数的联合显著性(即在方程中排除变量 unrate),其卡方统计量为 14.702,相应的  $p$  值为 0.012,故可认为 unrate 是 inflation 的格兰杰原因。类似地,如果检验变量 fedfunds 系数的联合显著性(即在方程中排除变量 fedfunds),其卡方统计量为 2.5547,相应的  $p$  值为 0.768,故可认为 fedfunds 不是 inflation 的格兰杰原因。如果同时检验变量 unrate 与 fedfunds 系数的联合显著性(即在方程中同时排除变量 inflation 与 unrate),其卡方统计量为 31.159,相应的  $p$  值为 0.001,强烈拒绝“unrate 与 fedfunds 都不是 inflation 的格兰杰原因”的原假设。

上表中部汇报了以 unrate 为被解释变量的检验结果。结果表明,inflation 不是 unrate 的格兰杰原因,而 fedfunds 是 unrate 的格兰杰原因。上表下部汇报了以 fedfunds 为被解释变量的检验结果。结果表明,inflation 与 unrate 都是 fedfunds 的格兰杰原因。显然,格兰杰因果关系并未给出唯一的变量作用次序。为此,进一步考察交叉相关图,结果参见图 20.6。

```
. xcorr inflation unrate if date <= tq(2002q1),name(iu)
. xcorr inflation fedfunds if date <= tq(2002q1),name(ir)
. xcorr unrate fedfunds if date <= tq(2002q1),name(ur)
. graph combine iu ir ur
```

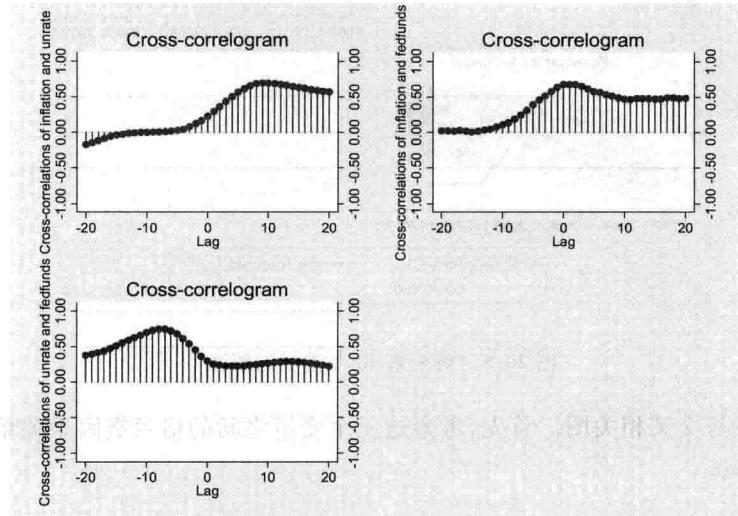


图 20.6 三个变量之间的交叉相关图

从图 20.6 左上部可看出, inflation 与滞后 10 个季度的 unrate 最相关;右上部显示,inflation 与滞后 1 个季度的 fedfunds 最相关;左下部显示,unrate 与提前 8 个季度的 fedfunds 最相关。如果无法从图 20.6 准确地看出交叉相关系数的最大值,可通过命令“`xcorr,table`”将交叉相关系数列表,例如:

```
. xcorr inflation unrate if date <= tq(2002q1),table
```

LAG	CORR	-1	0	1
		[Cross-correlation]		
-20	-0.1696			
-19	-0.1461			
-18	-0.1189			
-17	-0.0857			
-16	-0.0573			
-15	-0.0353			
-14	-0.0239			
-13	-0.0144			
-12	-0.0054			
-11	-0.0020			
-10	0.0018			
-9	0.0051			
-8	0.0066			
-7	0.0134			
-6	0.0186			
-5	0.0289			
-4	0.0438			
-3	0.0809			
-2	0.1238			
-1	0.1673			
0	0.2251			
1	0.2936			
2	0.3639			
3	0.4359			
4	0.5031			
5	0.5630			
6	0.6128			
7	0.6559			
8	0.6852			
9	0.6957			
10	0.6968			
11	0.6860			
12	0.6747			
13	0.6603			
14	0.6457			
15	0.6322			
16	0.6230			
17	0.6031			
18	0.5867			
19	0.5755			
20	0.5717			

总之,交叉相关图提示的变量次序为:inflation→fedfunds→unrate。从经济理论上,这可以解释为对通货膨胀的正冲击引起美联储提高基准利率,而后者导致失业率上升。当然,这并非唯一的理论解释。比如,根据宏观经济学中的菲利普斯曲线,失业率下降会引起通货膨胀上升,即 unrate→inflation;而根据泰勒规则(the Taylor rule),美联储在制定利率时,也会考虑失业率或产出缺口(output gap),即 unrate→fedfunds。并且,上文的格兰杰因果检验表明,inflation 不是 unrate 的格兰杰原因,与交叉相关图提示的变量次序相悖。因此,在计算脉冲响应函数与预测方差分解时,仍需变换变量次序以考察结果的稳健性。

下面考察脉冲响应函数,结果如图 20.7。

```
. irf create iuf, set(macrovar) step(20)
(file macrovar.irf created)
```

```
(file macrovar.irf now active)
(file macrovar.irf updated)
```

其中,命令“irf create”将计算所有与脉冲响应函数有关的变量与统计量。此命令将计算结果命名为“iuf”(此命名提示变量次序为 inflation,unrate,fedfunds),存入新建立的脉冲文件 macrovar.irf,将此脉冲文件激活为当前文件,并更新它。选择项“step(20)”意味着计算 20 期的脉冲响应函数,默认为“step(8)”。在此命令中,由于没有使用选择项“order(varlist)”来指定变量次序,故默认为上文使用命令 var 进行估计的变量次序,即 inflation,unrate,fedfunds(此次序为 Stock and Watson,2001 所采用)。

脉冲文件 macrovar.irf 其实就是普通的 Stata 数据文件,只不过文件扩展名为 irf。比如,看一下该文件中存储了哪些变量:

```
. describe using macrovar.irf
```

Contains data				
obs:	189			1 Sep 2013 10:32
vars:	23			
size:	36,099			
variable name	storage type	display format	value label	variable label
irf	double	%10.0g		impulse response function (irf)
step	int	%10.0g		step
cirf	double	%10.0g		cumulative irf
oircf	double	%10.0g		orthogonalized irf
coircf	double	%10.0g		cumulative orthogonalized irf
sirf	double	%10.0g		structural irf
dm	double	%10.0g		dynamic multipliers
cdm	double	%10.0g		cumulative dynamic multipliers
stdirf	double	%10.0g		std error of irf
stdcirf	double	%10.0g		std error of cirf
stdoircf	double	%10.0g		std error of oircf
stdcoircf	double	%10.0g		std error of coircf
stdsirf	double	%10.0g		std error of sirf
stddm	double	%10.0g		std error of dm
stdcdm	double	%10.0g		std error of cdm
fevd	double	%10.0g		fraction of mse due to impulse
sfevd	double	%10.0g		(structural) fraction of mse due to impulse
mse	double	%10.0g		SE of forecast of response variable
stdfevd	double	%10.0g		std error of fevd
stdsfevd	double	%10.0g		std error of sfevd
irfname	str15	%15s		name of results
response	str9	%9s		response variable
impulse	str9	%9s		impulse variable

Sorted by:

上表中的“dm”为“动态乘数”(dynamic multipliers),指的是当 VAR 模型中的外生变量变动一单位时对内生变量的动态影响(本例无外生变量,故变量 dm 为空,无观测值)。也可用命令“use macro.irf”打开此脉冲文件(正如打开 Stata 数据文件),但这将覆盖内存中的原有数据。

下面根据此脉冲文件来画正交脉冲响应图(未正交的脉冲响应没有太大意义)。

```
. irf graph oircf,yline(0)
```

其中,选择项“`yline(0)`”表示在纵轴  $y=0$  的位置画一条水平线,作为正交脉冲响应函数的参照线。

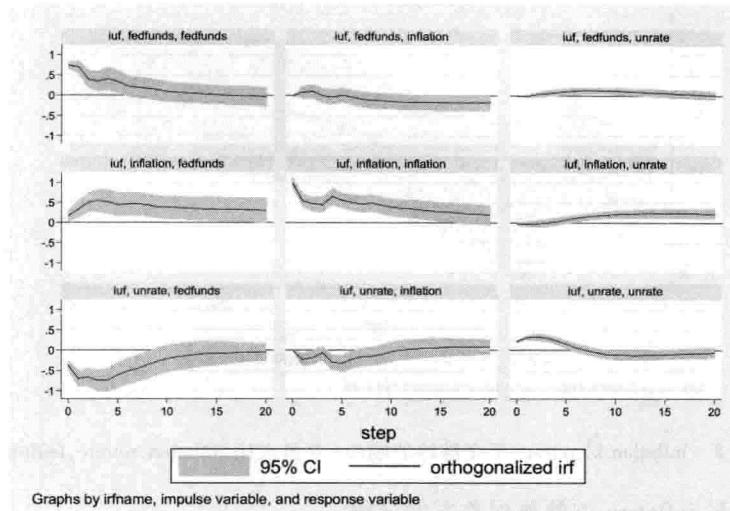


图 20.7 正交化脉冲响应图(变量次序 inflation,unrate,fedfunds)

图 20.7 包含 9 个小图,每个图的标题依次为“irfname”(此处为 `iuf`) ,“impulse variable”(脉冲变量),以及“response variable”(响应变量)。比如,第一行的三个小图均以 `fedfunds` 为脉冲变量,分别描绘 `fedfunds` 对 `fedfunds`, `inflation` 与 `unrate` 的动态效应。从第一行可以看出,联邦基金利率对于通货膨胀与失业率几乎没有作用,与预期的货币政策效应不符;这可能是因为乔利斯基分解将 `fedfunds` 排在变量次序的最后(参见下文变换次序的稳健性检验)。

类似地,第二行的三个小图分别描绘 `inflation` 对 `fedfunds`, `inflation` 与 `unrate` 的动态效应。从第二行可以看出,通货膨胀上升会引起美联储长久地提高联邦基金利率,并导致失业率上升。而且,最初的通胀冲击会引起通胀长期存在,即通胀的惯性(可能由于通胀预期);换言之,通胀一旦形成则不容易根除。

第三行的三个小图分别描绘 `unrate` 对 `fedfunds`, `inflation` 与 `unrate` 的动态效应。失业率上升将引起美联储在未来三年里(约 12 个月)保持较低的联邦基金利率以刺激经济,而失业率上升还将通过菲利普斯曲线的作用使得通货膨胀降低。

也可以显示单个脉冲响应图。比如,想单独看 `inflation` 对 `unrate` 正交脉冲的响应,可输入如下命令(结果如图 20.8):

```
. irf graph oirf,yline(0) i(unrate) r(inflation)
```

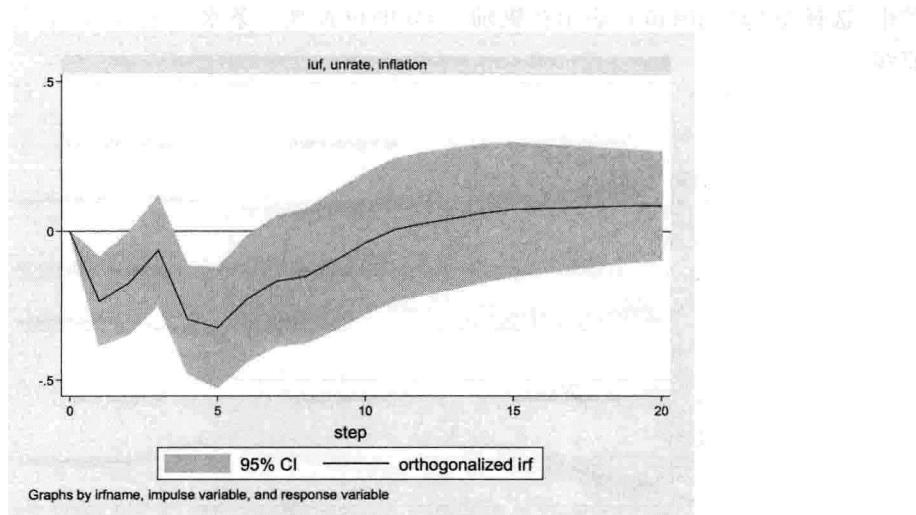


图 20.8 inflation 对 unrate 正交脉冲的响应(变量次序 inflation, unrate, fedfunds)

下面看一下变量 inflation 的预测误差方差分解。

```
. irf table fevd,r(inflation) noci
```

其中,选择项“r(inflation)”表示以 inflation 为响应变量,而选择项“noci”表示不显示置信区间(为了使结果更简洁)。

Results from iuf			
step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	1	0	0
2	.953093	.04199	.004917
3	.934426	.054805	.010769
4	.939476	.050951	.009573
5	.913917	.078127	.007957
6	.887961	.105279	.00676
7	.881257	.112039	.006703
8	.878773	.112222	.009005
9	.877767	.110112	.01221
10	.877838	.106494	.015668
11	.877551	.102269	.020179
12	.87583	.098461	.025709
13	.873281	.095157	.031563
14	.870045	.092718	.037237
15	.865953	.091104	.042943
16	.861174	.09009	.048736
17	.856246	.089351	.054403
18	.851252	.088909	.059838
19	.846121	.088754	.065125
20	.840902	.088808	.07029

(1) irfname = iuf, impulse = inflation, and response = inflation  
 (2) irfname = iuf, impulse = unrate, and response = inflation  
 (3) irfname = iuf, impulse = fedfunds, and response = inflation

Results from iuf			
step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	1	0	0
2	.953093	.04199	.004917
3	.934426	.054805	.010769
4	.939476	.050951	.009573
5	.913917	.078127	.007957
6	.887961	.105279	.00676
7	.881257	.112039	.006703
8	.878773	.112222	.009005
9	.87767	.11012	.01221
10	.877838	.106494	.015668
11	.877551	.102269	.020179
12	.87583	.098461	.025709
13	.873281	.095157	.031563
14	.870045	.092718	.037237
15	.865953	.091104	.042943
16	.861174	.09009	.048736
17	.856246	.089351	.054403
18	.851252	.088909	.059838
19	.846121	.088754	.065125
20	.840902	.088808	.07029

(1) irfname = iuf, impulse = inflation, and response = inflation  
(2) irfname = iuf, impulse = unrate, and response = inflation  
(3) irfname = iuf, impulse = fedfunds, and response = inflation

上表显示,对 inflation 进行向前 1 个季度的预测,其预测方差完全来自于 inflation 本身;即使向前作 20 季度的预测,也依然有 84.1% 的预测方差来自 inflation 本身,其余的 8.9% 与 7% 分别来自 unrate 与 fedfunds。这意味着,inflation 主要受自身的影响,变量 unrate 与 fedfunds 的作用很小(至少在目前的变量排序下)。也可以直观地将上述结果通过图示来表达,参见图 20.9。

. irf graph fevd,r(inflation)

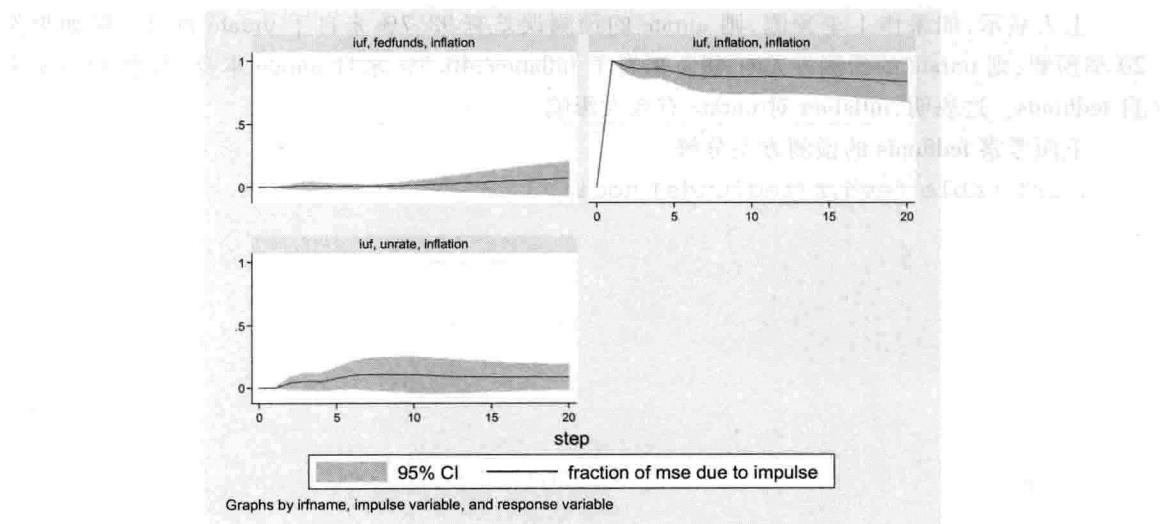


图 20.9 inflation 的预测方差分解图(变量次序 inflation,unrate,fedfunds)

下面考察 unrate 的预测方差分解。

```
. irf table fevd,r(unrate) noci
```

Results from iuf			
step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	.01264	.98736	0
2	.009345	.990471	.000184
3	.007585	.988049	.004366
4	.00551	.980447	.014042
5	.007125	.963736	.029139
6	.018249	.9292	.052551
7	.043133	.876104	.080764
8	.080901	.811523	.107576
9	.126238	.745167	.128595
10	.174162	.684145	.141693
11	.219919	.632196	.147886
12	.261628	.589172	.149199
13	.299129	.553461	.147409
14	.332917	.523301	.143782
15	.363443	.497355	.139202
16	.391168	.474643	.134189
17	.416454	.454492	.129054
18	.439605	.436412	.123984
19	.460828	.42007	.119102
20	.480289	.405224	.114487

(1) irfname = iuf, impulse = inflation, and response = unrate  
(2) irfname = iuf, impulse = unrate, and response = unrate  
(3) irfname = iuf, impulse = fedfunds, and response = unrate

上表显示,如果作1季预测,则unrate的预测误差有98.7%来自于unrate自身。但如果作20季预测,则unrate的预测方差有48%来自于inflation,40.5%来自unrate本身,其余11.4%来自fedfunds。这表明,inflation对unrate有较大影响。

下面考察 fedfunds 的预测方差分解。

```
. irf table fevd,r(fedfunds) noci
```

Results from iuf			
step	(1) fevd	(2) fevd	(3) fevd
0	0	0	0
1	.036696	.200998	.762306
2	.075248	.354469	.570283
3	.146628	.407824	.445548
4	.187531	.453521	.358948
5	.204253	.476606	.319141
6	.216086	.485133	.298782
7	.232608	.487134	.280258
8	.24991	.484575	.265515
9	.264633	.479535	.255832
10	.276911	.473724	.249365
11	.289319	.466917	.243764
12	.301628	.459708	.238664
13	.312878	.452849	.234273
14	.32306	.446469	.230471
15	.332846	.440272	.226882
16	.342299	.434262	.223438
17	.351197	.428579	.220224
18	.359497	.423241	.217263
19	.367347	.418152	.2145
20	.374771	.4133	.211929

(1) irfname = iuf, impulse = inflation, and response = fedfunds  
(2) irfname = iuf, impulse = unrate, and response = fedfunds  
(3) irfname = iuf, impulse = fedfunds, and response = fedfunds

上表显示,如果作 20 季预测,则 fedfunds 的预测方差大部分来自 unrate 与 inflation(分别占 41.3% 与 37.5%),只有少部分源于 fedfunds 自身(占 21.2%)。这表明,fedfunds 受 unrate 与 inflation 的影响较大。

下面将所有的预测方差分解画图,结果参见图 20.10。

. irf graph fevd

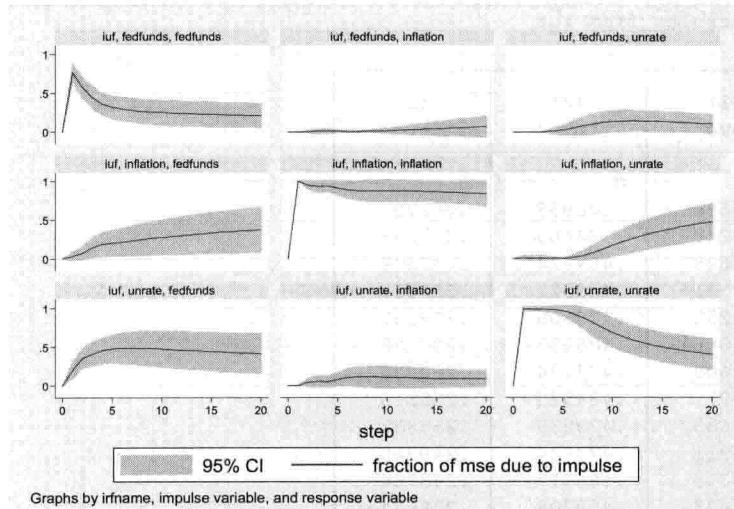


图 20.10 所有变量的预测方差分解图(变量次序 inflation,unrate,fedfunds)

由于以上的脉冲响应与方差分解结果依赖于变量次序,故下面将变量次序变为交叉相关图所推荐的次序 inflation,fedfunds,unrate,以考察结果的稳健性(对于三个变量的 VAR 系统,共有 6 种可能排序,其他情形从略)。为此,建立一个新的脉冲响应结果,并命名为 ifu。

```
. irf create ifu,order(inflation fedfunds unrate) step(20)
(file macrovar.irf updated)
```

下面比较在以上两种变量排序下,inflation 对 fedfunds 脉冲的响应,结果参见图 20.11。

```
.irf graph oirf,i(fedfunds) r(inflation) yline(0) noci
```

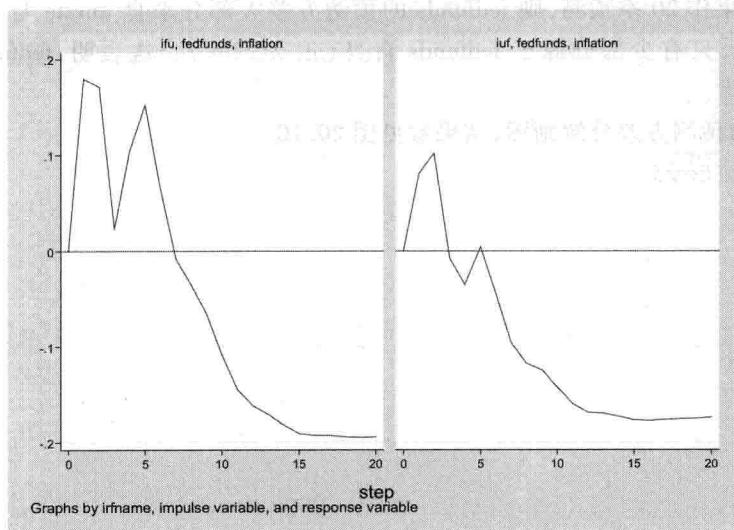


图 20.11 对比两种变量排序下,inflation 对 fedfunds 脉冲的响应

在图 20.11 中,左图为在变量排序 ifu 的情况下,fedfunds 对 inflation 的动态效应;而右图为在变量排序 iuf 的情况下,fedfunds 对 inflation 的动态效应。显然,在左图中,fedfunds 对 inflation

的作用幅度明显更大。

下面,比较在两种变量排序下, inflation 的预测方差来源于 fedfunds 的比重,结果参见图 20.12。

```
. irf graph fevd,i(fedfunds) r(inflation) noci
```

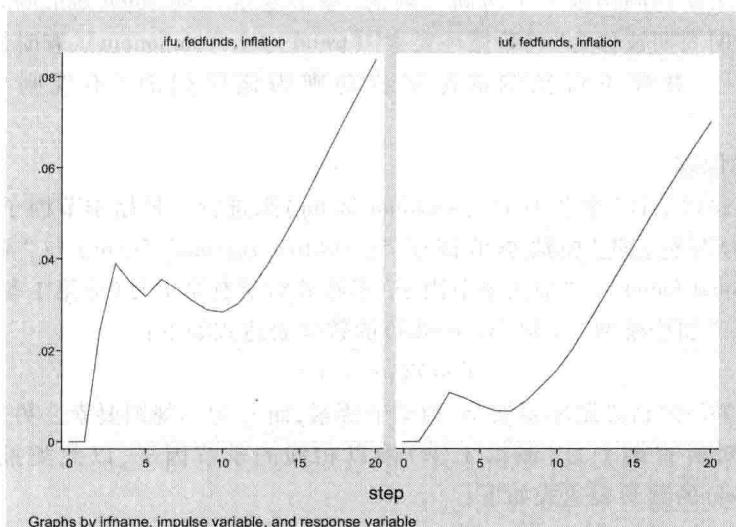


图 20.12 对比两种变量排序下,unrate 对 fedfunds 脉冲的响应

在图 20.12 中,左图为在变量排序 ifu 的情况下,inflation 预测方差中来源于 fedfunds 的比重,而右图则为在变量排序 iuf 的情况下,inflation 预测方差中来源于 fedfunds 的比重。显然,在左图中,fedfunds 对于 inflation 的作用明显大于右图。

总之,上述结果显示,无论脉冲响应函数还是预测方差分解,都在一定程度上依赖于变量排序。因此,在使用或解释其结果时应十分谨慎。对于  $n$  个变量的 VAR 系统,共有  $n!$  种可能的变量排序。如果无法确定唯一的变量排序,则应对可能的变量排序进行稳健性检验。

## 20.15 季节调整

### 1. 季节效应

对于月度或季度时间序列,常常需要对其进行“季节调整”(seasonal adjustment),去掉“季节效应”后才能使用。比如,考察中国的季度 GDP 数据。由于第一季度包含春节,故通常第一季度的 GDP 偏低。如果直接以第二季度 GDP 除以第一季度 GDP 来计算环比增长率,则会高估第二季度的 GDP 增长率;同样道理,将第一季度 GDP 除以上年第四季度 GDP 则会低估第一季度的 GDP 增长率。总之,包含季节效应的时间序列不能直接计算环比增长率。如果不进行季节调整,则只能计算同比增长率,即与去年同一季(月)相比。年度数据不需要进行季节调整。

可能导致季节效应的因素包括:

(1) 天气因素。比如,在冬季由于取暖而增加能源消耗。

(2) 行政因素。比如,学校开学与放假的日期对交通量的影响。

- (3) 固定假日。比如,十一国庆节对旅游与交通的影响。
- (4) 移动假日(moving holiday)。比如,春节期间,铁路运输量增加而 GDP 下降。
- (5) 日历因素。比如,闰年与闰月的影响。
- (6) 交易日效应。比如,五金店销售额在有五个周末的月份高于只有四个周末的月份。

所有这些季节因素共同构成一个时间序列的“季节要素”(seasonal component)。该时间序列的长期走势与中期周期被称为“趋势循环要素”(trend cycle component),有时简称“趋势要素”(trend component)<sup>①</sup>。其他不可预测的随机扰动则为该序列的“不规则要素”(irregular component)。

## 2. 季节调整的原理

季节调整通常通过估计“季节因子”(seasonal factor)来进行。根据季节因子起作用的方式,季节因子主要分为两种,即“加法季节因子”(additive seasonal factor)与“乘法季节因子”(multiplicative seasonal factor)。“加法季节因子”意味着对所有第 1 月(或第 1 季)加上相同的季节因子,以此类推。“加法模型”(additive model)的数学表达式如下:

$$Y_t = TC_t + S_t + I_t \quad (20.92)$$

其中, $Y_t$  为原序列, $TC_t$  为趋势循环要素, $S_t$  为季节要素,而  $I_t$  为不规则要素。另一方面,“乘法季节因子”则意味着对所有第 1 月(或第 1 季)乘以相同的季节因子,以此类推。“乘法模型”(multiplicative model)的数学表达式如下:

$$Y_t = TC_t \times S_t \times I_t \quad (20.93)$$

使用乘法模型要求  $Y_t$  序列中不包含零或负数。对方程(20.93)两边同时取对数可得:

$$\ln Y_t = \ln TC_t + \ln S_t + \ln I_t \quad (20.94)$$

方程(20.94)在形式上与加法模型相同,故称为“对数加法模型”。季节调整的目标就是将原序列  $Y_t$  分解为趋势循环要素、季节要素与不规则要素,然后去掉季节要素  $S_t$ ,得到季节调整序列(seasonally adjusted series)。季节调整的具体方法有多种,使用不同方法,会得到不同的季节调整序列,带有一定的主观性;这是季节调整的局限性。因此,国外的统计部门一般同时公布原序列与季节调整序列。下面以加法模型为例,介绍常用的回归法、移动平均比率法与 X12 方法。

(1) 回归法。<sup>②</sup> 回归法的基本步骤为,首先生成月度(或季度)虚拟变量,然后把时间序列对这些虚拟变量进行 OLS 回归,所得到的残差就是经季节调整后的序列。

例 Kennan(1985)在使用美国制造业的月度产量数据时,首先取对数,然后将其对时间、时间平方以及月度虚拟变量进行回归,所得残差即为经过季节调整并去掉时间趋势的产量对数。

下面以 turksales.dta 为例,该数据集包括 1990 年第 1 季至 1999 年第 4 季的火鸡销售额(sales)与时间变量(t)。首先看一下火鸡销售额的时间趋势,结果参见图 20.13。

```
. use turksales.dta,clear
. tsline sales
```

从图 20.13 可知,火鸡销售额有时明显的季节波动。其中,第四季度的火鸡销售最好,因为感恩节在第四季度,而火鸡是感恩节的传统食物;第一与第二季度的销售较为平淡。

为了生成季度虚拟变量,首先从时间变量 t 中提取季度信息,记为变量 quarter。

```
. gen quarter = quarter((dofq(t)))
```

<sup>①</sup> 事实上,很难区分“趋势”与“周期”,人为的区分总是带有主观性,故将此二者并称。

<sup>②</sup> 有关季节调整的回归法,主要借鉴了 Baum(2006)。

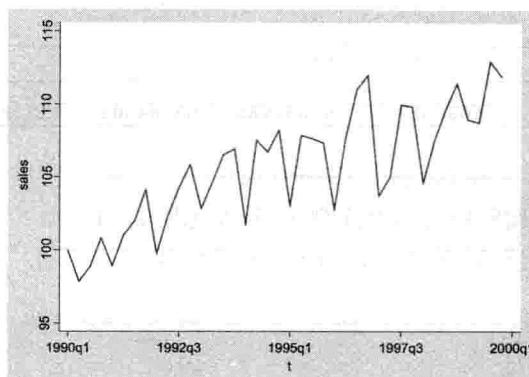


图 20.13 火鸡销售额的时间趋势

其次, 使用命令 tab 来生成季度虚拟变量。

. tab quarter,gen(q)

quarter	Freq.	Percent	Cum.
1	10	25.00	25.00
2	10	25.00	50.00
3	10	25.00	75.00
4	10	25.00	100.00
Total	40	100.00	

然后, 以第 1 季度为参照值, 把变量 sales 对常数项以及第 2 ~ 4 季度虚拟变量进行回归。

. reg sales q2 - q4

Source	SS	df	MS	Number of obs =	40
Model	161.370376	3	53.7901254	F( 3, 36) =	4.03
Residual	480.52796	36	13.3479989	Prob > F =	0.0143
Total	641.898336	39	16.4589317	R-squared =	0.2514
				Adj R-squared =	0.1890
				Root MSE =	3.6535
sales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
q2	2.389294	1.633891	1.46	0.152	- .9243908 5.702978
q3	4.33511	1.633891	2.65	0.012	1.021425 7.648794
q4	5.232047	1.633891	3.20	0.003	1.918362 8.545731
_cons	102.6287	1.155335	88.83	0.000	100.2856 104.9719

上表中的  $F$  统计量的  $p$  值为 0.0143, 这个回归方程高度显著。而且  $R^2 = 0.25$ , 表明季度虚拟变量对于变量 sales 有较强的解释力。为了获得经季度调整的序列, 下面使用命令 predict 来获得上述回归的残差项(记为 sales\_sa)。

. predict sales\_sa,r

然而, OLS 残差项的平均值一定为 0, 故需要把原序列的均值加回去, 并记季节调整序列为 sales\_sa\_reg。

```
. sum sales
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sales	40	105.6178	4.056961	97.84603	112.9617

```
. gen sales_sa_reg = sales_sa + r(mean)
```

下面,将回归法的季节调整序列与原序列画图,参见图 20.14。

```
. tsline sales_sa_reg sales,lpattern("_)")
```

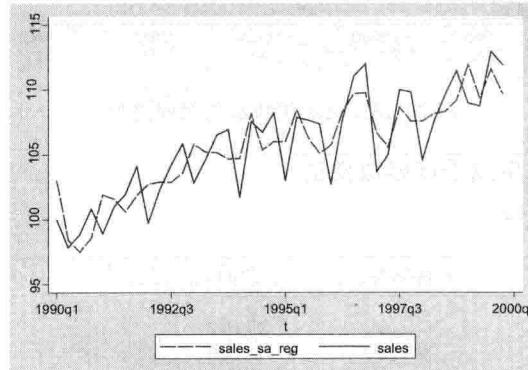


图 20.14 回归法的季节调整效果

从图 20.14 可以看出,经回归法进行季节调整后,序列变得更为光滑,季节性波动的特征已不太明显。回归法的缺点是,它假设每年的季节调整因子均不变(可以理解为季节固定效应),但对于移动假日(比如,春节有时落在 1 月,有时落在 2 月),每年的季节调整因子可能不尽相同。为此,下面转而介绍移动平均比率法。

(2) 移动平均比率法。移动平均比率法的基础是移动平均(Moving Average)。奇数项移动平均很容易进行。考虑数据  $\{y_1, \dots, y_T\}$ , 则时点  $t$  的 3 项移动平均为

$$MA_t = \frac{y_{t-1} + y_t + y_{t+1}}{3} \quad (t=2, \dots, T-1) \quad (20.95)$$

显然,进行 3 项移动平均后的序列  $MA_t$  缺失第一项与最后一项,可通过插值来补齐,比如,  $MA_1 = (2y_1 + y_2)/3$ , 而  $MA_T = (2y_T + y_{T-1})/3$ 。消除季节影响的简单方法为,对季度数据进行 4 项移动平均;而对月度数据进行 12 项移动平均。然而,由于 4 与 12 均为偶数,故进行以下“中心化移动平均”。下面以季度数据为例,求  $t=3$  时点上的 4 项平均。

$$MA_{2.5} = \frac{y_1 + y_2 + y_3 + y_4}{4} \quad (20.96)$$

$$MA_{3.5} = \frac{y_2 + y_3 + y_4 + y_5}{4} \quad (20.97)$$

定义相应的中心化移动平均值为

$$MA_3 = \frac{1}{2}(MA_{2.5} + MA_{3.5}) = \frac{1}{2}\left(\frac{y_1 + y_2 + y_3 + y_4}{4} + \frac{y_2 + y_3 + y_4 + y_5}{4}\right) \quad (20.98)$$

以此类推,中心化移动平均的一般公式为

$$\begin{aligned}
 MA_t &= \frac{1}{2} \left( \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1}}{4} + \frac{y_{t-1} + y_t + y_{t+1} + y_{t+2}}{4} \right) \\
 &= \frac{1}{4} \left( \frac{y_{t-2}}{2} + y_{t-1} + y_t + y_{t+1} + \frac{y_{t+2}}{2} \right)
 \end{aligned} \quad (20.99)$$

从公式(20.99)可知,中心化移动平均也可视为加权平均,即各期权重不完全相同。类似地,可以计算适用于月度数据的12项中心化移动平均。下面主要以加法模型为例,说明季节调整的“移动平均比率法”(ratio to moving average)。该法分为以下五个步骤:

- ① 对季度(月度)数据 $Y_t$ 进行4项(12项)中心化移动平均,得到趋势循环序列 $TC_t$ 。
- ② 计算季节要素 $S_t$ 与不规则要素 $I_t$ 之和: $SI_t \equiv S_t + I_t = Y_t - TC_t$ 。注:如果使用乘法模型,则 $SI_t \equiv S_t \times I_t = Y_t / TC_t$ 。
- ③ 对于季(月)度数据的 $SI_t$ ,分别计算其第 $j$ 季(月)的季(月)度平均值,得到季(月)节因子 $s_j$ 。例如,如果 $j=2$ ,则计算 $SI_t$ 序列在整个数据期间所有第2季(月)数据的平均值。
- ④ 调整季节因子,使得它们的和为0,得到标准化的季节因子 $S_t \equiv s_t - \frac{1}{k} \sum_{j=1}^k s_j$ ,其中 $k=4$ (季度数据),或 $k=12$ (月度数据)。注:如果使用乘法模型,则 $S_t \equiv \frac{s_t}{\sqrt{s_1 \times \dots \times s_k}}$ 。

⑤ 计算季节调整的最终结果: $TCI_t \equiv Y_t - S_t$ 。注:如果使用乘法模型,则 $TCI_t \equiv Y_t / S_t$ 。

其中,第③—④步的目的是将季节序列 $S_t$ 从序列 $SI_t \equiv S_t + I_t$ 中分离出来。

下面继续以数据集turksales.dta为例,使用加法模型进行移动平均比率法的季节调整。首先进行中心化移动平均,得到趋势循环序列 $TC_t$ 。

. tssmooth ma sales1 = sales, window(2 1 1)

此命令表示对序列sales进行移动平均,并将结果记为sales1;选择项“window(2 1 1)”表示在移动平均时,包括过去2项,当期1项,以及未来1项。

```
The smoother applied was
(1/4)*[x(t-2) + x(t-1) + 1*x(t) + x(t+1)]; x(t)= sales
```

. tssmooth ma sales2 = sales, window(1 1 2)

其中,选择项“window(1 1 2)”表示在移动平均时,包括过去1项,当期1项,以及未来2项。然后进行中心化。

```
The smoother applied was
(1/4)*[x(t-1) + 1*x(t) + x(t+1) + x(t+2)]; x(t)= sales
```

. gen sales\_tc = (sales1 + sales2) / 2

再将原序列减去趋势循环序列 $TC_t$ 。

. gen sales\_si = sales - sales\_tc

下面,通过循环语句forvalues来计算每一季度的平均值,即季节因子 $s_i$ <sup>①</sup>。

```
. forvalues i = 1 / 4 {
    quietly sum sales_si if q`i' == 1
    scalar s`i' = r(mean)
```

① 也可以不使用循环语句,而直接对每一季度计算一次。但如果是月度数据,需要重复12次,比较麻烦。

```
4. }
```

然后将这四个季度的平均值再进行平均。

```
. scalar s_mean = (s1 + s2 + s3 + s4) / 4
```

接下来,再以循环语句将每个季节因子减去季节因子的平均值,得到标准化的季节因子,然后将原序列减去标准化的季节因子,即得到季节调整序列。

```
. gen sales_sa_ma = 0 (初始化季节调整序列为 0,以便后面更新)
. forvalues i = 1 / 4 {
    2. scalar s`i'=s`i'-s_mean
    3. replace sales_sa_ma = sales - s`i' if q`i'= = 1
    4. }
```

最后,看一下移动平均比率法的季节调整序列与原序列的时间序列图,结果见图 20.15。

```
.tsline sales_sa_reg sales_sa_ma sales,lpattern("—" -")
```

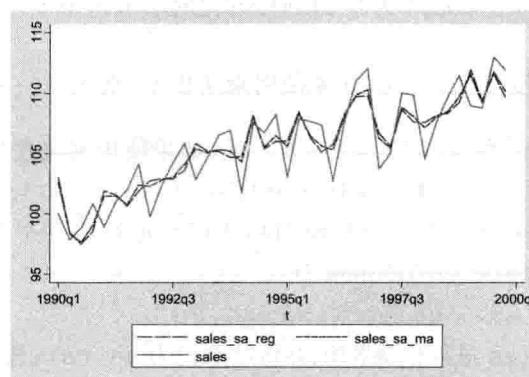


图 20.15 回归法与移动平均比率法的季节调整结果比较

从图 20.15 可以看出,经过回归法与移动平均比率法的季节调整序列十分接近。为了与加法模型对比,下面使用乘法模型(相关变量名加后缀 *m* 以示区别,*m* 表示 multiplicative)。

```
. gen sales_sim = sales / sales_tc
. forvalues i = 1 / 4 {
    2. quietly sum sales_sim if q`i'= = 1
    3. scalar sm`i'=r(mean)
    4. }
. scalar s_meanm = (sm1 * sm2 * sm3 * sm4)^0.25
. gen sales_sa_mam = 0 (初始化季节调整序列为 0,以便后面更新)
. forvalues i = 1 / 4 {
    2. scalar sm`i'=sm`i'/s_meanm
    3. replace sales_sa_mam = sales / sm`i' if q`i'= = 1
    4. }
```

下面,对比加法模型与乘法模型的季节调整结果,参见图 20.16。

```
.tsline sales_sa_ma sales_sa_mam,lpattern("—" -")
```

从图 20.16 可知,乘法模型与加法模型的季节调整结果十分接近,没有实质区别。但二者的

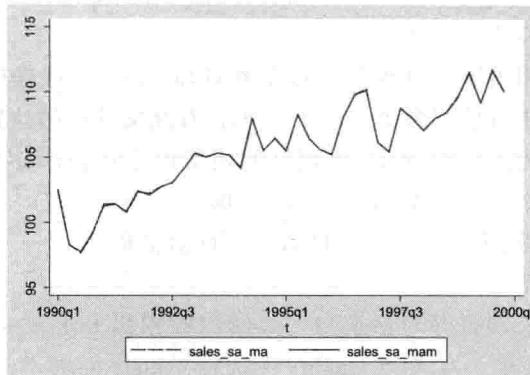


图 20.16 加法模型与乘法模型的季节调整结果对比

取值并不完全相同。下面看一下前 5 个观测值。

```
. list sales_sa_ma sales_sa_mam in 1 / 5
```

	sales~ma	sales~am
1.	102.4991	102.3909
2.	98.29412	98.24789
3.	97.64602	97.74304
4.	99.07458	99.17122
5.	101.4089	101.2747

(3) X12 方法。1954 年美国商务部人口普查局首次开发了可在计算机上运行的季节调整程序, 称为 X1<sup>①</sup>。以后该程序每次改进都以 X 加上序号表示。1965 年推出基于移动平均法的 X11, 很快成为全世界统计机构进行季节调整的标准方法。X11 的基本方法为通过多次移动平均来分解原序列, 并对极端值(outlier)进行自动调整<sup>②</sup>, 以得到每一步都更为准确的结果。对于时间序列的两端观测值, X11 进行单边移动平均(one-sided moving average), 但这种非对称滤波法(asymmetric filter)的质量不高。为此, 加拿大统计局于 1980 年提出 X11 - ARIMA, 使用 ARIMA 建模方法对原序列进行前推(forecast)与后推(backcast), 然后对加长的序列进行 X11 季节调整。

美国人口普查局于 1998 年推出 X12 - ARIMA, 基本方法与 X11 - ARIMA 一致, 主要增加了对交易日、节假日影响的调节功能, 以及对各种极端值的处理。这些极端值包括“单点异常值”(Additive Outlier, 简记 AO), 即发生于某时刻的异常值; “水平移动”(Level Shift, 简记 LS), 表示在瞬间变化到一个新水平并维持在此新水平上; 以及“暂时变化”(Temporary Change, 简记 TC), 表示在瞬间变化到一个新水平, 但逐渐恢复到原水平。X12 - ARIMA 分为两个模块, 即 regARIMA 与 X11。模块 regARIMA 使用线性回归的方法(带 ARIMA 扰动项)来处理交易日、移动假日与极端值的影响, 并对原序列进行前推与后推拓展。

在 Stata 12 中<sup>③</sup>, 可通过下载南开大学王群勇老师撰写的 sax12 命令来实现 X12 - ARIMA。

① X 表示 eXperimental。

② 极端值会影响季节调整的准确性。在季节调整后, 再把极端值加回去。

③ 在 Stata 11 中也可以运行命令 sax12, 但无法运行后续命令 sax12im。

下载方法为(或输入命令“findit sax12”,找到此命令后下载):

```
.net install st0255.pkg
```

命令 sax12 需要调用美国人口普查局的程序 x12a.exe。下载程序 x12a.exe 后<sup>①</sup>,将其放入 Stata 的 PLUS 文件夹即可(可使用命令 sysdir 来查看此文件夹的位置)。由于命令 sax12 包含很多选择项,且有些选项无法同时使用,故建议使用菜单方式(menu-driven)来执行此命令。有关此命令的详细说明与示例,参见 Wang and Wu(2012)。

下载命令 sax12 之后,输入以下命令即可打开此对话框。

```
.db sax12
```

其中,db 表示 dialog begin,即打开对话窗口。继续以数据集 turksales.dta 中的变量 sales 为例。对话窗口提供了一系列菜单。在菜单“main”中选择“single series”,然后在 Variable 栏中选择 sales,作为季节调整的对象。菜单“if/in”用于限制限制样本区间,在此忽略。菜单“prior”用于指定对原序列先进行某种变换(比如,取对数),再进行季节调整。为了与上文的做法一致,此处不进行任何变换,即在“Data transformation”栏中选择“none”<sup>②</sup>。在菜单“regression”中选择“constant”(常数项)与“trading day with leap year”以剔除贸易日与闰年的影响。在菜单“regression”之下,还允许设定“User defined variables”,比如用于捕捉中国数据中的春节效应;在本例中不作设定。在菜单“outlier”中选择所有的极端值种类,即 AO,LS 与 TC;并选择“Automatic selection given maximum lag and difference”。在菜单“adjustment”中选择“X11 mode:additive”<sup>③</sup>。菜单“other”提供了对季节调整结果的诊断方法。最后点击 OK。

此时,会发现 Stata 命令窗口出现如下命令:

```
. sax12 sales, satype (single) inpref (sales.spc) outpref (sales)
transfunc(none) regpre(const td) outauto (ao ls tc) outlsrun(0) ammaxlag(2
1) ammaxdiff(2 1) ammaxlead(12) x11mode (add) x11seas(x11default)
```

显然,此命令的选择项过多,故适合进行菜单操作。为了调用(import)季节调整序列,打开另一菜单对话框:

```
. db sax12 im
```

在此对话框中,点击 Browse,找到文件 sales.out 之后将其打开。在“Select the series or insert the suffix”选项中选择“seasonally adjusted series(d11)”,然后点击 OK。此时,会发现变量 sales\_d11 出现在变量窗口,这就是经过 X12-ARIMA 季节调整后的序列。下面,对比 X12-ARIMA 与移动平均比率法的季节调整结果,参见图 20.17。

```
.tsline sales_sa_ma sales_d11,lpattern(" - ")
```

从图 20.17 可知,经 X12-ARIMA 季节调整的序列(图中实线)与移动平均比率法(图中虚线)在某些区间十分类似,但在有些区间则前者明显比后者更为平滑,这可能是因为 X12-ARIMA 进行了多次移动平均,并且处理了极端值。

X12 季节调整也可以在 Eviews 中实现。假设数据存于 Excel 表中,基本步骤如下:File→Open→Eviews Workfile→文件类型选择“Excel file (\*.xls)”→打开→双击 Workfile 中需要进行季节调整的变量→点击 proc→Seasonal Adjustment→Census X12→根据对话框进行选项,包括指

<sup>①</sup> 下载地址为 [http://www.census.gov/srd/www/x12a/x12downv03\\_pc.html](http://www.census.gov/srd/www/x12a/x12downv03_pc.html)。

<sup>②</sup> 如果选择“Auto(log or none)”则将根据信息准则自动选择是否先对原序列取对数,再进行季节调整。

<sup>③</sup> 这也是为了与上文的加法模型相一致。如果使用乘法模型,则选择“multiplicative”。

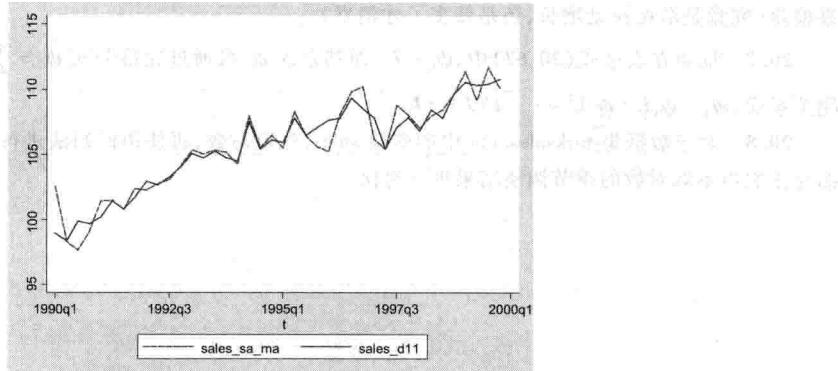


图 20.17 对比移动平均比率法与 X12-ARIMA

定需要保存的序列(component series to save)。详见高铁梅(2009)。

需要注意的是,Eviews 中的 X12 程序仅针对美国的节假日而设,并不完全适用于中国。为此,2009 年国家统计局与南开大学数量经济研究所组成联合研究小组,开发了针对中国节假日的 NBS-SA 季节调整软件。以此软件为基础,国家统计局于 2011 年 4 月开始发布 GDP 等指标的环比数据,结束了中国基本不生产环比统计数据的历史。Wang and Wu(2012)介绍了一个实例,利用 Stata 命令 sax12 来处理中国数据的春节效应。

总之,X12 依然是目前最权威的季节调整方法,为美国、英国、加拿大等统计局所采用。有关季节调整的更详细说明,参见栾惠德、张晓峒(2007),高铁梅(2009),以及《时间序列 X-12-ARIMA 季节调整——原理与方法》(中国人民银行调查统计司,2006)。

## 习 题

**20.1** (部分调整模型,partial adjustment model) 记  $y_t^*$  为  $y_t$  的最优值或理想值(比如, $y_t$  为资本存量,而  $y_t^*$  为最优资本存量),满足以下关系,  $y_t^* = \alpha + \beta x_t + u_t$ , 其中  $u_t$  为独立于  $\{x_t\}$  的扰动项。假设  $y_t$  不能立即调整到最优值  $y_t^*$ , 而只能进行部分调整, 即  $y_t - y_{t-1} = \theta(y_t^* - y_{t-1})$ , 其中  $0 < \theta < 1$ 。证明  $\{y_t\}$  为 ADL 模型。

**20.2** (使用 Yule-Walker equations 求解自协方差) 对于 AR(2) 模型,  $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t$ , 其中  $\varepsilon_t$  为白噪声, 对此方程两边分别求与  $y_{t-k}$  的协方差, 其中  $k=0,1,2$ , 得到关于  $\gamma_0, \gamma_1, \gamma_2$ (0,1,2 阶自协方差)的三元一次方程组。对于  $k>2$ , 给出  $\gamma_k$  的递推公式(将  $\gamma_k$  表示为  $\gamma_{k-1}$  与  $\gamma_{k-2}$  的函数)。

**20.3** 假设模型为  $y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \gamma_0 x_t + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \varepsilon_t$ , 推导其对应的误差修正模型。提示: 假设  $(x, y)$  之间的长期均衡关系为  $y = \phi + \theta x$ 。

**20.4** 使用美国 1960—2002 年季度数据集 macro\_swanson.dta, 估计如下的经验菲利普斯曲线 ADL(2,1) 模型:

$$\Delta \inf_t = \beta_0 + \beta_1 \Delta \inf_{t-1} + \beta_2 \Delta \inf_{t-2} + \beta_3 \text{unem}_{t-1} + \varepsilon_t \quad (20.100)$$

其中,  $\inf$  为通货膨胀率,  $\text{unem}$  为失业率。使用 BG 检验, 检验是否存在一阶自相关; 如果有, 则使用滞后一阶的 HAC 标准误。 $\hat{\beta}_3$  的符号是否与宏观理论相一致?

**20.5** 数据集 lutkepohl2.dta 包含了联邦德国 1960—1982 年的以下季度宏观变量:  $\text{inv}$ (投资),  $\text{inc}$ (收入),  $\text{consump}$ (消费),  $\ln_{\text{inv}}$ (投资之对数),  $\ln_{\text{inc}}$ (收入的对数),  $\ln_{\text{consump}}$ (消费的对数),  $\text{dln}_{\text{inv}}$ ( $\ln_{\text{inv}}$  的一阶差分),  $\text{dln}_{\text{inc}}$ ( $\ln_{\text{inc}}$  的一阶差分),  $\text{dln}_{\text{consump}}$ ( $\ln_{\text{consump}}$  的一阶差分)。估计一个关于( $\text{dln}_{\text{inv}}, \text{dln}_{\text{inc}}, \text{dln}_{\text{consump}}$ ) 的 VAR 模型, 并进行相关的检验。

**20.6** 使用数据集 consumption\_china.dta 估计一个有关( $c, y$ ) 的 VAR 模型, 画脉冲响应图, 并进行格兰杰因

果检验(究竟是消费拉动增长,还是钱多了才消费)。

**20.7** 证明在表达式(20.67)中,  $\psi_0 = I_n$ , 而其余的  $\psi_i$  可通过递推公式  $\psi_i = \sum_{j=1}^i \psi_{i-j} \Gamma_j$  来确定。提示: 使用关系式  $(\psi_0 + \psi_1 L + \psi_2 L^2 + \cdots) \Gamma(L) = I_n$ 。

**20.8** 对于数据集 turksales.dta 中的变量 sales, 先取对数, 再使用回归法进行季节调整, 并将此结果通过画图与正文中不取对数的季节调整结果进行对比。

# 第 21 章 单位根与协整

## 21.1 非平稳序列

如果一个时间序列不是平稳过程，则称为“非平稳序列”(non-stationary time series)。在以下几种情况下，都有可能出现非平稳序列。

(1) 确定性趋势：如果一个时间序列有一个“确定性趋势”(deterministic trend)，则为非平稳序列。比如， $y_t = \beta_0 + \beta_1 t + \varepsilon_t$ 。显然， $E(y_t) = \beta_0 + \beta_1 t$  随时间而改变，故不是平稳序列。对于这种非平稳序列，只要把时间趋势去掉，就变成平稳序列，故称为“趋势平稳”(trend stationary)序列。

(2) 结构变动(structural break)：如果一个时间序列存在结构变动，则为非平稳序列。对此，可用邹检验(Chow test)进行检验，参见第 9 章。

(3) 随机趋势：另一种导致非平稳的趋势为“随机趋势”(stochastic trend)。比如，随机游走模型(random walk)：

$$y_t = y_{t-1} + \varepsilon_t \quad (21.1)$$

其中， $\{\varepsilon_t\}$  为白噪声。由于  $\Delta y_t = \varepsilon_t$ ，故来自  $\{\varepsilon_t\}$  的任何扰动对  $\{y_t\}$  都具有永久性的冲击，其影响力不随时间而衰减，故称  $\{\varepsilon_t\}$  为这个模型的“随机趋势”。在上式中，如果包含常数项，则为“带漂移的随机游走”(random walk with drift)：

$$y_t = \beta_0 + y_{t-1} + \varepsilon_t, \quad \beta_0 \neq 0 \quad (21.2)$$

其中， $\beta_0$  为每个时期的平均“漂移”(drift)，因为  $E(y_t) = \beta_0 + E(y_{t-1})$ 。显然，随机游走是 AR(1) 的特例。对于 AR(1)， $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ ，如果  $\beta_1 = 1$ ，则为随机游走。对于随机游走，只要对其进行一阶差分，就可以得到平稳序列，故也称为“差分平稳”(difference stationary)序列。

**定义** 称平稳的时间序列为“零阶单整”(Integrated of order zero)<sup>①</sup>，记为 I(0)。如果时间序列的一阶差分为平稳过程，则称为“一阶单整”(Integrated of order one)，记为 I(1)，也称为“单位根过程”(unit root process)。更一般地，如果时间序列的  $d$  阶差分为平稳过程，则称为“ $d$  阶单整”(Integrated of order  $d$ )，记为 I( $d$ )。

对于 I(0) 序列，由于它是平稳的，故长期而言有回到其期望值的趋势。这种性质被称为“均值回复”(mean-reverting)。非平稳的 I(1) 序列则会“到处乱跑”(wander widely)，没有上述性质。另外，I(0) 序列对于其过去的行为只有有限的记忆，即发生在过去的扰动项对未来的影响随时间而衰减；而 I(1) 序列则对过去的行为具有无限长的记忆，即任何过去的冲击都将永久性地改变未来的整个序列。比如，假设  $\{y_t\}$  是 GDP 的时间序列，且为 I(1)，则任何对货币政策或财政政

① 由于只考虑单一变量，故称为“单整”，与后面的多变量“协整”概念相呼应。

策的调整都将对未来的 GDP 产生永久 (permanent) 的影响<sup>①</sup>。

**定义** 如果时间序列  $\{y_t\}$  的  $d$  阶差分为平稳的 ARMA( $p, q$ ) 过程, 则称  $\{y_t\}$  为 ARIMA( $p, d, q$ ) 过程。最常见的为 ARIMA( $p, 1, q$ ), 即经过一次差分就得到平稳的 ARMA( $p, q$ )。

## 21.2 ARMA 的平稳性

在什么情况下, ARMA( $p, q$ ) 才平稳呢? 显然, MA( $q$ ) 是平稳的, 因为它是有限个白噪声的线性组合。因此, ARMA( $p, q$ ) 的平稳性取决于其 AR( $p$ ) 的部分。从第 5 章已经知道, 对于 AR(1),  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ , 如果  $\varepsilon_t$  平稳且  $|\beta_1| < 1$ , 则为平稳过程。更一般地, 考虑 AR( $p$ ) 的平稳性, 即  $y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t$ 。

回顾 AR(1) 的情形, “ $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ ” 其实是一阶随机差分方程, 其稳定性与对应的确定性差分方程“ $y_t = \beta_0 + \beta_1 y_{t-1}$ ”是一样的。因此, 只要考虑一阶差分方程“ $y_t = \beta_0 + \beta_1 y_{t-1}$ ”是否有稳定解即可, 而这个非齐次(含常数项  $\beta_0$ ) 差分方程的解取决于对应的齐次(不含常数项) 差分方程“ $y_t = \beta_1 y_{t-1}$ ”的通解  $y_t = y_0 \beta_1^t$ (解的形式为指数函数)。因此, 其稳定条件为  $|\beta_1| < 1$ 。

对于 AR( $p$ ), 考虑其对应的确定性齐次差分方程:  $y_t = \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p}$ 。假设其解的形式仍为指数函数, 即  $y_t = z^{-t} = (1/z)^t$ , 其中  $z$  待定。将此解代入差分方程可得

$$z^{-t} - \beta_1 z^{-(t-1)} - \cdots - \beta_p z^{-(t-p)} = 0 \quad (21.3)$$

将上式两边同乘以  $z^t$  可得特征方程:

$$\phi(z) \equiv 1 - \beta_1 z - \cdots - \beta_p z^p = 0 \quad (21.4)$$

这个多项式方程在复数域中一定有  $p$  个根(包括重根)。与此对应, 齐次差分方程也有  $p$  个形如  $(1/z)^t$  的解<sup>②</sup>, 而其通解则是这  $p$  个解的线性组合。给定初始条件  $\{y_0, y_1, \dots, y_{p-1}\}$ , 则可求出此齐次差分方程的唯一特解。显然, 如果要求  $\{y_t\}$  收敛于一个稳定值, 则特征方程所有解的范数  $\|z\|$  (即在复平面上  $z$  离原点的距离)都必须大于 1, 故所有解必须都落在复平面上的单位圆之外, 参见图 21.1<sup>③</sup>。如果某个根正好落在单位圆之上, 则称为“单位根”(unit root), 比如随机

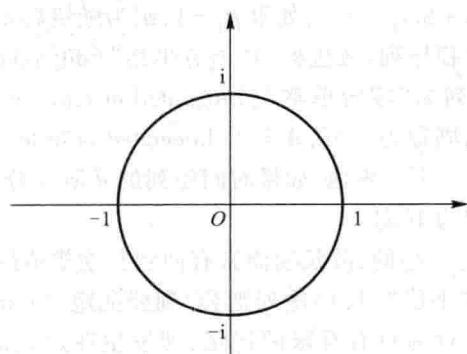


图 21.1 复平面上的单位圆

<sup>①</sup> 这是一个很强的结论, 因为通常我们认为货币政策或财政政策会对宏观经济产生持久 (persistent) 的影响(自回归系数  $\beta_1$  接近于 1), 即其作用将持续一段时间, 但不会是永久的 (permanent)。

<sup>②</sup> 如果有重根, 则其解的形式为  $t^m (1/z)^t$ , 其中  $m$  表示共有  $m$  重根。这并不影响有关 AR( $p$ ) 稳定性的结论。

<sup>③</sup> 如果将特征方程定义为 “ $z^p - \beta_1 z^{p-1} - \cdots - \beta_p = 0$ ”, 则结论与此相反。

游走的情形。如果特征方程的某个根落在单位圆之内，则为爆炸式(explosive)增长的非平稳过程。

**例** 对于 AR(1), 其特征方程为  $1 - \beta_1 z = 0$ , 故  $z = 1/\beta_1$ 。因此  $\|z\| = |z| > 1 \Leftrightarrow |\beta_1| < 1$ 。显然, 有关 AR( $p$ ) 稳定性的结论是对 AR(1) 情形的推广。

## 21.3 VAR 的平稳性

有关一维 AR( $p$ ) 的平稳性条件可以推广到多维 VAR( $p$ ) 的情形。考虑以下 VAR( $p$ ) 模型:

$$\mathbf{y}_t = \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (21.5)$$

其中,  $\{\boldsymbol{\varepsilon}_t\}$  为向量白噪声过程。可以证明(参见 Lutkepohl, 2005), 如果对于复数  $z$ , 特征方程

$$|\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}_1 z - \cdots - \boldsymbol{\Gamma}_p z^p| = 0 \quad (21.6)$$

的所有根都落在复平面的单位圆之外(即  $\|z\| > 1$ ), 则此 VAR( $p$ ) 为平稳过程。在方程(21.6)中,  $|\cdot|$  表示行列式。特别地, 当  $p=1$  时, VAR(1) 的平稳性要求  $|\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}_1 z| = 0$  的所有根都满足  $\|z\| > 1$ , 即  $\|1/z\| < 1$ 。由于  $|\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}_1 z| = |z| \cdot |(1/z)\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}_1|$ , 故  $|(1/z)\boldsymbol{\Gamma}_n - \boldsymbol{\Gamma}_1| = 0$ , 因此  $1/z$  为矩阵  $\boldsymbol{\Gamma}_1$  的特征值(根据特征值的定义)。这意味着, VAR(1) 的平稳性要求  $\boldsymbol{\Gamma}_1$  的所有特征值都落在单位圆之内(即  $\|1/z\| < 1$ )。由于矩阵的特征值求解方便, 故此判断条件容易应用。对于一般的 VAR( $p$ ), 可以先将其写为 VAR(1) 的形式, 然后再通过特征值来判断其平稳性。

首先, 定义如下三个  $np \times 1$  的列向量:

$$\tilde{\mathbf{y}}_t = \begin{pmatrix} \mathbf{y}_t \\ \mathbf{y}_{t-1} \\ \vdots \\ \mathbf{y}_{t-p+1} \end{pmatrix}_{np \times 1}, \quad \tilde{\boldsymbol{\Gamma}}_0 = \begin{pmatrix} \boldsymbol{\Gamma}_0 \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}_{np \times 1}, \quad \tilde{\boldsymbol{\varepsilon}}_t = \begin{pmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}_{np \times 1} \quad (21.7)$$

其次, 定义  $np \times np$  “伴随矩阵”(companion matrix):

$$\tilde{\boldsymbol{\Gamma}} = \begin{pmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \cdots & \boldsymbol{\Gamma}_p \\ \boldsymbol{\Gamma}_n & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Gamma}_n & \mathbf{0} \end{pmatrix}_{np \times np} \quad (21.8)$$

由此, 可将 VAR( $p$ ) 模型(21.5)写为以下 VAR(1) 的形式:

$$\tilde{\mathbf{y}}_t = \tilde{\boldsymbol{\Gamma}}_0 + \tilde{\boldsymbol{\Gamma}} \tilde{\mathbf{y}}_{t-1} + \tilde{\boldsymbol{\varepsilon}}_t \quad (21.9)$$

因此, VAR( $p$ ) 平稳性要求其伴随矩阵  $\tilde{\boldsymbol{\Gamma}}$  的所有特征值都落在单位圆之内。

## 21.4 单位根所带来的问题

对于 AR(1), 一般从理论上认为, 不太可能出现  $|\beta_1| > 1$  的情形, 否则任何对经济的扰动都将被无限放大。因此, 经济学家通常只担心存在单位根的情形, 即  $\beta_1 = 1$ 。如果时间序列存在单位根, 则为非平稳序列, 可能带来以下问题。

(1) 自回归系数的估计值向左偏向于 0。假设对于 AR(1),  $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$ , 其真实值为  $\beta_1 = 1$ 。然而,  $\hat{\beta}_1$  的 OLS 估计量  $\hat{\beta}_1$  却不是渐近正态分布, 甚至不是对称分布 (即使是在大样本中), 而是向左偏向于 0。这是因为, 由于  $\{y_t\}$  不是平稳序列, 中心极限定理不再适用。虽然  $\operatorname{plim}_{n \rightarrow \infty} \hat{\beta}_1 = \beta_1$  (仍为一致估计), 但在有限样本下可能存在较大偏差。使用蒙特卡罗法可以得到  $\hat{\beta}_1$  的大样本分布, 参见图 21.2 (产生该图的 Stata 程序参见附录)。

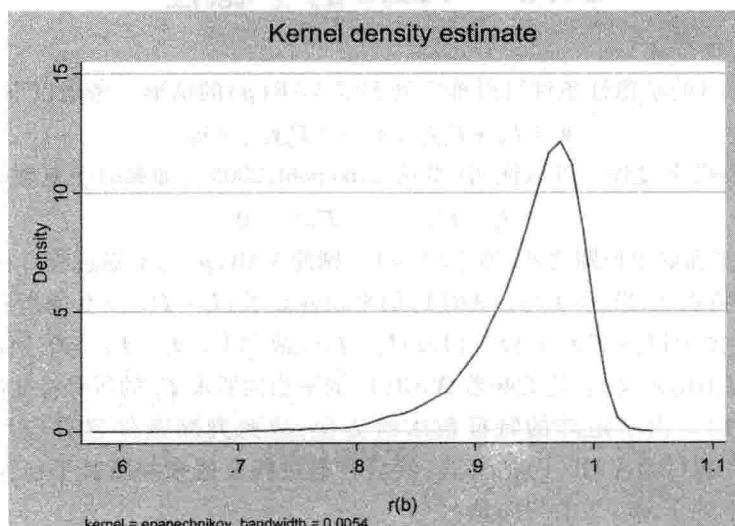


图 21.2 在单位根情况下  $\hat{\beta}_1$  的大样本分布 ( $n = 10000$ )

(2) 传统的  $t$  检验失效。由于  $\hat{\beta}_1$  不是渐近正态分布,  $t$  统计量也不服从渐近标准正态分布, 传统的区间估计与假设检验是无效的。更一般地, 建立于平稳性假设基础之上的大样本理论不再适用。

(3) 两个相互独立的单位根变量可能出现伪回归 (spurious regression) 或伪相关。比如, 假设  $y_t = y_{t-1} + u_t$ ,  $x_t = x_{t-1} + v_t$ , 其中,  $u_t, v_t$  均为独立同分布且相互独立。因此,  $y_t$  与  $x_t$  也是相互独立的。考虑 OLS 回归,  $y_t = \alpha + \beta x_t + \varepsilon_t$ 。由于  $y_t$  与  $x_t$  相互独立, 故真实参数  $\beta = 0$ 。如果样本容量足够大, 我们期待 OLS 估计值  $\hat{\beta} \approx 0$ ,  $R^2 \approx 0$ , 然而实际结果并非如此, 因为扰动项  $\varepsilon_t = y_t - \alpha - \beta x_t$  也是非平稳的。这一结论最初由 Granger 和 Newbold (1974) 通过蒙特卡罗模拟而发现。

下面, 在 Stata 中模拟此过程。假设  $y_0 = 0$ , 则  $y_1 = 0 + u_1 = u_1$ ,  $y_2 = y_1 + u_2 = u_1 + u_2$ ,  $y_3 = y_2 + u_3 = u_1 + u_2 + u_3$ ; 以此类推,  $y_t = \sum_{s=1}^t u_s = u_1 + \cdots + u_t$ 。根据此性质, 很容易在 Stata 中定义随机游走或单位根变量。

```
. drop _all          (删去内存中已有数据)
.set obs 10000      (确定样本容量为  $T = 10000$ )
.set seed 1234
.gen u=rnormal()    (产生服从标准正态分布的扰动项  $u_t$ )
.gen y=sum(u)        (定义随机游走  $y_t = \sum_{s=1}^t u_s$ )
.set seed 12345
```

```
.gen v = rnormal() (产生服从标准正态分布的扰动项  $v_t$ )
.gen x = sum(v) (定义随机游走  $x_t = \sum_{s=1}^t v_s$ )
.reg y x
```

Source	SS	df	MS	Number of obs = 10000 F( 1, 9998) = 8835.74 Prob > F = 0.0000 R-squared = 0.4691 Adj R-squared = 0.4691 Root MSE = 20.948			
Model	3877265.25	1	3877265.25				
Residual	4387283.57	9998	438.81612				
Total	8264548.82	9999	826.537536				
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
x	.5621503	.0059804	94.00	0.000	.5504275	.5738731	
_cons	-8.53695	.252395	-33.82	0.000	-9.031695	-8.042205	

从上表可知,尽管  $y$  与  $x$  相互独立(因为  $u$  与  $v$  相互独立),但  $y$  对  $x$  的回归系数在 1% 水平上显著,而且  $R^2$  高达 0.47,存在“伪回归”。进一步,把  $u$  对  $v$  回归。

```
.reg u v
```

Source	SS	df	MS	Number of obs = 10000 F( 1, 9998) = 0.05 Prob > F = 0.8157 R-squared = 0.0000 Adj R-squared = -0.0001 Root MSE = .99336			
Model	.0535855	1	.0535855				
Residual	9865.58665	9998	.986756016				
Total	9865.64024	9999	.98666269				
u	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]		
v	.0023061	.0098959	0.23	0.816	-.0170918	.021704	
_cons	-.0039404	.0099338	-0.40	0.692	-.0234125	.0155318	

由上表可知, $u$  与  $v$  的样本值高度不相关( $p$  值高达 0.816,  $R^2 = 0.0000$ );但由  $u$  与  $v$  所产生的随机游走过程  $y$  与  $x$  却显著相关。更直

观地,看一下  $y$  与  $x$  的时间趋势图。

```
.gen t = _n (定义时间变量 t)
.line y x t,lpattern(dash)
```

从图 21.3 可知,虽然  $y$  与  $x$  为相互独立的单位根变量,但二者的样本值却比较相关,存在“伪相关”。如何避免伪相关或伪回归?方法之一,先对变量作一阶差分,然后再回归。方法之二为“协整”(Cointegration),参见下文。但首先必须对是否存在单位根进行检验。

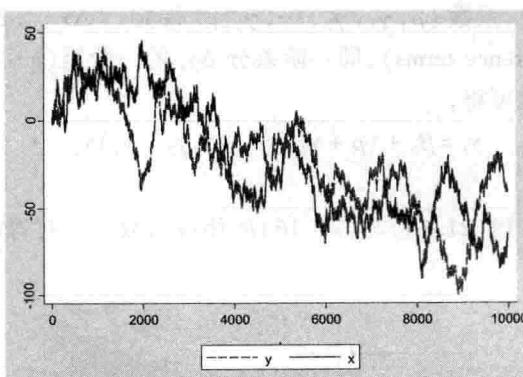


图 21.3 伪相关示意图

## 21.5 单位根检验与平稳性检验

### 1. Dickey-Fuller 单位根检验(DF 检验)

**定义** 如果  $\{\varepsilon_t\}$  独立同分布, 期望为 0, 方差有限(finite variance), 则称  $\{\varepsilon_t\}$  为“独立白噪声”(independent white noise)。换言之, 独立白噪声就是期望为 0, 方差存在的 iid 序列。

考虑以下 AR(1) 模型:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \gamma t + \varepsilon_t \quad (21.10)$$

其中,  $\gamma t$  为时间趋势(如果不含时间趋势, 可令  $\gamma = 0$ );  $\varepsilon_t$  为独立白噪声。考虑以下单边检验:

$$H_0: \beta_1 = 1 \quad vs \quad H_1: \beta_1 < 1 \quad (21.11)$$

其中, 替代假设为 “ $H_1: \beta_1 < 1$ ”, 因为我们从理论上认为  $\beta_1 > 1$  不可能<sup>①</sup>。如果  $H_0$  成立, 则  $y_t$  为带漂移项  $\beta_0$  的随机游走; 如果不带漂移项, 可令  $\beta_0 = 0$ 。在方程(21.10)两边同时减去  $y_{t-1}$  可得

$$\Delta y_t = \beta_0 + \delta y_{t-1} + \gamma t + \varepsilon_t \quad (21.12)$$

其中,  $\delta = \beta_1 - 1$ 。对应的原假设与替代假设变为

$$H_0: \delta = 0 \quad vs \quad H_1: \delta < 0 \quad (21.13)$$

对方程(21.13)作 OLS 回归, 可得估计量  $\hat{\delta}$  及相应的  $t$  统计量。此  $t$  统计量被称为“Dickey-Fuller 统计量”(简记 DF)<sup>②</sup>, 在 Stata 中记为  $Z(t)$ 。但  $Z(t)$  并不服从渐近正态分布, 其临界值须通过蒙特卡罗模拟来获得<sup>③</sup>。显然,  $Z(t)$  越小(绝对值很大的负数), 则越倾向于拒绝原假设。因此, DF 检验是左边单侧检验, 即其拒绝域只在分布的最左边。

### 2. Augmented Dickey-Fuller 单位根检验(ADF 检验)

DF 检验使用一阶自回归来检验单位根, 要求扰动项  $\{\varepsilon_t\}$  为独立白噪声, 故扰动项无自相关。如果  $\{\varepsilon_t\}$  存在自相关, 可以引入更高阶的滞后项来控制。假设选择了适当的滞后期  $p$ , 使得以下 AR( $p$ ) 模型的扰动项  $\{\varepsilon_t\}$  为独立白噪声:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \gamma t + \varepsilon_t \quad (21.14)$$

为了方便检验, 将上式转换为以下形式:

$$y_t = \beta_0 + \rho y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \cdots + \gamma_{p-1} \Delta y_{t-p+1} + \gamma t + \varepsilon_t \quad (21.15)$$

其中, 系数  $(\rho, \gamma_1, \gamma_2, \dots, \gamma_{p-1})$  待定,  $\{\Delta y_{t-1}, \Delta y_{t-2}, \dots, \Delta y_{t-p+1}\}$  称为“滞后差分项”(lagged difference terms), 即一阶差分  $\Delta y_t$  的一阶至  $(p-1)$  阶滞后项。将上式中的差分算子去掉, 合并同类项可得,

$$y_t = \beta_0 + (\rho + \gamma_1) y_{t-1} + (\gamma_2 - \gamma_1) y_{t-2} + \cdots + (\gamma_{p-1} - \gamma_{p-2}) y_{t-p+1} - \gamma_{p-1} y_{t-p} + \gamma t + \varepsilon_t \quad (21.16)$$

将方程(21.14)与(21.16)的相应系数一一对应起来可得

<sup>①</sup> 如果  $\beta_1 > 1$ , 将导致  $y_t$  呈指数增长。对于有指数增长趋势的经济变量, 如 GDP, 通常先取对数去掉其指数趋势。

<sup>②</sup> 参见 Dickey and Fuller(1979)。

<sup>③</sup> Phillips(1987) 证明, DF 统计量的渐近分布为布朗运动(Brownian motion)的函数, 但仍然没有解析解, 必须进行数值计算。

$$\begin{cases} \beta_1 = \rho + \gamma_1 \\ \beta_2 = \gamma_2 - \gamma_1 \\ \vdots \\ \beta_{p-1} = \gamma_{p-1} - \gamma_{p-2} \\ \beta_p = -\gamma_{p-1} \end{cases} \quad (21.17)$$

在方程组(21.17)中,把 $(\beta_1, \beta_2, \dots, \beta_p)$ 作为已知数,可以解出 $(\rho, \gamma_1, \gamma_2, \dots, \gamma_{p-1})$ 的表达式。由方程组(21.17)中的最后一个方程倒推上去可得

$$\begin{cases} \rho = \beta_1 + \dots + \beta_p \\ \gamma_1 = -(\beta_2 + \dots + \beta_p) \\ \gamma_{p-2} = -(\beta_{p-1} + \beta_p) \\ \gamma_{p-1} = -\beta_p \end{cases} \quad (21.18)$$

**命题** 如果 $\rho=1$ ,则 $AR(p)$ 有一个单位根。

**证明:**由于 $\rho=1$ ,故 $\phi(1)=1-\beta_1-\dots-\beta_p=1-\rho=0$ 。因此,1是特征方程 $\phi(z)=1-\beta_1z-\dots-\beta_pz^p=0$ 的一个根,正好落在单位圆上,故 $AR(p)$ 有一个单位根。

**命题** 如果 $\rho>1$ ,则特征方程至少有一个根在单位圆之内,故 $AR(p)$ 为非平稳。

**证明:**首先, $\phi(0)=1-\beta_1 \cdot 0 - \dots - \beta_p \cdot 0 = 1$ 。

其次,由于 $\rho>1$ ,故 $\phi(1)=1-\beta_1-\dots-\beta_p=1-\rho<0$ 。

然而 $\phi(z)$ 为连续函数,根据中值定理,存在复数 $z^*$ ,满足 $0<\|z^*\|<1$ ,使得 $\phi(z^*)=0$ ,参见图21.4<sup>①</sup>。由于 $\|z^*\|<1$ , $z^*$ 落在单位圆之内,故 $AR(p)$ 为非平稳。

综上所述,为了检验 $AR(p)$ 是否有单位根,可以考虑对方程(21.15)进行回归,并检验

$$H_0: \rho = 1 \quad vs \quad H_1: \rho < 1 \quad (21.19)$$

在方程(21.15)两边同时减去 $\gamma_{t-1}$ 可得

$$\Delta y_t = \beta_0 + \delta y_{t-1} + \gamma_1 \Delta y_{t-1} + \gamma_2 \Delta y_{t-2} + \dots +$$

$$\gamma_{p-1} \Delta y_{t-p+1} + \gamma t + \varepsilon_t$$

$$(21.20)$$

其中, $\delta=\rho-1$ 。则原假设与替代假设变为

$$H_0: \delta = 0 \quad vs \quad H_1: \delta < 0 \quad (21.21)$$

对方程(21.20)使用OLS可得估计量 $\hat{\delta}$ 及相应

$t$ 统计量。此 $t$ 统计量被称为“Augmented Dickey-Fuller统计量”(简记ADF),Stata仍记其为 $Z(t)$ 。ADF检验是最常用的单位根检验。Nelson and Plosser(1982)使用ADF检验考察美国14个年度宏观经济序列,结果发现只有其中一个变量可拒绝单位根的原假设。此文一出,引起经济学界对单位根的广泛关注。

与DF检验一样,ADF检验也是左边单侧检验,其拒绝域只在分布的最左边。ADF统计量的临界值也要通过蒙特卡罗模拟得到。具体而言,ADF统计量的临界值取决于真实模型( $H_0$ )是否

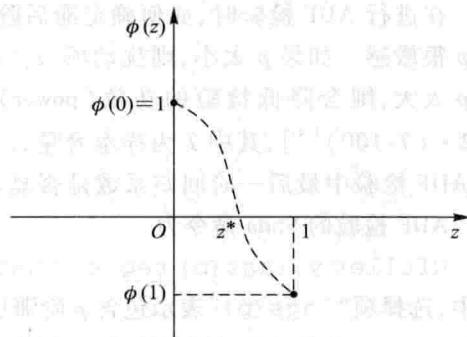


图 21.4  $\rho>1$  时  $AR(p)$  的非平稳性

<sup>①</sup> 由于 $z$ 为复数(在整个复平面上取值),故实际上应该为三维图,图21.4只是示意图。

带漂移项,以及回归方程(21.20)是否包含常数项或时间趋势。Stata 手册(Stata Manual)总结了四种情形,参见表 21.1。

表 21.1 ADF 检验的四种情形

情形	真实模型( $H_0$ )	对回归方程(21.20)的约束	Stata 选择项
1	不带漂移项	$\beta_0 = 0, \gamma = 0$	noconstant
2	不带漂移项	$\gamma = 0$	默认值
3	带漂移项	$\gamma = 0$	drift
4	不带或带漂移项	无约束	trend

表 21.1 中的情形 2,虽然真实模型不包含漂移项(即无常数项),但在 ADF 检验的回归方程(21.20)中依然包括了常数项。情形 2 与情形 3 对回归方程的约束相同,故检验统计量也相同,但临界值不同(因为真实模型不同)。

#### 关于常数项与时间趋势项

是否应该带常数项或时间趋势项,主要应从理论上考虑。比如,考察 GDP 之对数是否有单位根,一般应包含时间趋势项;而利率、汇率等则不应有时间趋势项。也可以通过画变量的时间序列图来大致判断有无长期增长趋势。另外,在作 ADF 检验时,使用选择项“regress”,可以看到常数项或时间趋势项是否显著。如果无从判断,为了稳健起见,可以把各种情况都进行检验,将结果以( $c, t, p$ )格式列表,其中“ $c = 1$ ”表示带常数项,“ $c = 0$ ”表示不带常数项;“ $t = 1$ ”表示带趋势项,“ $t = 0$ ”表示不带趋势项;而  $p$  表示滞后期数。

#### 关于滞后阶数 $p$ 的确定

在进行 ADF 检验时,如何确定滞后阶数  $p$  是个实际问题。ADF 检验的结果常常对滞后阶数  $p$  很敏感。如果  $p$  太小,则扰动项  $\{\varepsilon_t\}$  可能存在自相关,使得检验出现偏差。另一方面,如果  $p$  太大,则会降低检验的功效(power)。Schwert (1989) 建议,取最大滞后阶数为  $p_{\max} = [12 \cdot (T/100)^{1/4}]$ ,其中  $T$  为样本容量,[ · ] 表示整数部分,然后使用由大到小的序贯  $t$  规则,看 ADF 检验中最后一阶回归系数是否显著。也可使用信息准则,比如 AIC 或 BIC。

ADF 检验的 Stata 命令为

```
dfuller y, lags(p) regress noconstant drift trend
```

其中,选择项“lags(p)”表示包含  $p$  阶滞后差分项,默认为“lags(0)”,对应于 DF 检验。选择项“regress”表示显示回归结果。选择项“noconstant drift trend”的含义参见表 21.1。

#### 3. Phillips-Perron 单位根检验(PP 检验)

ADF 检验通过引入高阶滞后项来保证扰动项  $\{\varepsilon_t\}$  没有自相关。“Phillips-Perron 检验”(Phillips and Perron, 1988, 简记 PP)仍然使用一阶自回归,但使用异方差自相关稳健的标准误(Newey-West standard errors)对 DF 统计量进行修正:

$$y_t = \beta_0 + \rho y_{t-1} + \gamma t + \varepsilon_t \quad (21.22)$$

其中,  $\{\varepsilon_t\}$  可以存在异方差或自相关。经过修正的  $Z(t)$  统计量,其渐近分布与 DF 检验统计量相同,故临界值也相同,也是左边单侧检验。Phillips and Perron (1988) 还提供了另一检验统计量  $Z(\rho)$ 。使用 PP 检验须指定用于计算 Newey-West 标准误的滞后阶数(Newey-West lags)。Stata 默认的 Newey-West 滞后阶数为  $[4(T/100)^{2/9}]$ ,其中  $T$  为样本容量,而[ · ] 表示取整数。由于金融变量常存在异方差与自相关,故 PP 检验在金融数据中应用较广。在某种意义上,PP 检验相当于

异方差稳健的 ADF 检验。PP 检验的另一优点是,不必指定差分滞后项的滞后阶数。

PP 检验的 Stata 命令为

```
pperron y, noconstant trend regress lags (#)
```

其中,选择项“lags (#)”用来指定 Newey-West 滞后阶数。其他选择项的含义与 ADF 检验 (dfuller) 相同。命令 pperron 不提供选择项“drift”(对应于表 21.1 的情形 3),但表 21.1 的情形 3 只是情形 4(对应于选择项“trend”的特例),故没有太大妨碍。

#### 4. DF-GLS 单位根检验

ADF 检验与 PP 检验的共同缺点是,检验的功效较低(犯第 II 类错误的概率很大),尤其当样本容量不大,或真实模型接近于单位根的情形。具体来说,当一个时间序列为很接近于 I(1) 的 I(0) 序列时,即持续性很强的序列 (highly persistent)(比如,一阶自回归系数为 0.98),则单位根检验将很难根据数据区分其究竟为 I(1) 或 I(0)。如果在上述检验中加入常数项或时间趋势项,则将进一步降低检验的功效。

为此,Elliott, Rothenberg and Stock (1996) 提出以下的两步检验。第一步,用 GLS 估计原序列  $\{y_t\}$  的常数项与时间趋势项  $\hat{\delta}_0 + \hat{\delta}_1 t$ ,计算去势后 (detrended) 的序列  $\{y_t^d = y_t - \hat{\delta}_0 - \hat{\delta}_1 t\}$ <sup>①</sup>。第二步,对  $\{y_t^d\}$  使用 ADF 检验。这个检验被称为“DF-GLS 检验”,是目前最有功效的单位根检验。如果真实模型接近于单位根(自回归系数接近于 1),则应考虑使用 DF-GLS 检验。

DF-GLS 检验的 Stata 命令为

```
dfgls y, notrend maxlag (#) ers
```

其中,选择项“notrend”表示不带时间趋势项,即令  $\hat{\delta}_1 = 0$ ;默认带时间趋势项。选择项“maxlag (#)”用来确定 ADF 检验的最大滞后阶数,默认值为  $p_{\max} = [12 \cdot (T/100)^{1/4}]$  (Schwert, 1989)。该命令将汇报上至  $p_{\max}$  的各阶检验结果,并提供三种方法选择最优滞后阶数,即 Ng-Perron 序贯  $t$  准则 (Ng and Perron, 1995), SIC 信息准则,以及 MAIC 信息准则 (Modified AIC) (Ng and Perron, 2000)。选择项“ers”表示使用 Elliott, Rothenberg and Stock (1996) 提供的临界值,默认临界值来自 Cheung and Lai (1995)。

#### 5. KPSS 平稳性检验

以上单位根检验的原假设皆为“ $H_0$ : 有单位根”。由于单位根检验的功效不高,对于宏观经济变量,经常无法拒绝原假设,而被迫接受单位根的存在。此时,犯第 II 类错误(即在“ $H_1$ : 无单位根”为真的情况下却接受“ $H_0$ : 有单位根”)的概率较高。在某种意义上,单位根检验就相当于先假设“某人有罪”,如果他确实做了一辈子好事,才推翻此“有罪”之原假设。这种“假定有罪”的前提似乎有失公平,至少在法庭判案时,一般假设当事人无罪,除非找到有力的相反证据。

为此,Kwiatkowski, Phillips, Schmidt and Shin (1992) 提出的平稳性检验 (KPSS) 将原假设改为“ $H_0$ : 时间序列为平稳”,而替代假设变为“ $H_1$ : 有单位根”。假设时间序列  $y_t$  可分解为时间趋势、随机游走与平稳过程之和:

$$\begin{aligned} y_t &= \beta t + u_t + \varepsilon_t \\ u_t &= u_{t-1} + v_t, v_t \sim WN(0, \sigma_v^2) \end{aligned} \tag{21.23}$$

<sup>①</sup> 根据 Stock and Watson (2004, p. 549),在 DF-GLS 检验的第一步,使用广义最小二乘法估计常数项与趋势项。具体方法如下。记被检验变量为  $V_t$ 。首先计算 3 个新变量:  $V_1, X_{1t}, X_{2t}$ , 即  $V_1 = Y_1, V_t = Y_t - \alpha^* Y_{t-1}, t = 2, \dots, T, X_{1t} = 1, X_{2t} = 1 - \alpha^*, t = 2, \dots, T$ , 以及  $X_{21} = 1, X_{2t} = t - \alpha^*(t-1)$ , 其中,  $\alpha^* = 1 - 13.5/T$ 。然后将  $V_t$  对  $X_{1t}$  与  $X_{2t}$  回归,即用 OLS 来估计回归方程  $V_t = \delta_0 X_{1t} + \delta_1 X_{2t} + e_t (t = 1, \dots, T)$  的系数,其中  $e_t$  为扰动项。

其中,  $\beta t$  为时间趋势,  $u_t$  是随机游走, 而  $\varepsilon_t$  为平稳过程(允许存在异方差与自相关)。假设  $u_t$  的初始值  $u_0$  为固定, 则可视  $u_0$  为  $y_t$  的常数项。假设  $u_t$  的扰动项  $v_t$  为白噪声(White Noise, 简记 WN), 记其方差为  $\sigma_v^2$ 。在这个模型中, “ $y_t$  为趋势平稳”(trend stationary)的原假设等价于“ $H_0: \sigma_v^2 = 0$ ”(此时,  $v_t = 0$ , 故  $u_t$  为常数), 而替代假设为“ $H_1: \sigma_v^2 > 0$ ”。如果不含时间趋势( $\beta = 0$ ), 则原假设为“ $y_t$  为平稳过程”(level stationary)。对此原假设进行拉格朗日乘子检验(LM), 即得到 KPSS 统计量。KPSS 检验是单边右侧检验(正如  $\chi^2$  检验一样), 其临界值须通过蒙特卡罗模拟得到。Kwiatkowski, Phillips, Schmidt and Shin (1992) 使用 KPSS 检验考察 Nelson and Plosser (1982) 所使用的美国年度宏观时间序列, 结果发现许多变量均无法拒绝“趋势平稳”的原假设。因此, 究竟宏观经济变量是否为单位根过程仍然是存疑的。

KPSS 检验的 Stata 命令为

```
ssc install kpss (下载安装命令 kpss)
kpss y, notrend maxlag (#)
```

其中, 选择项“notrend”表示不包括时间趋势, 即在原模型中  $\beta = 0$ ; 默认包括时间趋势。选择项“maxlag (#)”用来确定最大滞后阶数, 默认值为  $p_{\max} = [12 \cdot (T/100)^{1/4}]$  (Schwert, 1989), 并汇报上至  $p_{\max}$  的各阶检验结果。

#### 6. 单整阶数(order of integration)的确定

对时间序列  $\{y_t\}$  进行单位根检验后, 如果认为  $\{y_t\}$  为非平稳, 则要进一步判断其为 I(1) 或 I(2)。可以对一阶差分  $\{\Delta y_t\}$  进行单位根检验, 如果  $\{\Delta y_t\}$  为平稳, 则  $\{y_t\}$  是 I(1)。否则, 要继续对二阶差分  $\{\Delta^2 y_t\}$  进行单位根检验。如果  $\{\Delta^2 y_t\}$  为平稳, 则  $\{y_t\}$  为 I(2), 以此类推。

## 21.6 单位根检验的 Stata 实例

下面以数据集 macro\_swanson.dta 为例<sup>①</sup>, 检验美国的季度通货膨胀率 inf 是否含有单位根。首先, 看一下 inf 的时间趋势图(如图 21.5):

```
.use macro_swanson.dta, clear
.line inf quarter
```

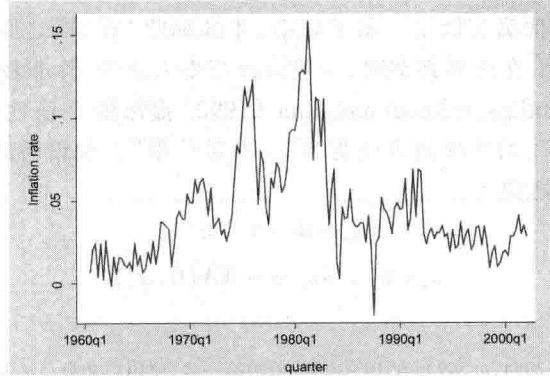


图 21.5 美国通货膨胀率的时间趋势

<sup>①</sup> 此数据集截取自 Stock and Watson (2004)。

从上图可以大致看出,通胀率 inf 应该有常数项,但没有明显的时间趋势。为此,首先考虑带常数项,但不带趋势项的 DF 检验:

```
.dfuller inf
```

Dickey-Fuller test for unit root		Number of obs = 166		
Test Statistic	Interpolated Dickey-Fuller			Value
	1% Critical Value	5% Critical Value	10% Critical Value	
Z(t)	-3.718	-3.488	-2.886	-2.576

MacKinnon approximate p-value for Z(t) = 0.0039

由于 DF 统计量为  $-3.718 < -3.488$ (左边单侧检验),故可以在 1% 的水平上拒绝“存在单位根”的原假设。由于 DF 检验中的扰动项可能存在自相关,故要考虑更高阶的 ADF 检验。首先,计算 Schwert (1989)建议的最大滞后阶数  $p_{\max} = [12 \cdot (T/100)]^{1/4}$ :

```
.di 12 * (167 / 100)^^(1 / 4)
```

13.641445

这表明, $p_{\max} = 13$ 。下面,令  $\hat{p} = 13$ ,进行 ADF 检验:

```
.dfuller inf, lags(12) reg
```

Augmented Dickey-Fuller test for unit root		Number of obs = 154		
Test Statistic	Interpolated Dickey-Fuller			Value
	1% Critical Value	5% Critical Value	10% Critical Value	
Z(t)	-1.782	-3.492	-2.886	-2.576

MacKinnon approximate p-value for Z(t) = 0.3893

D.inf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
inf	-.0874144	.0490495	-1.78	0.077	-.1843878 .009559
L1.	-.1754754	.0893062	-1.96	0.051	-.3520386 .0010877
LD.	-.2646741	.0894366	-2.96	0.004	-.441495 -.0878533
L2D.	.1934371	.0912662	2.12	0.036	.012999 .3738752
L3D.	-.0410425	.091365	-0.45	0.654	-.2216761 .1395911
L4D.	.0994029	.0884824	1.12	0.263	-.0755316 .2743374
L5D.	.079468	.0887313	0.90	0.372	-.0959585 .2548946
L6D.	.083617	.0885392	0.94	0.347	-.0914297 .2586638
L7D.	-.196626	.0885273	-2.22	0.028	-.3716493 -.0216028
L8D.	-.1043182	.0890171	-1.17	0.243	-.2803097 .0716733
L9D.	-.0788603	.0888823	-0.89	0.376	-.2545854 .0968648
L10D.	-.0765305	.0845429	-0.91	0.367	-.2436763 .0906153
L11D.	-.1506183	.0809574	-1.86	0.065	-.3106754 .0094388
_cons	.0041168	.0025237	1.63	0.105	-.0008727 .0091064

上表显示,最后一阶滞后项(L12D.)在5%的水平上并不显著。依次令 $\hat{p}=12, 11, 10$ ,进行ADF检验,最后一阶滞后项仍然不显著(过程略)。下面,令 $\hat{p}=9$ ,再进行ADF检验。

```
.dfuller inf, lags(8) reg
```

Augmented Dickey-Fuller test for unit root				Number of obs = 158	
Test Statistic	Interpolated Dickey-Fuller				
	1% Critical Value	5% Critical Value	10% Critical Value		
Z(t)	-2.607	-3.491	-2.886	-	-2.576

MacKinnon approximate p-value for Z(t) = 0.0915

D.inf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
inf					
L1.	-.1168873	.0448301	-2.61	0.010	-.2054771 -.0282974
LD.	-.1219152	.0824072	-1.48	0.141	-.2847619 .0409315
L2D.	-.2372779	.0826736	-2.87	0.005	-.4006511 -.0739047
L3D.	.2436394	.0841557	2.90	0.004	.0773374 .4099414
L4D.	-.0054838	.0863573	-0.06	0.949	-.1761365 .1651688
L5D.	.1182747	.0853734	1.39	0.168	-.0504336 .2869829
L6D.	.0637782	.0852927	0.75	0.456	-.1047706 .2323271
L7D.	.1232936	.0811298	1.52	0.131	-.0370289 .2836161
L8D.	-.1584475	.0785274	-2.02	0.045	-.3136273 -.0032678
cons	.0053964	.0023139	2.33	0.021	.0008239 .009969

上表显示,最后一阶滞后项在5%的水平上显著地不等于0。ADF统计量 $Z(t)$ 显示,无法在5%的水平上拒绝存在单位根的原假设( $-2.607 > -2.886$ ),即可认为通胀率inf含有单位根。“麦金农的近似 $p$ 值”(MacKinnon approximate  $p$ -value)为0.0915,与此结论一致。

解决DF检验中扰动项自相关问题的另一方法为PP检验:

```
.pperron inf
```

Phillips-Perron test for unit root				Number of obs = 166	
				Newey-West lags = 4	
Test Statistic	Interpolated Dickey-Fuller				
	1% Critical Value	5% Critical Value	10% Critical Value		
Z(rho)	-21.002	-20.020	-13.832	-	-11.088
Z(t)	-3.427	-3.488	-2.886	-	-2.576

MacKinnon approximate p-value for Z(t) = 0.0101

由于PP检验也是左边单侧检验,而检验统计量均小于5%的临界值,故可在5%的水平上拒绝“存在单位根”的原假设(此结论与ADF检验相反)。

下面进行目前最有功效的单位根检验,DF-GLS检验:

```
.dfgls inf
```

DF-GLS for inf		Number of obs = 153		
Maxlag = 13 chosen by Schwert criterion				
[lags]	DF-GLS tau Test Statistic	1% Critical Value	5% Critical Value	10% Critical Value
13	-1.259	-3.500	-2.797	-2.521
12	-1.250	-3.500	-2.813	-2.536
11	-1.537	-3.500	-2.829	-2.551
10	-1.638	-3.500	-2.845	-2.565
9	-1.724	-3.500	-2.860	-2.579
8	-2.004	-3.500	-2.874	-2.592
7	-2.474	-3.500	-2.888	-2.605
6	-2.158	-3.500	-2.901	-2.617
5	-2.025	-3.500	-2.914	-2.628
4	-2.057	-3.500	-2.926	-2.639
3	-2.109	-3.500	-2.937	-2.649
2	-1.671	-3.500	-2.947	-2.659
1	-2.483	-3.500	-2.957	-2.667

Opt Lag (Ng-Perron seq t) = 12 with RMSE .0142593  
Min SC = -8.269403 at lag 3 with RMSE .0149887  
Min MAIC = -8.310303 at lag 12 with RMSE .0142593

上表显示,从1阶至13阶滞后<sup>①</sup>,均无法在5%的水平上拒绝“存在单位根”的原假设(甚至无法在10%水平上拒绝原假设)。下面将原假设变为平稳序列,进行KPSS检验:

.kpss inf,notrend

KPSS test for inf	
Maxlag = 4 chosen by Schwert criterion	
Autocovariances weighted by Bartlett kernel	
Critical values for H0: inf is level stationary	
10%: 0.347 5% : 0.463 2.5%: 0.574 1% : 0.739	
Lag order	Test statistic
0	2.14
1	1.16
2	.81
3	.625
4	.514

上表显示,5%的临界值为0.463,而从0阶至4阶滞后,其检验统计量均大于0.463。由于KPSS检验为右边单侧检验,故可在5%的水平上拒绝“平稳序列”的原假设,即认为存在单位根。

总之,以上各种检验的证据似乎支持通胀率inf为单位根过程(尽管PP检验的结果支持平稳过程)<sup>②</sup>。为此,进一步检验是否通胀率之差分dinf为平稳过程。以DF-GLS检验与KPSS检验为例:

.dfgls dinf

① 根据序贯t规则(Ng-Perron seq t),应选择滞后12阶。但在此例中,无论使用哪阶滞后,均不改变DF-GLS检验的结果。

② 究竟应将inf视为平稳过程还是单位根过程,文献中并无共识。

[lags]	DF-GLS tau	1% Critical	5% Critical	10% Critical
	Test Statistic	Value	Value	Value
13	-2.477	-3.501	-2.796	-2.520
12	-2.692	-3.501	-2.813	-2.535
11	-3.060	-3.501	-2.829	-2.550
10	-2.980	-3.501	-2.844	-2.565
9	-3.177	-3.501	-2.860	-2.579
8	-3.428	-3.501	-2.874	-2.592
7	-3.366	-3.501	-2.888	-2.605
6	-3.050	-3.501	-2.901	-2.617
5	-3.752	-3.501	-2.914	-2.629
4	-4.514	-3.501	-2.926	-2.640
3	-5.217	-3.501	-2.937	-2.650
2	-6.150	-3.501	-2.948	-2.659
1	-12.079	-3.501	-2.958	-2.668

Opt Lag (Ng-Perron seq t) = 7 with RMSE .0153809  
Min SC = -8.179493 at lag 2 with RMSE .0159336  
Min MAIC = -7.206234 at lag 6 with RMSE .0155328

上表显示,信息准则或序贯  $t$  规则的最优滞后阶数介于 2 与 7 之间。在此区间,检验统计量均小于 5% 的临界值,故可在 5% 的水平上拒绝“存在单位根”的原假设,即认为  $dinf$  为平稳过程。下面进行 KPSS 检验:

```
.kpss dinf,notrend
```

KPSS test for dinf	
Maxlag = 4 chosen by Schwert criterion	
Autocovariances weighted by Bartlett kernel	
Critical values for H0: dinf is level stationary	
10%: 0.347 5% : 0.463 2.5%: 0.574 1% : 0.739	
Lag order	Test statistic
0	.0349
1	.0461
2	.0702
3	.0662
4	.0672

从上表可知,检验统计量均远小于 5% 的临界值(0.463),故可接受“平稳过程”的原假设。综上所述,可以接受  $inf$  为一阶单整 I(1) 过程。

## 21.7 面板单位根检验

如果样本容量较小,则对单个变量进行单位根检验的功效可能很弱,即如果接受“存在单位根”的原假设,则犯第 II 类错误的概率很大。此时,如果有面板数据,可以找到更有效的检验方法。比如,可同时检验多个国家的实际汇率是否含有单位根。

文献中已有一系列的面板单位根检验,可通过 Stata 命令 `xtunitroot` 来实现。为了检验  $\{y_{it}\}$  是否包含单位根,考虑如下面板自回归模型:

$$y_{it} = \rho_i y_{i,t-1} + z'_{it} \gamma_i + \varepsilon_{it} \quad (21.24)$$

其中,  $i = 1, \dots, n$  表示横截面单位,  $t = 1, \dots, T_i$  表示时间,而  $\varepsilon_{it}$  为平稳的扰动项。根据命令 `xtunitroot` 的默认设置,  $z'_{it} \gamma_i$  表示个体固定效应(即  $z_{it} = 1$ ),也称为“panel-specific means”;如果加上选择项“`trend`”,则  $z'_{it} \gamma_i$  表示个体固定效应与线性时间趋势(linear time trend),即  $z'_{it} = (1, t)$ ;如果加上选择项“`noconstant`”,则忽略  $z'_{it} \gamma_i$  这一项。下文的 IPS 检验、费雪式检验与 Hadri LM 检验允许非平衡面板;其他检验则要求平衡面板,即  $T_i = T, \forall i$ 。

面板单位根检验的原假设为“ $H_0: \rho_i = 1, \forall i$ ”,而替代假设为“ $H_1: \rho_i < 1$ ”。方程(21.24)可写为等价形式:

$$\Delta y_{it} = \delta_i y_{i,t-1} + z'_{it} \gamma_i + \varepsilon_{it} \quad (21.25)$$

其中,  $\delta_i = \rho_i - 1$ 。相应的原假设与替代假设变为

$$H_0: \delta_i = 0, \forall i \quad vs \quad H_1: \delta_i < 0 \quad (21.26)$$

有些面板单位根检验(LLC 检验、HT 检验与 Breitung 检验),假设各面板单位的自回归系数均相同,也称为“共同根”(common root),即在方程(21.24)中,  $\rho_i = \rho, \forall i$ 。其他检验则允许各面板单位的自回归系数不同。另外,为了导出检验统计量的大样本分布,这些检验对于横截面维度  $n$  或时间维度  $T$  是否固定,或趋于无穷的速度所作的渐近假定也不尽相同。因此,对于具体数据,究竟适用何种面板单位根检验,主要取决于样本容量。比如,基于  $n/T \rightarrow 0$  的检验,要求时间维度  $T$  增长速度快于横截面维度  $n$ ,故适用于长面板。又比如,基于  $T$  固定而  $n \rightarrow \infty$  的检验显然适用于短面板。Stata 手册将这些检验分类总结如表 21.2。

表 21.2 面板单位根检验的特征

检验	Stata 选择项	适用的渐近理论	允许不同的自回归系数	允许非平衡面板
LLC	noconstant	$\sqrt{n}/T \rightarrow 0$	否	否
LLC		$n/T \rightarrow 0$	否	否
LLC	trend	$n/T \rightarrow 0$	否	否
HT	noconstant	$n \rightarrow \infty, T$ 固定	否	否
HT		$n \rightarrow \infty, T$ 固定	否	否
HT	trend	$n \rightarrow \infty, T$ 固定	否	否
Breitung	noconstant	$(T, n) \rightarrow_{seq} \infty$	否	否
Breitung		$(T, n) \rightarrow_{seq} \infty$	否	否
Breitung	trend	$(T, n) \rightarrow_{seq} \infty$	否	否
IPS		$n \rightarrow \infty, T$ 固定; 或 $n, T$ 都固定	是	是
IPS	trend	$n \rightarrow \infty, T$ 固定; 或 $n, T$ 都固定	是	是
IPS	lags()	$(T, n) \rightarrow_{seq} \infty$	是	是
IPS	trend lags()	$(T, n) \rightarrow_{seq} \infty$	是	是
费雪式		$T \rightarrow \infty, n$ 有限或趋无穷	是	是
Hadri LM		$(T, n) \rightarrow_{seq} \infty$	—	否
Hadri LM	trend	$(T, n) \rightarrow_{seq} \infty$	—	否

在表 21.1 中,  $(T, n) \rightarrow_{\text{seq}} \infty$  表示“序贯极限”(sequential limit), 即首先给定  $n$ , 让  $T \rightarrow \infty$ , 然后再让  $n \rightarrow \infty$ 。在实践中, 这要求  $T$  较大(large), 而且  $n$  也不能太小(at least moderate)。另外, 由于 Hadri LM 检验为面板平稳性检验(原假设为平稳过程), 故不存在是否“允许不同的自回归系数”的问题。下面具体介绍这些面板单位根检验。

### 1. LLC 检验

由于方程(21.25)的扰动项可能存在自相关, Levin, Lin and Chu (2002)(简记 LLC)在方程(21.25)的基础上引入高阶差分滞后项(类似于 ADF 检验的形式):

$$\Delta y_{it} = \delta y_{i,t-1} + z'_{it} \gamma_i + \sum_{j=1}^{p_i} \theta_{ij} \Delta y_{i,t-j} + \varepsilon_{it} \quad (21.27)$$

其中,  $\delta$  为共同的自回归系数(共同根); 不同个体的滞后阶数  $p_i$  可以不同;  $\{\varepsilon_{it}\}$  为平稳的 ARMA 过程; 不同个体的  $\varepsilon_{it}$  相互独立(不存在截面相关), 但允许异方差。通过引入足够高阶的差分滞后项, 可以保证  $\varepsilon_{it}$  为白噪声。

由于方程(21.27)为动态模型且包含个体固定效应, 故存在动态面板偏差。因此, 如果直接进行 OLS 回归, 估计量  $\hat{\delta}$  及相应的  $t$  统计量将存在偏差, 且不服从渐近正态分布; 故需要对此  $t$  统计量进行校正。由于方程(21.27)中的  $p_i$  未知, LLC 检验具体分为以下三个步骤。第一步, 对于每个面板单位  $i$ , 把  $\Delta y_{it}$  对  $\left(z'_{it}, \sum_{j=1}^{p_i} \Delta y_{i,t-j}\right)$  回归(通过信息准则选择滞后阶数  $p_i$ ), 得到残差  $\hat{e}_{it}$ ; 把  $y_{i,t-1}$  对  $\left(z'_{it}, \sum_{j=1}^{p_i} \Delta y_{i,t-j}\right)$  回归, 得到残差  $\hat{\nu}_{i,t-1}$ 。第二步, 考虑到不同面板单位可能存在异方差, 故将第一步中的残差做如下标准化:

$$\tilde{e}_{it} = \hat{e}_{it} / \hat{\sigma}_{ei}, \quad \tilde{\nu}_{i,t-1} = \hat{\nu}_{i,t-1} / \hat{\sigma}_{ei} \quad (21.28)$$

其中,  $\hat{\sigma}_{ei}$  为  $\varepsilon_{it}$  的标准误(可通过  $\hat{e}_{it}$  对  $\hat{\nu}_{i,t-1}$  的回归残差来计算)。第三步, 使用全部数据, 进行以下混合回归:

$$\tilde{e}_{it} = \delta \tilde{\nu}_{i,t-1} + \tilde{\varepsilon}_{it} \quad (21.29)$$

即可得  $\hat{\delta}$  及相应的  $t$  统计量  $t_{\hat{\delta}}$ <sup>①</sup>, 但如果存在个体固定效应, 则  $t_{\hat{\delta}}$  发散至负无穷(diverges to negative infinity)。为此, Levin, Lin and Chu (2002) 提出“偏差校正  $t$  统计量”(bias-adjusted  $t$  statistic), 记为  $t_{\hat{\delta}}^*$ , 在大样本下服从标准正态分布。与 ADF 检验类似, LLC 检验也是左边单侧检验, 即拒绝域仅在分布的最左边。

LLC 检验假设不存在截面相关。如果此假设不成立, 则 LLC 检验将存在“显著性水平扭曲”(size distortion)(O'Connell, 1998)。为了缓解可能存在的截面相关, Levin, Lin and Chu (2002) 建议先将面板数据减去各截面单位的均值(cross-sectional means), 再进行 LLC 检验。

LLC 检验的 Stata 命令格式为

```
xtunitroot llc y, trend noconstant demean lags (#) lags (aic #) lags (bic #) lags (hqic #)
```

其中, “y”为进行检验的变量; 选择项“trend”表示加入个体固定效应与线性时间趋势, 选择项“noconstant”表示这两项都不加, 默认仅加入个体固定效应; 选择项“demean”表示先将面板数据减去各截面单位的均值, 再进行检验; 选择项“lags (#)”用于指定差分滞后项  $\Delta y_{i,t-j}$  的滞后阶数  $p$  (要求所有个体的滞后阶数都相同); 选择项“lags (aic #)”、“lags (bic #)”与“lags

<sup>①</sup> 此种偏回归的理论基础是“Frisch-Waugh-Lovell Theorem”(参见第 3 章)。

(hqic #)" 分别表示使用 AIC、BIC 或 HQIC 信息准则来选择  $p_i$  并指定其最大值 #, 且不同个体的滞后阶数  $p_i$  可以不同。

下面以 Stata 提供的数据集 pennxrate.dta 为例。该平衡面板来自 Penn World Table 6.2, 包含 151 个国家, 1970—2003 年的实际汇率数据。目标是检验“购买力平价”(Purchasing Power Parity, 简记 PPP)是否成立。购买力平价假说认为, 两国之间的名义汇率反映两国之间的物价水平, 经物价调整后的实际汇率在长期内趋于均衡值, 故应为平稳过程。因此, 检验 lnrxrate(实际汇率的对数)是否为单位根过程; 如果是, 则拒绝 PPP 假说。该数据集还包括两个虚拟变量 g7 与 oecd, 分别表示 G7 与 OECD 国家。另外, 由于选择美国作为参照国来考察世界各国的汇率, 故美国不在此数据集中。

在理论上, 没有理由认为 lnrxrate 有时间趋势, 故不使用选择项“trend”, 而使用默认设置, 即仅加入个体固定效应。在这种情况下, LLC 检验的前提是  $n/T \rightarrow 0$ , 即要求横截面维度小于时间维度。因此, 为了演示的目的, 仅使用 G7 中的六个国家(不含美国)进行检验。

```
. use pennxrate.dta, clear
. xtunitroot llc lnrxrate if g7, lags(aic 10)
```

其中, 选择项“lags(aic 10)”表示将差分滞后项的最大滞后阶数设为 10, 并根据 AIC 信息准则选择最优滞后阶数。

Levin-Lin-Chu unit-root test for lnrxrate		
Ho: Panels contain unit roots	Number of panels =	6
Ha: Panels are stationary	Number of periods =	34
AR parameter: Common	Asymptotics: N/T → 0	
Panel means: Included		
Time trend: Not included		
ADF regressions: 1.00 lags average (chosen by AIC)		
LR variance: Bartlett kernel, 10.00 lags average (chosen by LLC)		
Statistic	p-value	
Unadjusted t	-6.7538	
Adjusted t*	-4.0277	0.0000

上表显示, 根据 AIC 信息准则选择的平均滞后阶数为 1。偏差校正  $t^*$  统计量(Adjusted t\*) 为 -4.03, 显著为负( $p$  值为 0.0000), 故强烈拒绝面板包含单位根的原假设, 认为面板为平稳过程。显然, 此结论支持 PPP。上表中的“未校正  $t$  统计量”(Unadjusted t) 为传统的  $t$  统计量, 不能用于检验。

由于 G7 国家经济发展水平相近且联系密切, 故每个国家的扰动项可能存在截面相关。为此, 下面使用选择项“demean”来缓解此截面相关。

```
. xtunitroot llc lnrxrate if g7, lags(aic 10) demean
```

Levin-Lin-Chu unit-root test for lnrxrate	
Ho: Panels contain unit roots	Number of panels = 6
Ha: Panels are stationary	Number of periods = 34
AR parameter: Common	Asymptotics: N/T → 0
Panel means: Included	
Time trend: Not included	Cross-sectional means removed
ADF regressions: 1.50 lags average (chosen by AIC)	
LR variance: Bartlett kernel, 10.00 lags average (chosen by LLC)	
Statistic	p-value
Unadjusted t -5.5473	
Adjusted t* -2.0813	0.0187

上表显示, 使用选择项“demean”将面板数据减去截面均值后, 偏差校正  $t_s^*$  统计量的  $p$  值上升为 0.0187, 但依然在 5% 水平上显著为负。

## 2. HT 检验

LLC 检验仅适用于长面板, 而许多微观面板数据的时间维度  $T$  较小。为此, Harris and Tzavalis (1999) (简记 HT) 提出了基于  $T$  固定而  $n \rightarrow \infty$  的检验统计量。令方程(21.24)中的自回归系数均相等可得:

$$y_{it} = \rho y_{i,t-1} + z'_{it} \gamma_i + \varepsilon_{it} \quad (21.30)$$

其中,  $\rho$  为共同根;  $\varepsilon_{it}$  服从 iid 正态分布, 故为同方差。在  $H_0: \rho = 1$  成立的情况下, Harris and Tzavalis (1999) 推导出 OLS 估计量  $\hat{\rho}$  的期望  $\mu$  与方差  $\sigma^2$  的表达式(为  $T$  的函数)<sup>①</sup>, 并证明当  $T$  固定而  $n \rightarrow \infty$  时,  $z \equiv \frac{\hat{\rho} - \mu}{\sqrt{\sigma/n}} \xrightarrow{d} N(0, 1)$ 。基于此大样本分布, 然后进行左边单侧检验。

HT 检验的 Stata 句式为

```
xtunitroot ht y, trend noconstant demean
```

其中, 各选择项的含义如上。

继续以数据集 pennxrate.dta 为例演示。由于 HT 检验的前提为  $T$  固定而  $n \rightarrow \infty$ , 故使用全部 151 个国家进行检验。为了缓解截面相关, 依然使用选择项“demean”。

```
.use pennxrate.dta, clear
```

```
.xtunitroot ht lnrxrate, demean
```

Harris-Tzavalis unit-root test for lnrxrate			
Ho: Panels contain unit roots	Number of panels = 151		
Ha: Panels are stationary	Number of periods = 34		
AR parameter: Common	Asymptotics: N → Infinity		
Panel means: Included	T Fixed		
Time trend: Not included	Cross-sectional means removed		
Statistic	z	p-value	
rho	0.8184	-13.1239	0.0000

① 由于存在动态面板偏差, OLS 估计有偏差, 故在原假设成立的情况下,  $\hat{\rho}$  的期望值也不为 1。

从上表可知,  $\hat{\rho} = 0.82$ , 而  $z = -13.12$ , 相应的  $p$  值为 0.000 0, 故强烈拒绝面板单位根的原假设, 依然支持 PPP。

### 3. Breitung 检验

LLC 检验与 HT 检验的共同特点是直接用 OLS 估计回归方程, 然后再对自回归系数或  $t$  统计量进行校正, 以消除动态面板偏差。Breitung 检验(Breitung 2000)的基本思路与 LLC 检验类似; 主要区别在于, 首先对数据进行“向前正交变换”(forward orthogonalization), 即减去未来各期的平均值, 然后再进行回归, 使得回归后不再需要偏差校正。具体步骤参见 Stata 手册。所得检验统计量记为  $\lambda$ , 服从渐近标准正态分布, 然后进行左边单侧检验。

Breitung 检验假设数据生成过程为 AR(1)。如果存在更高阶的自回归项, 则应先进行“预白噪声化”(prewhitening), 以消除原序列的自相关, 即分别把  $\Delta y_{it}$  与  $y_{i,t-1}$  对  $(\Delta y_{i,t-1}, \dots, \Delta y_{i,t-p})$  进行回归, 然后以这两个回归的残差来替代  $\Delta y_{it}$  与  $y_{i,t-1}$  进行 Breitung 检验。

Breitung (2000) 的蒙特卡罗模拟结果显示, LLC 检验的偏差校正  $t_s^*$  统计量的功效(power)较低, 特别在存在个体固定效应且自回归系数接近 1 时。而在这些情况下, Breitung 检验的功效则高很多(much higher)。Breitung (2000) 假设不同个体的扰动项不存在截面相关, 而 Breitung and Das (2005) 则提出在截面相关情况下也成立的检验统计量。

Breitung 检验的 Stata 命令句式为

```
xtunitroot breitung y, trend noconstant demean robust lags(#)
```

其中, 选择项“robust”表示使用截面相关稳健的统计量(Breitung and Das, 2005); 选择项“lags(#)”用于指定进行预白噪声化的滞后阶数, 默认不进行预白噪声化; 其他选择项的含义如上。

继续以数据集 pennxrate.dta 为例。由于 Breitung 检验的渐近理论假设  $(T, n) \rightarrow_{\text{seq}} \infty$ , 故选择 OECD 国家作为样本数据。我们将使用选择项“robust”来控制截面相关, 故不再使用选择项“demean”。由于上文 LLC 检验选择的平均滞后期为 1, 故假设数据生成过程为 AR(1), 不进行预白噪声化处理。

```
.xtunitroot breitung lnrxrate if oecd, robust
```

Breitung unit-root test for lnrxrate		
Ho: Panels contain unit roots	Number of panels =	27
Ha: Panels are stationary	Number of periods =	34
AR parameter: Common	Asymptotics:	$T, N \rightarrow \text{Infinity}$
Panel means: Included		sequentially
Time trend: Not included	Preliminary:	Not performed
Statistic	p-value	
lambda*	-1.6794	0.0465
* Lambda robust to cross-sectional correlation		

上表显示, 检验统计量  $\lambda = -1.68$ , 相应的  $p$  值为 0.046 5, 故可在 5% 水平上拒绝面板单位根的原假设。

### 4. IPS 检验

LLC 检验、HT 检验与 Breitung 检验的共同局限在于, 它要求每位个体的自回归系数  $\delta$

都相等,此共同根假设在实践中可能过强。比如,不同国家由于制度与文化的原因,经济规律可能不同。为了克服此缺点,Im, Pesaran and Shin (2003)(简记 IPS)提出了如下面板单位根检验。假设面板数据中共有  $n$  个相互独立的个体,对每位个体分别进行如下 DF 式回归:

$$\Delta y_{it} = \delta_i y_{i,t-1} + z'_{it} \gamma_i + \varepsilon_{it} \quad (21.31)$$

其中,  $\delta_i$  为个体  $i$  的自回归系数;  $\varepsilon_{it}$  服从相互独立的正态分布(扰动项无自相关),但允许异方差。假设  $T$  固定,而  $n \rightarrow \infty$  或固定。面板单位根的原假设为“ $H_0: \delta_i = 0, \forall i$ ”,而替代假设为“服从平稳过程的个体比例大于零”,即当  $n \rightarrow \infty$  时,  $n_1/n$  收敛至某非零正数,其中  $n_1$  为服从平稳过程的个体数。

记个体  $i$  的  $t$  统计量(即 ADF 统计量)为  $t_i$ ,计算所有个体  $t$  统计量的样本均值  $\bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$

(Stata 称为 t-bar)。Im, Pesaran and Shin (2003) 给出了  $\bar{t}$  分布的临界值。Stata 还汇报另一统计量  $\tilde{t}$  (Stata 称为 t-tilde-bar)<sup>①</sup>,它与  $\bar{t}$  的区别仅在对方程(21.31)扰动项的方差估计量不同。进一步,可将  $\tilde{t}$  标准化,构造如下统计量  $Z_{\tilde{t}}$  (Stata 称为 Z-t-tilde-bar):

$$Z_{\tilde{t}} = \frac{\tilde{t} - E(\tilde{t})}{\sqrt{\text{Var}(\tilde{t})}/n} \xrightarrow{d} N(0, 1) \quad (21.32)$$

其中,  $E(\tilde{t})$  与  $\text{Var}(\tilde{t})$  为  $\tilde{t}$  的理论均值与方差,可通过查表获得。由于假设这  $n$  个时间序列相互独立,故适用中心极限定理,因此  $Z_{\tilde{t}}$  的渐近分布为标准正态。IPS 检验也是左边单侧检验。

如果方程(21.31)的扰动项  $\varepsilon_{it}$  存在自相关,可通过引入差分滞后项来消除,即对每位个体分别进行如下 ADF 式回归:

$$\Delta y_{it} = \delta_i y_{i,t-1} + z'_{it} \gamma_i + \sum_{j=1}^{p_i} \theta_{ij} \Delta y_{i,t-j} + \varepsilon_{it} \quad (21.33)$$

其中,不同个体的滞后阶数  $p_i$  可以不同(可通过信息准则来确定),且假设  $(T, n) \rightarrow_{\text{seq}} \infty$ 。其余检验步骤与扰动项无自相关的情形类似,记其统计量为  $W_i$  (Stata 称为 W-t-bar),对应于上文的  $Z_{\tilde{t}}$  统计量。

IPS 检验的 Stata 命令句式为

`xtunitroot ips y, trend demean lags(#)` lags(aic #) lags(bic #) lags(hqic #)

其中,各选择项的含义类似于 LLC 检验。下面继续以数据集 pennxrate.dta 为例,检验 OECD 国家是否符合 PPP 假说。首先假设扰动项无自相关,但使用选择项“demean”来缓解可能存在的自相关。

`. xtunitroot ips lnrxrate if oecd, demean`

① 统计量  $\tilde{t}$  中的波浪式符号读为“tilde”。

Im-Pesaran-Shin unit-root test for lnrxrate					
Ho: All panels contain unit roots	Number of panels =	27			
Ha: Some panels are stationary	Number of periods =	34			
AR parameter: Panel-specific	Asymptotics: T,N -> Infinity				
Panel means: Included	sequentially				
Time trend: Not included	Cross-sectional means removed				
ADF regressions: No lags included					
	Fixed-N exact critical values				
	Statistic	p-value	1%	5%	10%
t-bar	-3.1327		-1.810	-1.730	-1.680
t-tilde-bar	-2.5771				
Z-t-tilde-bar	-7.3911	0.0000			

上表显示,  $\bar{t}$  统计量为 -3.13, 小于 1% 水平的临界值 -1.81, 故拒绝面板单位根的原假设。而统计量  $Z_{\bar{t}}$  的  $p$  值为 0.0000, 同样拒绝原假设。下面, 考虑扰动项存在自相关的情形, 并引入差分滞后项。

```
.xtunitroot ips lnrxrate if oecd, lags(aic 8) demean
```

Im-Pesaran-Shin unit-root test for lnrxrate		
Ho: All panels contain unit roots	Number of panels =	27
Ha: Some panels are stationary	Number of periods =	34
AR parameter: Panel-specific	Asymptotics: T,N -> Infinity	
Panel means: Included	sequentially	
Time trend: Not included	Cross-sectional means removed	
ADF regressions: 1.48 lags average (chosen by AIC)		
	Statistic	p-value
W-t-bar	-7.3075	0.0000

上表显示, 根据 AIC 准则选择的平均滞后期为 1.48, 而统计量  $W_t$  的  $p$  值为 0.0000, 依然强烈拒绝原假设。

## 5. 费雪式检验

费雪式检验的基本思路类似于 IPS 检验, 即对每位个体分别进行检验, 然后再将这些信息综合起来。具体来说, 对面板数据中的每位个体分别进行单位根检验 (ADF 检验或 PP 检验), 得到  $n$  个检验统计量及相应的  $p$  值  $\{p_1, \dots, p_n\}$ 。Choi (2001) 提出以下四种方法将这些  $p$  值综合成“费雪式”(Fisher type) 统计量<sup>①</sup>。方法一为“逆卡方变换”(inverse chi-squared transformation):

$$P \equiv -2 \sum_{i=1}^n \ln p_i \xrightarrow{d} \chi^2(2n) \quad (T_i \rightarrow \infty) \quad (21.34)$$

其中,  $T_i$  为个体  $i$  的时间维度(因个体而异, 允许非平衡面板数据)。由于取了负号, 故这是一个单边右侧检验, 即统计量  $P$  越大, 则越倾向于拒绝“面板单位根”的原假设。方法二为“逆正态变

<sup>①</sup> Ronald Fisher 最早提出这类检验, 把几个统计检验的证据综合为一个检验统计量。

换”(inverse normal transformation)：

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Phi^{-1}(p_i) \xrightarrow{d} N(0,1) \quad (T_i \rightarrow \infty) \quad (21.35)$$

其中,  $\Phi^{-1}(\cdot)$  为标准正态累积分布函数的逆函数, 故名。如果使用方法二, 则为单边左侧检验。方法三为“逆逻辑变换”(inverse logit transformation)：

$$L^* \equiv \sqrt{k} \sum_{i=1}^n \ln\left(\frac{p_i}{1-p_i}\right) \xrightarrow{d} t(5n+4) \quad (T_i \rightarrow \infty) \quad (21.36)$$

其中,  $k = \frac{3(5n+4)}{\pi^2 n(5n+2)}$ 。方法三也是单边左侧检验。如果面板中的个体数  $n$  很大, 可使用“修正逆卡方变换”(modified inverse chi-squared transformation)：

$$P_m \equiv -\frac{1}{\sqrt{n}} \sum_{i=1}^n (\ln p_i + 1) \xrightarrow{d} N(0,1) \quad (T_i, n \rightarrow \infty) \quad (21.37)$$

费雪式检验的 Stata 命令句式为

```
xtunitroot fisher y, dfuller pperron demean lags(#)
```

其中, 选择项“dfuller”表示根据 ADF 检验获得  $p$  值, 选择项“pperron”表示根据 PP 检验获得  $p$  值。选择项“lags( $#$ )”如果与选择项“dfuller”同时使用, 表示 ADF 检验的滞后阶数; 如果与选择项“pperron”同时使用, 表示用于计算标准误的滞后阶数。进一步, 如果使用选择项“dfuller”, 则命令 dfuller 的所有选择项也都适用于此命令; 类似地, 如果使用选择项“pperron”, 则命令 pperron 的所有选择项也都适用于此命令。

继续以数据集 pennxrate.dta 为例, 使用滞后两期的 ADF 回归检验所有 151 个国家的实际汇率。由于许多国家的实际汇率对数的平均值都不为零, 故假设真实模型存在漂移项, 并加入命令 dfuller 的选择项“drift”。另外, 使用选择项“demean”来缓解可能存在的截面相关。

```
.xtunitroot fisher lnrxrate, dfuller drift lags(2) demean
```

Fisher-type unit-root test for lnrxrate			
Based on augmented Dickey-Fuller tests			
Ho: All panels contain unit roots	Number of panels =	151	
Ha: At least one panel is stationary	Number of periods =	34	
AR parameter: Panel-specific	Asymptotics: T -> Infinity		
Panel means: Included	Cross-sectional means removed		
Time trend: Not included	ADF regressions: 2 lags		
Drift term: Included			
		Statistic	p-value
Inverse chi-squared(302)	P	975.9130	0.0000
Inverse normal	Z	-19.6183	0.0000
Inverse logit t(759)	L*	-20.9768	0.0000
Modified inv. chi-squared	Pm	27.4211	0.0000
P statistic requires number of panels to be finite.			
Other statistics are suitable for finite or infinite number of panels.			

上表显示, 所有四个统计量均强烈拒绝面板单位根的原假设, 相应的  $p$  值为 0.000 0。

## 6. Hadri LM 检验

Hadri (2000) 把 KPSS 平稳性检验推广到面板数据, 提出了检验面板平稳性的 LM 检验(原假设为平稳过程)。考虑以下面板形式的 KPSS 检验模型:

$$\begin{aligned} y_{it} &= \beta_i t + u_{it} + \varepsilon_{it} \\ u_{it} &= u_{i,t-1} + \nu_{it} \end{aligned} \quad (21.38)$$

其中,  $\beta_i t$  为个体  $i$  的时间趋势(panel-specific time trend), 而扰动项  $\varepsilon_{it}$  与  $\nu_{it}$  均服从 iid 正态分布, 其方差分别为  $\sigma_\varepsilon^2$  与  $\sigma_\nu^2$ 。面板平稳性的原假设等价于 " $H_0: \lambda = \frac{\sigma_\nu^2}{\sigma_\varepsilon^2} = 0$ ", 而替代假设为 " $H_1: \lambda > 0$ "。

Hadri LM 检验的 Stata 命令句式为

```
xtunitroot hadri y, trend demean robust kernel (bartlett #) kernel (parzen #) kernel (quadraticspectral #)
```

其中, 选择项“robust”表示允许不同个体的  $\varepsilon_{it}$  存在异方差(不再是 iid, 但仍为正态); 选择项“kernel(bartlett #) kernel(parzen #) kernel(quadraticspectral #)”分别指定使用 bartlett, parzen 或 quadraticspectral 核函数<sup>①</sup>以及滞后阶数#, 来估计扰动项的长期方差(long-run variance), 即渐近方差。使用选择项“kernel()”可使得检验结果在存在异方差与自相关的情况下也成立。

仍以数据集 pennxrate.dta 为例, 对 OECD 国家进行检验。为了控制自相关, 使用滞后 5 阶的 bartlett 核函数。

```
.xtunitroot hadri lnrxrate if oecd, kernel(bartlett 5) demean
```

Hadri LM test for lnrxrate		
Ho: All panels are stationary		Number of panels = 27
Ha: Some panels contain unit roots		Number of periods = 34
Time trend:	Not included	Asymptotics: T, N -> Infinity
Heteroskedasticity:	Robust	sequentially
LR variance:	Bartlett kernel, 5 lags	Cross-sectional means removed
Statistic p-value		
z	9.6473	0.0000

上表显示, 可以拒绝“所有面板单位均为平稳过程”的原假设, 与上文的面板单位根检验结果有所不同。Banerjee et al (2005) 探讨了面板单位根检验对于 PPP 假说的适用性, 并指出由于不同国家之间的汇率可能存在协整或长期关系, 面板单位根检验经常在原假设正确的情况下也拒绝原假设。

以上介绍的面板单位根检验大多要求不同个体的扰动项相互独立, 在实践中可能很难满足。比如, 由于国际经济的互动, 跨国面板数据中的变量常存在“共同运动趋势”(comovement), 故不同个体的扰动项可能相关。近些年出现的新一代面板单位根检验开始允许不同个体的扰动项之间存在相关性(参见 Silva, 2009; 陈海燕, 2012), 但 Stata 12 尚未包含这些前沿结果。

<sup>①</sup> 有关核函数的介绍, 参见第 27 章。

## 21.8 协整的思想与初步检验

对于有单位根的变量,传统的处理方法是对其进行一阶差分而得到平稳序列。但是,一阶差分后变量的经济含义与原序列并不相同,而有时我们仍然希望使用原序列进行回归。如果多个单位根变量之间由于某种经济力量而存在“长期均衡关系”(long-run equilibrium),则有可能进行这种回归。其基本思想是,如果多个单位根序列拥有“共同的随机趋势”(common stochastic trend),则可以对这些变量作线性组合而消去此随机趋势。

**例** 短期利率与长期利率可能都是单位根过程,而二者的走势也很相似。从经济理论上来  
看<sup>①</sup>,长期利率是未来预期短期利率的平均值与“风险溢价”(risk premium)之和,故存在长期均  
衡关系。

**例(非正式)** 当你遛狗时,假设你与狗的每一步位置为随机游走过程(带漂移项),故均为  
单位根过程。由于你与狗之间有一根皮带相连(“长期均衡关系”),故你与狗的位置之间不会相  
离太远(尽管二者都是单位根过程)。

假设两个 I(1) 过程  $\{y_t\}, \{x_t\}$  可以分别表示为

$$\begin{cases} y_t = \alpha + \beta w_t + \varepsilon_t \\ x_t = \gamma + \delta w_t + u_t \end{cases} \quad (21.39)$$

其中,  $w_t$  为随机游走,  $w_t = w_{t-1} + v_t$ ; 而  $\varepsilon_t, u_t, v_t$  均为白噪声。由于  $\{y_t\}$  与  $\{x_t\}$  拥有共同的随机趋  
势  $w_t$ , 故二者的如下线性组合为平稳过程:

$$\delta y_t - \beta x_t = (\alpha\delta - \beta\gamma) + (\delta\varepsilon_t - \beta u_t) \quad (21.40)$$

在这种情况下,称  $\{y_t\}$  与  $\{x_t\}$  是“协整的”(cointegrated),而称向量  $(\delta, -\beta)$  为“协整向量”  
(cointegrating vector)或“协整系数”。显然,可以把  $(\delta, -\beta)$  标准化为  $(1, -\beta/\delta)$ 。从这个例子可  
以看出,对于两个 I(1) 变量,只可能存在一个协整关系;而对于  $n$  个 I(1) 变量,则最多可能存在  
 $(n-1)$  个协整关系。一组 I(1) 变量之间协整关系的个数被称为“协整秩”(cointegration rank),  
即线性无关的协整向量的个数。

如何判断一组 I(1) 变量间是否存在协整关系呢?首先,这些变量必须在理论上可能存在长  
期均衡关系;否则,进行协整分析将没有意义。其次,如果只有两个变量,则可以直接画图,看二  
者的时间趋势是否有相似性。但这个方法不严格,也不适用于两个以上的变量。作为正式的协  
整检验,Engle and Granger (1987) 提出了如下的“EG-ADF 检验”。

原假设为  $\{y_t, x_t\}$  存在协整关系,且协整系数为  $\{1, -\theta\}$ ,则  $\{z_t \equiv y_t - \theta x_t\}$  为平稳过程。如果  
 $\theta$  已知(比如,通过经济理论而知),则可以用 ADF 检验来确定  $\{z_t\}$  是否平稳。如果接受“ $\{z_t\}$  为  
平稳”,则认为  $\{y_t, x_t\}$  存在协整关系。然而,通常我们并不知道  $\theta$ ,故“EG-ADF 方法”分两步进  
行。

**第一步** 用 OLS 估计协整系数  $\theta$ ,即  $y_t = \phi + \theta x_t + z_t$ 。在“ $\{y_t, x_t\}$  存在协整关系”的原假设  
下,虽然  $\{y_t, x_t\}$  为非平稳的 I(1) 过程,但  $\{z_t\}$  为平稳过程。在这种情况下,OLS 的估计量  $\hat{\phi}$  与  $\hat{\theta}$   
都是一致估计量。

<sup>①</sup> 根据“利率期限结构的预期理论”(The expectation theory of the term structure of interest rates)。

**第二步** 对残差序列  $\{\hat{z}_t \equiv y_t - \hat{\phi} - \hat{\theta}x_t\}$  进行 ADF 检验, 确定其是否平稳。由于协整系数  $\hat{\theta}$  是估计出来的, 不一定是最真实的协整系数, 故 EG-ADF 统计量的临界值与普通的 ADF 检验不同, 参见 Hayashi (2000, p. 646) 或 Stock and Watson (2004, p. 557)。

如果检验结果确认  $\{\hat{z}_t\}$  为平稳, 则接受“ $\{y_t, x_t\}$  存在协整关系”的原假设。此时, 估计出的协整关系 “ $y_t = \hat{\phi} + \hat{\theta}x_t$ ” 即为  $\{y_t, x_t\}$  之间的长期均衡关系 ( $\hat{\phi}, \hat{\theta}$  为长期参数)。如果要估计  $\{y_t, x_t\}$  之间的短期关系 (短期参数), 则需要使用误差修正模型 (ECM)。

假设  $\{y_t, x_t\}$  之间的关系可由一个 ADL 模型来表示, 比如,  $y_t = \beta_0 + \beta_1 y_{t-1} + \gamma_0 x_t + \gamma_1 x_{t-1} + \varepsilon_t$ , 则其对应的 ECM 模型为 (参见第 20 章)

$$\Delta y_t = \gamma_0 \Delta x_t + (\beta_1 - 1)(y_{t-1} - \phi - \theta x_{t-1}) + \varepsilon_t \quad (21.41)$$

显然, 方程 (21.41) 左边的  $\Delta y_t$  为平稳过程。如果  $\{y_t, x_t\}$  存在协整关系, 则方程右边的误差修正项  $(y_{t-1} - \phi - \theta x_{t-1})$  为平稳, 而  $\Delta x_t$  也为平稳, 故方程右边整体上也是平稳的。反之, 如果  $\{y_t, x_t\}$  不存在协整关系, 则方程右边的误差修正项  $(y_{t-1} - \phi - \theta x_{t-1})$  为非平稳, 而方程左边依然平稳, 故 ECM 模型不能成立<sup>①</sup>。在  $\{y_t, x_t\}$  存在协整关系的前提下, 将残差  $\{\hat{z}_t \equiv y_t - \hat{\phi} - \hat{\theta}x_t\}$  代入 ECM 模型可得

$$\Delta y_t = \gamma_0 \Delta x_t + (\beta_1 - 1)\hat{z}_{t-1} + error_t \quad (21.42)$$

使用 OLS 估计上式, 即可得到对短期参数的估计。

EG-ADF 方法的一个缺点是, 它不能处理同时存在多个协整关系 (即协整秩大于 1) 的情形。另外, 由于 EG-ADF 方法分两步进行, 第一步估计的误差会被带到第二步中, 故不是最有效率的方法。比 EG-ADF 方法更有效率的方法是, 用 MLE 同时估计长期与短期参数 (也是目前最流行的方法)<sup>②</sup>。为此, 必须更深入地研究 I(1) 过程, 并给出协整的严格定义。由于最简单的 I(1) 过程就是随机游走, 故下面首先把一般的 I(1) 过程分解为随机游走、时间趋势与平稳过程之和。

## 21.9 Beveridge-Nelson 分解公式

**定义** 称时间序列  $\{y_t\}$  为“线性 I(0) 过程”, 如果  $y_t = \delta + u_t$ , 其中  $\delta$  为常数,  $u_t = \psi(L)\varepsilon_t$ ,  $\{\varepsilon_t\}$  为独立白噪声, 滤波  $\psi(L) \equiv \psi_0 + \psi_1 L + \psi_2 L^2 + \dots$ , 满足  $\sum_{j=0}^{\infty} j |\psi_j| < \infty$  (称为“一可加总”, one-summable, 简记为 OS),  $\psi(1) = \psi_0 + \psi_1 + \psi_2 + \dots \neq 0$ 。

显然, OS 比 AS (绝对值可加总) 的假定更强, 前者是后者的充分条件。“ $\psi(1) \neq 0$ ”是一个技术性条件, 防止出现退化的情形 (详见下文)。本章下面讨论的 I(0) 皆为线性 I(0) 过程。假设序列  $\{y_t\}$  为 I(1), 则其差分为 I(0), 故可表示为  $\Delta y_t = \delta + u_t$ , 其中  $u_t$  为线性 I(0) 过程。假设时间从  $t=0$  开始, 则  $y_1 = y_0 + \delta + u_1$ ,  $y_2 = \underbrace{\delta + y_0 + \delta + u_1 + u_2}_{=y_1} + y_1 + u_2 = y_0 + 2\delta + u_1 + u_2, \dots, y_t = y_0 +$

$$\delta t + \sum_{s=1}^t u_s$$

① 如果  $\{y_t, x_t\}$  为平稳过程, 则 ECM 模型总能成立。

② 由于 EG-ADF 方法不如 MLE 有效率, 故在 Stata 中没有现成的命令, 但可以较方便地手工进行 (需要查表以获得临界值)。

由于  $u_t$  为线性  $I(0)$  过程, 故可写为  $u_t = \psi(L)\varepsilon_t$ , 其中

$$\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \dots$$

可以将滤波  $\psi(L)$  分解为

$$\begin{aligned}\psi(L) &= \psi_0 + \psi_1 L + \psi_2 L^2 + \dots \\ &= (\psi_0 + \psi_1 + \psi_2 + \dots) + (\psi_1 L - \psi_1) + (\psi_2 L^2 - \psi_2) + (\psi_3 L^3 - \psi_3) + \dots \\ &= \psi(1) + [-\psi_1(1-L) - \psi_2(1-L^2) - \psi_3(1-L^3) - \dots] \\ &= \psi(1) + (1-L)[- -\psi_1 - \psi_2(1+L) - \psi_3(1+L+L^2) - \dots] \\ &= \psi(1) + (1-L)[- (\psi_1 + \psi_2 + \dots) - (\psi_2 + \psi_3 + \dots)L - (\psi_3 + \psi_4 + \dots)L^2 - \dots] \\ &= \psi(1) + (1-L)[\alpha_0 + \alpha_1 L + \alpha_2 L^2 + \dots], \text{ 其中 } \alpha_j \equiv -(\psi_{j+1} + \psi_{j+2} + \dots) \\ &= \psi(1) + (1-L)\alpha(L)\end{aligned}$$

其中,  $\alpha(L) \equiv \alpha_0 + \alpha_1 L + \alpha_2 L^2 + \dots$ 。由于  $\psi(L)$  为 OS, 故可以证明(参见附录),  $\alpha(L)$  为 AS。

定义  $\eta_t \equiv \alpha(L)\varepsilon_t$ 。由于  $\alpha(L)$  为 AS, 故  $\eta_t$  为平稳过程。因此,

$$u_t = \psi(L)\varepsilon_t = [\psi(1) + (1-L)\alpha(L)]\varepsilon_t = \psi(1)\varepsilon_t + (1-L)\eta_t = \psi(1)\varepsilon_t + \eta_t - \eta_{t-1}$$

(21.43)

$$y_t = y_0 + \delta t + \sum_{s=1}^t u_s = \underbrace{\delta t}_{\text{time trend}} + \underbrace{\psi(1) \sum_{s=1}^t \varepsilon_s}_{\text{random walk}} + \underbrace{\eta_t}_{\text{stationary}} + \underbrace{(y_0 - \eta_0)}_{\text{initial condition}} \quad (21.44)$$

上式就是著名的“Beveridge-Nelson 分解公式”。它将  $I(1)$  过程分解为时间趋势  $\delta t$ 、随机游走  $\psi(1) \sum_{s=1}^t \varepsilon_s$ <sup>①</sup>、平稳序列  $\eta_t$  及初始条件  $(y_0 - \eta_0)$  之和。由于协整关系涉及多个变量之间的关系, 下面把 Beveridge-Nelson 分解公式向多维推广。

定义 称  $\{y_t\}$  为  $n$  维“线性向量  $I(0)$  过程”, 如果  $y_t = \delta + u_t$ , 其中  $\delta$  为常数向量,  $u_t = \psi(L)\varepsilon_t$ , 多维滤波  $\psi(L) \equiv I_n + \psi_1 L + \psi_2 L^2 + \dots$ ,  $\{\varepsilon_t\}$  为独立同分布的,  $E(\varepsilon_t) = \mathbf{0}$ ,  $E(\varepsilon_t \varepsilon_t')$  为正定矩阵,  $\{\psi_j\}$  为 OS(即矩阵的每个元素均为 OS), 而且

$$\psi(1) = I_n + \psi_1 + \psi_2 + \dots \neq \mathbf{0}_{n \times n}$$

因此, 可将  $n$  维向量  $I(1)$  过程  $y_t$  表示为

$$\Delta y_t = \delta + u_t = \delta + \psi(L)\varepsilon_t \quad (21.45)$$

这称为  $I(1)$  系统的“向量移动平均表示法”(VMA Representation)。将其写成绝对水平的形式:

$$y_t = y_0 + \delta t + \sum_{s=1}^t u_s \quad (21.46)$$

依据类似的推导, 有如下向量形式的 Beveridge-Nelson 分解公式:

$$y_t = \underbrace{\delta t}_{\text{time trend}} + \underbrace{\psi(1) \sum_{s=1}^t \varepsilon_s}_{\text{random walk}} + \underbrace{\eta_t}_{\text{stationary}} + \underbrace{(y_0 - \eta_0)}_{\text{initial condition}} \quad (21.47)$$

## 21.10 协整的定义与最大似然估计

一般来说, 由于存在随机趋势项  $\psi(1) \sum_{s=1}^t \varepsilon_s$ , 上述  $I(1)$  系统  $\{y_t\}$  为非平稳。然而, 如果在方程(21.47)两边同时乘以某  $1 \times n$  的非零行向量  $a'$ , 则可能将此随机趋势项消去:

① 如果  $\psi(1) = 0$ , 则这个随机游走部分退化为 0, 故要求  $\psi(1) \neq 0$ 。

$$\underbrace{\mathbf{a}' \mathbf{y}_t}_{\substack{1 \times n \\ 1 \times n}} = \underbrace{\mathbf{a}' \boldsymbol{\delta} t}_{\substack{n \times n \\ 1 \times 1}} + \underbrace{\mathbf{a}' \boldsymbol{\psi}(1) \sum_{s=1}^t \boldsymbol{\varepsilon}_s}_{\substack{1 \times n \\ 1 \times n}} + \mathbf{a}' \boldsymbol{\eta}_t + \mathbf{a}' (\mathbf{y}_0 - \boldsymbol{\eta}_0) \quad (21.48)$$

如果  $\mathbf{a}' \boldsymbol{\psi}(1) = \mathbf{0}'$ , 则上式可简化为

$$\mathbf{a}' \mathbf{y}_t = \mathbf{a}' \boldsymbol{\delta} t + \mathbf{a}' \boldsymbol{\eta}_t + \mathbf{a}' (\mathbf{y}_0 - \boldsymbol{\eta}_0) \quad (21.49)$$

$\{\mathbf{a}' \mathbf{y}_t\}$  是一个趋势平稳 (trend stationary) 的过程, 即只要将时间趋势项  $\mathbf{a}' \boldsymbol{\delta} t$  去掉, 就是平稳过程。此时, 称  $\mathbf{y}_t$  是“协整的” (cointegrated), 而称  $\mathbf{a}$  为“协整向量” (cointegrating vector)。

问题的关键在于, 是否存在  $n$  维非零列向量  $\mathbf{a}$ , 使得  $\mathbf{a}' \boldsymbol{\psi}(1) = \mathbf{0}'$ , 即  $\boldsymbol{\psi}(1)' \mathbf{a} = \mathbf{0}$ 。

考虑关于  $\mathbf{a}$  的  $n$  元方程组  $\boldsymbol{\psi}(1)' \mathbf{a} = \mathbf{0}$ , 称其线性无关解的个数为  $\{\mathbf{y}_t\}$  的“协整秩” (cointegration rank), 即线性无关的协整向量的个数, 记为  $h$ 。根据线性代数知识,  $h = n - \text{rank}[\boldsymbol{\psi}(1)]$ 。由于  $1 \leq \text{rank}[\boldsymbol{\psi}(1)] \leq n$ , 故  $0 \leq h \leq n - 1$ 。如果  $h = 0$ , 则不存在协整关系。如果  $h = 1$ , 则存在一个协整关系, 可以解释为长期均衡关系 (long-run equilibrium)。如果  $h > 1$ , 则存在多个协整关系, 通常需要用经济理论来剔除不合理的协整向量, 而将最合理的协整向量解释为长期均衡关系。

不失一般性, 假设协整向量  $\mathbf{a}$  的第一个分量不为 0, 并将其标准化为 1, 即  $\mathbf{a} = \begin{pmatrix} 1 \\ \gamma' \\ \vdots \\ 0 \end{pmatrix}$ 。将

$\mathbf{y}_t$  也同样地分块, 即  $\mathbf{y}_t = \begin{pmatrix} \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \\ \vdots \\ \mathbf{y}_{nt} \end{pmatrix}$ 。假设  $\mathbf{a}' \boldsymbol{\delta} = 0$  (不存在时间趋势项), 则可以将方程 (21.49)

写成

$$\mathbf{y}_{1t} = \alpha + \gamma' \mathbf{y}_{2t} + z_t \quad (21.50)$$

其中,  $\alpha \equiv (1 - \gamma')(\mathbf{y}_0 - \boldsymbol{\eta}_0)$  为初始条件, 可以视为截距项; 而  $z_t \equiv (1 - \gamma') \boldsymbol{\eta}_t$  为平稳过程, 可以视为扰动项。这个回归被称为“协整回归” (cointegrating regression)。在  $\mathbf{y}_t$  存在协整关系的前提下, OLS 估计是一致的。

以上讨论的是 I(1) 系统的 VMA 表示法。但由于扰动项不可观测, 为了进行 MLE 估计, 考虑以下的“向量自回归表示法” (VAR Representation):

$$\mathbf{y}_t = \alpha + \boldsymbol{\delta} t + \boldsymbol{\Phi}_1 \mathbf{y}_{t-1} + \boldsymbol{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t \quad (21.51)$$

如何知道此 VAR ( $p$ ) 何时为协整系统呢? 为此, 先导出其对应的 VMA 表示法。定义  $\boldsymbol{\Phi}(L) \equiv \mathbf{I}_n - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p$ , 则

$$\boldsymbol{\Phi}(L) \mathbf{y}_t = \alpha + \boldsymbol{\delta} t + \boldsymbol{\varepsilon}_t \quad (21.52)$$

在方程 (21.52) 两边同时左乘  $\boldsymbol{\Phi}(L)^{-1}$  可得

$$\mathbf{y}_t = \boldsymbol{\Phi}(L)^{-1} \alpha + \boldsymbol{\Phi}(L)^{-1} \boldsymbol{\delta} t + \boldsymbol{\Phi}(L)^{-1} \boldsymbol{\varepsilon}_t \quad (21.53)$$

对方程 (21.53) 两边同时进行差分可得

$$\Delta \mathbf{y}_t = \boldsymbol{\Phi}(L)^{-1} \boldsymbol{\delta} (1 - L) t + \boldsymbol{\Phi}(L)^{-1} (1 - L) \boldsymbol{\varepsilon}_t = \boldsymbol{\delta}^* + \boldsymbol{\psi}(L) \boldsymbol{\varepsilon}_t \quad (21.54)$$

其中,  $(1 - L)t = t - (t - 1) = 1$ ,  $\boldsymbol{\delta}^* \equiv \boldsymbol{\Phi}(1)^{-1} \boldsymbol{\delta}$ ,  $\boldsymbol{\psi}(L) \equiv \boldsymbol{\Phi}(L)^{-1} (1 - L)$ 。显然, 如果此 VAR 系统为协整秩为  $h$  的协整系统, 则  $\boldsymbol{\psi}(L)$  必须为 OS, 而且  $\text{rank}[\boldsymbol{\psi}(1)] = n - h$ 。

由于  $\boldsymbol{\psi}(L) \equiv \boldsymbol{\Phi}(L)^{-1} (1 - L)$ , 故  $\boldsymbol{\Phi}(L) \boldsymbol{\psi}(L) = (1 - L) \mathbf{I}_n$ 。令  $L = 1$ , 则有

$$\boldsymbol{\Phi}(1)_{n \times n} \boldsymbol{\psi}(1)_{n \times n} = \mathbf{0}_{n \times n} \quad (21.55)$$

可以证明, 如果  $\{\mathbf{y}_t\}$  的协整秩为  $h$ , 则  $\text{rank}[\boldsymbol{\Phi}(1)] = h$ 。根据线性代数知识, 可以将  $\boldsymbol{\Phi}(1)$  分

解为

$$\boldsymbol{\Phi}(1)_{n \times n} = \mathbf{B}_{n \times h} (\mathbf{A}')_{h \times n} \quad (21.56)$$

其中,  $\mathbf{B}, \mathbf{A}$  为两个  $n \times h$  的满列秩矩阵 ( $\mathbf{B}, \mathbf{A}$  不唯一)。此条件称为“降秩条件”(reduced rank condition)。因此

$$\mathbf{B}_{n \times h} (\mathbf{A}')_{h \times n} \boldsymbol{\Psi}(1)_{n \times n} = \mathbf{0}_{n \times n} \quad (21.57)$$

因为  $\mathbf{B}_{n \times h}$  满列秩, 故

$$\mathbf{A}'_{h \times n} \boldsymbol{\Psi}(1)_{n \times n} = \mathbf{0}_{h \times n} \quad (21.58)$$

由此可见, 矩阵  $\mathbf{A}_{n \times h}$  中的列向量都是协整向量(满足协整向量的定义)。从 VAR 方程(21.51)出发, 可以导出其对应的“向量误差修正表示法”(VECM Representation)。根据与推导 ADF 检验类似的方法可得

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\delta} t + (\boldsymbol{\rho} - \mathbf{I}_n) \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t \quad (21.59)$$

其中,  $\boldsymbol{\rho} \equiv \boldsymbol{\Phi}_1 + \boldsymbol{\Phi}_2 + \cdots + \boldsymbol{\Phi}_p$ ,  $\boldsymbol{\Gamma}_s \equiv -(\boldsymbol{\Phi}_{s+1} + \cdots + \boldsymbol{\Phi}_p)$ ,  $s = 1, 2, \dots, p-1$ 。

由于  $\mathbf{I}_n - \boldsymbol{\rho} = \mathbf{I}_n - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p = \boldsymbol{\Phi}(1) = \mathbf{BA}'$ , 故

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\delta} t - \mathbf{B} \underbrace{\mathbf{A}' \mathbf{y}_{t-1}}_{\mathbf{z}_{t-1}} + \boldsymbol{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_2 \Delta \mathbf{y}_{t-2} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t \quad (21.60)$$

其中,  $\mathbf{z}_{t-1} \equiv \mathbf{A}' \mathbf{y}_{t-1}$  为误差修正项(因为  $\mathbf{A}$  中的列向量皆为协整向量)。因此, 一个协整的 I(1) 系统同时有 VMA, VAR 与 VECM 表示法, 此结论称为“格兰杰表示法定理”(Granger Representation Theorem)。

定义  $\boldsymbol{\Gamma}_0 \equiv -\boldsymbol{\Phi}(1) = -\mathbf{BA}'$ 。Johansen(1988) 使用 MLE 来估计如下 VECM 模型:

$$\Delta \mathbf{y}_t = \boldsymbol{\alpha} + \boldsymbol{\Gamma}_0 \mathbf{y}_{t-1} + \boldsymbol{\Gamma}_1 \Delta \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_{p-1} \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t \quad (21.61)$$

其中, 为了简化推导, 假定没有时间趋势项。假设样本容量为  $T+p$ , 即观测数据为  $\{\mathbf{y}_{-p+1}, \mathbf{y}_{-p+2}, \dots, \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_T\}$ 。显然, 只有当  $\mathbf{y}_t$  存在协整关系时, 这个 VECM 模型才能成立; 否则, 方程(21.61)左边为平稳过程, 而右边为非平稳过程(因为  $\boldsymbol{\Gamma}_0 \mathbf{y}_{t-1}$  不平稳)。假设协整秩为  $h$ , 则系数矩阵  $\boldsymbol{\Gamma}_0$  必须满足约束条件 “ $\text{rank}(\boldsymbol{\Gamma}_0) = h$ ”。Johansen 的方法是, 在满足 “ $\text{rank}(\boldsymbol{\Gamma}_0) = h$ ” 以及给定  $\{\mathbf{y}_{-p+1}, \mathbf{y}_{-p+2}, \dots, \mathbf{y}_0\}$  的条件下, 最大化  $\{\mathbf{y}_1, \dots, \mathbf{y}_T\}$  的对数似然函数(即条件 MLE)。

假设扰动项为  $n$  维正态分布, 即  $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega})$ , 而且  $\{\boldsymbol{\varepsilon}_t\}$  为独立同分布的。 $n$  维随机向量  $\boldsymbol{\varepsilon}_t$  的联合密度为(参见第 2 章)

$$\frac{1}{(2\pi)^{n/2} |\boldsymbol{\Omega}|^{1/2}} \exp\left\{-\frac{1}{2} \boldsymbol{\varepsilon}_t' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}_t\right\} \quad (21.62)$$

将(21.62)式取对数可得

$$-\frac{n}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Omega}| - \frac{1}{2} \boldsymbol{\varepsilon}_t' \boldsymbol{\Omega}^{-1} \boldsymbol{\varepsilon}_t \quad (21.63)$$

记  $\boldsymbol{\Pi} \equiv (\boldsymbol{\alpha} \ \boldsymbol{\Gamma}_0 \ \boldsymbol{\Gamma}_1 \ \cdots \ \boldsymbol{\Gamma}_{p-1})'$ ,  $\mathbf{x}_t \equiv (1 \ \mathbf{y}_{t-1} \ \Delta \mathbf{y}_{t-1} \ \cdots \ \Delta \mathbf{y}_{t-p+1})'$ , 则  $\boldsymbol{\varepsilon}_t = \Delta \mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t$ , 故  $\{\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \dots, \boldsymbol{\varepsilon}_T\}$  所对应的对数条件似然函数为

$$\max -\frac{nT}{2} \ln 2\pi - \frac{T}{2} \ln |\boldsymbol{\Omega}| - \frac{1}{2} \sum_{t=1}^T [(\Delta \mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t)' \boldsymbol{\Omega}^{-1} (\Delta \mathbf{y}_t - \boldsymbol{\Pi}' \mathbf{x}_t)] \\ s. t. \text{rank}(\boldsymbol{\Gamma}_0) = h \quad (21.64)$$

其中,  $n$  为该系统中变量的个数,  $T$  为样本容量。为了求解以上约束极值问题, 必须确定协整秩  $h$ 。当协整秩为  $h$  时, 系数矩阵  $\boldsymbol{\Gamma}_0$  有  $h$  个自由(线性无关)的列向量。因此, 协整秩  $h$  越大, 则对矩阵  $\boldsymbol{\Gamma}_0$  的约束越少, 其对应的似然函数最大值应该越大。因此, 可以进行以下似然比检验:

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = 0 \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) > 0 \quad (21.65)$$

由于矩阵  $\boldsymbol{\Gamma}_0$  的秩取决于其非零特征值的个数, 故检验统计量被称为“迹统计量”, 记为  $\lambda_{\text{trace}}$ 。由于“迹检验”(trace test)是似然比检验, 故为单边右侧检验, 即  $\lambda_{\text{trace}}$  越大(加上  $H_0$  约束后, 似然函数的最大值下降越多), 则越倾向于拒绝原假设。如果接受 “ $H_0: \text{rank}(\boldsymbol{\Gamma}_0) = 0$ ”, 则认为不存在协整关系。反之, 则继续检验是否存在多个协整关系:

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = 1 \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) > 1 \quad (21.66)$$

依此顺序不断进行检验, 直到接受  $H_0$ , 确认协整秩  $h$  为止。确认协整秩  $h$  后, 就可以用条件 MLE 来估计 VECM 模型中的所有参数, 包括长期参数(协整系数)与短期参数。

Johansen 还考虑了另一类检验:

$$H_0: \text{rank}(\boldsymbol{\Gamma}_0) = p \quad vs \quad H_1: \text{rank}(\boldsymbol{\Gamma}_0) = p + 1 \quad (21.67)$$

其检验统计量为“最大特征值统计量”(maximum eigenvalue statistics), 记为  $\lambda_{\max}$ , 称为“最大特征值检验”。一般认为, 迹检验的效果比特征值检验更好, 故前者为 Stata 检验协整秩的默认方法。

## 21.11 协整分析的 Stata 实例

检验协整秩的 Stata 命令为

`vecrank y1 y2 ... yn, lags(#)` (默认包括常数项, 但不包括时间趋势)

`vecrank y1 y2 ... yn, lags(#) trend(none)` (不包括常数项或时间趋势)

`vecrank y1 y2 ... yn, lags(#) trend(trend)` (包括常数项与时间趋势)

`vecrank y1 y2 ... yn, max` (显示最大特征值统计量及其临界值)

其中, 选择项“lags (#)”表示对应的 VAR 模型中滞后的阶数, 默认为“lags (2)”。命令 `vecrank` 的输出结果将列出“ $h = 0, 1, \dots, n - 1$ ”的一系列检验, 并以星号(\*)标出所接受的  $h$  值。

在作完协整秩检验, 并确定  $h \geq 1$  后, 就可以对 VECM 模型进行最大似然估计。

`vec y1 y2 ... yn, lags(#) rank(#)` (默认设置包括常数项, 但不包括时间趋势)

`vec y1 y2 ... yn, lags(#) rank(#) trend(none)` (不包括常数项, 也不包括时间趋势)

`vec y1 y2 ... yn, lags(#) rank(#) trend(trend)` (包括常数项, 也包括时间趋势)

其中, 选择项“lags (#)”表示对应的 VAR 模型中滞后的阶数, 默认值为“lags (2)”; “rank (#)”表示协整秩的阶数, 默认值为“rank (1)”。以上命令将同时估计短期的 VECM 模型, 以及长期的协整回归。

与 VAR 模型类似, 估计完 VECM 模型后, 应对模型的假设进行检验(diagnostic checking), 然后考察 VECM 模型的脉冲响应函数, 并进行预测。有关的 Stata 命令包括

`veclmar` (估计 VECM 后, 对残差是否存在自相关进行 LM 检验)

`vecnorm` (估计 VECM 后, 检验残差是否服从正态分布)

`vecstable, graph` (估计 VECM 后, 通过特征值检验该 VECM 系统是否为平稳过程, 如果所有特征值都在单位圆内部, 则为平稳过程。选择项“graph”将画出特征值的几何分布图)

`irf create irfname, set(filename) step(#) replace`

估计 VECM 后, 将有关脉冲响应的结果存为“`irfname`”(可自行命名)。选择项“`set`

(filename)" 表示建立脉冲文件 "filename", 使之成为当前的脉冲文件 (make filename active), 并将脉冲结果 "irfname" 存入此脉冲文件 "filename" (若未使用选择项 "set (filename)" 指定脉冲文件, 则将脉冲响应结果存入当前的脉冲文件); 选择项 "step (#)" 表示考察 # 期的脉冲响应函数, 默认值为 "step (8)"; 选择项 "replace" 表示替代已有的同名脉冲响应结果 irfname (如果有)。一个脉冲文件 "filename" 可存储多个脉冲响应结果 "irfname"。

```
irf graph irf, impulse(varname) response(varname) noci
```

画脉冲响应图(未正交化)。其中, 选择项 "impulse(varname)" 用于指定脉冲变量, 而选择项 "response(varname)" 用来指定反应变量, 默认画出所有变量的脉冲响应图。选择项 "noci" 表示不画置信区间, 默认画置信区间。

```
irf graph cirf, impulse(varname) response(varname)
```

画累积脉冲响应图(未正交化)

```
irf graph oirf, impulse(varname) response(varname)
```

画正交化的脉冲响应图

```
irf graph coirf, impulse(varname) response(varname)
```

画正交化的累积脉冲响应图

```
irf graph fevd, impulse(varname) response(varname)
```

画预测方差分解图

如果将以上命令中的 "irf graph" 改为 "irf table", 则将相应信息列表而非画图。

```
fcast compute prefix, step (#)
```

估计 VECM 后, 计算被解释变量未来 # 期的预测值, 并把预测值赋予被解释变量加上前缀 "prefix" (自行确定) 的变量名。

```
fcast graph varlist, observed
```

运行命令 "fcast compute" 后, 将变量 "varlist" 的预测值画图, 其中选择项 "observed" 表示与实际观测值相比较。

下面以数据集 "mpyr.dta" 为例<sup>①</sup>, 对美国的货币需求函数进行协整分析。该数据集包含了美国 1900—1989 年的以下年度宏观变量: logp (价格水平的对数), logy (名义国民生产总值<sup>②</sup>的对数), logm1 (M<sub>1</sub> 的对数), logmr (实际货币<sup>③</sup>的对数, 即 logm1-logp), r (名义利率<sup>④</sup>), logv (货币流通速度的对数, 即 logmr-logy)。

从经济理论出发, 通常将货币需求函数写为

$$\logmr_t = \logm1_t - \logp_t = \beta_0 + \beta_1 \logy_t + \beta_2 r_t + \varepsilon_t \quad (21.68)$$

其中,  $\beta_1$  为货币需求的收入弹性, 一般认为接近于 1; 而  $\beta_2$  为货币需求的利率 "半弹性" (semielasticity), 一般为负。假定以上变量均为单整 I(1) 过程 (参见习题), 故应进行协整分析。首先, 从图形上大致考察 ( $\logmr_t$ ,  $\logy_t$ ,  $r_t$ ) 是否存在协整关系 (如图 21.6):

```
.use mpyr.dta, clear
```

<sup>①</sup> 此例来自 Hayashi (2000), 原始数据来自 Stock and Watson (1993)。

<sup>②</sup> 即 Net National Product (NNP)。

<sup>③</sup> 实际货币的定义为  $M_1/p$ 。

<sup>④</sup> 6 月商业票据的年利率, 6-month commercial paper rate in percentage at an annual rate.

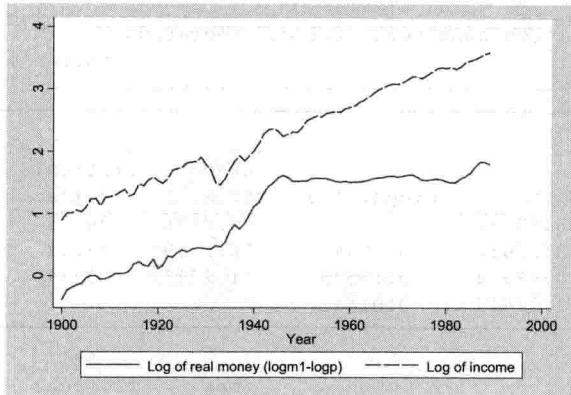


图 21.6 实际货币与收入的时间趋势

```
.line logmr logy year,lpattern("1" "_")
```

从图 21.6 可以看出,实际货币对数与收入对数的时间走势比较接近。在货币需求函数中,如果令  $\beta_1 = 1$  并把  $\log y_t$  移项可得

$$\log m_r - \log y_t = \beta_0 + \beta_2 r_t + \varepsilon_t \quad (21.69)$$

上式左边可以写为,  $\log m_r - \log y_t = \log(M_1/p_y) \equiv \log v$ , 其中  $v = M_1/p_y$  为  $M_1$  的货币流通速度 ( $M_1$  velocity)。这意味着,货币流通速度之对数与名义利率存在线性关系。由于货币流通速度小于 1,故其对数小于 0,为此考虑其负数:

```
.gen _logv = -logv
```

另外,为了在数量级上更匹配,将名义利率除以 10:

```
.gen r10 = r/10
```

然后,考察  $_logv$  与  $r10$  的走势图(如图 21.7):

```
.tsline _logv r10,lpattern("—" "-")
```

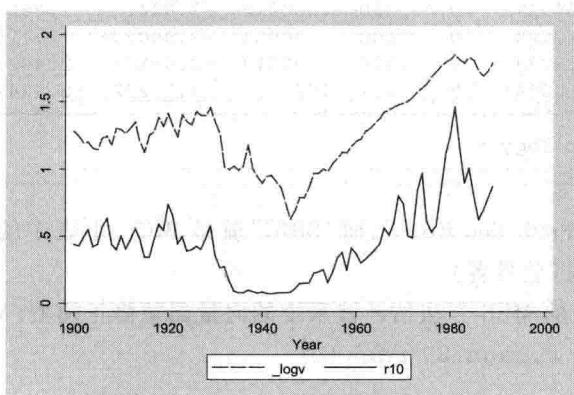


图 21.7 货币流通速度与名义利率的时间趋势

从图 21.7 可以看出,名义利率与货币流通速度对数的升降性具有一定的联动性。以上两图表明,  $(\log m_r, \log y_t, r_t)$  很可能存在长期均衡关系,即为协整系统。为此,首先要确定协整秩,即共有多少个线性无关的协整向量:

```
.vecrank logmr logy r,trend(trend) max
```

Johansen tests for cointegration						
Trend:	trend				Number of obs =	88
Sample:	1902 - 1989				Lags =	2
5%						
maximum			trace	critical		
rank	parms	LL	eigenvalue	statistic	value	
0	15	138.03791	.	46.3731	34.55	
1	20	153.13651	0.29047	16.1759*	18.17	
2	23	160.58579	0.15575	1.2773	3.74	
3	24	161.22445	0.01441			
5%						
maximum			max	critical		
rank	parms	LL	eigenvalue	statistic	value	
0	15	138.03791	.	30.1972	23.78	
1	20	153.13651	0.29047	14.8985	16.87	
2	23	160.58579	0.15575	1.2773	3.74	
3	24	161.22445	0.01441			

包含常数项与时间趋势项的协整秩迹检验 (trace statistic) 结果表明, 只有一个线性无关的协整向量 (上表中打星号者)。而最大特征值检验 (max statistic) 也表明, 可以在 5% 的水平上拒绝“协整秩为 0”的原假设, 但无法拒绝“协整秩为 1”的原假设。

其次, 检验该系统所对应的 VAR 表示法 (VAR representation) 的滞后阶数:

```
.varsoc logmr logy r
```

Selection-order criteria							
Sample:	1904 - 1989	Number of obs = 86					
lag	LL	LR	df	p	FPE	AIC	HQIC
0	-251.056				.073876	5.90827	5.94272
1	132.578	767.27	9	0.000	.000012	-2.80415	-2.66632
2	148.293	31.429	9	0.000	.00001*	-2.96029*	-2.7191*
3	151.979	7.3723	9	0.598	.000012	-2.83672	-2.49215
4	162.506	21.054*	9	0.012	.000011	-2.87222	-2.42429

Endogenous: logmr logy r  
Exogenous: cons

其中, “FPE”表示“Final Prediction Error”, 而“SBIC”就是 BIC。上表中的大多数准则表明 (包括 AIC), 应选择滞后二阶 (打星号者)。

下面, 使用 Johansen 的 MLE 方法估计该系统的向量误差修正模型 (VECM):

```
.vec logmr logy r,lags(2) rank(1)
```

Vector error-correction model						
Sample:	1902 - 1989		No. of obs	=	88	
Log likelihood =	150.6503		AIC	=	-3.037506	
Det(Sigma_ml) =	6.54e-06		HQIC	=	-2.8447	
			SBIC	=	-2.55893	
Equation	Parms	RMSE	R-sq	chi2	P>chi2	
D_logmr	5	.050841	0.2758	31.61646	0.0000	
D_logy	5	.056773	0.3483	44.35204	0.0000	
D_r	5	1.1287	0.2133	22.50845	0.0004	
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
D_logmr						
_cel						
L1.	-.0533601	.0410678	-1.30	0.194	-.1338515	.0271314
logmr						
LD.	.2079032	.1107256	1.88	0.060	-.0091149	.4249214
logy						
LD.	.0086587	.101984	0.08	0.932	-.1912263	.2085438
r						
LD.	-.0063968	.0052854	-1.21	0.226	-.0167559	.0039624
_cons	.0186333	.0064766	2.88	0.004	.0059393	.0313273
D_logy						
_cel						
L1.	.0298268	.0458591	0.65	0.515	-.0600554	.119709
logmr						
LD.	.2666361	.1236437	2.16	0.031	.0242989	.5089733
logy						
LD.	.2330244	.1138823	2.05	0.041	.0098191	.4562296
r						
LD.	-.0145323	.005902	-2.46	0.014	-.0261001	-.0029646
_cons	.0157173	.0072323	2.17	0.030	.0015424	.0298923
D_r						
_cel						
L1.	-3.482578	.9117297	-3.82	0.000	-5.269536	-1.695621
logmr						
LD.	2.663613	2.458173	1.08	0.279	-2.154318	7.481544
logy						
LD.	.6533844	2.264106	0.29	0.773	-3.784182	5.09095
r						
LD.	.294868	.1173386	2.51	0.012	.0648885	.5248475
_cons	-.0001509	.1437852	-0.00	0.999	-.2819648	.281663
Cointegrating equations						
Equation	Parms	chi2	P>chi2			
_cel	2	794.1155	0.0000			
Identification: beta is exactly identified						
Johansen normalization restriction imposed						
beta	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
_cel						
logmr	1	.	.	.	.	.
logy	-.9754246	.0346169	-28.18	0.000	-1.043273	-.9075767
r	.1124051	.0097191	11.57	0.000	.093356	.1314542
_cons	.7299535	.	.	.	.	.

上表上部为误差修正模型,下部则为协整方程(cointegrating equation),以“\_ce1”来表示。我们主要对货币需求函数感兴趣,即上表中协整方程所代表的长期均衡关系。根据表中信息,可将估计的货币需求函数写为

$$\widehat{\logmr}_t = -0.73 + 0.98\logy_t - 0.11r_t \quad (21.70)$$

其中,货币需求的收入弹性为 0.98,而货币需求的利率半弹性为 -0.11,符合经济理论的预期。作为对比,下面直接用 OLS 估计此长期均衡关系(即 EG-ADF 两步法):

```
.reg logmr logy r
```

Source	SS	df	MS	Number of obs = 90 F( 2, 87) = 1169.93 Prob > F = 0.0000 R-squared = 0.9642 Adj R-squared = 0.9633 Root MSE = .13272		
Model	41.216226	2	20.608113			
Residual	1.53248421	87	.017614761			
Total	42.7487102	89	.480322587			
logmr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
logy	.9418376	.0196178	48.01	0.000	.9028451	.9808302
r	-.0832229	.0053829	-15.46	0.000	-.0939219	-.0725239
_cons	-.7737089	.0426886	-18.12	0.000	-.8585572	-.6888606

从上表可以看出,OLS 系数估计值与 Johansen 的 MLE 估计结果比较接近。当然,从理论来说,MLE 估计更有效率。

下面检验 VECM 模型的残差是否存在自相关。如果存在自相关,则预示着需要增加滞后阶数:

```
.qui vec logmr logy r, lags(2) rank(1)
.veclmar
```

Lagrange-multiplier test				
lag	chi2	df	Prob > chi2	
1	6.6260	9	0.67599	
2	12.5541	9	0.18384	
H0: no autocorrelation at lag order				

结果显示,可以接受“无自相关”的原假设。进一步检验残差的正态性:

```
.vecnorm
```

## Jarque-Bera test

Equation	chi2	df	Prob > chi2
D_logmr	15.620	2	0.00041
D_logy	23.598	2	0.00001
D_r	6.147	2	0.04626
ALL	45.365	6	0.00000

## Skewness test

Equation	Skewness	chi2	df	Prob > chi2
D_logmr	.04087	0.024	1	0.87564
D_logy	-.90936	12.128	1	0.00050
D_r	.04973	0.036	1	0.84897
ALL	12.189	3	0.00676	

## Kurtosis test

Equation	Kurtosis	chi2	df	Prob > chi2
D_logmr	5.0624	15.595	1	0.00008
D_logy	4.7687	11.470	1	0.00071
D_r	4.291	6.111	1	0.01344
ALL	33.176	3	0.00000	

上表显示,可在 5% 的显著性水平上拒绝 D. logmr, D. logy 与 D. r 的残差项服从正态分布的原假设,尽管对 D. logmr 与 D. r 残差的偏度检验并不拒绝正态性(说明这两个变量的残差在对称性方面与正态分布较接近)。尽管 Johansen 的 MLE 估计是基于扰动项服从正态分布而推导出来的,但作为准最大似然估计量(QMLE),在更弱的非正态条件下也成立(参见 Lutkepohl, 2005, p. 297)。因此,残差的非正态性对 VECM 模型的估计影响不大(但可能对区间预测产生影响)。

下面检验此 VECM 系统是否稳定,结果如图 21.8。

.vecstable, graph

## Eigenvalue stability condition

Eigenvalue	Modulus
1	1
1	1
.4092107 + .4061819i	.576574
.4092107 - .4061819i	.576574
.2217304 + .07266624i	.233334
.2217304 - .07266624i	.233334

The VECM specification imposes 2 unit moduli.

结果显示,除了 VECM 模型本身所假设的单位根之外,伴随矩阵的所有特征值均落在单位圆之内,故是稳定的。

下面考察此 VECM 模型的正交化脉冲响应函数(将脉冲文件命名为 money)(结果如图 21.9)。

```
.irf create money, set(money)
step(10) replace
(file money.irf created)
(file money.irf now active)
(file money.irf updated)
.irf graph oirf
```

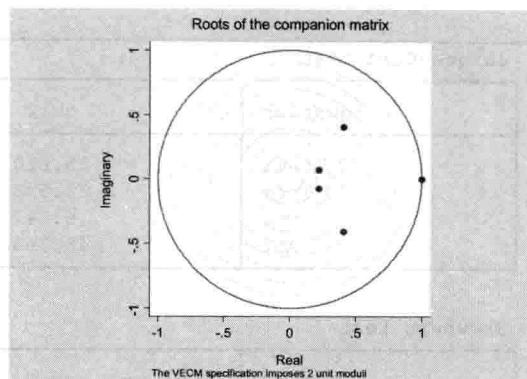


图 21.8 VECM 系统稳定性的判别图

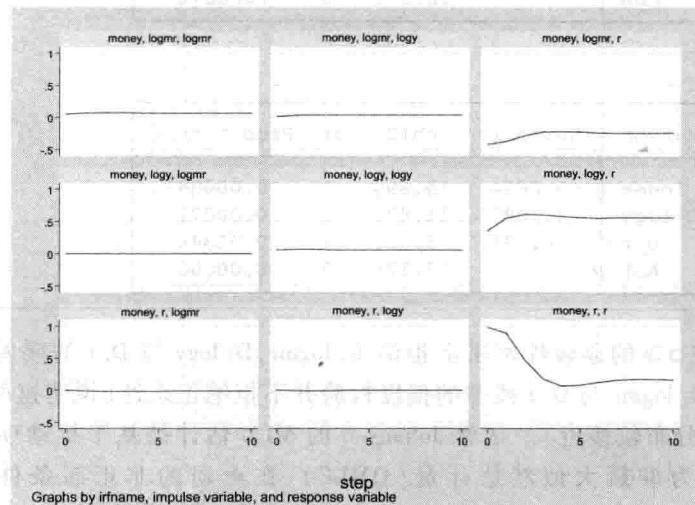


图 21.9 正交化的脉冲响应图

从图 21.9 可知,对于一个含单位根的协整系统,对一个变量的冲击可能对其自身与其他变量具有持久的影响(平稳 VAR 过程的脉冲响应不具有此性质)。估计 VECM 模型后,可以用它进行预测。假设我们仅用 1980 年以前的数据来估计 VECM 模型,然后预测 1980—1989 年的十年数据,并与实际观测值比较,结果如图 21.10。

```
.quietly vec logmr logy r if year < 1980, lags(2) rank(1)
.fcast compute f_, step(10)
.fcast graph f_logmr f_logy f_r, observed lpattern("_")
```

图 21.10 似乎表明,对国民收入(logy)的预测最为准确,对货币供给(logmr)的预测次之,而利率(r)最难预测(但要注意它们的纵轴坐标单位与绝对位置都不同)。其中,利率(r)的实际观测值曾一度落在预测值 95% 的置信区间之外。

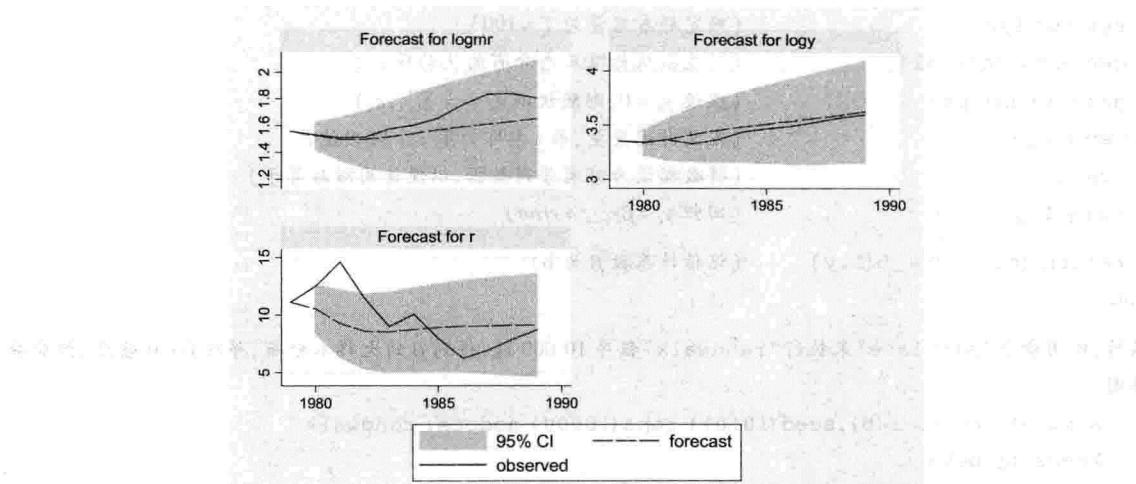


图 21.10 对未来 10 年的预测

## 习题

- 21.1** 对于 AR(2),  $y_t = 2 + \frac{5}{2}y_{t-1} - y_{t-2} + \varepsilon_t$ , 写出其特征方程  $\varphi(z)$ , 并确定其稳定性。
- 21.2** 考虑 AR(1) 模型,  $y_t = \rho y_{t-1} + \varepsilon_t, t = 2, \dots, T$ , 其中  $|\rho| < 1$ ,  $\varepsilon_t$  为白噪声且独立于  $\{y_t\}$ 。记  $\rho$  的 OLS 估计量为  $\hat{\rho}$ 。
- (1) 计算  $E(y_t^2)$ 。
  - (2) 根据定理 5.1,  $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{d} N(0, \text{Avar}(\hat{\rho}))$ 。证明  $\text{Avar}(\hat{\rho}) = 1 - \rho^2$  (提示: 使用同方差情形下  $\text{Avar}(\hat{\rho})$  的简化公式)。
  - (3) 假设在  $\rho = 1$  时, (2) 的结论依然成立。证明  $\sqrt{T}(\hat{\rho} - \rho) \xrightarrow{P} 0$ 。
  - (4) (此题不必做) 可以证明, 当  $\rho = 1$  时,  $T(\hat{\rho} - \rho) \xrightarrow{d}$  非退化分布。因此, 在单位根情形下,  $\hat{\rho}$  的收敛速度为  $T$ , 快于平稳情况下 OLS 估计量的收敛速度  $\sqrt{T}$ , 称为“超一致”(superconsistency), 详见 Hamilton (1994, p. 486–488)。
- 21.3** 使用 ADF, PP, DF-GLS 及 KPSS 检验, 检验数据集 mpyr.dta 中各主要变量是否含有单位根。
- 21.4** 使用 ADF, PP, DF-GLS 及 KPSS 检验, 检验数据集 lutkepohl2.dta 中以下变量是否含有单位根: ln\_inc (收入的对数), ln\_consump (消费的对数), 以及 ln\_inv (投资的对数)。
- 21.5** 假设数据集 lutkepohl2.dta 中变量 ln\_inv, ln\_inc 及 ln\_consump 均为一阶单整。使用 Johansen 方法对 (ln\_inc, ln\_consump, ln\_inv) 进行协整分析。长期均衡关系的系数符号合理吗?

## 附录

### A21.1 产生图 21.2 的 Stata 程序

首先定义一个叫“randwalk”的程序来产生不带漂移项的随机游走, 并进行一阶自回归的 OLS 估计:

```
program randwalk, rclass
drop _all
        (删去内存中已有数据)
```

```

set obs 100          (确定样本容量为  $T=100$ )
gen eps = rnormal() (产生服从标准正态分布的扰动项  $\varepsilon_t$ )
gen y = sum(eps)    (假设  $y_0=0$ , 则随机游走  $y_t = \sum_{s=1}^t \varepsilon_s$ )
gen t = _n           (定义时间变量, 第  $t$  期即为第  $i$  个观测值)
tsset t              (将数据设为时间序列数据, 以便使用滞后算子)
reg y L.y            (回归  $y_t = \beta y_{t-1} + error$ )
return scalar b=_b[L.y] (记估计系数  $\hat{\beta}$  为 b)
end

```

然后, 使用命令“simulate”来执行“randwalk”程序 10 000 遍, 得到  $\hat{\beta}$  的大样本分布, 并画其(核密度)经验分布图:

```

simulate beta=r(b), seed(10101) reps(10000) nodots: randwalk
kdensity beta

```

### A21.2 $\alpha(L)$ 为绝对值可加总(AS)

证明: 由于  $\alpha_j = -(\psi_{j+1} + \psi_{j+2} + \dots) = -\sum_{i=j+1}^{\infty} \psi_i$ , 故

$$\begin{aligned}
\sum_{j=0}^{\infty} |\alpha_j| &= \sum_{j=0}^{\infty} \left| -\sum_{i=j+1}^{\infty} \psi_i \right| \\
&= \sum_{j=0}^{\infty} \left| \sum_{i=j+1}^{\infty} \psi_i \right| \\
&\leq \sum_{j=0}^{\infty} \sum_{i=j+1}^{\infty} |\psi_i| \quad (\text{根据三角不等式 } |a+b| \leq |a| + |b|) \\
&= \sum_{j=0}^{\infty} (|\psi_{j+1}| + |\psi_{j+2}| + \dots) \\
&= |\psi_1| + |\psi_2| + |\psi_3| + \dots \quad (j=0) \\
&\quad |\psi_2| + |\psi_3| + \dots \quad (j=1) \\
&\quad \quad |\psi_3| + \dots \quad (j=2) \\
&\quad \quad \quad \vdots \\
&= \sum_{i=1}^{\infty} i |\psi_i| \\
&= \sum_{i=0}^{\infty} i |\psi_i| < \infty \quad (\text{因为 } \psi(L) \text{ 为 OS})
\end{aligned}$$

因此,  $\sum_{j=0}^{\infty} |\alpha_j| < \infty$ , 故  $\alpha(L)$  为 AS。

# 第 22 章 自回归条件异方差模型

## 22.1 条件异方差模型的例子

通常认为,横截面数据容易存在异方差,而时间序列数据常存在自相关。然而,Engle (1982)指出,时间序列数据也常存在一种特殊的异方差,即“自回归条件异方差”(Autoregressive Conditional Heteroskedasticity,简记 ARCH)<sup>①</sup>。Bollerslev (1986)对 ARCH 进行了推广,称为“Generalized ARCH”,简记 GARCH。

考察美国道琼斯股指在 1953—1990 年期间日收益率的波动,参见图 22.1。

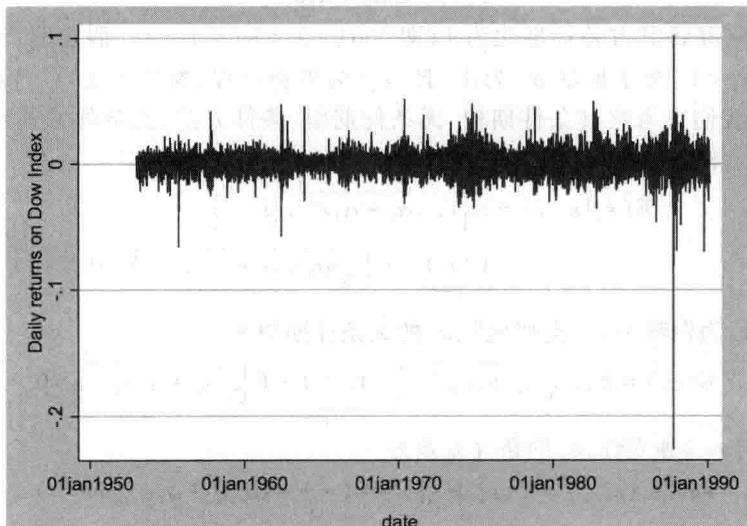


图 22.1 美国道琼斯股指 1953—1990 的日收益率<sup>②</sup>

从图 22.1 可以看出,股指日收益率在某一段时间内剧烈波动,而在另一段时间内风平浪静。从理论上,这可以抽象为,当本期或过去若干期的波动(方差)较大时,未来几期的波动(方差)很可能也较大;反之亦然。换言之,方差大的观测值似乎集聚在一起,而方差小的观测值似乎也集聚在一起。这被称为“波动性集聚”(volatility clustering)或“扎堆”。

在 Engle (1982) 的论文发表之前,由于缺乏更好的度量,经济学家一直假设时间序列的方差是恒定的。由于 ARCH 模型考虑了方差的波动性,故可以更好地预测方差(variance forecast),在

① Robert Engle 因此于 2003 年获诺贝尔经济学奖。

② 根据数据集 dow1.dta 计算,参见习题。

金融领域有着重要的应用价值。比如,金融学中使用“VaR 方法”(Value-at-Risk)来度量金融资产所面临的风险,就依赖于对未来收益率方差的预测。因此,Engle 的贡献是一个重要突破。

## 22.2 ARCH 模型的性质

考虑一般的线性回归模型:

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t \quad (22.1)$$

记扰动项  $\varepsilon_t$  的条件方差为  $\sigma_t^2 \equiv \text{Var}(\varepsilon_t | \varepsilon_{t-1}, \dots)$ , 其中  $\sigma_t^2$  的下标  $t$  表示条件方差可以随时而变。受到波动性集聚现象的启发,假设  $\sigma_t^2$  取决于上一期扰动项之平方:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \quad (22.2)$$

这就是“ARCH(1) 扰动项”。更一般地,假设  $\sigma_t^2$  依赖于前  $p$  期扰动项之平方:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_p \varepsilon_{t-p}^2 \quad (22.3)$$

这就是“ARCH( $p$ ) 扰动项”。不失一般性,以 ARCH(1) 为例来考察 ARCH 扰动项的性质。假设扰动项  $\varepsilon_t$  的生成过程为

$$\varepsilon_t = v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \quad (22.4)$$

其中,  $v_t$  为白噪声,并将其方差标准化为 1,即  $\text{Var}(v_t) = E(v_t^2) = 1$ 。假定  $v_t$  与  $\varepsilon_{t-1}$  相互独立,而且  $\alpha_0 > 0, 0 < \alpha_1 < 1$ (为了保证  $\sigma_t^2$  为正,且  $\{\varepsilon_t\}$  为平稳过程,参见下文)。序列  $\{\varepsilon_t\}$  具有怎样的性质呢?下面我们来考察其条件期望、无条件期望、条件方差、无条件方差及序列相关。

由于  $v_t$  与  $\varepsilon_{t-1}$  相互独立,  $\varepsilon_t$  的条件期望为

$$\begin{aligned} E(\varepsilon_t | \varepsilon_{t-1}) &= E\left\{v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \mid \varepsilon_{t-1}\right\} \\ &= \underbrace{E(v_t)}_{=0} \cdot E\left\{\sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \mid \varepsilon_{t-1}\right\} = 0 \end{aligned} \quad (22.5)$$

其中,  $E(v_t) = 0$ ( $v_t$  为白噪声)。类似地<sup>①</sup>,  $\varepsilon_t$  的无条件期望为

$$E(\varepsilon_t) = E\left\{v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}\right\} = \underbrace{E(v_t)}_{=0} \cdot E\left\{\sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}\right\} = 0 \quad (22.6)$$

同样地,根据  $v_t$  与  $\varepsilon_{t-1}$  独立性,  $\varepsilon_t$  的条件方差为

$$\begin{aligned} \text{Var}(\varepsilon_t | \varepsilon_{t-1}) &= E(\varepsilon_t^2 | \varepsilon_{t-1}) = E(v_t^2) \cdot E(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 | \varepsilon_{t-1}) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2 | \varepsilon_{t-1}) = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 \end{aligned} \quad (22.7)$$

其中,  $E(v_t^2) = 1$ 。上式就是 ARCH(1) 的定义式“ $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$ ”。 $\alpha_1$  越大,则说明上一期扰动项之平方对条件方差  $\sigma_t^2$  的冲击越大。进一步考察  $\varepsilon_t$  的无条件方差:

$$\begin{aligned} \text{Var}(\varepsilon_t) &= E(\varepsilon_t^2) = E[v_t^2 (\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)] \\ &= E(v_t^2) \cdot E(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) = \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) \end{aligned} \quad (22.8)$$

对于差分方程“ $E(\varepsilon_t^2) = \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2)$ ”,由于  $0 < \alpha_1 < 1$ ,故该差分方程有稳定解。令  $E(\varepsilon_t^2) = E(\varepsilon_{t-1}^2)$ ,可得  $E(\varepsilon_t^2) = \frac{\alpha_0}{1 - \alpha_1}$ 。因此,ARCH 扰动项的无条件方差为常数,不随时间而

<sup>①</sup> 也可以使用迭代期望定律来证明。

变化。再来看  $\varepsilon_t$  与  $\varepsilon_{t-i}$  ( $i \neq 0$ ) 的序列相关:

$$\begin{aligned} E(\varepsilon_t \varepsilon_{t-i}) &= E\{v_t v_{t-i} \sqrt{(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)(\alpha_0 + \alpha_1 \varepsilon_{t-i}^2)}\} \\ &= \underbrace{E(v_t v_{t-i})}_{=0} \cdot E\{\sqrt{(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)(\alpha_0 + \alpha_1 \varepsilon_{t-i}^2)}\} = 0 \end{aligned} \quad (22.9)$$

其中,由于  $v_t$  为白噪声,故  $E(v_t v_{t-i}) = 0$ 。从上面的推导可以看出,扰动项  $\{\varepsilon_t\}$  完全满足古典模型关于“同方差”<sup>①</sup>与“无自相关”的假定。事实上,虽然  $\{\varepsilon_t\}$  存在条件异方差,却是白噪声!因此,高斯-马尔可夫定理成立,OLS 是最佳线性无偏估计(BLUE)。然而,OLS 显然忽略了条件异方差这一重要信息。如果我们跳出线性估计的范围,则可以找到更优的非线性估计,即最大似然估计。

## 22.3 ARCH 模型的 MLE 估计

对于 ARCH(1) 模型,为了保证条件方差  $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$  始终为非负,必须限制参数  $\alpha_0, \alpha_1$  均为正数。如果  $\alpha_0 < 0$  或  $\alpha_1 < 0$ ,则可能出现 “ $\sigma_t^2 < 0$ ” 的情形,参见图 22.2。另外,  $\alpha_1 < 1$  是为了保证  $\{\varepsilon_t\}$  为平稳过程。如果  $\alpha_1 > 1$ ,则  $\text{Var}(\varepsilon_t)$  将随时间而增大,不是平稳过程。

假设样本容量为  $T$ 。显然,在 ARCH(1) 模型中,  $\{\varepsilon_t\}$  并非独立同分布的,因为相邻的扰动项通过公式 “ $\varepsilon_t = v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2}$ ” 而联系在一起。此时,如何计算样本的似然函数呢?由于  $\varepsilon_t$  仅依赖于  $\varepsilon_{t-1}$ ,而不依赖于  $\{\varepsilon_{t-2}, \varepsilon_{t-3}, \dots\}$ ,故可以将  $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$  的联合密度函数分解如下:

$$\begin{aligned} f(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T) &= f(\varepsilon_1)f(\varepsilon_2 | \varepsilon_1)f(\varepsilon_3 | \varepsilon_2, \varepsilon_1) \cdots f(\varepsilon_T | \varepsilon_{T-1}, \dots) \\ &= f(\varepsilon_1)f(\varepsilon_2 | \varepsilon_1)f(\varepsilon_3 | \varepsilon_2) \cdots f(\varepsilon_T | \varepsilon_{T-1}) \end{aligned} \quad (22.10)$$

由于无条件密度函数  $f(\varepsilon_1)$  不易计算(要用到  $\varepsilon_1$  的无条件方差的表达式,导致在似然函数中出现非线性项),常将  $f(\varepsilon_1)$  忽略不计,即考虑在  $\varepsilon_1$  给定情况下的条件最大似然估计法(conditional MLE)。假设  $\varepsilon_t \sim N(0, \sigma_t^2)$ ,而  $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$ ,可得似然函数:

$$L = \prod_{t=2}^T \frac{1}{\sqrt{2\pi(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)}} \exp\left\{-\frac{\varepsilon_t^2}{2(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2)}\right\} \quad (22.11)$$

$$\ln L = -\frac{T-1}{2} \ln 2\pi - \frac{1}{2} \sum_{t=2}^T \ln(\alpha_0 + \alpha_1 \varepsilon_{t-1}^2) - \frac{1}{2} \sum_{t=2}^T \frac{\varepsilon_t^2}{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2} \quad (22.12)$$

根据方程(22.1),将 “ $\varepsilon_t = y_t - \mathbf{x}_t' \boldsymbol{\beta}$ ” 代入上式,则对数似然函数  $\ln L$  成为参数  $(\alpha_0, \alpha_1, \beta_0, \beta_1)$

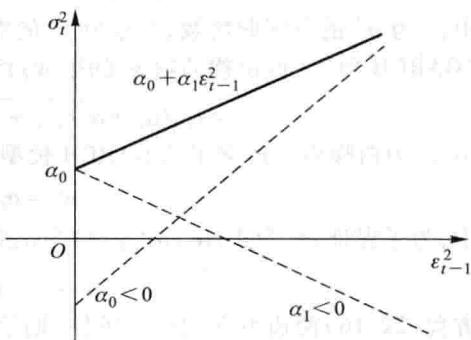


图 22.2 对 ARCH(1) 参数的正约束

<sup>①</sup> 指的是“无条件方差”相同。需要注意的是,本章的条件方差指的是,给定扰动项的滞后值与解释变量  $X$ ,即  $\text{Var}(\varepsilon_t | \varepsilon_{t-1}, \dots, X)$ ;而第 3 章的条件方差指的是,给定解释变量  $X$ ,即  $\text{Var}(\varepsilon_t | X)$ 。因此,本章的“无条件方差”就是第 3 章的“条件方差”,故适用高斯-马尔可夫定理。

的函数。可以对  $(\alpha_0, \alpha_1, \beta_0, \beta_1)$  求偏导得到  $\ln L$  的最大值, 通常由计算机数值计算来进行。因此, 对 ARCH 模型进行 MLE 估计的特点是, 对原方程 ( $y_t = \mathbf{x}'_t \boldsymbol{\beta} + \varepsilon_t$ ) 与条件方差方程 ( $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2$ ) 同时进行估计。

类似地, 如果要估计 ARCH( $p$ ), 则将  $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p\}$  (即前  $p$  个观测值) 视为给定, 然后使用条件 MLE。即使扰动项不服从正态分布, 作为准最大似然估计量 (QMLE), 仍可能是一致的。

## 22.4 GARCH 模型

在 ARCH( $p$ ) 模型中, 如果  $p$  很大, 则要估计很多参数, 会损失样本容量。Bollerslev (1986) 提出 GARCH, 使得待估计参数减少, 而对未来条件方差的预测更加准确。其基本思想是, 在 ARCH 模型的基础上, 再加上  $\sigma_t^2$  的自回归部分, 即  $\sigma_t^2$  还是  $\{\sigma_{t-1}^2, \dots, \sigma_{t-p}^2\}$  的函数。GARCH( $p, q$ ) 的模型设定为

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 + \gamma_1 \sigma_{t-1}^2 + \dots + \gamma_p \sigma_{t-p}^2 \quad (22.13)$$

其中,  $p$  为  $\sigma_t^2$  的自回归阶数, 而  $q$  为  $\varepsilon_t^2$  的滞后阶数。在 Stata 中, 称  $\varepsilon_{t-i}^2$  为“ARCH 项”, 而称  $\sigma_{t-i}^2$  为“GARCH 项”。假定扰动项  $\varepsilon_t$  的生成过程为

$$\varepsilon_t = v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \dots + \alpha_q \varepsilon_{t-q}^2 + \gamma_1 \sigma_{t-1}^2 + \dots + \gamma_p \sigma_{t-p}^2} \quad (22.14)$$

其中,  $v_t$  为白噪声。最常用的 GARCH 模型为 GARCH(1,1):

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2 \quad (22.15)$$

其中, 为了保证  $\sigma_t^2$  为正,  $\alpha_0, \alpha_1, \gamma_1$  均为正数。GARCH(1,1) 扰动项  $\varepsilon_t$  的生成过程为

$$\varepsilon_t = v_t \sqrt{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2} \quad (22.16)$$

将方程(22.16)两边平方, 取(无条件)期望可得

$$\begin{aligned} E(\varepsilon_t^2) &= E(v_t^2) \cdot E(\underbrace{\alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2}_{=1}) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) + \gamma_1 E(\sigma_{t-1}^2) \quad (\sigma_{t-1}^2 \equiv E(\varepsilon_{t-1}^2 | \varepsilon_{t-2})) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) + \gamma_1 E[E(\varepsilon_{t-1}^2 | \varepsilon_{t-2})] \quad (\text{迭代期望定律}) \\ &= \alpha_0 + \alpha_1 E(\varepsilon_{t-1}^2) + \gamma_1 E(\varepsilon_{t-1}^2) \\ &= \alpha_0 + (\alpha_1 + \gamma_1) E(\varepsilon_{t-1}^2) \end{aligned} \quad (22.17)$$

在上式的推导中使用了迭代期望定律。由此可知, 为了保证  $\{\varepsilon_t\}$  为平稳过程(无条件方差不发散), 必须要求  $\alpha_1 + \gamma_1 < 1$ 。

为何使用 GARCH 模型能节省待估参数? 直观来说, 因为  $\sigma_{t-1}^2$  中已经包含了  $\{\varepsilon_{t-2}^2, \dots, \varepsilon_{t-p-1}^2\}$  的信息。比如, 对 GARCH(1,1) 使用迭代法可得

$$\begin{aligned} \sigma_t^2 &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 \sigma_{t-1}^2 \\ &= \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \gamma_1 (\alpha_0 + \alpha_1 \varepsilon_{t-2}^2 + \gamma_1 \sigma_{t-2}^2) \\ &= \alpha_0 + \alpha_0 \gamma_1 + \alpha_1 \varepsilon_{t-1}^2 + \alpha_1 \gamma_1 \varepsilon_{t-2}^2 + \gamma_1^2 \sigma_{t-2}^2 \\ &= \dots \\ &= \alpha_0 (1 + \gamma_1 + \gamma_1^2 + \dots) + \alpha_1 (\varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-2}^2 + \gamma_1^2 \varepsilon_{t-3}^2 + \dots) \\ &= \frac{\alpha_0}{1 - \gamma_1} + \alpha_1 (\varepsilon_{t-1}^2 + \gamma_1 \varepsilon_{t-2}^2 + \gamma_1^2 \varepsilon_{t-3}^2 + \dots) \end{aligned} \quad (22.18)$$

由此可见,在某种意义上<sup>①</sup>,GARCH(1,1)等价于无穷阶 ARCH 模型。因此,如果将  $\sigma_{t-1}^2$  作为解释变量引入,常可把高阶 ARCH( $p$ )模型简化为 GARCH(1,1)。对 GARCH 模型可同样使用 MLE 估计。

## 22.5 何时使用 ARCH 或 GARCH 模型

只有在扰动项存在条件异方差时,才需要使用 ARCH 或 GARCH 模型。那么,如何判断扰动项是否存在条件异方差呢?初步的方法可以画时间序列图,看看是否存在“波动性集聚”。

严格的统计检验包括以下三种方法。

**方法一** 首先,用 OLS 估计原方程“ $y_t = \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t$ ”,得到残差序列  $\{e_t\}$ 。其次,用 OLS 估计辅助回归, $e_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2 + \text{error}_t$ ,并检验原假设“ $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$ ”(不存在条件异方差)。Engle (1982) 提出进行 LM 检验,其检验统计量为  $TR^2 \xrightarrow{d} \chi^2(p)$ ,其中  $T$  为样本容量,  $R^2$  为上述辅助回归的可决系数。如果拒绝  $H_0$ ,则认为应使用 ARCH 或 GARCH 模型。

在 Stata 中,此 LM 检验可通过命令 reg 的“后估计命令”(postestimation command) estat archlm 来实现。

**方法二** 可以对残差平方序列  $\{e_t^2\}$  进行 Q 检验,检验其序列相关性。如果  $\{e_t^2\}$  存在自相关,则认为  $\varepsilon_t$  存在条件异方差。

**方法三** 最为直接的方法是,在估计 ARCH 或 GARCH 模型之后,看条件方差方程中的系数(即所有  $\alpha$  与  $\gamma$ )是否显著。

## 22.6 ARCH 与 GARCH 模型的扩展

### 1. ARCH-M

金融理论认为,金融资产的风险越高,其期望收益率也应该越高,这样才会有人愿意持有它。超出正常期望收益率的部分,称为“风险溢价”(risk premium)。但在标准的 ARCH 模型中,变量的均值与条件方差却没有关系。Engle, Lilien and Robins (1987) 提出了如下“ARCH-in-Mean 模型”(简记 ARCH-M)。

假设金融资产的超额收益率满足以下方程:

$$y_t = \beta + \underbrace{\delta \sigma_t^2}_{\text{risk premium}} + \varepsilon_t \quad (22.19)$$

其中, $y_t$  为“超额收益率”(excess return),即超出无风险的国库券收益率的部分; $\varepsilon_t$  为对超额收益率不可预见的冲击;而  $(\beta + \delta \sigma_t^2)$  为风险溢价,是条件方差  $\sigma_t^2$  的增函数,即  $\delta > 0$ 。假设  $\varepsilon_t$  服从 ARCH( $p$ ) 过程, $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2$ 。可以使用 MLE 对方程 (22.19) 与条件方差方程同时进行估计。由于超额收益率  $y_t$  的期望(mean)中包含一个条件方差项( $\delta \sigma_t^2$ ),故名“ARCH-in-Mean”。

① 要求此 ARCH( $\infty$ ) 的系数呈几何级数递减。

## 2. TARCH

“坏消息”对资产价格波动性的影响可能大于好消息的影响。Glosten, Jagannathan and Runkle (1993) 提出了非对称 (asymmetric) 的“门限 GARCH”模型 (Threshold GARCH, 简记 TARCH)。

假设条件方差方程为

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \lambda_1 \underbrace{\varepsilon_{t-1}^2 \cdot \mathbf{1}(\varepsilon_{t-1} > 0)}_{\text{TARCH}} + \beta_1 \sigma_{t-1}^2 \quad (22.20)$$

其中,  $\mathbf{1}(\cdot)$  为示性函数, 即当  $\varepsilon_{t-1} > 0$  时, 取值为 1; 反之, 则为 0。Stata 称 “ $\varepsilon_{t-1}^2 \cdot \mathbf{1}(\varepsilon_{t-1} > 0)$ ” 为 “TARCH” 项。

## 3. EGARCH

在标准的 GARCH 模型中, 对参数的取值有所限制, 给 MLE 估计带来不便。为此, 考虑以下对数形式的条件方差方程:

$$\ln \sigma_t^2 = \alpha_0 + \alpha_1 \underbrace{(\varepsilon_{t-1}/\sigma_{t-1})}_{\text{EARCH}} + \lambda_1 \underbrace{|\varepsilon_{t-1}/\sigma_{t-1}|}_{\text{EARCH\_a}} + \beta_1 \underbrace{\ln \sigma_{t-1}^2}_{\text{EGARCH}} \quad (22.21)$$

其中,  $(\varepsilon_{t-1}/\sigma_{t-1})$  为  $\varepsilon_{t-1}$  的标准化(除以自身的标准差), Stata 称为 “EARCH” 项。只要  $\alpha_1 \neq 0$ , 则这个模型也包括了非对称效应(类似于 TARCH)。 $|\varepsilon_{t-1}/\sigma_{t-1}|$  表示对称效应, Stata 称为 “EARCH\_a” 项 (a 表示 “absolute value”, 即绝对值)。由于  $\sigma_t^2$  为指数形式, 故称为 “指数 GARCH” (Exponential GARCH, 简记 EGARCH)。Stata 称  $\ln \sigma_{t-1}^2$  为 “EGARCH” 项。EGARCH 的优点在于, 无论  $\ln \sigma_t^2$  取何值, 都有  $\sigma_t^2 = \exp(\ln \sigma_t^2) > 0$ , 故对方程 (22.21) 中的所有参数都没有任何限制。

## 4. 带 ARMA 的 GARCH

考虑如下线性回归模型:

$$y_t = \mathbf{x}' \boldsymbol{\beta} + u_t \quad (22.22)$$

其中, 扰动项  $u_t$  为 ARMA( $p, q$ ) 过程:

$$u_t = \sum_{i=1}^p \rho_i u_{t-i} + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (22.23)$$

其中,  $\varepsilon_t$  为 GARCH(或 ARCH) 扰动项。将方程 (22.23) 代入 (22.22) 可得

$$y_t = \mathbf{x}' \boldsymbol{\beta} + \sum_{i=1}^p \rho_i (y_{t-i} - \mathbf{x}'_{t-i} \boldsymbol{\beta}) + \sum_{j=1}^q \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (22.24)$$

方程 (22.24) 被称为 “带 ARMA 的 GARCH” (GARCH with ARMA terms)。

## 5. 在条件方差方程中引入解释变量

例如, 为了考虑在 “9·11” 恐怖袭击后是否波动性增大了, 可以引入虚拟变量

$$D_t = \begin{cases} 1, & t \geq 2001/09/11 \\ 0, & t < 2001/09/11 \end{cases} \quad (22.25)$$

然后考虑以下 GARCH(1,1) 模型:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \gamma D_t \quad (22.26)$$

## 6. 使用非正态扰动项

如果被解释变量(比如, 某些金融变量)的分布函数存在厚尾, 则小概率事件比在正态分布情况下更容易发生。此时, 可以选择让扰动项服从  $t(k)$  分布而非正态分布来估计 ARCH 或 GARCH 模型。在进行 MLE 估计时, 将  $t$  分布的自由度  $k$  也作为待估参数。

## 22.7 ARCH 与 GARCH 的 Stata 命令及实例

有关 ARCH 与 GARCH 的 Stata 命令包括

```
arch y x1 x2, arch(1 / 3)          (ARCH(3))
arch y x1 x2, arch(1) garch(1)      (GARCH(1,1))
arch y x1 x2, ar(1) ma(1) arch(1) garch(1) (带 ARMA(1,1)的 GARCH(1,1))
arch y x1 x2, arch(1) dist(t)       (ARCH(1),扰动项服从 t 分布)
arch y x1 x2, arch(1) het(z1 z2)    (ARCH(1),将 z1,z2 加入条件方差方程)
arch y x1 x2, arch(1) garch(1) tarch(1) (GARCH(1,1)加上 TARCH(1))
arch y x1 x2, earch(1) egarch(1)     (EGARCH(1,1))
arch y x1 x2, arch(1 / 3) archm      (ARCH(3)加上 ARCH - M)
```

Stata 允许对 ARCH 或 GARCH 的模型设定进行更多的变化,详见“help arch”。Stata 手册提示,对 GARCH 模型进行 MLE 估计,通常须花费较长时间进行迭代计算,甚至会出现不收敛的情形。

下面以数据集 sp500.dta 为例,对 1981 年 1 月至 1991 年 4 月美国标准普尔股指(S&P 500)的日收益率进行 ARCH/GARCH 分析。该数据集包含以下变量: $r$ (股指日收益率), $t$ (日期, $1 \leq t \leq 2783$ )。

首先,看日收益率的时间趋势图,结果如图 22.3:

```
.use sp500.dta, clear
.line r t
```

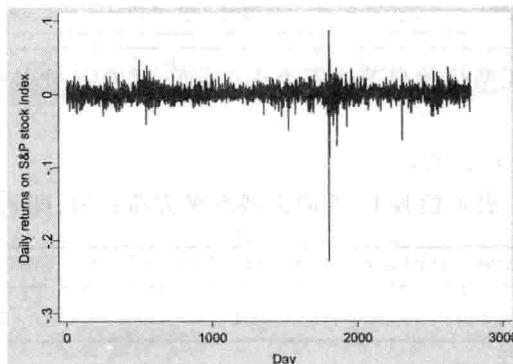


图 22.3 股指日收益率的时间趋势

从图 22.3 可以比较明显地看出,存在波动性集聚。作为对照,先考虑一个自回归模型。为此,用信息准则来确定自回归模型的阶数。将  $AR(p)$  视为 1 维  $VAR(p)$ ,则可使用 VAR 系列的命令:

```
.varsoc r,maxlag(8)
```

Selection-order criteria								
					Number of obs = 2775			
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	8612.3				.000118	-6.20634	-6.20557	-6.2042
1	8616.31	8.0317	1	0.005	.000118	-6.20851	-6.20697	-6.20424*
2	8619.54	6.443	1	0.011	.000118	-6.21012	-6.2078	-6.20371
3	8620.26	1.4422	1	0.230	.000118	-6.20991	-6.20683	-6.20137
4	8622.07	3.6252	1	0.057	.000118	-6.2105	-6.20664	-6.19982
5	8625.86	7.5913*	1	0.006	.000117*	-6.21251*	-6.20789*	-6.1997
6	8625.9	.0624	1	0.803	.000117	-6.21182	-6.20642	-6.19686
7	8626.13	.46369	1	0.496	.000117	-6.21126	-6.20509	-6.19417
8	8626.62	.9867	1	0.321	.000118	-6.2109	-6.20395	-6.19167

Endogenous: r  
Exogenous: cons

上表显示,大多数准则均选择 AR(5) 模型。因此,用 OLS 估计 AR(5) 模型<sup>①</sup>:

.reg r L(1/5).r

其中,“L(1/5)”表示 1~5 阶滞后。

Source	SS	df	MS	Number of obs = 2778		
Model	.00321088	5	.000642176	F( 5, 2772) = 5.49		
Residual	.324428204	2772	.000117038	Prob > F = 0.0000		
Total	.327639084	2777	.000117983	R-squared = 0.0098		
				Adj R-squared = 0.0080		
				Root MSE = .01082		

r	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
r					
L1.	.0562165	.0189661	2.96	0.003	.0190273 .0934056
L2.	-.047366	.0189723	-2.50	0.013	-.0845672 -.0101647
L3.	-.0191588	.0189899	-1.01	0.313	-.0563945 .0180769
L4.	-.0389725	.0189717	-2.05	0.040	-.0761726 -.0017723
L5.	.0524083	.018952	2.77	0.006	.0152468 .0895697
cons	.0004126	.000206	2.00	0.045	8.66e-06 .0008166

上表显示,5 阶滞后的系数依然显著地不为 0。下面,对 OLS 残差是否存在 ARCH 效应进行 LM 检验。

.estat archlm, lags(1/5)

其中,选择项“lags(1/5)”表示检验 1~5 阶的残差平方滞后项,即  $e_{t-1}^2, \dots, e_{t-5}^2$ 。

LM test for autoregressive conditional heteroskedasticity (ARCH)			
lags(p)	chi2	df	Prob > chi2
1	45.415	1	0.0000
2	72.001	2	0.0000
3	80.514	3	0.0000
4	80.693	4	0.0000
5	103.418	5	0.0000

H0: no ARCH effects vs. H1: ARCH(p) disturbance

<sup>①</sup> 也可使用命令“var r, lags(1/5)”来得到同样的结果。此处为了使用后估计命令“estat archlm”,故使用命令“reg r L(1/5).r”。

上表显示,对 ARCH(1)—ARCH(5)的检验结果均表明,存在显著的 ARCH 效应。

下面,通过画图更直观地考察 OLS 的残差平方是否存在自相关:

```
.predict e1,res  
(5 missing values generated)  
.g e2 = e1^2  
(5 missing values generated)  
.ac e2(画残差平方的自相关图,结果如图 22.4)  
.pac e2 (画残差平方的偏自相关图,结果如图 22.5)
```

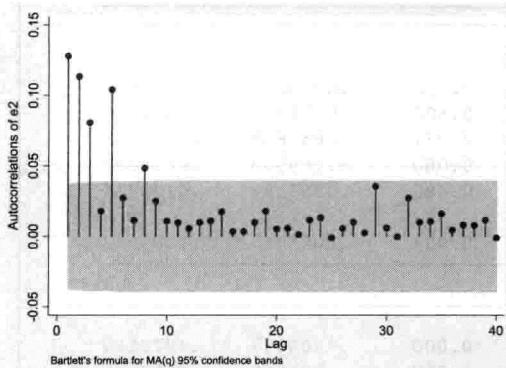


图 22.4 残差平方的自相关图

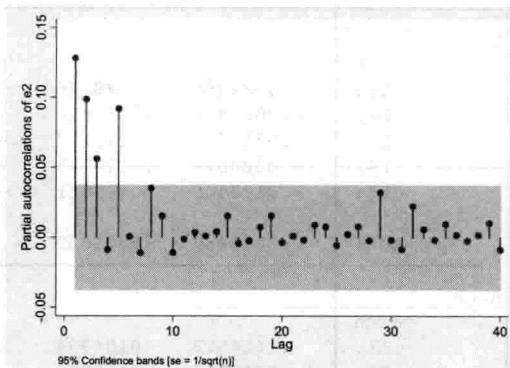


图 22.5 残差平方的偏自相关图

```
.corrgram e2, lags(10)
```

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.1279	0.1279	45.48	0.0000		
2	0.1134	0.0987	81.28	0.0000		
3	0.0804	0.0562	99.291	0.0000		
4	0.0179	-0.0088	100.18	0.0000		
5	0.1041	0.0919	130.35	0.0000		
6	0.0273	0.0007	132.42	0.0000		
7	0.0113	-0.0111	132.77	0.0000		
8	0.0486	0.0353	139.36	0.0000		
9	0.0250	0.0155	141.11	0.0000		
10	0.0107	-0.0107	141.43	0.0000		

从以上结果可以看出,无论是自相关图、偏自相关图,还是 Q 检验,均显示 OLS 残差之平方序列  $\{e_i^2\}$  存在自相关,故扰动项存在条件异方差,即波动性集聚。此结论与 LM 检验的结果相一致。为此,考察 ARCH( $p$ )模型。为了确定  $p$ ,估计序列  $\{e_i^2\}$  的自回归阶数:

```
.varsoc e2
```

Selection-order criteria							
Sample: 10 - 2783				Number of obs = 2774			
lag	LL	LR	df	p	FPE	AIC	HQIC
0	15210.3				1.0e-06	-10.9656	-10.9649
1	15233.2	45.742	1	0.000	1.0e-06	-10.9814	-10.9798
2	15246.8	27.155	1	0.000	9.9e-07	-10.9905	-10.9881
3	15251.2	8.7769*	1	0.003	9.8e-07*	-10.9929*	-10.9898*
4	15251.3	.21651	1	0.642	9.9e-07	-10.9923	-10.9884

Endogenous: e2  
Exogenous: \_cons

所有的准则均显示(上表中打星号者),应考虑 ARCH(3) 模型:

.arch r L(1/5).r, arch(1/3) nolog

ARCH family regression						
		OPG				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
r	r					
	L1.	.0522458	.0228176	2.29	0.022	.0075241 .0969676
	L2.	-.0094867	.0181812	-0.52	0.602	-.0451213 .0261478
	L3.	-.0311578	.0176172	-1.77	0.077	-.0656869 .0033713
	L4.	-.0320676	.0170775	-1.88	0.060	-.065539 .0014037
	L5.	-.0103408	.0149861	-0.69	0.490	-.0397131 .0190314
	_cons	.0005433	.0001832	2.97	0.003	.0001843 .0009023
ARCH	arch					
	L1.	.1606503	.0104934	15.31	0.000	.1400837 .1812169
	L2.	.0359253	.0150815	2.38	0.017	.006366 .0654845
	L3.	.118723	.0192852	6.16	0.000	.0809247 .1565214
	_cons	.0000687	1.67e-06	41.10	0.000	.0000654 .000072

上表显示,所有 ARCH 项均很显著。下面估计更为简洁的 GARCH(1,1) 模型:

.arch r L(1/5).r, arch(1) garch(1) nolog

ARCH family regression						
		OPG				
		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
r	r					
	L1.	.0666247	.0231129	2.88	0.004	.0213243 .1119252
	L2.	-.0153185	.0226133	-0.68	0.498	-.0596398 .0290027
	L3.	-.0119301	.0212921	-0.56	0.575	-.0536618 .0298016
	L4.	-.0434335	.0208996	-2.08	0.038	-.0843959 -.0024711
	L5.	-.0011428	.019924	-0.06	0.954	-.0401931 .0379075
	_cons	.0005794	.0001715	3.38	0.001	.0002433 .0009155
ARCH	arch					
	L1.	.0893397	.0033301	26.83	0.000	.0828128 .0958666
	garch					
	L1.	.8642674	.0084824	101.89	0.000	.8476422 .8808925
	_cons	4.94e-06	5.10e-07	9.69	0.000	3.94e-06 5.94e-06

上表显示, ARCH(1) 与 GARCH(1) 项均很显著。

考虑到股市中坏消息与好消息的效应可能不对称, 下面在 GARCH(1,1) 模型中加入一个 TARCH 项:

```
.arch r L(1/5).r, arch(1) garch(1) tarch(1) nolog
```

ARCH family regression						
	Number of obs = 2778					
	Wald chi2(5) = 12.30					
	Prob > chi2 = 0.0309					
r	OPG					
	Coef.	Std. Err.	z	P> z	[ 95% Conf. Interval]	
r						
L1.	.06502	.0226963	2.86	0.004	.0205361	.109504
L2.	-.006217	.0223478	-0.28	0.781	-.0500179	.0375839
L3.	-.0101253	.0206436	-0.49	0.624	-.0505861	.0303354
L4.	-.0373281	.0200826	-1.86	0.063	-.0766892	.0020329
L5.	.0001177	.0194354	0.01	0.995	-.0379749	.0382103
_cons	.0003849	.0001735	2.22	0.027	.0000448	.000725
ARCH						
arch						
L1.	.126726	.0049348	25.68	0.000	.117054	.136398
tarch						
L1.	-.0915795	.008728	-10.49	0.000	-.108686	-.074473
garch						
L1.	.8681879	.0094734	91.65	0.000	.8496205	.8867554
_cons	5.20e-06	4.86e-07	10.70	0.000	4.25e-06	6.15e-06

上表显示, TARCH 项很显著, 即存在不对称效应。而且, 不对称效应的规模 (-0.09) 几乎接近对称效应(0.13)。TARCH 项的负号表明, “好消息”对资产价格波动性的影响小于“坏消息”。

考虑到收益率中可能包含风险溢价, 下面估计“ARCH - in - Mean”模型。

```
.arch r L(1/5).r, arch(1) archm nolog
```

ARCH family regression						
	Number of obs = 2778					
	Wald chi2(6) = 452.49					
	Prob > chi2 = 0.0000					
r	OPG					
	Coef.	Std. Err.	z	P> z	[ 95% Conf. Interval]	
r						
L1.	.0387874	.0241306	1.61	0.108	-.0085077	.0860825
L2.	.0078756	.0158632	0.50	0.620	-.0232156	.0389668
L3.	-.0193001	.0122494	-1.58	0.115	-.0433085	.0047082
L4.	-.1074494	.0089398	-12.02	0.000	-.1249712	-.0899277
L5.	.0043577	.0148722	0.29	0.770	-.0247913	.0335068
_cons	.0006864	.0002508	2.74	0.006	.0001949	.001178
ARCHM						
sigma2	-1.18805	1.739309	-0.68	0.495	-4.597033	2.220933
ARCH						
arch						
L1.	.2360421	.0095681	24.67	0.000	.217289	.2547952
_cons	.0000794	1.56e-06	50.75	0.000	.0000763	.0000825

上表显示, ARCHM 项并不显著, 且符号为负(风险越高, 则收益率越低)。之所以出现这种反常现象, 可能因为这里考察的是股指(S&P 500), 而非个股。

下面考虑 EGARCH(1,1) 模型。但输入命令“`arch r L(1/5).r, earch(1) egarch(1) nolog`”却无法得到收敛的结果。为了演示目的, 转而使用以下命令:

```
. arch r L(1/3).r, earch(1) egarch(1) nolog
```

ARCH family regression						
	OPG					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
r						
r						
L1.	.059919	.0217624	2.75	0.006	.0172654	.1025726
L2.	.0068176	.0209431	0.33	0.745	-.0342302	.0478653
L3.	-.0165195	.0201316	-0.82	0.412	-.0559768	.0229378
_cons	.000252	.0001718	1.47	0.142	-.0000848	.0005889
ARCH						
earch						
L1.	-.0781144	.0070599	-11.06	0.000	-.0919515	-.0642773
earch_a						
L1.	.1558427	.0117046	13.31	0.000	.1329021	.1787834
egarch						
L1.	.9635746	.0042339	227.58	0.000	.9552763	.971873
_cons	-.3295685	.0393753	-8.37	0.000	-.4067427	-.2523942

上表显示, 非对称效应(earch)与对称效应(earch\_a)均十分显著。而且, 前者的规模约为后者的一半。非对称效应(earch)的符号为负, 表明“坏消息”的作用更大。

在以上的 ARCH, GARCH, TARCH, ARCHM, EGARCH 估计中, 均假设扰动项服从正态分布。但股指收益率可能存在厚尾。为此, 将日收益率的核密度图(参见第 27 章)与正态分布对比(如图 22.6):

```
. kdensity r,normopt(lpattern (" - "))
```

其中, 选择项“`normopt(lpattern (" - "))`”表示以虚线作为比较的正态密度。

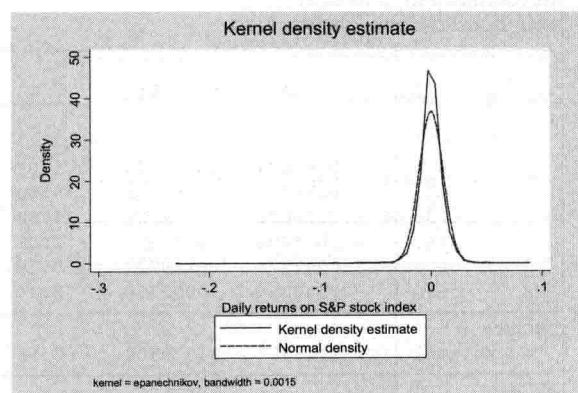


图 22.6 日收益率的核密度与正态密度

上图显示, 日收益率的核密度图很可能存在厚尾, 尤其在分布的左端。

下面对扰动项的正态性进行严格的统计检验:

```
. quietly var r, lags(1/5)
```

```
. varnorm
```

Jarque-Bera test					
Equation		chi2	df	Prob > chi2	
r		6.1e+05	2	0.00000	
ALL		6.1e+05	2	0.00000	

Skewness test					
Equation	Skewness	chi2	df	Prob > chi2	
r	-3.4412	5482.932	1	0.00000	
ALL		5482.932	1	0.00000	

Kurtosis test					
Equation	Kurtosis	chi2	df	Prob > chi2	
r	75.047	6.0e+05	1	0.00000	
ALL		6.0e+05	1	0.00000	

以上各检验均强烈拒绝“扰动项服从正态分布”的原假设。为此, 假设扰动项服从  $t$  分布, 重新用 GARCH(1,1) 进行估计:

```
. arch r L(1/5), r, arch(1) garch(1) dist(t) nolog
```

ARCH family regression						
	Number of obs = 2778					
	Wald chi2(5) = 12.82					
	Prob > chi2 = 0.0251					
<hr/>						
r	OPG					
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
r						
L1.	.0487162	.0188628	2.58	0.010	.0117457	.0856866
L2.	-.0194403	.0188372	-1.03	0.302	-.0563606	.01748
L3.	-.027825	.0184064	-1.51	0.131	-.0639009	.008251
L4.	-.0263707	.0183287	-1.44	0.150	-.0622943	.0095528
L5.	-.0080478	.0179622	-0.45	0.654	-.0432531	.0271576
_cons	.0005129	.0001579	3.25	0.001	.0002035	.0008224
ARCH						
arch						
L1.	.0354574	.0070274	5.05	0.000	.021684	.0492308
garch						
L1.	.9391238	.0107548	87.32	0.000	.9180448	.9602028
_cons	2.21e-06	5.67e-07	3.90	0.000	1.10e-06	3.32e-06
/lndfm2						
	1.35163	.1309665	10.32	0.000	1.094941	1.60832
df						
	5.86372	.506018			4.989006	6.994414

最后,对 GARCH(1,1)模型的条件方差进行预测:

```
. quietly arch r L(1/5).r, arch(1) garch(1) nolog  
. predict h, variance  
. line h t (画条件方差的时间趋势,结果如图 22.7 所示)
```

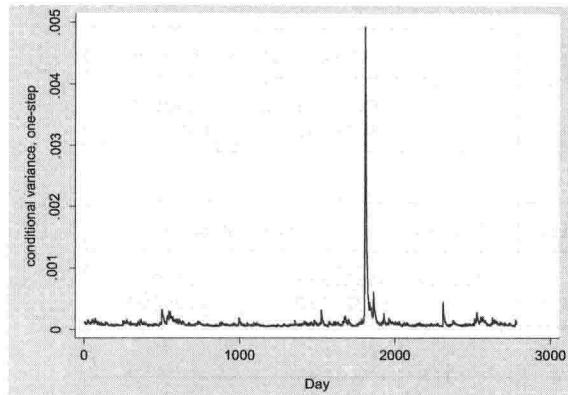


图 22.7 条件方差的时间趋势

图 22.7 显示,日收益率的条件方差时有波动,有时甚至急剧上升。如果使用 OLS 估计,则无法得到这些信息(OLS 将方差假定为常数,即一条水平线)。

## 22.8 多维 GARCH 模型(选读)

从概念上看,单变量的 GARCH 模型不难推广到多维 GARCH,即让当期扰动项  $\varepsilon_t$  的条件协方差矩阵  $\mathbf{H}_t$  依赖于上一期扰动项的“平方项” $\varepsilon_{t-1}\varepsilon'_{t-1}$  与上一期的条件协方差矩阵  $\mathbf{H}_{t-1}$ 。考虑最简单的双变量情形:

$$\mathbf{y}_t \equiv \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix} \begin{pmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \equiv \mathbf{C}\mathbf{x}_t + \boldsymbol{\varepsilon}_t \quad (22.27)$$

其中,  $\mathbf{C} \equiv \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{pmatrix}$  为待估系数矩阵,  $\mathbf{x}_t \equiv (x_{1t} \ x_{2t} \ x_{3t})'$  为解释变量(可以包含  $y_t$  的滞后项)。记扰动项  $\boldsymbol{\varepsilon}_t$  的条件协方差矩阵为  $\mathbf{H}_t$ :

$$\mathbf{H}_t \equiv \text{Var} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} = \begin{pmatrix} \sigma_{11,t}^2 & \sigma_{12,t}^2 \\ \sigma_{21,t}^2 & \sigma_{22,t}^2 \end{pmatrix} \quad (22.28)$$

由于  $\mathbf{H}_t$  为对称矩阵,故所有信息均包含于主对角线及以下的元素中。为此,将此对称矩阵的下三角部分取出,按顺序叠放成列向量:

$$\text{vech} \begin{pmatrix} \sigma_{11,t}^2 & \sigma_{12,t}^2 \\ \sigma_{21,t}^2 & \sigma_{22,t}^2 \end{pmatrix} \equiv \begin{pmatrix} \sigma_{11,t}^2 \\ \sigma_{21,t}^2 \\ \sigma_{22,t}^2 \end{pmatrix} \quad (22.29)$$

其中, vech 称为“半向量化算子”(half vectorization)。类似地, 可将“平方项”矩阵  $\boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1}$  半向量化:

$$\text{vech}(\boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1}) = \text{vech} \left[ \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1} \\ \boldsymbol{\varepsilon}_{2,t-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1} & \boldsymbol{\varepsilon}_{2,t-1} \end{pmatrix} \right] = \text{vech} \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1}^2 & \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} \\ \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} & \boldsymbol{\varepsilon}_{2,t-1}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1}^2 \\ \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} \\ \boldsymbol{\varepsilon}_{2,t-1}^2 \end{pmatrix} \quad (22.30)$$

记  $\text{vech}(\mathbf{H}_t) = \mathbf{h}_t$ , 借助 vech 算子, Bollerslev, Engle and Wooldridge (1988) 提出以下一般的多维 GARCH( $p, q$ ) 模型, 简记 MGARCH( $p, q$ ):

$$\mathbf{h}_t = \boldsymbol{\gamma} + \sum_{i=1}^p \mathbf{A}_i \text{vech}(\boldsymbol{\varepsilon}_{t-1} \boldsymbol{\varepsilon}'_{t-1}) + \sum_{j=1}^q \mathbf{B}_j \mathbf{h}_{t-j} \quad (22.31)$$

其中,  $\boldsymbol{\gamma}$  为待估常数向量,  $\mathbf{A}_i$  与  $\mathbf{B}_j$  分别为待估系数矩阵。方程(22.31)也称为“General Vech GARCH”模型。回到双变量的例子, 并令  $p=1, q=1$ , 则方程(22.31)可以写为:

$$\begin{pmatrix} \sigma_{11,t}^2 \\ \sigma_{21,t}^2 \\ \sigma_{22,t}^2 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1}^2 \\ \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} \\ \boldsymbol{\varepsilon}_{2,t-1}^2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} \sigma_{11,t-1}^2 \\ \sigma_{21,t-1}^2 \\ \sigma_{22,t-1}^2 \end{pmatrix} \quad (22.32)$$

根据方程(22.32)可知:

$$\sigma_{11,t}^2 = \gamma_1 + a_{11} \boldsymbol{\varepsilon}_{1,t-1}^2 + a_{12} \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} + a_{13} \boldsymbol{\varepsilon}_{2,t-1}^2 + b_{11} \sigma_{11,t-1}^2 + b_{12} \sigma_{21,t-1}^2 + b_{13} \sigma_{22,t-1}^2 \quad (22.33)$$

对于  $\sigma_{21,t}^2$  与  $\sigma_{22,t}^2$  也有类似的表达式。显然, 即使对于最简单的双变量 MGARCH(1,1) 模型(22.32), 也有很多待估参数。因此, MGARCH( $p, q$ ) 模型可能过于灵活, 不便拟合数据。为此, Bollerslev, Engle and Wooldridge (1988) 提出限制所有  $\mathbf{A}_i$  与  $\mathbf{B}_j$  矩阵为对角矩阵, 称为“对角半向量化 GARCH”模型(Diagonal Vech GARCH), 简记 DVECH。此时, 方程(22.32)简化为

$$\begin{pmatrix} \sigma_{11,t}^2 \\ \sigma_{21,t}^2 \\ \sigma_{22,t}^2 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} + \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix} \begin{pmatrix} \boldsymbol{\varepsilon}_{1,t-1}^2 \\ \boldsymbol{\varepsilon}_{1,t-1} \boldsymbol{\varepsilon}_{2,t-1} \\ \boldsymbol{\varepsilon}_{2,t-1}^2 \end{pmatrix} + \begin{pmatrix} b_{11} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & b_{33} \end{pmatrix} \begin{pmatrix} \sigma_{11,t-1}^2 \\ \sigma_{21,t-1}^2 \\ \sigma_{22,t-1}^2 \end{pmatrix} \quad (22.34)$$

方程(22.33)也相应地简化为

$$\sigma_{11,t}^2 = \gamma_1 + a_{11} \boldsymbol{\varepsilon}_{1,t-1}^2 + b_{11} \sigma_{11,t-1}^2 \quad (22.35)$$

估计 DVECH 模型的困难之处在于, 在任何时期  $t$ , 都需要保证条件协方差矩阵  $\mathbf{H}_t$  为正定矩阵; 且待估参数依然较多。为此, 学者提出了一类“条件相关模型”(conditional correlation model)<sup>①</sup>。由于其构建方法, 条件相关模型可自然地保证条件协方差矩阵  $\mathbf{H}_t$  为正定矩阵, 而且可以减少待估参数。 $\mathbf{H}_t$  的  $(i,j)$  元素, 即变量  $i$  与变量  $j$  在时期  $t$  的条件协方差  $\sigma_{ij,t}$  可写为:

$$\sigma_{ij,t} = \rho_{ij,t} \sqrt{\sigma_{ii,t} \sigma_{jj,t}} \quad (22.36)$$

其中,  $\rho_{ij,t}$  为变量  $i$  与变量  $j$  在时期  $t$  的条件相关系数, 而  $\sigma_{ii,t}, \sigma_{jj,t}$  分别为变量  $i$  与变量  $j$  在时期  $t$  的条件方差。条件相关模型让条件方差  $\sigma_{ii,t}$  ( $\forall i = 1, \dots, m$ ) (假设共有  $m$  个变量) 服从一维的 GARCH 过程<sup>②</sup>, 而集中刻画条件相关系数  $\rho_{ij,t}$  的行为。如果假设条件相关系数为常数, 即  $\rho_{ij,t} \equiv$

① 其他多维 GARCH 模型包括 BEKK 模型(Baba, Engle, Kraft and Kroner, 1990; Engle and Kroner, 1995), 但尚无法在 Stata 12 中实现。

② 可在此一维的条件方差方程中加入额外的解释变量。在 Stata 命令 mgarch 中, 可通过选择项 het(varlist) 来指定这些解释变量。

$\rho_{ij}$ , 不随时间而变, 则称为“常条件相关模型”(Constant Conditional Correlation, 简记 CCC) (Bollerslev, 1990)。但常条件相关的假设较强, 在实践中未必满足。于是, Engle (2002) 与 Tse and Tsui (2002) 分别提出了“动态条件相关模型”(Dynamic Conditional Correlation, 简记 DCC) 与“可变条件相关模型”(Varying Conditional Correlation, 简记 VCC)。

将方程(22.36)写为矩阵形式可得:

$$\mathbf{H}_t = \mathbf{D}_t^{1/2} \mathbf{R}_t \mathbf{D}_t^{1/2} \quad (22.37)$$

其中,  $\mathbf{D}_t$  为条件方差对角矩阵, 而  $\mathbf{R}_t$  为条件相关系数矩阵:

$$\mathbf{D}_t = \begin{pmatrix} \sigma_{11,t} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \sigma_{mm,t} \end{pmatrix}, \quad \mathbf{R}_t = \begin{pmatrix} 1 & \rho_{12,t} & \cdots & \rho_{1m,t} \\ \rho_{12,t} & 1 & \cdots & \rho_{2m,t} \\ \vdots & \vdots & & \vdots \\ \rho_{1m,t} & \rho_{2m,t} & \cdots & 1 \end{pmatrix} \quad (22.38)$$

显然, 如果  $\mathbf{R}_t$  为常数矩阵, 则为 CCC 模型。Engle (2002) 的 DCC 模型假定,  $\rho_{ij,t}$  由标准化扰动项的几何加权平均 (geometrically weighted average of standardized residuals) 来决定:

$$\rho_{ij,t} = \frac{\sum_{s=1}^{t-1} \lambda^s \tilde{\varepsilon}_{i,t-s} \tilde{\varepsilon}_{j,t-s}}{\sqrt{\left(\sum_{s=1}^{t-1} \lambda^s \tilde{\varepsilon}_{i,t-s}^2\right) \left(\sum_{s=1}^{t-1} \lambda^s \tilde{\varepsilon}_{j,t-s}^2\right)}} \quad (22.39)$$

其中,  $\tilde{\varepsilon}_{i,t-s}$  为标准化扰动项 (方差标准化为 1), 而  $\lambda^s$  为几何权重 (离时期  $t$  越远, 则权重依几何级数减少)。写为矩阵形式,  $\mathbf{R}_t$  的动态过程由以下两个方程来决定:

$$\mathbf{R}_t = \text{diag}(\mathbf{Q}_t)^{-1/2} \mathbf{Q}_t \text{diag}(\mathbf{Q}_t)^{-1/2}, \quad (22.40)$$

$$\mathbf{Q}_t = (1 - \lambda_1 - \lambda_2) \mathbf{R} + \lambda_1 \tilde{\varepsilon}_{t-1} \tilde{\varepsilon}'_{t-1} + \lambda_2 \mathbf{Q}_{t-1}$$

其中,  $\tilde{\varepsilon}_t$  为标准化的扰动项, 即  $\tilde{\varepsilon}_t = \mathbf{D}_t^{-1/2} \varepsilon_t$ ; 参数  $\lambda_1$  与  $\lambda_2$  皆为非负, 而且  $0 \leq \lambda_1 + \lambda_2 < 1$ 。显然,  $\lambda_1$  与  $\lambda_2$  决定了的动态过程。特别地, 如果  $\lambda_1 = \lambda_2 = 0$ , 则 DCC 模型简化为 CCC 模型。因此, 通过检验联合假设 “ $\lambda_1 = \lambda_2 = 0$ ”, 可在 DCC 与 CCC 模型之间进行选择。由于  $\mathbf{R}_t$  中的元素不一定介于 -1 与 1 之间, 故也称为“条件准相关系数矩阵”(matrix of conditional quasicorrelations)。

Tse and Tsui (2002) 的 VCC 模型则假设  $\mathbf{R}_t$  服从类似于 ARMA 的过程:

$$\mathbf{R}_t = (1 - \lambda_1 - \lambda_2) \mathbf{R} + \lambda_1 \Psi_{t-1} + \lambda_2 \mathbf{R}_{t-1} \quad (22.41)$$

其中,  $\mathbf{R}$  为  $\mathbf{R}_t$  的均值;  $\Psi_{t-1}$  为对标准化扰动项  $\tilde{\varepsilon}_t$  的相关系数矩阵的滚动估计量 (rolling estimator), 即在方程(22.39)中, 令  $\lambda = 1$  (简单平均, 而非加权平均); 参数  $\lambda_1$  与  $\lambda_2$  皆非负, 且  $0 \leq \lambda_1 + \lambda_2 < 1$ 。

对于多维 GARCH 模型, 通常假设扰动项  $\varepsilon_t$  服从多维正态分布, 然后进行 MLE 估计。如果扰动项不服从正态分布, 则为 QMLE 估计, 在大样本下依然是一致且渐近正态的。QMLE 与 MLE 的估计系数相同, 只是对协方差矩阵 (标准误) 的估计不同。也可以假设扰动项  $\varepsilon_t$  来自多维  $t$  分布, 然后进行 MLE 估计。如果扰动项  $\varepsilon_t$  确实服从多维  $t$  分布, 则基于多维  $t$  分布假设的 MLE 估计是一致且有效的; 而基于多维正态分布假设的 MLE 依然是一致的, 但并非最有效。反之, 如果扰动项  $\varepsilon_t$  并非来自多维  $t$  分布或多维正态分布, 则基于多维  $t$  分布假设的 MLE 估计是不一致的, 而基于多维正态分布假设的 QMLE 估计依然是一致的 (尽管并非最有效)。

在 Stata 12 中, 估计多维 GARCH 模型的命令句式为

```
mgarch model eq ... eq, arch(#) garch(#) constraints(#) het(varlist) dist(t) robust noconstant
```

其中,“model”可以是 dvech,ccc,dcc 或 vcc,分别表示 DVECH,CCC,DCC 与 VCC 模型。选择项“arch(#)”表示 ARCH 项,“garch(#)”表示 GARCH 项;选择项“constraints(#)”用于定义对参数的约束;选择项“het(varlist)”用来指定条件方差方程中的额外解释变量;选择项“dist(t)”假设扰动项服从多维 t 分布,而默认假设为多维正态分布;选择项“robust”表示使用稳健标准误(进行 QMLE 估计);选择项“noconstant”表示忽略常数项;而每个“eq”表示一个方程,其一般格式为 (depvars = indepvars,noconstant)。

下面以 Stata 提供的数据集 irates4.dta 为例。该数据集包含美国 6 月期国库券二级市场利率(tbill)与 AAA 级公司债券收益率(bond)。这两个变量的一阶差分,进行 VAR(1) 建模,并带 ARCH(1) 扰动项<sup>①</sup>。首先估计 DVECH 模型。

```
.use irates4.dta,clear
.mgarch dvech(D.bond D.tbill = LD.bond LD.tbill),arch(1) nolog
```

Diagonal vech MGARCH model						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Sample: 3 - 2456					Number of obs = 2454	
Distribution: Gaussian					Wald chi2(4) = 1183.52	
Log likelihood = 4221.658					Prob > chi2 = 0.0000	
D.bond						
bond						
LD.	.2967674	.0247149	12.01	0.000	.2483271	.3452077
tbill						
LD.	.0947949	.0098683	9.61	0.000	.0754533	.1141364
_cons	.0003991	.00143	0.28	0.780	-.0024036	.0032019
D.tbill						
bond						
LD.	.0108373	.0301501	0.36	0.719	-.0482558	.0699304
tbill						
LD.	.4344747	.0176497	24.62	0.000	.3998819	.4690675
_cons	.0011611	.0021033	0.55	0.581	-.0029612	.0052835
Sigma0						
1_1	.004894	.0002006	24.40	0.000	.0045008	.0052871
2_1	.0040986	.0002396	17.10	0.000	.0036289	.0045683
2_2	.0115149	.0005227	22.03	0.000	.0104904	.0125395
L.ARCH						
1_1	.4514942	.0456835	9.88	0.000	.3619562	.5410323
2_1	.2518879	.036736	6.86	0.000	.1798866	.3238893
2_2	.843368	.0608055	13.87	0.000	.7241914	.9625446

上表下部的 Sigma0 表示方程(22.31)中的常数向量  $\gamma$ ,而 L.ARCH 表示 ARCH(1) 项的估计结果,1\_1,2\_1 与 2\_2 分别表示  $\varepsilon_{1,t-1}^2, \varepsilon_{1,t-1}\varepsilon_{2,t-1}$  与  $\varepsilon_{2,t-1}^2$  的相应系数,均显著为正。上表显示,在以 D.tbill 为被解释变量的方程中,解释变量 LD.bond(即 bond 的差分滞后项)并不显著。而且,两个方程的常数项均不显著。下面,为了去掉这些不显著项,在命令中将两个方程分开写。

<sup>①</sup> 在此例中,如果加上 GARCH(1) 项则很难收敛,这是估计多维 GARCH 模型经常碰到的问题。

```
.mgarch dvech (D.bond = LD.bond LD.tbill, noconstant) (D.tbill = LD.tbill,noconstant), arch(1) nolog
```

Diagonal vech MGARCH model						
		Coef.		Std. Err.	z	P> z  [95% Conf. Interval]
D.bond	bond	.2941649	.0234734	12.53	0.000	.2481579 .3401718
	LD.					
	tbill	.0953158	.0098077	9.72	0.000	.076093 .1145386
	LD.					
D.tbill	tbill	.4385945	.0136672	32.09	0.000	.4118072 .4653817
	LD.					
Sigma0						
	1_1	.0048922	.0002005	24.40	0.000	.0044993 .0052851
	2_1	.0040949	.0002394	17.10	0.000	.0036256 .0045641
	2_2	.0115043	.0005184	22.19	0.000	.0104883 .0125203
L.ARCH						
	1_1	.4519233	.045671	9.90	0.000	.3624099 .5414368
	2_1	.2515474	.0366701	6.86	0.000	.1796752 .3234195
	2_2	.8437212	.0600839	14.04	0.000	.7259589 .9614836

上表显示,所有解释变量的系数都很显著了。下面,使用 CCC 模型来估计带 GARCH(1,1) 扰动项的原模型。

```
.mgarch ccc (D.bond D.tbill = LD.bond LD.tbill), arch(1) nolog
```

Constant conditional correlation MGARCH model						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Sample: 1 - 2456					Number of obs = 2454	
Distribution: Gaussian					Wald chi2(4) = 1516.13	
Log likelihood = 4257.064					Prob > chi2 = 0.0000	
D.bond						
bond						
LD.	.2683797	.0227155	11.81	0.000	.2238581	.3129012
tbill						
LD.	.0822348	.0090725	9.06	0.000	.064453	.1000167
_cons	.0005556	.0014135	0.39	0.694	-.0022147	.003326
ARCH_D.bond						
arch						
L1.	.5233444	.049989	10.47	0.000	.4253678	.621321
_cons	.0046334	.000191	24.26	0.000	.004259	.0050077
D.tbill						
bond						
LD.	-.0118455	.0272737	-0.43	0.664	-.065301	.04161
tbill						
LD.	.4393633	.0155774	28.21	0.000	.408832	.4698945
_cons	.0009441	.0020204	0.47	0.640	-.0030159	.0049041
ARCH_D.tbill						
arch						
L1.	.951192	.0663333	14.34	0.000	.8211811	1.081203
_cons	.0106312	.0004776	22.26	0.000	.0096952	.0115673
Correlation						
D.bond						
D.tbill	.4575932	.0161694	28.30	0.000	.4259017	.4892847

上表显示, ARCH 项显著为正。上表最后一行显示, 国库券收益率差分与债券收益率差分的相关系数为 0.46, 且在 1% 水平上显著。这意味着, 国库券收益率上涨(下跌)时, 企业债券收益率也倾向于上涨(下跌)。但常相关模型假设二者的相关系数为常数, 未必现实, 故下面转向 VCC 模型。

```
.mgarch dcc (D.bond D.tbill = LD.bond LD.tbill), arch(1) nolog
```

Dynamic conditional correlation MGARCH model						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Sample: 1 - 2456					Number of obs = 2454	
Distribution: Gaussian					Wald chi2(4) = 1563.32	
Log likelihood = 4267.592					Prob > chi2 = 0.0000	
D.bond						
bond						
LD.	.2663008	.0226708	11.75	0.000	.2218668	.3107348
tbill						
LD.	.082495	.0090404	9.13	0.000	.0647762	.1002139
_cons	.0003347	.0014168	0.24	0.813	-.0024422	.0031116
ARCH_D.bond						
arch						
L1.	.5115416	.0489595	10.45	0.000	.4155827	.6075005
_cons	.0045801	.0001892	24.20	0.000	.0042092	.0049511
D.tbill						
bond						
LD.	-.0107307	.0269353	-0.40	0.690	-.0635229	.0420614
tbill						
LD.	.4385706	.0153425	28.59	0.000	.4084999	.4686413
_cons	.0009644	.0020006	0.48	0.630	-.0029568	.0048856
ARCH_D.tbill						
arch						
L1.	.9250511	.0640242	14.45	0.000	.7995659	1.050536
_cons	.0103025	.0004579	22.50	0.000	.009405	.0112
Correlation						
D.bond						
D.tbill	.4057807	.0237458	17.09	0.000	.3592397	.4523217
Adjustment						
lambda1	.0040023	.0014576	2.75	0.006	.0011456	.0068591
lambda2	.9795958	.0052797	185.54	0.000	.9692478	.9899438

上表显示,国库券收益率差分与企业债券收益率差分的条件准相关系数为 0.41,且在 1% 水平上显著。上表最后两行显示, $\lambda_1$  与  $\lambda_2$  均显著为正,故应使用 DCC 而非 CCC 模型。下面,估计 VCC 模型。

```
.mgarch vcc (D.bond D.tbill = LD.bond LD.tbill), arch(1) nolog
```

Varying conditional correlation MGARCH model						
	Coef.	Std. Err.	z	P> z	[ 95% Conf. Interval]	
Sample: 1 - 2456					Number of obs = 2454	
Distribution: Gaussian					Wald chi2(4) = 1629.49	
Log likelihood = 4273.118					Prob > chi2 = 0.0000	
D.bond						
bond	.2633396	.0227876	11.56	0.000	.2186767	.3080024
LD.						
tbill	.0841621	.0095937	8.77	0.000	.0653589	.1029654
LD.						
_cons	.00055	.001428	0.39	0.700	-.0022488	.0033488
ARCH_D.bond						
arch	.5212864	.0501374	10.40	0.000	.4230189	.619554
L1.						
_cons	.0047296	.0001995	23.71	0.000	.0043386	.0051206
D.tbill						
bond	-.0306537	.0233759	-1.31	0.190	-.0764696	.0151622
LD.						
tbill	.447364	.0145085	30.83	0.000	.4189278	.4758002
LD.						
_cons	.0012041	.0019613	0.61	0.539	-.0026401	.0050482
ARCH_D.tbill						
arch	.9315652	.0637373	14.62	0.000	.8066423	1.056488
L1.						
_cons	.010033	.0004448	22.55	0.000	.0091611	.0109049
Correlation						
D.bond						
D.tbill	.5708136	.1365947	4.18	0.000	.303093	.8385342
Adjustment						
lambda1	.0030634	.0007849	3.90	0.000	.001525	.0046019
lambda2	.9959393	.001035	962.31	0.000	.9939108	.9979678

上表显示,国库券收益率差分与企业债券收益率差分的条件准相关系数为 0.57,且在 1% 水平上显著。上表最后两行显示,  $\lambda_1$  与  $\lambda_2$  均显著为正,故应使用 VCC 而非 CCC 模型。

## 习 题

**22.1** 考虑  $ARCH(p)$  模型,  $\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2$ 。为了保证  $\sigma_t^2$  为正,以及  $\{\varepsilon_t\}$  为平稳过程,需要对参数  $\{\alpha_0, \alpha_1, \dots, \alpha_p\}$  进行哪些限制?

**22.2** 数据集 dow1.dta 包含了 1953—1990 年美国道琼斯股指的收盘价。计算道琼斯股指的日收益率,画其时间趋势图,并参照本章实例进行  $ARCH/GARCH$  估计。

# 第 23 章 似不相关回归

## 23.1 单一方程估计与系统估计

到目前为止,我们仅考虑对单一方程的估计,但有时候会出现多个方程的情形。如果多个方程之间有某种联系,那么将这些方程同时进行联合估计有可能提高估计的效率,这被称为“系统估计”(system estimation)。

有时多个方程是从同一个最大化问题推导而来(比如,从企业的利润最大化问题导出对资本与劳动力的需求),故在理论上存在“跨方程的参数约束”(cross-equation restrictions)。多方程联合估计为检验这些跨方程约束提供了可能。也可以在加上这些约束条件后再进行系统估计。

对多方程系统进行联合估计的缺点是,如果某个方程的误差较大,则系统估计会将这一方程的误差带入其他方程中,进而污染(contaminate)整个方程系统。在某种意义上,选择单一方程估计或系统估计,也是在“有效性”与“稳健性”之间的权衡。

多方程系统主要分为两类。一类为“联立方程组”(simultaneous equations),即不同方程之间的变量存在内在的联系,一个方程的解释变量是另一方程的被解释变量。另一类为“似不相关回归”(Seemingly Unrelated Regression Estimation,简记 SUR 或 SURE),即各方程的变量之间没有内在联系,但各方程的扰动项之间存在相关性。本章介绍似不相关回归,下一章介绍联立方程模型。

**例(似不相关回归)** 以研一学生的计量成绩与英语成绩作为两个被解释变量。这两个方程所包含的解释变量可以不同,比如,第一个方程可以包括虚拟变量“是否学过本科计量学”,而第二个方程可以包括“考研英语成绩”。这两个方程的变量之间貌似没有联系,但由于同一学生的不可观测因素同时对计量成绩与英语成绩造成影响,故两个方程的扰动项应该是相关的。如果将这两个方程同时进行联合估计,则可以提高估计效率。

## 23.2 似不相关回归的假定

似不相关回归模型的设定如下。假设共有  $n$  个方程( $n$  个被解释变量),每个方程共有  $T$  个观测值<sup>①</sup>, $T > n$ 。在第  $i$  个方程中,共有  $K_i$  个解释变量。第  $i$  个方程可以写为

$$\underset{T \times 1}{y_i} = \underset{T \times K_i}{X_i} \underset{K_i \times 1}{\beta_i} + \underset{T \times 1}{\varepsilon_i} \quad (i=1,2,\dots,n) \quad (23.1)$$

将所有的方程叠放在一起可得

① 此处的  $T$  不一定指的是时间。但为了叙述方便,我们认为它是时间。

$$\mathbf{y} = \underbrace{\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}}_{nT \times 1} = \underbrace{\begin{pmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X_n \end{pmatrix}}_{nT \times \sum_{i=1}^n K_i} \underbrace{\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_n \end{pmatrix}}_{\sum_{i=1}^n K_i \times 1} + \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}}_{nT \times 1} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (23.2)$$

可以想象,在计算机中,也把数据排成这个样子(需要加上许多零)。考察“大”扰动项  $\boldsymbol{\varepsilon}$  之协方差矩阵

$$\boldsymbol{\Omega} \equiv \text{Var} \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} = E \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_n \end{pmatrix} (\boldsymbol{\varepsilon}_1' \boldsymbol{\varepsilon}_2' \cdots \boldsymbol{\varepsilon}_n') = E \begin{pmatrix} \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_n' \\ \boldsymbol{\varepsilon}_2 \boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_2 \boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_2 \boldsymbol{\varepsilon}_n' \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_1' & \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_2' & \cdots & \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n' \end{pmatrix}_{nT \times nT} \quad (23.3)$$

假设同一方程不同期的扰动项不存在自相关,且方差也相同,记第  $i$  个方程的方差为  $\sigma_{ii}$ 。则协方差阵  $\boldsymbol{\Omega}$  中主对角线上的第  $(i,i)$  个矩阵为

$$E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i') = \sigma_{ii} \mathbf{I}_T \quad (23.4)$$

假设不同方程的扰动项之间存在同期相关<sup>①</sup>,即

$$E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j) = \begin{cases} \sigma_{ij}, & t=s \\ 0, & t \neq s \end{cases} \quad (23.5)$$

则协方差阵  $\boldsymbol{\Omega}$  中的第  $(i,j)$  个矩阵( $i \neq j$ )为

$$E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j) = \sigma_{ij} \mathbf{I}_T \quad (23.6)$$

综合以上结果可知

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_{11} \mathbf{I}_T & \sigma_{12} \mathbf{I}_T & \cdots & \sigma_{1n} \mathbf{I}_T \\ \sigma_{21} \mathbf{I}_T & \sigma_{22} \mathbf{I}_T & \cdots & \sigma_{2n} \mathbf{I}_T \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} \mathbf{I}_T & \sigma_{n2} \mathbf{I}_T & \cdots & \sigma_{nn} \mathbf{I}_T \end{pmatrix} \quad (23.7)$$

由于  $\boldsymbol{\Omega}$  中的每个小块矩阵都有共同的因子  $\mathbf{I}_T$ ,我们自然想把  $\mathbf{I}_T$  从右边提取出来。这可以通过矩阵的“克罗内克尔乘积”(Kronecker product)来实现。

**定义** 对于任意两个矩阵  $\mathbf{A}_{m \times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$  与  $\mathbf{B}_{p \times q}$ (矩阵  $\mathbf{A}, \mathbf{B}$  的维度可以完全不同),

克罗内克尔乘积为  $\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11} \mathbf{B} & \cdots & a_{1n} \mathbf{B} \\ \vdots & & \vdots \\ a_{m1} \mathbf{B} & \cdots & a_{mn} \mathbf{B} \end{pmatrix}_{mp \times nq}$

容易看出,对于任意矩阵  $\mathbf{A}, \mathbf{B}$ ,其克罗内克尔乘积  $\mathbf{A} \otimes \mathbf{B}$  总是有定义的。可以证明(参见习题),克罗内克尔乘积具有以下性质:

$$(1) (\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD});$$

$$(2) (\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}';$$

① 如果是横截面数据,则指的是同一个人在不同回归方程中的对应扰动项之间存在相关性。

$$(3) (\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$$

使用克罗内克尔乘积,可以将扰动项  $\boldsymbol{\varepsilon}$  的协方差矩阵简化为

$$\boldsymbol{\Omega} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \otimes \mathbf{I}_T \equiv \boldsymbol{\Sigma} \otimes \mathbf{I}_T \quad (23.8)$$

其中,  $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix}$  为同期协方差矩阵。根据克罗内克尔乘积的性质,  $\boldsymbol{\Omega}$  的逆矩阵可

以写为

$$\boldsymbol{\Omega}^{-1} = \boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T \quad (23.9)$$

### 23.3 SUR 的 FGLS 估计

由于  $\boldsymbol{\Omega}$  不是单位矩阵,故用 OLS 估计这个多方程系统  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  不是最有效率的。假设  $\boldsymbol{\Omega}$  已知,则 GLS 是最有效率的估计方法:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = (\mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\Omega}^{-1} \mathbf{y} = [\mathbf{X}' (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T) \mathbf{X}]^{-1} \mathbf{X}' (\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T) \mathbf{y} \quad (23.10)$$

一般来说,此 GLS 估计量与单一方程 OLS 估计量不同。但如果出现以下两种情形之一,则 GLS 与单一方程 OLS 的结果完全相同,使用 GLS 并不会提高效率。

(1) 各方程的扰动项互不相关。在似不相关回归模型中,各方程间唯一的联系就是扰动项之间的相关性。如果扰动项互不相关,则  $\boldsymbol{\Omega}$  是单位矩阵,那么系统估计与单一方程估计并无区别(参见习题)。

(2) 每个方程包含的解释变量完全相同(证明参见附录)。比如,向量自回归模型(VAR)中的每个方程包含完全相同的解释变量,故使用单一方程 OLS 估计 VAR 就够了。

除了以上两种特殊情形外,通常来说,各方程的扰动项之间的相关性越大,则 GLS 所能带来的效率改进就越大;各方程的数据矩阵  $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  之间的相关性越小,则 GLS 所能带来的效率改进也越大(如果数据矩阵完全相同,则 GLS 还原为单一方程 OLS)。

然而,在现实中  $\boldsymbol{\Omega}$  一般是未知的,故首先需要估计  $\hat{\boldsymbol{\Omega}}$ ,然后进行 FGLS 估计。由于对每个方程分别进行 OLS 回归也是一致的,故可以使用单一方程 OLS 的残差来一致地估计  $\sigma_{ij}$ 。假设第  $i$  个方程的 OLS 残差向量为  $\mathbf{e}_i$ ,则  $\sigma_{ij}$  的一致估计量为

$$\hat{\sigma}_{ij} = \frac{1}{T} \mathbf{e}_i' \mathbf{e}_j = \frac{1}{T} \sum_{t=1}^T e_{it} e_{jt} \quad (23.11)$$

因此,  $\hat{\boldsymbol{\Omega}} = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1n} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots & \hat{\sigma}_{2n} \\ \vdots & \vdots & & \vdots \\ \hat{\sigma}_{n1} & \hat{\sigma}_{n2} & \cdots & \hat{\sigma}_{nn} \end{pmatrix} \otimes \mathbf{I}_T$ 。将  $\hat{\boldsymbol{\Omega}}$  代入方程(23.10)可得

$$\hat{\beta}_{\text{SUR}} = (\mathbf{X}' \hat{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Omega}^{-1} \mathbf{y} \quad (23.12)$$

这就是“似不相关估计量”(Zellner, 1962), 记为  $\hat{\beta}_{\text{SUR}}$ 。使用 FGLS 后得到新的残差, 可以再一次计算  $\hat{\Omega}$ , 不断迭代直至系数估计值  $\hat{\beta}_{\text{SUR}}$  收敛为止。

## 23.4 SUR 的假设检验

在对多方程系统进行 SUR 估计后, 对线性假设 “ $H_0: \mathbf{R}\beta = \mathbf{r}$ ” 的检验可以照常进行。由于  $\beta$  包含了所有方程的参数, 故可以检验跨方程的参数约束。如果接受 “ $H_0: \mathbf{R}\beta = \mathbf{r}$ ”, 则可把 “ $\mathbf{R}\beta = \mathbf{r}$ ” 作为约束条件, 进行有约束的 FGLS 估计。

需要指出的是, 即使各方程的解释变量完全相同<sup>①</sup>, 有时也使用 SUR 而不使用单一方程 OLS, 以便检验跨方程的参数约束。如果存在跨方程的参数约束, 则即使各方程的解释变量完全相同, SUR 估计与单一方程 OLS 也不再等价。

SUR 模型的基本假设是, 各方程的扰动项之间存在同期相关。为此, 需要检验原假设 “ $H_0$ : 各方程的扰动项无同期相关”, 即 “ $H_0$ :  $\Sigma$  为对角矩阵”。Breusch and Pagan(1980)建议使用以下 LM 统计量:

$$\lambda_{\text{LM}} = T \sum_{i=2}^n \sum_{j=1}^{i-1} r_{ij}^2 \xrightarrow{d} \chi^2(n(n-1)/2) \quad (23.13)$$

其中,  $r_{ij} = \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}\hat{\sigma}_{jj}}}$  为根据残差计算的扰动项  $\epsilon_i$  与  $\epsilon_j$  之间的同期相关系数, 而  $\sum_{i=2}^n \sum_{j=1}^{i-1} r_{ij}^2$  为

同期相关系数矩阵  $\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$  主对角线以下各项之平方和(该矩阵为对称矩阵)。

## 23.5 似不相关回归的 Stata 命令及实例

似不相关回归的 Stata 命令基本格式为

```
sureg (depvar1 varlist1)(depvar2 varlist2)...(depvarn  
varlistn),isure corr
```

其中, 选择项 “isure” 表示迭代至收敛(默认值为不迭代), 选择项 “corr” 表示汇报对无同期相关的检验。

如果要检验联合假设 “第 1 个方程中变量  $x$  的系数与第 2 个方程中变量  $z$ , 以及第 3 个方程中变量  $w$  的系数相等”, 可使用如下命令:

<sup>①</sup> 这种情况在经济学中常常出现。比如, 对一组商品的需求取决于这组商品的价格及收入; 对一组资产的需求份额(portfolio share of assets)取决于这组资产的回报率及总财富。当需求份额必须加总为 1 时, 可以选择将其中一个方程去掉, 否则扰动项的协方差矩阵将为不可逆的退化矩阵。

```
test ([depvar1]x = [depvar2]z) ([depvar1]x = [depvar3]w)
```

如果要在满足以上两个约束条件下进行 SUR 估计,则可使用以下命令:

```
constraint 1 [depvar1]x = [depvar2]z (定义第 1 个约束条件)
```

```
constraint 2 [depvar1]x = [depvar2]w (定义第 2 个约束条件)
```

```
sureg (depvar1 varlist1) (depvar2 varlist2) ... (depvarn varlistn),c(1 2)
```

下面以来自美国 High School & Beyond Study 的数据集 hsb2.dta 为例。该数据集包含 200 名高中生的以下变量,read(阅读成绩),math(数学成绩),science(自然科学成绩),socst(社会科学成绩),female(是否女性),schtyp(是否私立学校),以及 ses(社会经济地位,social economic status,low = 1,middle = 2,high = 3)。考虑以下两个回归方程:

$$\text{read}_i = \alpha_0 + \alpha_1 \text{female}_i + \alpha_2 \text{ses}_i + \alpha_3 \text{schtyp}_i + \alpha_4 \text{socst}_i + u_i \quad (23.14)$$

$$\text{math}_i = \beta_0 + \beta_1 \text{female}_i + \beta_2 \text{ses}_i + \beta_3 \text{schtyp}_i + \beta_4 \text{science}_i + v_i \quad (23.15)$$

作为参照系,首先对以上两个方程进行单一方程 OLS 估计:

```
. use hsb2.dta,clear  
. reg3 (read female ses schtyp socst) (math female ses schtyp science),ols  
. estimates store OLS
```

Multivariate regression						
Equation	Obs	Parms	RMSE	"R-sq"	F-Stat	P
read	200	4	8.020526	0.4004	32.55	0.0000
math	200	4	7.250452	0.4131	34.31	0.0000
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
read	female	-1.521471	1.154157	-1.32	0.188	-3.790619 .747678
	ses	1.186993	.8473698	1.40	0.162	-.4789918 2.852977
	schtyp	.5071815	1.56462	0.32	0.746	-2.568964 3.583327
	socst	.5689789	.0565024	10.07	0.000	.4578916 .6800662
	_cons	20.21426	3.316836	6.09	0.000	13.69314 26.73538
math	female	1.131112	1.043281	1.08	0.279	-.9200471 3.182272
	ses	1.324103	.7491005	1.77	0.078	-.1486777 2.796883
	schtyp	1.126373	1.413143	0.80	0.426	-1.651959 3.904706
	science	.573999	.0544043	10.55	0.000	.4670367 .6809614
	_cons	18.23907	3.28543	5.55	0.000	11.7797 24.69844

其中,命令 reg3 的默认估计法为三阶段最小二乘法(3SLS)<sup>①</sup>,但加上选择项“ols”后则进行单一方程 OLS 估计。由于被解释变量 read 与 math 为同一学生的阅读与数学成绩,故这两个方程的扰动项在理论上很可能存在相关性。为此,进行 SUR 估计:

```
. sureg (read female ses schtyp socst) (math female ses schtyp
```

① 参见第 24 章。

science),corr<sup>①</sup>

其中,选择项“corr”表示,汇报对各方程扰动项之间“无同期相关”的检验结果。

. estimates store SUR

Seemingly unrelated regression						
Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
read	200	4	7.938329	0.3975	117.51	0.0000
math	200	4	7.189084	0.4082	117.78	0.0000
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
read						
female	-1.412142	1.139458	-1.24	0.215	-3.64544	.8211552
ses	1.457827	.8351991	1.75	0.081	-.1791329	3.094788
schtyp	.5854049	1.544871	0.38	0.705	-2.442486	3.613295
socst	.514735	.0548756	9.38	0.000	.4071808	.6222891
_cons	22.35003	3.251032	6.87	0.000	15.97812	28.72193
math						
female	1.000534	1.029996	0.97	0.331	-1.018222	3.01929
ses	1.577474	.7388309	2.14	0.033	.1293925	3.025556
schtyp	1.181735	1.395346	0.85	0.397	-1.553093	3.916562
science	.5045673	.0528379	9.55	0.000	.4010069	.6081277
_cons	21.32537	3.215353	6.63	0.000	15.02339	27.62735
Correlation matrix of residuals:						
read	math					
read	1.0000					
math	0.1985	1.0000				
Breusch-Pagan test of independence: chi2(1) = 7.880, Pr = 0.0050						

上表最后一行显示,各方程扰动项之间“无同期相关”的检验  $p$  值为 0.005,故可以在 1% 的显著性水平上拒绝各方程的扰动项相互独立的原假设。因此,使用 SUR 进行系统估计可以提高估计的效率。下面进行迭代式 SUR 估计:

. sureg (read female ses schtyp socst)(math female ses schtyp science),  
i nolog  
. estimates store SUR\_i

Seemingly unrelated regression, iterated						
Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
read	200	4	7.965347	0.3934	109.40	0.0000
math	200	4	7.229906	0.4014	105.94	0.0000
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
read						
female	-1.350362	1.145814	-1.18	0.239	-3.596118	.8953927
ses	1.610872	.8381693	1.92	0.055	-.03191	3.253653
schtyp	.6296076	1.553704	0.41	0.685	-2.415596	3.674812
socst	.4840826	.0540519	8.96	0.000	.3781427	.5900224
_cons	23.55691	3.24013	7.27	0.000	17.20638	29.90745
math						
female	.9298813	1.039964	0.89	0.371	-1.108411	2.968173
ses	1.714568	.7450829	2.30	0.021	.2542327	3.174904
schtyp	1.21169	1.409091	0.86	0.390	-1.550077	3.973457
science	.4669993	.0522569	8.94	0.000	.3645777	.5694209
_cons	22.9953	3.211463	7.16	0.000	16.70095	29.28965

① 此命令也可用“reg3 (read female ses schtyp socst)(math female ses schtyp science),sur”来实现,二者的输出结果相同。

为了便于比较,将以上各种方法的系数估计值及标准误列表:

```
. esttab OLS SUR SUR_i, se mttitle r2 star(* 0.1 ** 0.05 *** 0.01)
```

	(1) OLS	(2) SUR	(3) SUR_i
read			
female	-1.521 (1.154)	-1.412 (1.139)	-1.350 (1.146)
ses	1.187 (0.847)	1.458* (0.835)	1.611* (0.838)
schtyp	0.507 (1.565)	0.585 (1.545)	0.630 (1.554)
socst	0.569*** (0.0565)	0.515*** (0.0549)	0.484*** (0.0541)
_cons	20.21*** (3.317)	22.35*** (3.251)	23.56*** (3.240)
math			
female	1.131 (1.043)	1.001 (1.030)	0.930 (1.040)
ses	1.324* (0.749)	1.577** (0.739)	1.715** (0.745)
schtyp	1.126 (1.413)	1.182 (1.395)	1.212 (1.409)
science	0.574*** (0.0544)	0.505*** (0.0528)	0.467*** (0.0523)
_cons	18.24*** (3.285)	21.33*** (3.215)	23.00*** (3.211)
N	200	200	200
R-sq	0.400	0.398	0.393
Standard errors in parentheses			
* p<0.1, ** p<0.05, *** p<0.01			

从上表可知,SUR 与迭代 SUR 的估计结果比较接近,但与 OLS 有明显差别。这是因为,SUR 比单一方程 OLS 更有效率。

由于在两个方程中,变量 ses 的系数估计值比较接近,下面检验这两个系数是否相等:

```
. quietly sureg (read female ses schtyp socst) (math female ses schtyp science)
. test ([read]ses = [math]ses)
```

```
(1) [read]ses - [math]ses = 0
      chi2( 1) =     0.01
      Prob > chi2 =  0.9053
```

由于  $p$  值为 0.91,故可以接受“这两个系数相等”的原假设。下面,我们加上这个跨方程约束条件,再进行 SUR 估计:

```
. constraint 1 [read]ses = [math]ses
. sureg (read female ses schtyp socst) (math female ses schtyp science),
c(1)
```

Seemingly unrelated regression						
Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
read	200	4	7.939578	0.3973	119.44	0.0000
math	200	4	7.188205	0.4083	119.06	0.0000
( 1 ) [read]ses - [math]ses = 0						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
read						
female	-1.397722	1.132922	-1.23	0.217	-3.618209	.8227652
ses	1.526381	.602663	2.53	0.011	.3451827	2.707578
schtyp	.571229	1.540347	0.37	0.711	-2.447795	3.590253
socst	.5131128	.0530443	9.67	0.000	.4091479	.6170776
_cons	22.30275	3.22883	6.91	0.000	15.97436	28.63114
math						
female	.9939827	1.028492	0.97	0.334	-1.021824	3.009789
ses	1.526381	.602663	2.53	0.011	.3451827	2.707578
schtyp	1.193675	1.391732	0.86	0.391	-1.53407	3.921419
science	.5056923	.0520817	9.71	0.000	.4036141	.6077706
_cons	21.36176	3.198553	6.68	0.000	15.09271	27.63081

## 23.6 变系数面板数据的 SUR 估计

在第 16 章曾讨论过变系数面板数据模型：

$$y_{it} = \mathbf{x}'_i \boldsymbol{\beta}_i + \varepsilon_{it} \quad (23.16)$$

其中,  $\boldsymbol{\beta}_i$  为个体  $i$  的系数(常数)。虽然不同个体拥有自己的系数(含截距项与斜率), 但不同个体的扰动项却可能相关, 故应使用 SUR 进行系统估计。

以数据集 mus08cigar.dta 为例(参见第 16 章)。由于该面板数据集为长形(long form), 数据排列形式如第 15 章表 15.1 所示, 故首先要把它转换为宽形(wide form), 参见表 23.1。

表 23.1 面板数据的宽形排列

	$y^1$	$x_1^1$	$x_2^1$	$x_3^1$	...	$y^n$	$x_1^n$	$x_2^n$	$x_3^n$
$t = 1$									
$t = 2$									
$t = 3$									
...									
$t = T$									

其中,  $y^i$  表示个体  $i$  的被解释变量( $i = 1, \dots, n$ ),  $x_k^i$  表示个体  $i$  的第  $k$  个解释变量( $i = 1, \dots, n; k = 1, \dots, K$ )。使用命令 reshape 来完成面板数据在长形与宽形之间的转换<sup>①</sup>:

① 还可以使用命令 reshape 从面板数据中提取横截面数据。

```
. use mus08cigar.dta, clear
. reshape wide lnc lnp lnpmi lny, i(year) j(state)
```

(note: j = 1 2 3 4 5 6 7 8 9 10)						
Data		long	->	wide		
Number of obs.		300	->	30		
Number of variables		6	->	41		
j variable (10 values)		state	->	(dropped)		
xij variables:						
		lnc	->	lnc1 lnc2 ... lnc10		
		lnp	->	lnp1 lnp2 ... lnp10		
		lnpmi	->	lnpmi1 lnpmi2 ... lnpmi10		
		lny	->	lny1 lny2 ... lny10		

从上表可知,转换成宽形后,样本容量变为 30,即每位个体都有 30 年的观测数据。另一方面,变量的数量则大大增加。比如,变量 lnc 变为 lnc1,lnc2,...,lnc10,分别表示第 1 个州的 lnc,第 2 个州的 lnc,...,第 10 个州的 lnc;依此类推。

列出此宽形数据的前两个观测值:

```
. list in 1/2
```

1.	year	lnp1	lnpmi1	lnc1	lny1	lnp2	lnpmi2
	63	4.537577	4.446105	4.542231	7.351354	4.358048	4.399038
	lnc2 4.828314	lny2 7.572886	lnp3 4.480007	lnpmi3 4.434545	lnc3 4.638605	lny3 7.300023	lnp4 4.414975
	lnpmi4 4.358048	lnc4 4.955827	lny4 7.928801	lnp5 4.472572	lnpmi5 4.44993	lnc5 5.051137	lny5 7.977172
	lnp6 4.472572	lnpmi6 4.472572	lnc6 5.110782	lny6 7.723767	lnp7 4.336906	lnpmi7 4.390974	lnc7 5.506956
	lny7 7.913238	lnp8 4.453739	lnpmi8 4.476296	lnc8 4.923624	lny8 7.591199	lnp9 4.476296	lnpmi9 4.446105
	lnc9 4.684905	lny9 7.463902	lnp10 4.516375	lnpmi10 4.399038	lnc10 4.574711	lny10 7.569832	
2.	year	lnp1	lnpmi1	lnc1	lny1	lnp2	lnpmi2
	64	4.565691	4.485369	4.558079	7.428971	4.349237	4.409862
	lnc2 4.795791	lny2 7.632167	lnp3 4.47807	lnpmi3 4.417674	lnc3 4.630838	lny3 7.384039	lnp4 4.409862
	lnpmi4 4.349237	lnc4 4.929425	lny4 7.997976	lnp5 4.496219	lnpmi5 4.463309	lnc5 4.966335	lny5 8.040493
	lnp6 4.470717	lnpmi6 4.521087	lnc6 5.141078	lny6 7.751622	lnp7 4.345061	lnpmi7 4.398027	lnc7 5.460861
	lny7 7.970397	lnp8 4.56233	lnpmi8 4.485369	lnc8 4.832306	lny8 7.660546	lnp9 4.485369	lnpmi9 4.448328
	lnc9 4.678421	lny9 7.531511	lnp10 4.552177	lnpmi10 4.436943	lnc10 4.50092	lny10 7.611648	

下面进行 SUR 回归：

```
. sureg (lnc1 lnp1 lnpm1 lny1) (lnc2 lnp2 lnpm2 lny2) (lnc3 lnp3
lnpm3 lny3) (lnc4 lnp4 lnpm4 lny4) (lnc5 lnp5 lnpm5 lny5) (lnc6 lnp6
lnpm6 lny6) (lnc7 lnp7 lnpm7 lny7) (lnc8 lnp8 lnpm8 lny8) (lnc9 lnp9
lnpm9 lny9) (lnc10 lnp10 lnpm10 lny10), corr
```

Seemingly unrelated regression						
Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
lnc1	30	3	.0361064	0.8696	250.40	0.0000
lnc2	30	3	.0576629	0.7873	143.72	0.0000
lnc3	30	3	.031615	0.8852	249.10	0.0000
lnc4	30	3	.0409479	0.9578	727.41	0.0000
lnc5	30	3	.0460177	0.9072	359.52	0.0000
lnc6	30	3	.0310659	0.8903	241.90	0.0000
lnc7	30	3	.0849558	0.9447	525.99	0.0000
lnc8	30	3	.0398698	0.7859	179.43	0.0000
lnc9	30	3	.033639	0.8369	211.75	0.0000
lnc10	30	3	.0646011	0.7739	118.50	0.0000

		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
lnc1	lnp1	-.7439516	.0926778	-8.03	0.000	-.9255968 -.5623064
	lnpm1	.2611873	.1049528	2.49	0.013	.0554835 .466891
	lny1	.0954658	.0097772	9.76	0.000	.0763029 .1146287
	_cons	6.066224	.2589029	23.43	0.000	5.558784 6.573664
lnc2	lnp2	-.3412215	.0842515	-4.05	0.000	-.5063514 -.1760916
	lnpm2	-.2761468	.0863833	-3.20	0.001	-.445455 -.1068386
	lny2	-.0758077	.0159596	-4.75	0.000	-.1070879 -.0445274
	_cons	8.134837	.2955115	27.53	0.000	7.555645 8.714029
lnc3	lnp3	-.1949179	.0954677	-2.04	0.041	-.3820312 -.0078046
	lnpm3	-.1936758	.1022655	-1.89	0.058	-.3941125 .0067609
	lny3	.1272329	.0084641	15.03	0.000	.1106436 .1438222
	_cons	5.401923	.2095664	25.78	0.000	4.99118 5.812665
lnc4	lnp4	-.5295479	.0650439	-8.14	0.000	-.6570316 -.4020642
	lnpm4	-.1832105	.0870083	-2.11	0.035	-.3537436 -.0126774
	lny4	-.1915741	.0127586	-15.02	0.000	-.2165806 -.1665677
	_cons	9.645321	.2376458	40.59	0.000	9.179544 10.11111
lnc5	lnp5	-.8264925	.0857025	-9.64	0.000	-.9944662 -.6585187
	lnpm5	.2775828	.0893441	3.11	0.002	.1024715 .4526941
	lny5	-.1236889	.0137497	-9.00	0.000	-.1506378 -.0967401
	_cons	8.441914	.2686479	31.42	0.000	7.915373 8.968454
lnc6	lnp6	-.1827293	.0654166	-2.79	0.005	-.3109435 -.0545152
	lnpm6	-.0800041	.0637778	-1.25	0.210	-.2050062 .0449981
	lny6	-.1152345	.0088816	-12.97	0.000	-.132642 -.097827
	_cons	7.199805	.217034	33.17	0.000	6.774426 7.625184

lnc7	lnp7 lnpmin7 lny7 _cons	-.3231466 -.0444938 -.4557646 10.76327	.2431009 .148016 .0434102 .4644809	-1.33 -0.30 -10.50 23.17	0.184 0.764 0.000 0.000	-.7996156 -.3345998 -.540847 9.852906	.1533224 .2456122 .3706821 11.67364
lnc8	lnp8 lnpmin8 lny8 _cons	-.4363926 -.130435 -.0394183 7.771211	.0599955 .0690307 .0110457 .2481084	-7.27 -1.89 -3.57 31.32	0.000 0.059 0.000 0.000	-.5539817 -.2657326 -.0610675 7.284928	-.3188035 .0048626 -.0177691 8.257495
lnc9	lnp9 lnpmin9 lny9 _cons	-.2814291 -.2136093 .0723731 6.349793	.0596311 .0535691 .0085299 .1943468	-4.72 -3.99 8.48 32.67	0.000 0.000 0.000 0.000	-.3983038 -.3186029 .0556549 5.968881	-.1645544 -.1086157 .0890913 6.730706
lnc10	lnp10 lnpmin10 lny10 _cons	-1.062045 .4134775 -.0121777 7.662863	.1564136 .1739583 .0182374 .3629212	-6.79 2.38 -0.67 21.11	0.000 0.017 0.504 0.000	-1.36861 .0725255 -.0479223 6.951551	-.7554796 .7544294 .0235668 8.374176

Correlation matrix of residuals:

	lnc1	lnc2	lnc3	lnc4	lnc5	lnc6	lnc7	lnc8
lnc1	1.0000							
lnc2	0.1650	1.0000						
lnc3	0.5892	0.2456	1.0000					
lnc4	0.3870	0.6052	0.3259	1.0000				
lnc5	-0.2228	0.1637	-0.0553	0.0186	1.0000			
lnc6	-0.0216	0.3286	0.1553	0.3569	0.4797	1.0000		
lnc7	0.0642	0.2279	-0.0223	0.0084	0.5667	0.2576	1.0000	
lnc8	0.2957	0.8255	0.1734	0.5548	0.1800	0.1973	0.3246	1.0000
lnc9	0.5874	0.7296	0.5889	0.6558	0.1733	0.3400	0.1681	0.6903
lnc10	0.6368	0.3761	0.2247	0.4361	-0.1687	-0.2482	-0.0390	0.4449
		lnc9	lnc10					
lnc9	1.0000							
lnc10	0.4798	1.0000						

Breusch-Pagan test of independence: chi2(45) = 208.986, Pr = 0.0000

上表最后一行显示, Breusch-Pagan LM 检验结果强烈拒绝无同期相关的原假设, 故 SUR 比单一方程 OLS 更有效率, 应该使用 SUR 来估计此变系数面板模型。

## 习 题

**23.1** 证明: 如果  $\Omega$  是单位矩阵, 则系统 GLS 估计等价于单一方程 OLS 估计。

**23.2** 证明克罗内克尔乘积的三个性质。

**23.3** 面板数据集 grunfeld.dta 包含了 10 个公司 1935—1954 年的以下变量: invest(投资额), mvalue(公司市场市值), kstock(公司资本存量)。考虑以下投资函数:

$$\text{invest}_i = \beta_0 + \beta_1 \text{mvalue}_i + \beta_2 \text{kstock}_i + u_i + \varepsilon_i \quad (23.17)$$

使用 SUR 估计一个变系数面板模型, 即允许  $\beta_0, \beta_1, \beta_2$  随公司的不同而不同。

## 附录

**A23.1** 如果每个方程包含完全相同的解释变量, 则系统 GLS 还原为单一方程 OLS。

证明: 考虑将  $\hat{\boldsymbol{\beta}}_{\text{GLS}} = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}X'(\Sigma^{-1} \otimes I_T)\mathbf{y}$  按矩阵分块展开。

记  $\Sigma^{-1}$  的第  $(i,j)$  个元素为  $\sigma^{ij}$ 。由于  $X = \begin{pmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X_n \end{pmatrix}$ , 故

$$\begin{aligned} X'(\Sigma^{-1} \otimes I_T)X &= \begin{pmatrix} X'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X'_n \end{pmatrix} \begin{pmatrix} \sigma^{11}I_T & \sigma^{12}I_T & \cdots & \sigma^{1n}I_T \\ \sigma^{21}I_T & \sigma^{22}I_T & \cdots & \sigma^{2n}I_T \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}I_T & \sigma^{n2}I_T & \cdots & \sigma^{nn}I_T \end{pmatrix} \begin{pmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X_n \end{pmatrix} \\ &= \begin{pmatrix} X'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X'_n \end{pmatrix} \begin{pmatrix} \sigma^{11}X_1 & \sigma^{12}X_2 & \cdots & \sigma^{1n}X_n \\ \sigma^{21}X_1 & \sigma^{22}X_2 & \cdots & \sigma^{2n}X_n \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}X_1 & \sigma^{n2}X_2 & \cdots & \sigma^{nn}X_n \end{pmatrix} \\ &= \begin{pmatrix} \sigma^{11}X'_1X_1 & \sigma^{12}X'_1X_2 & \cdots & \sigma^{1n}X'_1X_n \\ \sigma^{21}X'_2X_1 & \sigma^{22}X'_2X_2 & \cdots & \sigma^{2n}X'_2X_n \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}X'_nX_1 & \sigma^{n2}X'_nX_2 & \cdots & \sigma^{nn}X'_nX_n \end{pmatrix} \\ X'(\Sigma^{-1} \otimes I_T)\mathbf{y} &= \begin{pmatrix} X'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X'_n \end{pmatrix} \begin{pmatrix} \sigma^{11}I_T & \sigma^{12}I_T & \cdots & \sigma^{1n}I_T \\ \sigma^{21}I_T & \sigma^{22}I_T & \cdots & \sigma^{2n}I_T \\ \vdots & \vdots & & \vdots \\ \sigma^{n1}I_T & \sigma^{n2}I_T & \cdots & \sigma^{nn}I_T \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{pmatrix} \\ &= \begin{pmatrix} X'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & X'_n \end{pmatrix} \begin{pmatrix} \sigma^{11}\mathbf{y}_1 + \sigma^{12}\mathbf{y}_2 + \cdots + \sigma^{1n}\mathbf{y}_n \\ \sigma^{21}\mathbf{y}_1 + \sigma^{22}\mathbf{y}_2 + \cdots + \sigma^{2n}\mathbf{y}_n \\ \vdots \\ \sigma^{n1}\mathbf{y}_1 + \sigma^{n2}\mathbf{y}_2 + \cdots + \sigma^{nn}\mathbf{y}_n \end{pmatrix} \\ &= \begin{pmatrix} \sigma^{11}X'_1\mathbf{y}_1 + \sigma^{12}X'_1\mathbf{y}_2 + \cdots + \sigma^{1n}X'_1\mathbf{y}_n \\ \sigma^{21}X'_2\mathbf{y}_1 + \sigma^{22}X'_2\mathbf{y}_2 + \cdots + \sigma^{2n}X'_2\mathbf{y}_n \\ \vdots \\ \sigma^{n1}X'_n\mathbf{y}_1 + \sigma^{n2}X'_n\mathbf{y}_2 + \cdots + \sigma^{nn}X'_n\mathbf{y}_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^n \sigma^{1j}X'_1\mathbf{y}_j \\ \sum_{j=1}^n \sigma^{2j}X'_2\mathbf{y}_j \\ \vdots \\ \sum_{j=1}^n \sigma^{nj}X'_n\mathbf{y}_j \end{pmatrix} \end{aligned}$$

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \begin{pmatrix} \sigma^{11} X'_1 X_1 & \sigma^{12} X'_1 X_2 & \cdots & \sigma^{1n} X'_1 X_n \\ \sigma^{21} X'_2 X_1 & \sigma^{22} X'_2 X_2 & \cdots & \sigma^{2n} X'_2 X_n \\ \vdots & \vdots & & \vdots \\ \sigma^{n1} X'_n X_1 & \sigma^{n2} X'_n X_2 & \cdots & \sigma^{nn} X'_n X_n \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^n \sigma^{1j} X'_j y_j \\ \sum_{j=1}^n \sigma^{2j} X'_j y_j \\ \vdots \\ \sum_{j=1}^n \sigma^{nj} X'_j y_j \end{pmatrix} \quad (23.18)$$

当每个方程包含完全相同的解释变量时, 则  $X_1 = X_2 = \cdots = X_n \equiv X$ <sup>①</sup>。因此,

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \begin{pmatrix} \sigma^{11} X' X & \sigma^{12} X' X & \cdots & \sigma^{1n} X' X \\ \sigma^{21} X' X & \sigma^{22} X' X & \cdots & \sigma^{2n} X' X \\ \vdots & \vdots & & \vdots \\ \sigma^{n1} X' X & \sigma^{n2} X' X & \cdots & \sigma^{nn} X' X \end{pmatrix}^{-1} \begin{pmatrix} \sum_{j=1}^n \sigma^{1j} X' y_j \\ \sum_{j=1}^n \sigma^{2j} X' y_j \\ \vdots \\ \sum_{j=1}^n \sigma^{nj} X' y_j \end{pmatrix} \quad (23.19)$$

记第  $j$  个方程的单一方程 OLS 估计值向量、拟合值向量与残差向量分别为  $\mathbf{b}_j, \hat{\mathbf{y}}_j, \mathbf{e}_j$ , 则  $\mathbf{y}_j = \hat{\mathbf{y}}_j + \mathbf{e}_j$ 。因此,

$$X' \mathbf{y}_j = X' (\hat{\mathbf{y}}_j + \mathbf{e}_j) = X' \hat{\mathbf{y}}_j + X' \mathbf{e}_j = X' \hat{\mathbf{y}}_j = X' X \mathbf{b}_j$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &= \left[ \begin{pmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n} \\ \vdots & \vdots & & \vdots \\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn} \end{pmatrix} \otimes (X' X) \right]^{-1} \begin{pmatrix} \sum_{j=1}^n \sigma^{1j} X' X \mathbf{b}_j \\ \sum_{j=1}^n \sigma^{2j} X' X \mathbf{b}_j \\ \vdots \\ \sum_{j=1}^n \sigma^{nj} X' X \mathbf{b}_j \end{pmatrix} \\ &= [\Sigma^{-1} \otimes (X' X)]^{-1} \begin{pmatrix} (X' X) \sum_{j=1}^n \sigma^{1j} \mathbf{b}_j \\ (X' X) \sum_{j=1}^n \sigma^{2j} \mathbf{b}_j \\ \vdots \\ (X' X) \sum_{j=1}^n \sigma^{nj} \mathbf{b}_j \end{pmatrix} \\ &= [\Sigma \otimes (X' X)^{-1}] \begin{pmatrix} (X' X) \sum_{j=1}^n \sigma^{1j} \mathbf{b}_j \\ (X' X) \sum_{j=1}^n \sigma^{2j} \mathbf{b}_j \\ \vdots \\ (X' X) \sum_{j=1}^n \sigma^{nj} \mathbf{b}_j \end{pmatrix} \\ &= \begin{pmatrix} \sigma_{11} (X' X)^{-1} & \sigma_{12} (X' X)^{-1} & \cdots & \sigma_{1n} (X' X)^{-1} \\ \sigma_{21} (X' X)^{-1} & \sigma_{22} (X' X)^{-1} & \cdots & \sigma_{2n} (X' X)^{-1} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} (X' X)^{-1} & \sigma_{n2} (X' X)^{-1} & \cdots & \sigma_{nn} (X' X)^{-1} \end{pmatrix} \begin{pmatrix} (X' X) \sum_{j=1}^n \sigma^{1j} \mathbf{b}_j \\ (X' X) \sum_{j=1}^n \sigma^{2j} \mathbf{b}_j \\ \vdots \\ (X' X) \sum_{j=1}^n \sigma^{nj} \mathbf{b}_j \end{pmatrix} \end{aligned}$$

上式中左边矩阵的第一行与右边矩阵之乘积为

① 为了书写方便, 此证明的  $X$  与前面章节中的  $X$  定义不同。

$$\begin{aligned}
 \hat{\beta}_{i,\text{GLS}} &= \sigma_{11} \sum_{j=1}^n \sigma^{ij} \mathbf{b}_j + \sigma_{12} \sum_{j=1}^n \sigma^{2j} \mathbf{b}_j + \cdots + \sigma_{1n} \sum_{j=1}^n \sigma^{nj} \mathbf{b}_j \\
 &= (\sigma_{11}\sigma^{11} + \sigma_{12}\sigma^{21} + \cdots + \sigma_{1n}\sigma^{n1})\mathbf{b}_1 + (\sigma_{11}\sigma^{12} + \sigma_{12}\sigma^{22} + \cdots + \sigma_{1n}\sigma^{n2})\mathbf{b}_2 \\
 &\quad + \cdots + (\sigma_{11}\sigma^{1n} + \sigma_{12}\sigma^{2n} + \cdots + \sigma_{1n}\sigma^{nn})\mathbf{b}_n \quad (\text{合并同类项}) \\
 &= 1 \cdot \mathbf{b}_1 + 0 \cdot \mathbf{b}_2 + \cdots + 0 \cdot \mathbf{b}_n = \mathbf{b}_1
 \end{aligned}$$

最后一步能成立是因为

$$\Sigma \cdot \Sigma^{-1} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \begin{pmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn} \end{pmatrix} = I_n$$

同理可证,  $\hat{\beta}_{i,\text{GLS}} = \mathbf{b}_i \quad (i=1, \dots, n)$ 。

# 第 24 章 联立方程模型

## 24.1 联立方程模型的结构式与简化式

迄今为止,我们主要关注单一方程。即使在上一章使用 SUR 对多个方程同时进行估计,各方程的变量之间也并没有内在关系(除了各方程扰动项的相关性以外)。然而,经济理论常常推导出一组相互联系的方程,其中一个方程的解释变量是另一方程的被解释变量,这就是联立方程组。

例 农产品市场均衡模型,由需求函数、供给函数及市场均衡条件组成,参见第 10 章。

例 简单的宏观经济模型<sup>①</sup>,参见第 10 章。

即使我们只关心单个方程,但如果该方程包含内生解释变量,则完整的模型仍然是联立方程组(另一方程以此内生解释变量为被解释变量)。为此,有必要研究如何估计联立方程模型。

由  $M$  个方程构成的联立方程模型的“结构式”(structural form)可以表示为

$$\left\{ \begin{array}{l} \gamma_{11}y_{t1} + \gamma_{21}y_{t2} + \cdots + \gamma_{M1}y_{tM} + \beta_{11}x_{t1} + \cdots + \beta_{K1}x_{tK} = \varepsilon_{t1} \\ \gamma_{12}y_{t1} + \gamma_{22}y_{t2} + \cdots + \gamma_{M2}y_{tM} + \beta_{12}x_{t1} + \cdots + \beta_{K2}x_{tK} = \varepsilon_{t2} \\ \cdots \cdots \cdots \\ \gamma_{1M}y_{t1} + \gamma_{2M}y_{t2} + \cdots + \gamma_{MM}y_{tM} + \beta_{1M}x_{t1} + \cdots + \beta_{KM}x_{tK} = \varepsilon_{tM} \end{array} \right. \quad (24.1)$$

其中,  $\{y_{it}\}$  为内生变量,  $\{x_{ij}\}$  为外生变量<sup>②</sup>, 其第一个下标表示第  $t$  个观测值 ( $t = 1, \dots, T$ )<sup>③</sup>, 第二个下标表示第  $i$  个内生变量 ( $i = 1, \dots, M$ ), 或第  $j$  个外生变量 ( $j = 1, \dots, K$ )。内生变量的系数为  $\{\gamma_{ik}\}$ , 其第一个下标表示它是第  $i$  个内生变量的系数, 而第二个下标表示它在第  $k$  个方程中 ( $k = 1, \dots, M$ )。类似地, 外生变量的系数为  $\{\beta_{jk}\}$ , 其第一个下标表示它是第  $j$  个外生变量的系数, 而第二个下标表示它在第  $k$  个方程中。结构方程的扰动项为  $\{\varepsilon_{tk}\}$ , 其第一个下标表示第  $t$  个观测值 ( $t = 1, \dots, T$ ), 而第二个下标表示它在第  $k$  个方程中。“完整的方程系统”(complete system of equations)要求, 内生变量个数等于方程个数  $M$ 。

将上述方程组写成更简洁的“横排”矩阵形式

$$(y_{t1} \ y_{t2} \ \cdots \ y_{tM}) \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1M} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2M} \\ \vdots & \vdots & & \vdots \\ \gamma_{M1} & \gamma_{M2} & \cdots & \gamma_{MM} \end{pmatrix} + (x_{t1} \ x_{t2} \ \cdots \ x_{tK}) \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2M} \\ \vdots & \vdots & & \vdots \\ \beta_{K1} & \beta_{K2} & \cdots & \beta_{KM} \end{pmatrix} = (\varepsilon_{t1} \ \varepsilon_{t2} \ \cdots \ \varepsilon_{tM}) \quad (24.2)$$

① 联立方程形式的宏观经济模型在 20 世纪 70 年代较流行, 但由于方程的稳定性及预测能力不理想, 现已基本为时间序列模型所替代, 比如 VAR 模型。

② 外生变量指的是, 由系统外部决定的变量(包括常数项); 而内生变量则在系统内部决定。

③ 数据形式既可以是时间序列(则  $t$  表示时间), 也可以是横截面数据(则  $t$  表示个体)。

用矩阵来表示即

$$\mathbf{y}'\boldsymbol{\Gamma} + \mathbf{x}'\mathbf{B} = \boldsymbol{\varepsilon}' \quad (24.3)$$

其中, 系数矩阵  $\boldsymbol{\Gamma}_{M \times M}$  与  $\mathbf{B}_{K \times M}$  的每一列对应于一个方程。比如, 第一个方程为

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1M} \end{pmatrix} + \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \end{pmatrix} \begin{pmatrix} \beta_{11} \\ \beta_{21} \\ \vdots \\ \beta_{K1} \end{pmatrix} = \varepsilon_{1t} \quad (24.4)$$

扰动项  $\boldsymbol{\varepsilon}_t$  由第  $t$  期各方程的扰动项所构成。假设扰动项  $\boldsymbol{\varepsilon}_t$  满足  $E(\boldsymbol{\varepsilon}_t | \mathbf{x}_t) = \mathbf{0}$  (因为  $\mathbf{x}_t$  是外生变量), 记其协方差矩阵为,

$$\boldsymbol{\Sigma} = E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' | \mathbf{x}_t) \quad (24.5)$$

由于存在内生变量, 如果直接用 OLS 估计联立方程组中的每一方程, 将导致内生变量偏差 (endogeneity bias) 或联立方程偏差 (simultaneity bias), 得不到一致估计。为此, 求解联立方程组 (24.3), 将  $\mathbf{y}_t$  表示为  $\mathbf{x}_t$  与  $\boldsymbol{\varepsilon}_t$  的函数:

$$\mathbf{y}'_t \boldsymbol{\Gamma} = -\mathbf{x}'_t \mathbf{B} + \boldsymbol{\varepsilon}'_t \quad (24.6)$$

假设  $\boldsymbol{\Gamma}$  非退化, 在上式两边同时右乘  $\boldsymbol{\Gamma}^{-1}$ ,

$$\mathbf{y}'_t = -\mathbf{x}'_t \mathbf{B} \boldsymbol{\Gamma}^{-1} + \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1} \quad (24.7)$$

$$\mathbf{y}'_t = \mathbf{x}'_t \boldsymbol{\Pi} + \mathbf{v}'_t \quad (24.8)$$

方程(24.8)被称为“简化式”(reduced form), 其系数矩阵为  $\boldsymbol{\Pi} = -\underbrace{\mathbf{B}}_{K \times M} \underbrace{\boldsymbol{\Gamma}^{-1}}_{M \times M}$ , 其扰动项为

$\mathbf{v}'_t = \boldsymbol{\varepsilon}'_t \boldsymbol{\Gamma}^{-1}$ , 故  $\mathbf{v}_t = \boldsymbol{\Gamma}^{-1} \boldsymbol{\varepsilon}_t$ 。简化式扰动项  $\mathbf{v}_t$  仍然与外生变量  $\mathbf{x}_t$  不相关, 因为

$$E(\mathbf{v}_t | \mathbf{x}_t) = E(\boldsymbol{\Gamma}^{-1} \boldsymbol{\varepsilon}_t | \mathbf{x}_t) = \boldsymbol{\Gamma}^{-1} E(\boldsymbol{\varepsilon}_t | \mathbf{x}_t) = \mathbf{0} \quad (24.9)$$

简化式扰动项  $\mathbf{v}_t$  的协方差矩阵为

$$\boldsymbol{\Omega} = E(\mathbf{v}_t \mathbf{v}_t' | \mathbf{x}_t) = E(\boldsymbol{\Gamma}^{-1} \boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' \boldsymbol{\Gamma}^{-1} | \mathbf{x}_t) = \boldsymbol{\Gamma}^{-1} E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t' | \mathbf{x}_t) \boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Gamma}^{-1} \quad (24.10)$$

由于每个简化式方程仅包含一个内生变量(即方程左边的被解释变量), 而方程右边的解释变量全部为外生变量  $\mathbf{x}_t$ , 故可以使用 OLS 得到简化式参数  $\boldsymbol{\Pi}$  与  $\boldsymbol{\Omega}$  的一致估计。然而, 通常我们最终关心的是结构式参数。那么, 在什么情况下, 才能从简化式参数( $\boldsymbol{\Pi}, \boldsymbol{\Omega}$ )反推出结构式参数( $\boldsymbol{\Gamma}, \mathbf{B}, \boldsymbol{\Sigma}$ )呢? 这就涉及联立方程模型的“识别问题”(problem of identification)。

## 24.2 联立方程模型的识别

在对模型的总体参数进行估计之前, 其参数必须“可识别”(identified)。可识别是进行参数估计的前提条件。如果一个总体参数可识别, 则该参数的任意两个不同取值, 都会在随机样本中显示出系统差异, 即如果样本容量足够大, 则应该能够在统计意义上区分这两个不同的参数值。反之, 如果无论多大的样本都区分不开, 即由不同参数值的总体产生的观测数据在统计意义上是一样的, 则存在“观测等价”(observational equivalence), 该参数“不可识别”(unidentified)。

下面以一个简单例子来说明可识别的概念。考虑以下回归模型:

$$y_i = \alpha_1 + \alpha_2 + \beta x_i + \varepsilon_i \quad (24.11)$$

显然, 仅通过样本数据  $\{y_i, x_i\}_{i=1}^n$  是无法对  $\alpha_1$  与  $\alpha_2$  分别进行识别的, 但可以识别二者之和

$(\alpha_1 + \alpha_2)$ 。回到联立方程模型的情形，“可识别”意味着，可以从简化式参数  $(\Pi, \Omega)$  求出结构式参数  $(\Gamma, B, \Sigma)$  的唯一解 (unique solution)。从上面的推导可知，这两组参数之间的关系如下：

$$\Pi = -B\Gamma^{-1} \quad (24.12)$$

$$\Omega = \Gamma^{-1'} \Sigma \Gamma^{-1} \quad (24.13)$$

显然，如果  $\Gamma$  已知，则可以通过  $\Pi$  与  $\Omega$  求得  $B$  与  $\Sigma$ 。然而，一般来说， $\Gamma$  是由未知参数组成的矩阵。事实上，结构式的参数个数比简化式的参数个数多出  $M^2$  个。这是因为，简化式参数  $(\Pi, \Omega)$  的总个数为  $[K \times M + M(M+1)/2]$  (其中， $\Pi_{K \times M}$  含  $K \times M$  个参数，而对称矩阵  $\Omega_{M \times M}$  含  $M(M+1)/2$  个参数)；而结构式参数  $(\Gamma, B, \Sigma)$  的总个数为  $[M^2 + K \times M + M(M+1)/2]$  (其中， $\Gamma_{M \times M}$  含  $M^2$  个参数， $B_{K \times M}$  含  $K \times M$  个参数，对称矩阵  $\Sigma$  含  $M(M+1)/2$  个参数)。这表明，在一般情况下，不可能从  $(\Pi, \Omega)$  求出  $(\Gamma, B, \Sigma)$  的唯一解。因此，如果不对结构式参数进行约束，将不可能从简化式参数得到结构式参数的唯一解，结构方程也就不可识别。为了识别结构方程，常对结构参数施加如下约束。

(1) 标准化 (normalization)。在每个结构方程中，可以将一个内生变量视为被解释变量，并将其系数标准化为 1。即使这样，结构式参数仍然比简化式参数多  $(M^2 - M)$  个。

(2) 恒等式 (identity)。比如，供需相等的均衡条件、会计恒等式、定义式。恒等式中每个变量的系数均为已知，不需要识别或估计。

(3) 排斥约束 (exclusion restrictions)。在结构方程中排斥某些内生或外生变量，这相当于对结构矩阵  $(\Gamma, B)$  施以“零约束” (zero restrictions)，即让  $(\Gamma, B)$  中的某些元素为 0。

(4) 线性约束 (linear restriction)。比如，在理论上可以假设生产函数为规模报酬不变 (constant returns to scale)，则资本的产出弹性与劳动力的产出弹性之和为 1。

(5) 对扰动项协方差矩阵的约束 (restrictions on the disturbance covariance matrix)。比如，在某些情况下，可以假设不同方程的扰动项之间不相关。

在实践中，最重要的约束方法是“排斥变量” (即零约束)。对于线性约束，可以通过重新定义变量而将其转化为“排斥变量”约束。那么，究竟需要多少零约束才可以保证结构方程可识别呢？

由于系数矩阵  $\Gamma$  与  $B$  的每一列对应于一个方程，故矩阵  $\Gamma$  的第  $k$  列  $\Gamma_k$  与矩阵  $B$  的第  $k$  列  $B_k$  包含了第  $k$  个方程的系数。由于  $\Pi = -B\Gamma^{-1}$ ，故  $\Pi\Gamma = -B$ 。将  $\Gamma$  与  $B$  进行矩阵分块可得

$$\underbrace{\Pi}_{K \times M} \underbrace{(\Gamma_1 \cdots \Gamma_M)}_{M \times M} = \underbrace{(\Pi\Gamma_1 \cdots \Pi\Gamma_M)}_{K \times M} = -\underbrace{(B_1 \cdots B_M)}_{K \times M} \quad (24.14)$$

由此可知， $\Gamma_k$  与  $B_k$  满足以下方程组：

$$\Pi\Gamma_k = -B_k \quad (24.15)$$

不失一般性，我们仅考虑第一个结构方程，即  $k=1$ <sup>①</sup>。假设在第一个方程中，内生变量  $y_1$  的系数已被标准化为 1，另有  $M_1$  个内生变量 (以  $y_1^*$  来表示) 也包括在此方程中，而其余  $M_1^*$  个内生变量 (以  $y_1^{**}$  来表示) 则被排斥在此方程之外，故  $1 + M_1 + M_1^* = M$ 。不失一般性，将  $y_1$  排在  $y_1^*$  之后、 $y_1^{**}$  之前，即  $(y_1, y_1^*, y_1^{**})'$ 。同时假设，第一个方程包含  $K_1$  个外生变量 (以  $x_1$  来表示)，而其余  $K_1^*$  个外生变量 (以  $x_1^*$  来表示) 则被排斥在此方程之外，故  $K_1 + K_1^* = K$ 。不失一般性，将  $x_1$  排在  $x_1^*$  之前，即  $(x_1, x_1^*)'$ 。以右上角的 \* 表示被排斥的变量及其对应的系数，则第一个方程可以写为

$$y_1 = y_1' \gamma_1 + y_1^{*'} \gamma_1^* + x_1' \beta_1 + x_1^{*'} \beta_1^* + \varepsilon_1 \quad (24.16)$$

① 如果  $k > 1$ ，可以通过调整方程的排序而使得  $k=1$ 。

显然,根据“排斥变量”约束,  $\gamma_1^* = \mathbf{0}, \beta_1^* = \mathbf{0}$ 。因此,矩阵  $\Gamma$  的第 1 列为  $\Gamma_1 = \begin{pmatrix} 1 \\ -\gamma_1 \\ \mathbf{0} \end{pmatrix}$ , 矩阵  $B$

的第 1 列为  $B_1 = \begin{pmatrix} -\beta_1 \\ \mathbf{0} \end{pmatrix}$ , 故  $\Pi_{K \times M} \cdot \begin{pmatrix} 1 \\ -\gamma_1 \\ \mathbf{0} \end{pmatrix}_{M \times 1} = \begin{pmatrix} -\beta_1 \\ \mathbf{0} \end{pmatrix}_{K \times 1}$ 。

将简化式矩阵  $\Pi$  进行相应的分块可得(当将  $y_1$  排在  $y_1^*$  之前,  $x_1$  排在  $x_1^*$  之前时,对矩阵  $\Pi$  的相应行与列也进行相应的调整)

$$\begin{array}{c} 1 \text{ 列} \quad M_1 \text{ 列} \quad M_1^* \text{ 列} \\ K_1 \text{ 行} \quad \left( \begin{array}{ccc} \pi_1 & \underline{\Pi}_1 & \overline{\Pi}_1 \\ \pi_1^* & \underline{\Pi}_1^* & \overline{\Pi}_1^* \end{array} \right)_{K \times M} \quad \left( \begin{array}{c} 1 \\ -\gamma_1 \\ \mathbf{0} \end{array} \right)_{M \times 1} \\ K_1^* \text{ 行} \end{array} = \begin{array}{c} K_1 \text{ 行} \\ K_1^* \text{ 行} \end{array} \quad (24.17)$$

将上式左边的两个矩阵相乘后,可以得到以下两个方程组

$$\underbrace{\pi_1}_{K_1 \times 1} - \underbrace{\underline{\Pi}_1}_{K_1 \times M_1} \underbrace{\gamma_1}_{M_1 \times 1} = -\underbrace{\beta_1}_{K_1 \times 1} \quad (24.18)$$

$$\underbrace{\underline{\Pi}_1^*}_{K_1^* \times M_1} \underbrace{\gamma_1}_{M_1 \times 1} = \underbrace{\pi_1^*}_{K_1^* \times 1} \quad (24.19)$$

如果可以根据  $\Pi$  求出  $(\gamma_1, \beta_1)$  的唯一解,则第一个结构方程可识别。假如知道  $\gamma_1$ ,就可以通过方程组(24.18)求解  $\beta_1$ 。因此,问题的关键在于求解  $\gamma_1$ 。第一个结构方程可识别的充分必要条件是,在方程组  $\underline{\Pi}_1^* \gamma_1 = \pi_1^*$  中,  $\gamma_1$  有唯一解。根据线性代数的知识,这要求满足以下秩条件,

$$\text{rank}(\underline{\Pi}_1^*) = \text{rank}(\underline{\Pi}_1^* \pi_1^*) = M_1 \quad (24.20)$$

即系数矩阵  $\underline{\Pi}_1^*$  的秩等于增广矩阵  $(\underline{\Pi}_1^*, \pi_1^*)$  的秩,且等于未知数的个数  $M_1$ 。秩条件是可识别的充分必要条件。如果秩条件不满足,则不可识别。由于矩阵的秩一定小于或等于矩阵的行数,而  $\underline{\Pi}_1^*$  为  $K_1^* \times M_1$  级矩阵,故  $K_1^* \geq \text{rank}(\underline{\Pi}_1^*) = M_1$ 。因此,秩条件成立的必要条件为

$$K_1^* \geq M_1 \quad (24.21)$$

这被称为“阶条件”(order condition)。它意味着,结构方程所排斥的外生变量的个数( $K_1^*$ )应大于或等于该方程所包含的内生解释变量的个数( $M_1$ )。显然,阶条件只是可识别的必要条件,而非充分条件。

从工具变量法的角度(这是估计联立方程组的主要方法,参见下文),如果有  $M_1$  个或更多的有效工具变量,则可以对含有  $M_1$  个内生解释变量的结构方程进行一致估计。容易看出,被第一个结构方程排斥的所有外生变量都是有效工具变量,因为根据外生变量的定义,它们与扰动项不相关(外生性);另一方面,根据简化式,内生变量可以表示为外生变量的函数,故它们与内生解释变量相关(相关性)。因此,阶条件“ $K_1^* \geq M_1$ ”的含义是,需要有足够的工具变量,才能对该方程进行一致估计。由于阶条件很容易检验,而秩条件不容易检验(矩阵  $\Pi$  包含未知参数,如何计算它的秩呢?),并且阶条件满足而秩条件不满足的情况很少见,故在实践中,常常只检验阶条件(只要数变量的个数即可),而忽略秩条件。

在可识别(即秩条件满足)的情况下,如果恰好  $K_1^* = M_1$ ,则称该结构方程“恰好识别”(just identified),即工具变量个数正好相等内生解释变量的个数;如果  $K_1^* > M_1$ ,则称该结构方程“过

度识别”(overidentified),即工具变量个数大于内生解释变量的个数。在过度识别的情况下,方程组  $\mathbf{H}_1^* \boldsymbol{\gamma}_1 = \boldsymbol{\pi}_1^*$  存在多余的方程。在理论上,无论使用  $K_1^*$  方程中的任意  $M_1$  个方程,都能得到  $\boldsymbol{\gamma}_1$  相同的唯一解。

## 24.3 单一方程估计法

估计联立方程组的方法可以分为两类,即“单一方程估计法”(single equation estimation),也称“有限信息估计法”(limited information estimation);以及系统估计法,也称“全信息估计法”(full information estimation)。前者对联立方程组中的每一个方程分别进行估计,而后者则将其作为一个系统进行联合估计。单一方程估计法主要包括,普通最小二乘法、间接最小二乘法、二阶段最小二乘法以及广义矩估计法。下一节考虑系统估计法。

### 1. 普通最小二乘法

由于存在内生解释变量,一般来说,OLS 是不一致的。但由于其计算简单,OLS 仍然可以作为一种参照系。另外,对于一种特殊的递归模型(recursive model)<sup>①</sup>,即  $\boldsymbol{\Gamma}$  为下三角矩阵(lower triangular matrix)而协方差矩阵  $\boldsymbol{\Sigma}$  为对角矩阵(不同方程之间的扰动项不相关)的情形,OLS 依然是一致的。以一个三方程的系统为例:

$$\begin{cases} y_1 = \mathbf{x}'\boldsymbol{\beta}_1 + \varepsilon_1 \\ y_2 = \mathbf{x}'\boldsymbol{\beta}_2 + \gamma_{12}y_1 + \varepsilon_2 \\ y_3 = \mathbf{x}'\boldsymbol{\beta}_3 + \gamma_{13}y_1 + \gamma_{23}y_2 + \varepsilon_3 \end{cases} \quad (24.22)$$

显然,第一个方程不含内生解释变量,可以用 OLS 得到一致估计。在第二个方程中,唯一的内生解释变量为  $y_1$ ,而且与扰动项不相关,因为

$$\text{Cov}(y_1, \varepsilon_2) = \text{Cov}(\mathbf{x}'\boldsymbol{\beta}_1 + \varepsilon_1, \varepsilon_2) = \underbrace{\text{Cov}(\mathbf{x}'\boldsymbol{\beta}_1, \varepsilon_2)}_{=0} + \underbrace{\text{Cov}(\varepsilon_1, \varepsilon_2)}_{=0} = 0 \quad (24.23)$$

因此,可以用 OLS 来估计第二个方程。同理,在第三个方程中,内生解释变量为  $(y_1, y_2)$ ,而且  $\text{Cov}(y_1, \varepsilon_3) = \text{Cov}(y_2, \varepsilon_3) = 0$ ,故也可以用 OLS 来估计。

### 2. 间接最小二乘法

在恰好识别的情况下,可以先用 OLS 来一致地估计简化式参数,然后通过结构式参数与简化式参数的关系来求解结构式参数,这被称为“间接最小二乘法”(Indirect Least Square,简记 ILS)。然而,虽然 ILS 在恰好识别的情况下是一致的,但却不是最有效率的,故不常用。

ILS 的另一缺点是,在过度识别的情况下无法使用。这是因为,在过度识别时,根据简化式参数求解结构式参数的联立方程组  $\mathbf{H}_1^* \boldsymbol{\gamma}_1 = \boldsymbol{\pi}_1^*$  有多余的方程。理论上,无论使用哪些方程,都能得到相同的唯一解。但在实践中,却会得到不同的估计量,而我们不知道如何取舍。以简单的商品供需模型为例,

$$\begin{cases} q = \beta p + \gamma y & (\text{需求}) \\ q = \delta p & (\text{供给}) \end{cases} \quad (24.24)$$

其中,内生变量  $q$  与  $p$  分别表示商品的数量与价格,外生变量  $y$  表示收入,为了简便略去了扰动项。由于供给方程包含一个内生解释变量  $p$ ,而排斥了一个外生变量  $y$ ,故为恰好识别;但需求方

<sup>①</sup> 也称为“三角系统”(triangular system)。

程不可识别。这个联立方程模型的简化式为

$$\begin{cases} p = \frac{\gamma}{\delta - \beta} y \equiv \pi_1 y \\ q = \frac{\gamma\delta}{\delta - \beta} y \equiv \pi_2 y \end{cases} \quad (24.25)$$

其中,  $\pi_1 \equiv \frac{\gamma}{\delta - \beta}$ ,  $\pi_2 \equiv \frac{\gamma\delta}{\delta - \beta}$ 。显然,  $\pi_2/\pi_1 = \delta$ 。记  $\hat{\pi}_1, \hat{\pi}_2$  为简化式参数  $\pi_1, \pi_2$  的 OLS 估计量, 则结构参数  $\delta$  的 ILS 估计量为  $\hat{\delta}_{ILS} = \hat{\pi}_2/\hat{\pi}_1$ 。ILS 在恰好识别的情况下成立。

现在假设需求方程中还有另一外生变量  $x$ , 而供给方程不变:

$$\begin{cases} q = \beta p + \gamma y + \theta x & (\text{需求}) \\ q = \delta p & (\text{供给}) \end{cases} \quad (24.26)$$

此时, 供给方程包含一个内生解释变量  $p$ , 排斥两个外生变量( $y, x$ ), 故为过度识别。需求方程仍然不可识别。这个新模型的简化式为

$$\begin{cases} p = \frac{\gamma}{\delta - \beta} y + \frac{\theta}{\delta - \beta} x \equiv \pi_1 y + \pi_3 x \\ q = \frac{\gamma\delta}{\delta - \beta} y + \frac{\theta\delta}{\delta - \beta} x \equiv \pi_2 y + \pi_4 x \end{cases} \quad (24.27)$$

其中,  $\pi_1 \equiv \frac{\gamma}{\delta - \beta}$ ,  $\pi_2 \equiv \frac{\gamma\delta}{\delta - \beta}$ ,  $\pi_3 \equiv \frac{\theta}{\delta - \beta}$ ,  $\pi_4 \equiv \frac{\theta\delta}{\delta - \beta}$ 。显然,  $\pi_2/\pi_1 = \delta$ ,  $\pi_4/\pi_3 = \delta$ 。记  $\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \hat{\pi}_4$  为简化式参数  $\pi_1, \pi_2, \pi_3, \pi_4$  的 OLS 估计量, 则存在对结构参数  $\delta$  的两个 ILS 估计, 即  $\hat{\pi}_2/\hat{\pi}_1$  与  $\hat{\pi}_4/\hat{\pi}_3$ 。一般来说,  $\hat{\pi}_2/\hat{\pi}_1 \neq \hat{\pi}_4/\hat{\pi}_3$ , 而我们不知道究竟该用何者作为  $\hat{\delta}_{ILS}$ , 故 ILS 不适用<sup>①</sup>。

### 3. 二阶段最小二乘法

在结构方程可识别的情况下, 其排斥的外生变量个数大于或等于包含的内生解释变量个数, 而所有排斥的外生变量都是有效工具变量, 故可以用工具变量法来估计。如果结构方程的扰动项满足同方差、无自相关的古典假定, 则二阶段最小二乘法(2SLS)是最有效率的工具变量法, 故也是最常见的单一方程估计法, 参见第 10 章。

### 4. 广义矩估计法

在过度识别的情况下, 如果结构方程的扰动项存在异方差或自相关, 则 GMM 比 2SLS 更有效率。在恰好识别的情况下, GMM 等价于 2SLS, 参见第 10 章。

### 5. 有限信息最大似然估计法

如果假定结构方程的扰动项服从正态分布, 则可以使用最大似然估计法对单一方程进行估计, 这被称为“有限信息最大似然估计法”(Limited Information Maximum Likelihood Estimation, 简记 LIML)。LIML 与 2SLS 在大样本下是渐近等价的, 但小样本性质似乎不如 2SLS。然而, 如果存在弱工具变量(weak instruments), 则 LIML 比 2SLS 更稳健。

## 24.4 三阶段最小二乘法

在使用单一方程估计法时, 由于忽略了各方程之间的联系(包括各方程扰动项之间的联

<sup>①</sup> 除非在用 OLS 估计简化式参数  $\pi_1, \pi_2, \pi_3, \pi_4$  时加上约束条件“ $\pi_2/\pi_1 = \pi_4/\pi_3$ ”。从此例还可看出, 结构参数的过度识别相当于给简化式参数附加了约束条件。

系),故不如将所有方程作为一个整体进行估计(即系统估计法)更有效率。系统估计法的缺点是,如果其中的某个方程估计得不准确,则可能影响系统中其他方程的估计。

最常见的系统估计法为“三阶段最小二乘法”(Three Stage Least Square,简记3SLS)<sup>①</sup>。在某种意义上,3SLS是将2SLS与SUR相结合的一种估计方法。对于一个多方程的系统,如果各方程中都不包含内生解释变量,则对每个方程进行OLS估计是一致的;但却不是最有效率的,因为单一方程OLS忽略了不同方程的扰动项之间可能存在相关性。此时,用SUR对整个方程系统同时进行估计是有效率的。

对于一个多方程系统,如果方程中包含内生解释变量,则对每个方程进行2SLS估计是一致的;但却不是最有效率的,因为单一方程2SLS忽略了不同方程的扰动项之间可能存在相关性。此时,用3SLS对整个联立方程系统同时进行估计是有效率的。

3SLS的基本步骤如下。

前两步:对每个方程进行2SLS估计。

第三步:根据前两步的估计,得到对整个系统的扰动项之协方差矩阵的估计。然后,据此对整个系统进行GLS估计(类似于SUR的做法)。具体操作如下。

对于联立方程模型的第 $j$ 个方程,忽略不在方程中的内生变量 $y_j^*$ 与外生变量 $x_j^*$ ,并同时考虑所有 $T$ 个观测值,则可以将第 $j$ 个方程写为

$$\underbrace{y_j}_{T \times 1} = \underbrace{Y_j}_{T \times M_j} \underbrace{\gamma_j}_{M_j \times 1} + \underbrace{X_j}_{T \times K_j} \underbrace{\beta_j}_{K_j \times 1} + \varepsilon_j \equiv \underbrace{Z_j}_{T \times 1} \underbrace{\delta_j}_{(M_j+K_j) \times 1} + \varepsilon_j \quad (j=1, \dots, M) \quad (24.28)$$

其中,第 $j$ 个方程中的解释变量为 $Z_j \equiv (Y_j X_j)$ (包含内生解释变量 $Y_j$ 与外生变量 $X_j$ ), $\delta_j \equiv (\gamma_j \beta_j)$ (同时包含内生变量与外生变量的系数)。将所有 $M$ 个方程叠放在一起可得

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{pmatrix} = \begin{pmatrix} Z_1 & 0 & \cdots & 0 \\ 0 & Z_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Z_M \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_M \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{pmatrix} = Z\delta + \varepsilon \quad (24.29)$$

假设 $E(\varepsilon|X)=0, E(\varepsilon\varepsilon'|X)=\Sigma \otimes I$ ,其中 $X$ 包含整个方程系统中所有的外生变量(都可以作为工具变量), $\otimes$ 为克罗内克尔乘积(参见第23章)。记 $\hat{Z}_j \equiv X(X'X)^{-1}X'Z_j$ 为第 $j$ 个方程解释变量 $Z_j$ 对所有外生变量(工具变量) $X$ 进行回归的拟合值(第一阶段回归),则第 $j$ 个方程的2SLS估计量为

$$\hat{\delta}_{j,2SLS} \equiv (\hat{Z}_j'\hat{Z}_j)^{-1}\hat{Z}_j'y_j \quad (24.30)$$

定义 $\hat{Z} \equiv \begin{pmatrix} \hat{Z}_1 & 0 & \cdots & 0 \\ 0 & \hat{Z}_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \hat{Z}_M \end{pmatrix}$ ,则可以将所有方程的单一方程2SLS估计量简洁地写在一起

(参见习题):

<sup>①</sup> 参见 Zellner and Theil(1962)。

$$\hat{\boldsymbol{\delta}}_{2SLS} = \begin{pmatrix} \hat{\boldsymbol{\delta}}_{1,2SLS} \\ \hat{\boldsymbol{\delta}}_{2,2SLS} \\ \vdots \\ \hat{\boldsymbol{\delta}}_{M,2SLS} \end{pmatrix} = (\hat{\mathbf{Z}}' \hat{\mathbf{Z}})^{-1} \hat{\mathbf{Z}}' \hat{\mathbf{y}} \quad (24.31)$$

为了进行 3SLS 估计, 必须先得到对协方差矩阵  $\Sigma$  的估计值  $\hat{\Sigma}$ 。记矩阵  $\hat{\Sigma}$  的  $(i,j)$  元素为  $\hat{\sigma}_{ij}$ , 利用单一方程 2SLS 估计的残差可得

$$\hat{\sigma}_{ij} = \frac{1}{T} (\mathbf{y}_i - \mathbf{Z}_i \hat{\boldsymbol{\delta}}_{i,2SLS})' (\mathbf{y}_j - \mathbf{Z}_j \hat{\boldsymbol{\delta}}_{j,2SLS}) \quad (24.32)$$

一般来说, 使用 GLS 会比单一方程 2SLS 更有效率。类比 SUR, 可定义 3SLS 估计量为

$$\hat{\boldsymbol{\delta}}_{3SLS} = [\hat{\mathbf{Z}}' (\hat{\Sigma}^{-1} \otimes \mathbf{I}) \hat{\mathbf{Z}}]^{-1} \hat{\mathbf{Z}}' (\hat{\Sigma}^{-1} \otimes \mathbf{I}) \mathbf{y} \quad (24.33)$$

对于 3SLS, 也可以进行迭代, 即用 3SLS 的残差重新估计协方差矩阵  $\Sigma$ , 然后再使用 GLS, 如此反复, 直至收敛。但迭代 3SLS 并不能提高其渐近效率。

除了 3SLS 外, 系统估计法还包括, “系统广义矩估计”(System GMM), 以及“全信息最大似然估计法”(Full Information Maximum Likelihood Estimation, 简记 FIML), 参见 Greene (2012) 或 Hayashi (2000)。

## 24.5 三阶段最小二乘法的 Stata 实例

3SLS 的 Stata 命令格式为

```
reg3 (depvar1 varlist1) (depvar2 varlist2) ... (depvarN
varlistN), ols 2sls sure ireg3 exog (varlist) endog (varlist) inst
(varlist)
```

其中, 选择项“ols”表示进行 OLS 估计, “2sls”表示进行 2SLS 估计, “sure”表示进行 SUR 估计, 而默认值为进行 3SLS 估计。选择项“ireg3”表示进行迭代式 3SLS 估计。而选择项“exog (varlist)”、“endog (varlist)”以及“inst (varlist)”则用于指定额外的外生变量, 内生变量(除被解释变量以外), 以及工具变量(除方程组系统自带的工具变量以外)。详见“help reg3”。

下面以数据集 klein.dta 为例。该数据集包含 1920—1941 年的以下宏观经济变量, consump(消费)、wagepriv(私企工资)、wagegovt(政府工资)、govt(政府开支)、capital1(资本存量的滞后值)。考虑以下联立方程模型:

$$\begin{cases} \text{consump}_t = \alpha_0 + \alpha_1 \text{wagepriv}_t + \alpha_2 \text{wagegovt}_t + u_t \\ \text{wagepriv}_t = \beta_0 + \beta_1 \text{consump}_t + \beta_2 \text{govt}_t + \beta_3 \text{capital1}_t + v_t \end{cases} \quad (24.34)$$

其中, 第一个方程以私企工资与政府工资(相当于收入)来解释消费, 即消费函数; 而第二个方程以消费、政府开支、资本存量的滞后值(相当于总需求)来解释私企工资。这个联立方程系统共有两个内生变量, 即 consump 与 wagepriv; 三个外生变量, 即 wagegovt, govt 与 capital1。每个方程均包含一个内生解释变量。第一个方程排斥了两个外生变量 govt 与 capital1, 即有两个工具变量可用, 故为过度识别; 而第二个方程排斥了一个外生变量 wagegovt, 即有一个工具变量可用, 故为

恰好识别。

作为参照系,首先用 OLS 对每个方程进行单一方程估计:

```
. use klein.dta,clear
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1),ols
. estimates store OLS
```

Multivariate regression						
Equation	Obs	Parms	RMSE	"R-sq"	F-Stat	P
consump	22	2	1.60651	0.9567	210.15	0.0000
wagepriv	22	3	1.553524	0.9489	111.33	0.0000
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
consump						
wagepriv	.9918122	.0678073	14.63	0.000	.8544216	1.129203
wagegovt	.6780964	.2147332	3.16	0.003	.2430055	1.113187
_cons	14.24549	2.045098	6.97	0.000	10.10173	18.38925
wagepriv						
consump	.7742524	.0654305	11.83	0.000	.6416777	.9068272
govt	.4048119	.1969143	2.06	0.047	.0058257	.8037981
capital1	-.0443646	.0356482	-1.24	0.221	-.1165947	.0278656
_cons	1.668479	6.744839	0.25	0.806	-11.99786	15.33482

其中,命令 reg3 的默认估计法为 3SLS,但加上选择项“ols”后则进行单一方程 OLS 估计。

下面进行单一方程 2SLS 估计:

```
. reg3 ( consump wagepriv wagegovt ) ( wagepriv consump govt
capital1 ),2sls
. estimates store Two_SLS
```

Two-stage least-squares regression						
Equation	Obs	Parms	RMSE	"R-sq"	F-Stat	P
consump	22	2	1.911394	0.9388	89.83	0.0000
wagepriv	22	3	2.720166	0.8432	21.67	0.0000
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
consump						
wagepriv	.8012754	.1376629	5.82	0.000	.5223438	1.080207
wagegovt	1.029531	.3280273	3.14	0.003	.3648848	1.694178
_cons	19.3559	3.856336	5.02	0.000	11.54222	27.16958
wagepriv						
consump	.3752562	.2848669	1.32	0.196	-.2019389	.9524514
govt	1.155399	.5996727	1.93	0.062	-.0596528	2.370452
capital1	.0107234	.072061	0.15	0.883	-.1352862	.1567329
_cons	8.443596	12.61305	0.67	0.507	-17.11287	34.00007
Endogenous variables: consump wagepriv						
Exogenous variables: wagegovt govt capital1						

其中,选择项“2sls”表示进行单一方程 2SLS 估计。下面进行 3SLS 估计。

```
. reg3 ( consump wagepriv wagegovt ) ( wagepriv consump govt  
capital1),first
```

. estimates store Three\_SLS

其中,选择项“first”表示显示第一阶段回归的结果(即内生解释变量对工具变量的回归)。

上表显示,第一阶段的两个回归方程均很显著( $F$ 统计量的 $p$ 值很小)。

下面进行迭代式3SLS估计:

```
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1),  
i  
ireg3  
. estimates store Three_SLS_iter
```

Iteration 1:	tolerance =	.655117
Iteration 2:	tolerance =	.00433981
Iteration 3:	tolerance =	.00004779
Iteration 4:	tolerance =	5.240e-07

Three-stage least-squares regression, iterated

Equation	Obs	Parms	RMSE	"R-sq"	chi2	P
consump	22	2	1.776297	0.9388	208.02	0.0000
wagepriv	22	3	2.373113	0.8542	86.04	0.0000

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
consump					
wagepriv	.8012754	.1279329	6.26	0.000	.5505314 1.052019
wagegovt	1.029531	.3048424	3.38	0.001	.432051 1.627011
_cons	19.3559	3.583772	5.40	0.000	12.33184 26.37996
wagepriv					
consump	.402311	.2475678	1.63	0.104	-.0829131 .887535
govt	1.177549	.5228599	2.25	0.024	.1527627 2.202336
capital1	-.0276932	.054752	-0.51	0.613	-.1350052 .0796188
_cons	14.56316	9.841946	1.48	0.139	-4.726701 33.85302

Endogenous variables: consump wagepriv

Exogenous variables: wagegovt govt capital1

为了便于比较,将以上方法的系数估计值及标准误列表:

```
. esttab OLS Two_SLS Three_SLS Three_SLS_iter, r2 mtitles star (*0.1  
**0.05 ***0.01)
```

	(1) OLS	(2) Two_SLS	(3) Three_SLS	(4) Three_SLS_~r
consump				
wagepriv	0.992*** (14.63)	0.801*** (5.82)	0.801*** (6.26)	0.801*** (6.26)
wagegovt	0.678*** (3.16)	1.030*** (3.14)	1.030*** (3.38)	1.030*** (3.38)
_cons	14.25*** (6.97)	19.36*** (5.02)	19.36*** (5.40)	19.36*** (5.40)
wagepriv				
consump	0.774*** (11.83)	0.375 (1.32)	0.403 (1.57)	0.402 (1.63)
govt	0.405** (2.06)	1.155* (1.93)	1.178** (2.17)	1.178** (2.25)
capital1	-0.0444 (-1.24)	0.0107 (0.15)	-0.0281 (-0.49)	-0.0277 (-0.51)
_cons	1.668 (0.25)	8.444 (0.67)	14.63 (1.42)	14.56 (1.48)
N	22	22	22	22
R-sq	0.957	0.939	0.939	0.939
t statistics in parentheses				
* p<0.1, ** p<0.05, *** p<0.01				

从上表可以看出,单一方程 2SLS,3SLS 与迭代 3SLS 的估计结果很接近,但与单一方程 OLS 的估计结果差别较大。

## 24.6 结构 VAR

在 20 世纪 80 年代之前,传统的联立方程模型流行一时,尤其在宏观经济领域。这些结构模型越建越大,似乎能很好地拟合样本数据;但对于样本外数据的预测能力却较弱。而且,为了识别结构方程组,常需要附加很多难以置信(incredible)的约束条件。作为解决方法,Sims(1980)提出了 VAR 模型(参见第 20 章)。然而,简化式 VAR 的脉冲响应函数并不唯一(依赖于变量次序),而且简化式 VAR 无法揭示经济结构(变量之间没有当期影响)。

为此,经济学家又试图将结构重新纳入 VAR 模型中,允许变量之间存在当期影响,形成“结构 VAR”的方法(Sims, 1981, 1986; Bernanke, 1986; Shapiro and Watson, 1988; Blanchard and Quah, 1989)。作为例子,考虑如下二元动态联立方程组(为叙述方便,忽略常数项):

$$\begin{cases} y_{1t} = -a_{12}y_{2t} + \gamma_{11}y_{1,t-1} + \gamma_{12}y_{2,t-1} + \varepsilon_{1t} \\ y_{2t} = -a_{21}y_{1t} + \gamma_{21}y_{1,t-1} + \gamma_{22}y_{2,t-1} + \varepsilon_{2t} \end{cases} \quad (24.35)$$

其中,扰动项的分布满足

$$\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \sim \text{iid} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right] \quad (24.36)$$

方程组(24.35)的显著特征是在方程右边的解释变量中包含了当期变量,即 $y_{1t}$ 的解释变量包括 $y_{2t}$ ,而 $y_{2t}$ 的解释变量也包括 $y_{1t}$ 。一般认为,方程组(24.35)来自于经济理论对于经济结构的建模,故称为“结构 VAR”(Structural VAR,简记 SVAR)。结构方程的扰动项 $\varepsilon_{1t}$ 与 $\varepsilon_{2t}$ 相互独立,称为“结构新息”(structural innovation)。比如, $y_{1t}$ 为去势(detrended)的实际 GDP 对数, $y_{2t}$ 为去势的名义货币供给对数;则结构新息的假设意味着,对产出的意外冲击(unexpected shocks to output)与对货币供给的意外冲击不相关。又比如, $y_{1t}$ 为实际 GDP 增长率, $y_{2t}$ 为失业率;则 $\varepsilon_{1t}$ 与 $\varepsilon_{2t}$ 可分别解释为需求冲击(demand shock)与供给冲击(supply shock),而需求冲击(例如消费者偏好变化)与供给冲击(例如石油价格波动)不相关(Blanchard and Quah, 1989)。

将方程组(24.35)写为矩阵形式可得

$$\underbrace{\begin{pmatrix} 1 & a_{12} \\ a_{21} & 1 \end{pmatrix}}_A \underbrace{\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}}_{\mathbf{y}_t} = \underbrace{\begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}}_{\Gamma_1} \underbrace{\begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix}}_{\mathbf{y}_{t-1}} + \underbrace{\begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}}_{\boldsymbol{\varepsilon}_t} \quad (24.37)$$

更简洁地,上式可写为

$$\mathbf{A} \mathbf{y}_t = \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \quad (24.38)$$

其中,矩阵 $\mathbf{A}$ 反映了 $y_{1t}$ 与 $y_{2t}$ 的当期互动,即内生性。假设矩阵 $\mathbf{A}$ 非退化,在方程(24.38)两边同时左乘 $\mathbf{A}^{-1}$ ,即可得到其相应的简化式 VAR(reduced-form VAR):

$$\mathbf{y}_t = \mathbf{A}^{-1} \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \mathbf{A}^{-1} \boldsymbol{\varepsilon}_t \quad (24.39)$$

其中,简化式 VAR 的扰动项 $\mathbf{u}_t \equiv \mathbf{A}^{-1} \boldsymbol{\varepsilon}_t$ 为结构式 VAR 扰动项 $\boldsymbol{\varepsilon}_t$ 的线性组合。简化式扰动项的协方差矩阵为

$$\text{Var}(\mathbf{u}_t) = \text{Var}(\mathbf{A}^{-1} \boldsymbol{\varepsilon}_t) = \mathbf{A}^{-1} \text{Var}(\boldsymbol{\varepsilon}_t) \mathbf{A}^{-1'} \quad (24.40)$$

其中,Var( $\boldsymbol{\varepsilon}_t$ )为对角矩阵;但Var( $\mathbf{u}_t$ )不再是对角矩阵,故包含3个参数。方程(24.38)可识别的必要条件(阶条件)是,结构 VAR(24.38)的待估参数个数小于或等于简化 VAR(24.39)的待估参数个数。在本例中,SVAR 的待估参数为8个(6个系数,2个方差),而 VAR 的待估参数为7个(4个系数,3个协方差)。因此,为了识别此 SVAR,至少需要对方程(24.38)施加一个约束,比如 $a_{12}=0$ (这意味着 $y_{2t}$ 对 $y_{1t}$ 无直接影响)。

下面,考虑一般形式的 SVAR。从 $p$ 阶简化 VAR 出发:

$$\mathbf{y}_t = \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \mathbf{u}_t \quad (24.41)$$

其中, $\mathbf{y}_t$ 为 $M \times 1$ 向量; $\mathbf{u}_t$ 为简化式扰动项,允许存在同期相关(contemporaneous correlation)。在方程(24.41)两边同时左乘某非退化矩阵 $\mathbf{A}$ :

$$\mathbf{A} \mathbf{y}_t = \mathbf{A} \boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \mathbf{A} \boldsymbol{\Gamma}_p \mathbf{y}_{t-p} + \mathbf{A} \mathbf{u}_t \quad (24.42)$$

经移项整理可得:

$$\mathbf{A} (\mathbf{I} - \boldsymbol{\Gamma}_1 L - \cdots - \boldsymbol{\Gamma}_p L^p) \mathbf{y}_t = \mathbf{A} \mathbf{u}_t \quad (24.43)$$

我们希望 SVAR 的扰动项正交,一种简单的作法为令 $\mathbf{A} \mathbf{u}_t = \boldsymbol{\varepsilon}_t$ ,其中 $\boldsymbol{\varepsilon}_t$ 为 SVAR 的结构扰动项,不存在同期相关;但此假定可能过强(矩阵 $\mathbf{A}$ 来自经济理论对经济结构的建模,未必能同时使得 $\mathbf{A} \mathbf{u}_t$ 同期不相关)。更一般地,假设 $\mathbf{A} \mathbf{u}_t = \mathbf{B} \boldsymbol{\varepsilon}_t$ ,其中 $\mathbf{B}$ 为 $M \times M$ 矩阵;则方程(24.43)可写为

$$\mathbf{A} (\mathbf{I} - \boldsymbol{\Gamma}_1 L - \cdots - \boldsymbol{\Gamma}_p L^p) \mathbf{y}_t = \mathbf{A} \mathbf{u}_t = \mathbf{B} \boldsymbol{\varepsilon}_t \quad (24.44)$$

其中,结构扰动项  $\boldsymbol{\varepsilon}_t$  的协方差矩阵被标准化为单位矩阵  $\mathbf{I}_M$ 。方程(24.44)称为 SVAR 的“AB 模型”(AB - Model)(Amisano and Giannini, 1997)。对于传统的联立方程模型,分析的重点在于解释变量的边际效应,故一般不要求结构扰动项正交;而对于 AB 模型,分析的重点在于正交化冲击的效应,故一般假设结构扰动项  $\boldsymbol{\varepsilon}_t$  正交。如果令  $\mathbf{A} = \mathbf{I}_M$ , 则为 B 模型;如果令  $\mathbf{B} = \mathbf{I}_M$ , 则为 A 模型。A 模型与 B 模型都是 AB 模型的特例。在方程(24.44)两边同时左乘  $\mathbf{A}^{-1}$ , 即可得到相应的简化 VAR:

$$\mathbf{y}_t = \underbrace{\boldsymbol{\Gamma}_1 \mathbf{y}_{t-1} + \cdots + \boldsymbol{\Gamma}_p \mathbf{y}_{t-p}}_{\mathbf{u}_t} + \mathbf{A}^{-1} \mathbf{B} \boldsymbol{\varepsilon}_t \quad (24.45)$$

由于  $\mathbf{u}_t = \mathbf{A}^{-1} \mathbf{B} \boldsymbol{\varepsilon}_t$ , 故简化式扰动项  $\mathbf{u}_t$  的协方差矩阵为

$$\text{Var}(\mathbf{u}_t) = \mathbf{A}^{-1} \mathbf{B} \mathbf{B}' \mathbf{A}^{-1} \quad (24.46)$$

对于结构 VAR 模型(24.44),其待估参数总数为“ $M^2$ (矩阵  $\mathbf{A}$  的参数个数) +  $M^2$ (矩阵  $\mathbf{B}$  的参数个数) +  $pM^2$ (矩阵  $\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_p$  的参数个数)”,即  $2M^2 + pM^2$ 。另一方面,对于简化 VAR 模型(24.45),其待估计参数总数为“ $M(M+1)/2$ (对称协方差矩阵  $\text{Var}(\mathbf{u}_t)$  的参数个数) +  $pM^2$ (矩阵  $\boldsymbol{\Gamma}_1, \dots, \boldsymbol{\Gamma}_p$  的参数个数)”,即  $[M(M+1)/2] + pM^2$ 。因此,在一般情况下,SVAR 的参数比 VAR 的参数多出  $[2M^2 - M(M+1)/2]$  个。

因此,为了识别 AB 模型(24.44),至少需要对矩阵  $\mathbf{A}$  与  $\mathbf{B}$  中元素施加  $[2M^2 - M(M+1)/2]$  个约束。即使将矩阵  $\mathbf{A}$  的主对角线元素都标准化为 1,也还需要附加  $[2M^2 - M - M(M+1)/2]$  个约束条件。如果正好施加如此多约束,则为恰好识别;如果施加更多约束,则为过度识别。此阶条件(order condition)为识别 AB 模型的必要条件<sup>①</sup>。

为了估计 SVAR 模型,一般假设结构扰动项  $\boldsymbol{\varepsilon}_t$  服从多维正态分布,即  $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I}_M)$ ,然后进行带约束条件的最大似然估计(通常无解析解,须数值计算)。虽然此 MLE 估计量在多维正态的假设下导出,但在更弱的条件下,QMLE 估计量依然是一致的。

在具体操作上,如何施加约束条件,是一门艺术。一般来说,应从经济理论或对简化式 VAR 的估计结果出发,来设置约束条件。比较常用的方法沿用了乔利斯基分解的思路,将矩阵  $\mathbf{A}$  设为下三角矩阵且主对角线元素全部为 1,并将矩阵  $\mathbf{B}$  设为对角矩阵,称为“乔利斯基约束”(Cholesky restrictions)。以  $M=3$  为例,约束条件可写为

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ \cdot & 1 & 0 \\ \cdot & \cdot & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \cdot & 0 & 0 \\ 0 & \cdot & 0 \\ 0 & 0 & \cdot \end{pmatrix} \quad (24.47)$$

其中,缺失值“.”表示自由参数(即没有约束)。在上式中,从矩阵  $\mathbf{A}$  的第一行可以看出,  $y_{2t}$  与  $y_{3t}$  对  $y_{1t}$  无直接影响。类似地,从矩阵  $\mathbf{A}$  的第二行可看出,  $y_{1t}$  对  $y_{2t}$  有直接影响,但  $y_{3t}$  对  $y_{2t}$  无直接影响。最后,从矩阵  $\mathbf{A}$  的第三行可看出,  $y_{1t}$  与  $y_{2t}$  对  $y_{3t}$  都有直接影响。显然,使用乔利斯基约束来识别 SVAR,其估计结果依赖于变量的次序。因此,对于所选择的特定变量次序,需要从理论上进行说明,或进行敏感度分析,即变换变量次序,并对比其结果。

针对矩阵  $\mathbf{A}$  与  $\mathbf{B}$  所施加的约束也称为“短期约束”(short-run restrictions),其 SVAR 模型称为“短期 SVAR”(short-run SVAR)。另一类约束为“长期约束”(long-run restrictions),即对结构冲击  $\boldsymbol{\varepsilon}_t$  对于  $\mathbf{y}_t$  的长期效应进行约束,由 Blanchard and Quah(1989)所首倡;其 SVAR 模型称为

<sup>①</sup> 充分必要条件为秩条件(rank condition),在此从略;详见 Lutkepohl(2005, p. 365)。

“长期 SVAR”(long-run SVAR)。比如,根据货币中性假说(money neutrality hypothesis),货币在长期内是中性的,即货币供给在长期内对于实际产出的累积影响为零。从简化式 VAR(24.45)出发,可推导出 SVAR 模型的脉冲响应函数。进一步,根据与第 20 章类似的推导可知,结构冲击  $\varepsilon_t$  对  $y_t$  的长期效应为

$$C \equiv (I - \Gamma_1 - \cdots - \Gamma_p)^{-1} A^{-1} B \quad (24.48)$$

因此,在长期内可将 SVAR 模型简洁地写为

$$y_t = C\varepsilon_t \quad (24.49)$$

例如,假设  $M=2$ ,第一个变量为实际 GDP,第二个变量为货币供给;则可约束长期效应矩阵  $C$  的(1,2)元素为 0,即对第二个变量(货币供给)的结构冲击在长期内对第一个变量(实际 GDP)无作用。

## 24.7 SVAR 的 Stata 实例

短期 SVAR 的 Stata 命令格式为

```
svar y1 y2 y3, aconstraints(constraints_a) aeq(matrix_aeq) acns(matrix_acns) bconstraints(constraints_a) beq(matrix_aeq) bcns(matrix_acns) varconstraints(constraints_v) lags(numlist)
```

此命令的前三个选择项提供了对矩阵  $A$  施加约束的三种方法,其中“aconstraints(constraints\_a)”表示由命令 constraint 所定义的约束;“aeq(matrix\_aeq)”表示由命令 matrix 所定义的约束,该矩阵的元素取值或为实数(可以为 0),或为缺失值“.”(表示自由参数);“acns(matrix\_acns)”表示由命令 matrix 所定义的跨参数约束,其矩阵元素取值或为缺失值“.”(表示自由参数),或为 0,或为正整数,而且每个正整数至少出现两次,比如,所有取值为“1”的元素都相等,而所有取值为“2”的元素也都相等,但取值为 1 的元素可以不同于取值为 2 的元素。类似地,选择项“bconstraints(constraints\_a)”,“beq(matrix\_aeq)”与“bcns(matrix\_acns)”表示对矩阵  $B$  施加这三种约束。选择项“varconstraints(constraints\_v)”表示对简化 VAR 的参数进行约束。最后,选择项“lags(numlist)”表示简化 VAR 的阶数,比如“lags(1/4)”表示 VAR(4);默认为“lags(1/2)”,即 VAR(2)。

下面使用 Stata 提供的数据集 lutkepohl2.dta 进行演示。该数据集包含了联邦德国的季度宏观变量(已经过季节调整):inv(投资),inc(收入),consump(消费),ln\_inv(投资的对数),ln\_inc(收入的对数),ln\_consump(消费的对数),dln\_inv(ln\_inv 的一阶差分),dln\_inc(ln\_inc 的一阶差分),dln\_consump(ln\_consump 的一阶差分)。下面估计一个关于(dln\_inv, dln\_inc, dln\_consump)的 SVAR 模型(对数一阶差分可解释为百分比变化),并对矩阵  $A$  与  $B$  施以乔利斯基约束,参见表达式(24.47)。这意味着,我们假定投资不受收入与消费的当期影响;收入受到投资的当期影响,但不受消费的当期影响;而消费受到投资与收入的当期影响<sup>①</sup>。

<sup>①</sup> 此假定并无太多理论支持,故应变换变量次序,进行敏感度分析(比如,比较不同变量排序的脉冲响应图),在此从略。

首先, 定义矩阵  $A$  与  $B$ 。

```
. use lutkepohl2.dta, clear
. matrix A = (1,0,0 \ . ,1,0 \ . . ,1)
. matrix B = (. ,0,0 \ 0 ,.,0 \ 0,0 ..)
```

看一下定义矩阵  $A$  与  $B$  的结果。

```
. matrix list A
```

A[ 3, 3]			
	c1	c2	c3
r1	1	0	0
r2	.	1	0
r3	.	.	1

```
. matrix list B
```

symmetric B[3,3]			
	c1	c2	c3
r1	.	.	.
r2	0	.	.
r3	0	0	.

其次, 确定 VAR 模型的滞后阶数。

```
. varsoc dln_inv dln_inc dln_consump
```

Selection-order criteria							
Sample: 1961q2 - 1982q4				Number of obs = 87			
lag	LL	LR	df	p	FPE	AIC	HQIC

0	696.398				2.4e-11	-15.9402	-15.9059	-15.8552*
1	711.682	30.568	9	0.000	2.1e-11	-16.0846	-15.9477*	-15.7445
2	724.696	26.028	9	0.002	1.9e-11*	-16.1769*	-15.9372	-15.5817
3	729.124	8.8557	9	0.451	2.1e-11	-16.0718	-15.7294	-15.2215
4	738.353	18.458*	9	0.030	2.1e-11	-16.0771	-15.632	-14.9717

Endogenous: dln\_inv dln\_inc dln\_consump

Exogenous: \_cons

根据 AIC 准则, 可使用默认的二阶 VAR<sup>①</sup>。下面, 使用对矩阵  $A$  与  $B$  的上述约束, 估计 SVAR 模型。

```
. svar dln_inv dln_inc dln_consump, aeq(A) beq(B) nolog
```

<sup>①</sup> 根据 BIC 准则, 可使用一阶 VAR。为了保证扰动项为白噪声, 此处使用二阶 VAR。

Estimating short-run parameters						
Structural vector autoregression						
(1)	[a_1_1]_cons	=	1			
(2)	[a_1_2]_cons	=	0			
(3)	[a_1_3]_cons	=	0			
(4)	[a_2_2]_cons	=	1			
(5)	[a_2_3]_cons	=	0			
(6)	[a_3_3]_cons	=	1			
(7)	[b_1_2]_cons	=	0			
(8)	[b_1_3]_cons	=	0			
(9)	[b_2_1]_cons	=	0			
(10)	[b_2_3]_cons	=	0			
(11)	[b_3_1]_cons	=	0			
(12)	[b_3_2]_cons	=	0			
Sample:	1960q4 - 1982q4			No. of obs	=	89
Exactly identified model				Log likelihood	=	742.2131
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
/a_1_1	1	(constrained)				
/a_2_1	-.031361	.0266532	-1.18	0.239	-.0836003	.0208782
/a_3_1	-.0566791	.0187076	-3.03	0.002	-.0933453	-.0200128
/a_1_2	0	(omitted)				
/a_2_2	1	(constrained)				
/a_3_2	-.4792407	.0738282	-6.49	0.000	-.6239414	-.3345399
/a_1_3	0	(omitted)				
/a_2_3	0	(omitted)				
/a_3_3	1	(constrained)				
/b_1_1	.0425173	.0031868	13.34	0.000	.0362712	.0487633
/b_2_1	0	(omitted)				
/b_3_1	0	(omitted)				
/b_1_2	0	(omitted)				
/b_2_2	.0106908	.0008013	13.34	0.000	.0091202	.0122613
/b_3_2	0	(omitted)				
/b_1_3	0	(omitted)				
/b_2_3	0	(omitted)				
/b_3_3	.0074461	.0005581	13.34	0.000	.0063522	.0085399

上表上部列出了所有的约束条件,共有12个,为恰好识别(exactly identified)。上表下部列出了对矩阵A与B自由元素的估计值及标准误。其中,对 $a_{21}, a_{31}, a_{32}$ 的估计值均为负,经移项后可知,这些当期效应均为正(与理论预期相符)。估计SVAR的主要兴趣通常在于考察其脉冲响应函数。为此,输入如下命令,结果参见图24.1。

```
. irf create germany, set(germany)
(file germany.irf created)
(file germany.irf now active)
(file germany.irf updated)
.irf graph sirf
```

其中,“sirf”表示“结构脉冲响应函数”(structural irf),等价于“正交化的脉冲响应函数”(如果

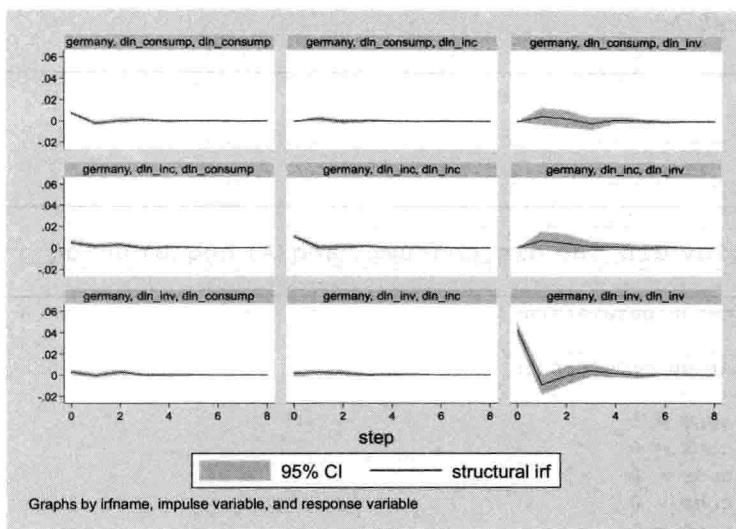


图 24.1 SVAR 的结构脉冲响应函数

输入命令“`irf graph oirf`”，可得到完全相同的结果）。

下面，考察 SVAR 模型的预测误差方差分解，结果参见图 24.2。

. irf graph sfevd

其中，“`sfevd`”表示“结构预测误差方差分解”(structural FEVD)。此命令等价于“`irf graph fevd`”。

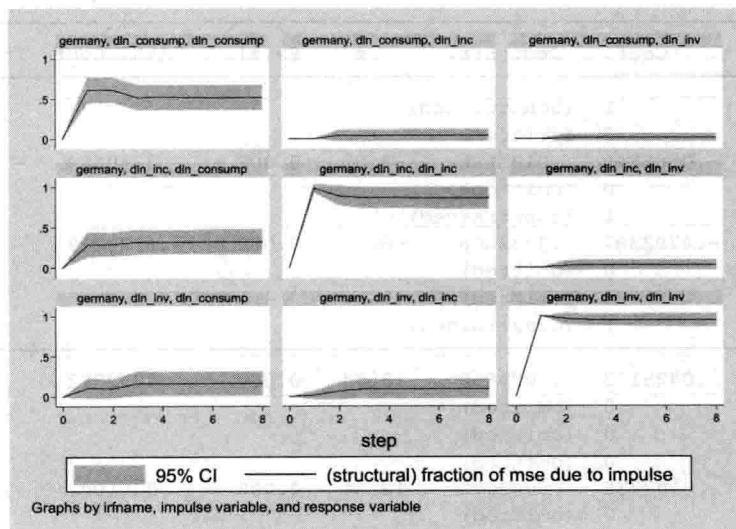


图 24.2 SVAR 的结构预测误差方差分解

以上为恰好识别的 SVAR 模型。注意到，矩阵  $A$  的  $a_{21}$  元素并不显著 ( $p$  值为 0.239)。为了演示过度识别的 SVAR 模型，下面约束  $a_{21} = 0$ ，重新定义矩阵  $A$  如下（矩阵  $B$  的定义不变）。

. matrix A = (1, 0, 0 \ 0, 1, 0 \ ., ., 1)

```
. matrix B = (.,0,0\0,.,0\0,0,.)
. matrix list A
```

A[3,3]
c1 c2 c3
r1 1 0 0
r2 0 1 0
r3 . . 1

```
. svar dln_inv dln_inc dln_consump, aeq(A) beq(B) nolog
```

Estimating short-run parameters

Structural vector autoregression

- ( 1) [a\_1\_1]\_cons = 1
- ( 2) [a\_1\_2]\_cons = 0
- ( 3) [a\_1\_3]\_cons = 0
- ( 4) [a\_2\_1]\_cons = 0
- ( 5) [a\_2\_2]\_cons = 1
- ( 6) [a\_2\_3]\_cons = 0
- ( 7) [a\_3\_3]\_cons = 1
- ( 8) [b\_1\_2]\_cons = 0
- ( 9) [b\_1\_3]\_cons = 0
- (10) [b\_2\_1]\_cons = 0
- (11) [b\_2\_3]\_cons = 0
- (12) [b\_3\_1]\_cons = 0
- (13) [b\_3\_2]\_cons = 0

Sample: 1960q4 - 1982q4

No. of obs = 89

Overidentified model

Log likelihood = 741.5262

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/a_1_1	1	(constrained)			
/a_2_1	0	(omitted)			
/a_3_1	-.0566846	.0185638	-3.05	0.002	-.0930689 -.0203002
/a_1_2	0	(omitted)			
/a_2_2	1	(constrained)			
/a_3_2	-.4792397	.0732606	-6.54	0.000	-.6228279 -.3356515
/a_1_3	0	(omitted)			
/a_2_3	0	(omitted)			
/a_3_3	1	(constrained)			
/b_1_1	.0425173	.0031868	13.34	0.000	.0362712 .0487633
/b_2_1	0	(omitted)			
/b_3_1	0	(omitted)			
/b_1_2	0	(omitted)			
/b_2_2	.0107736	.0008075	13.34	0.000	.0091909 .0123563
/b_3_2	0	(omitted)			
/b_1_3	0	(omitted)			
/b_2_3	0	(omitted)			
/b_3_3	.0074461	.0005581	13.34	0.000	.0063522 .0085399

LR test of identifying restrictions: chi2( 1)= 1.374 Prob > chi2 = 0.241

上表显示,共有 13 个约束条件(比恰好识别多出 1 个约束条件)。参数估计值及显著性变化不大。上表底部提供了对过度识别的似然比检验结果,该检验的原假设为“所有约束条件均成立”,结果无法拒绝此原假设。

长期 SVAR 的 Stata 命令基本格式为

```
svar y1 y2 y3, lrconstraints (constraints_lr) lreq (matrix_lreq) lrcns  
(matrix_lrcns) varconstraints (constraints_v) lags (numlist)
```

此命令的前三个选择项提供了对长期效应矩阵  $C$  施加约束的三种方法,其含义与短期 SVAR 命令相同。其余选择项的含义也类似。下面以 Stata 提供的季度数据集 m1gdp.dta 为例。该数据集包括的主要变量为 ln\_gdp(经季节调整的实际 GDP 对数)与 ln\_m1(经季节调整的货币供给量 M1 对数)。根据货币中性假说,货币供给对实际产出的长期效应为 0。因此,对长期效应矩阵  $C$  作如下约束:

$$C = \begin{pmatrix} \cdot & 0 \\ \cdot & \cdot \end{pmatrix} \quad (24.50)$$

首先,定义矩阵  $C$ 。

```
. use m1gdp.dta, clear  
. matrix C = (. , 0 \ . , .)  
. matrix list C
```

C[2,2]
c1 c2
r1 . 0
r2 . .

其次,根据信息准则,确定简化 VAR 的滞后阶数。

```
. varsoc ln_gdp ln_m1
```

Selection-order criteria								
Number of obs = 170								
lag	LL	LR	df	p	FPE	AIC	HQIC	SBIC
0	13.412				.002997	-.134258	-.119288	-.097367
1	1084.15	2141.5	4	0.000	1.1e-08	-12.6841	-12.6392	-12.5734
2	1142.65	117.01	4	0.000	5.6e-09	-13.3253	-13.2504	-13.1408*
3	1149.76	14.217*	4	0.007	5.4e-09*	-13.3619*	-13.2571*	-13.1036
4	1153.27	7.0269	4	0.134	5.4e-09	-13.3561	-13.2214	-13.0241

Endogenous: ln\_gdp ln\_m1

Exogenous: \_cons

上表显示,大多数信息准则支持 VAR(3)。下面,估计长期 SVAR 模型。

```
. svar ln_gdp ln_m1, lreq(C) lags(1/3) nolog
```

Estimating long-run parameters						
Structural vector autoregression						
(1)	[c_1_2]_cons = 0					
Sample:	1959q4 - 2002q2			No. of obs	=	171
Exactly identified model				Log likelihood	=	1154.855
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
/c_1_1	3.853233	.208359	18.49	0.000	3.444857	4.261609
/c_2_1	7.419362	.4066705	18.24	0.000	6.622303	8.216422
/c_1_2	0	(omitted)				
/c_2_2	.8698669	.047037	18.49	0.000	.7776761	.9620578

上表显示,此长期 SVAR 模型为恰好识别,且  $C$  矩阵中的所有元素都很显著。

## 习 题

24.1 证明方程(24.31)。

24.2 对于数据集 supDem.dta, 考虑以下商品市场的供需均衡模型:

$$\begin{cases} \text{quantity}_t = \alpha_0 + \alpha_1 \text{ price}_t + \alpha_2 \text{ pcompete}_t + u_t \\ \text{quantity}_t = \beta_0 + \beta_1 \text{ price}_t + \beta_2 \text{ praw}_t + v_t \end{cases} \quad (24.51)$$

其中,第一个方程为需求函数,第二个方程为供给函数, quantity 为商品数量, price 为商品价格, pcompete 为替代商品的价格, praw 为原材料价格。

- (1) 需求方程与供给方程为不可识别、恰好识别或过度识别?
- (2) 使用 OLS 分别估计需求与供给方程;
- (3) 使用 2SLS 分别估计需求与供给方程;
- (4) 使用 3SLS 估计整个方程系统(提示:由于两个方程的被解释变量均为 quantity, 故在使用 Stata 命令 reg3 时,应加上选择项“endog(price)”,指定 price 也为内生变量)。

# 第25章 非线性回归与门限回归

## 25.1 非线性最小二乘法

对于非线性回归模型,除了使用最大似然估计法(MLE),还可以使用“非线性最小二乘法”(Nonlinear Least Square,简记NLS)。考虑以下非线性回归模型:

$$y_i = g(x_i, \beta) + \varepsilon_i \quad (i=1, \dots, n) \quad (25.1)$$

其中,  $\beta$  为  $K$  维未知(真实)参数向量,而  $g(\cdot)$  是  $\beta$  的非线性函数,且无法通过变量转换变为  $\beta$  的线性函数。如果  $g(x_i, \beta) = x'_i \beta$ , 则回到古典线性回归模型。记  $\tilde{\beta}$  为  $\beta$  的一个假想值(hypothetical value),其对应的残差为  $e_i = y_i - g(x_i, \tilde{\beta})$ 。非线性最小二乘法通过选择  $\tilde{\beta}$ ,使得残差平方和最小:

$$\min_{\tilde{\beta}} \text{SSR}(\tilde{\beta}) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n [y_i - g(x_i, \tilde{\beta})]^2 \quad (25.2)$$

最小化的一阶条件为

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}} = -2 \sum_{i=1}^n [y_i - g(x_i, \tilde{\beta})] \frac{\partial g(x_i, \tilde{\beta})}{\partial \tilde{\beta}} = \mathbf{0} \quad (25.3)$$

可以简化为

$$\sum_{i=1}^n [y_i - g(x_i, \tilde{\beta})] \frac{\partial g(x_i, \tilde{\beta})}{\partial \tilde{\beta}} = \mathbf{0} \quad (25.4)$$

$$\sum_{i=1}^n e_i \frac{\partial g(x_i, \tilde{\beta})}{\partial \tilde{\beta}} = \mathbf{0} \quad (25.5)$$

这是一个  $K$  个方程、 $K$  个未知数的非线性方程组。满足这个非线性方程组的估计量被称为“非线性最小二乘估计量”,记为  $\hat{\beta}_{\text{NLS}}$ 。方程(25.5)表明,残差向量  $e$  与  $\frac{\partial g(x, \tilde{\beta})}{\partial \tilde{\beta}}$  正交,而不是与  $x$  正交(线性回归的情形)。这个非线性方程组通常没有解析解,要使用数值迭代方法来求解,比如牛顿-拉弗森法。

例① 考虑如下非线性回归模型:

$$y_i = \beta_1 + \beta_2 \exp(\beta_3 x_i) + \varepsilon_i \quad (25.6)$$

① 此例来自 Greene(2003, p. 165)。

显然,这个模型含有三个未知参数( $\beta_1, \beta_2, \beta_3$ ),即  $K=3$ 。使用 NLS 进行估计,残差平方和为

$$\min_{\tilde{\beta}} \text{SSR}(\tilde{\beta}) = \sum_{i=1}^n [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)]^2 \quad (25.7)$$

NLS 估计量的一阶条件为

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_1} = -2 \sum_{i=1}^n [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] = 0 \quad (25.8)$$

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_2} = -2 \sum_{i=1}^n [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] \exp(\tilde{\beta}_3 x_i) = 0 \quad (25.9)$$

$$\frac{\partial \text{SSR}(\tilde{\beta})}{\partial \tilde{\beta}_3} = -2 \sum_{i=1}^n [y_i - \tilde{\beta}_1 - \tilde{\beta}_2 \exp(\tilde{\beta}_3 x_i)] \tilde{\beta}_2 x_i \exp(\tilde{\beta}_3 x_i) = 0 \quad (25.10)$$

这是一个含有三个方程、三个未知数的非线性方程组,没有解析解,只能用数值方法求解。

#### NLS 的大样本性质

可以证明,如果  $E(\varepsilon_i | x_i) = 0$ (即扰动项与当期解释变量不相关),再加上一些技术性条件,则  $\hat{\beta}_{\text{NLS}}$  为真实参数  $\beta$  的一致估计量,且  $\hat{\beta}_{\text{NLS}}$  服从渐近正态分布<sup>①</sup>。进一步地,如果扰动项为球形扰动项(即满足同方差与无自相关的假设),则  $\hat{\beta}_{\text{NLS}}$  是渐近有效的(asymptotically efficient)。

## 25.2 非线性回归的 Stata 命令及实例

NLS 的 Stata 命令格式为

`n1 (depvar = <sexp> [,options])`

其中,“sexp”表示“substitutable expression”,用来定义非线性回归模型。比如,对于方程(25.6),可以使用以下命令:

`n1 (y = {beta1} + {beta2} * exp({beta3 = 1} * x)), robust`

其中,括弧“{}”中表示的是待估参数,“{beta3 = 1}”表示  $\beta_3$  的初始值为 1。如果担心存在异方差,可以使用选择项“robust”以得到稳健标准误。

下面以数据集 `consumption_china.dta` 为例,来估计以下非线性消费函数:

$$c_t = \beta_1 + \beta_2 y_t^{\beta_3} + \varepsilon_t \quad (25.11)$$

由于参数  $\beta_3$  未知,故这是非线性回归。如果  $\beta_3 = 1$ ,则为线性回归。因此,选择  $\beta_3 = 1$  作为迭代计算的初始值:

```
. use consumption_china.dta, clear
. n1 (c = {beta1} + {beta2} * y^{beta3 = 1}), nolog
```

① 参见 Cameron and Trivedi(2005), Proposition 5.6, p. 153。

Source	SS	df	MS			
Model	93414757.2	2	46707378.6	Number of obs =	29	
	184339.472	26	7089.97971	R-squared =	0.9980	
				Adj R-squared =	0.9979	
Total	93599096.7	28	3342824.88	Root MSE =	84.20202	
				Res. dev.	=	336.2584
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
/beta1	-235.1029	55.98706	-4.20	0.000	-350.186	-120.0199
/beta2	3.942834	.8282788	4.76	0.000	2.240283	5.645386
/beta3	.7634422	.0215016	35.51	0.000	.7192451	.8076393

Parameter beta1 taken as constant term in model & ANOVA table

从上表可知,非线性参数  $\beta_3$  的  $p$  值为 0.000,这说明非线性消费函数的模型设定有其合理性<sup>①</sup>,而线性消费函数过于简化了。在线性消费函数的假设下,边际消费倾向为常数。事实上,由于  $\hat{\beta}_3 = 0.76$ ,故边际消费倾向随着收入的增加而递减。

下面用稳健标准误重新进行估计:

```
. nl(c = {beta1} + {beta2} * y^{beta3 = 1}), r nolog
```

Nonlinear regression			Number of obs = 29 R-squared = 0.9980 Adj R-squared = 0.9979 Root MSE = 84.20202 Res. dev. = 336.2584			
c	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
/beta1	-235.1029	47.5145	-4.95	0.000	-332.7704	-137.4355
/beta2	3.942834	.88375	4.46	0.000	2.12626	5.759408
/beta3	.7634422	.0229917	33.21	0.000	.7161821	.8107023

Parameter beta1 taken as constant term in model

总的来说,稳健标准误与普通标准误相差不大。这说明,大概不存在异方差问题。

## 25.3 门限回归

在回归分析中,我们常常关心系数估计值是否稳定,即如果将整个样本分成若干个子样本(subsample)分别进行回归,是否还能得到大致相同的估计系数。对于时间序列数据,这意味着经济结构是否随着时间的推移而改变(参见第9章)。对于横截面数据,比如,样本中有男性与女性,则可以根据性别将样本一分为二,分别估计男性样本与女性样本。如果用来划分样本的变量不是离散型变量而是连续型变量,比如,企业规模、人均国民收入,则需要给出一个划分的标

① 也可以引入收入的平方项,而仍然使用线性回归。

准,即“门限(门槛)值”(threshold level)。

在应用研究中,人们常常怀疑大企业与小企业的投资行为不同,那么如何区分大企业与小企业呢?另外,受到流动性约束(liquidity constraint)的企业与没有流动性约束企业的投资行为也可能不同,如何通过债务股本比(debt to equity ratio)或其他指标来区分这两类企业?再比如,发达国家与发展中国家的经济增长规律可能不同,如何通过人均国民收入这一指标来区分一个国家发达与否?总之,经济规律可能是非线性的,其函数形式可能依赖于某个变量(称为“门限变量”)而改变。

传统的做法是,由研究者主观(随意)地确定一个门限值,然后根据此门限值把样本一分为二(或分成更多子样本),既不对门限值进行参数估计,也不对其显著性进行统计检验。显然,这样得到的结果并不可靠。为此,Hansen(2000)提出“门限(门槛)回归”(threshold regression),以严格的统计推断方法对门限值进行参数估计与假设检验。

假设样本数据为 $\{y_i, \mathbf{x}_i, q_i\}_{i=1}^n$ ,其中 $q_i$ 为用来划分样本的“门限变量”(threshold variable), $q_i$ 可以是解释变量 $\mathbf{x}_i$ 的一部分。考虑以下门限回归模型:

$$\begin{cases} y_i = \boldsymbol{\beta}'_1 \mathbf{x}_i + \varepsilon_i, & \text{若 } q_i \leq \gamma \\ y_i = \boldsymbol{\beta}'_2 \mathbf{x}_i + \varepsilon_i, & \text{若 } q_i > \gamma \end{cases} \quad (25.12)$$

其中, $\gamma$ 为待估计的门限值, $\mathbf{x}_i$ 为外生解释变量,与扰动项 $\varepsilon_i$ 不相关。可以将上面这个分段函数合并写为

$$y_i = \underbrace{\boldsymbol{\beta}'_1 \mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma)}_{=z_{i1}} + \underbrace{\boldsymbol{\beta}'_2 \mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma)}_{=z_{i2}} + \varepsilon_i \quad (25.13)$$

其中, $\mathbf{1}(\cdot)$ 为示性函数,即如果括号中的表达式为真,则取值为1;反之,取值为0。显然,这是一个非线性回归,因为它无法写成参数( $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \gamma$ )的线性函数。可以用非线性最小二乘法(NLS)来估计,即最小化残差平方和。事实上,如果 $\gamma$ 的取值已知,则可以通过定义 $z_{i1} \equiv \mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma)$ 与 $z_{i2} \equiv \mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma)$ ,将方程(25.13)转化为参数为( $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ )的线性回归模型:

$$y_i = \boldsymbol{\beta}'_1 z_{i1} + \boldsymbol{\beta}'_2 z_{i2} + \varepsilon_i \quad (25.14)$$

因此,在实际计算上,常分两步来最小化残差平方和。首先,给定 $\gamma$ 的取值,对方程(25.14)使用OLS来估计 $\hat{\boldsymbol{\beta}}_1(\gamma)$ 与 $\hat{\boldsymbol{\beta}}_2(\gamma)$ (显然 $\hat{\boldsymbol{\beta}}_1$ 与 $\hat{\boldsymbol{\beta}}_2$ 依赖于 $\gamma$ ),并计算残差平方和 $\text{SSR}(\gamma)$ (称为Concentrated Sum of Squared Residuals),也是 $\gamma$ 的函数。其次,选择 $\gamma$ 使得 $\text{SSR}(\gamma)$ 最小化。注意到,给定 $q_i$ ,由于示性函数 $\mathbf{1}(q_i \leq \gamma)$ 与 $\mathbf{1}(q_i > \gamma)$ 只能取值0或1,故是 $\gamma$ 的阶梯函数,而“阶梯的升降点”正好是 $q_i$ (只有一级“台阶”)。由此可知, $\text{SSR}(\gamma)$ 也是 $\gamma$ 的阶梯函数,而阶梯的升降点恰好在 $\{q_i\}_{i=1}^n$ 不重叠的观测值上,因为如果 $\gamma$ 取 $\{q_i\}_{i=1}^n$ 以外的其他值,不会对子样本的划分产生影响,故不改变 $\text{SSR}(\gamma)$ 。因此,最多只需要考虑 $\gamma$ 取 $n$ 个值即可,即 $\gamma \in \{q_1, q_2, \dots, q_n\}$ 。这使得 $\text{SSR}(\gamma)$ 的最小化计算得以简化。记最后的参数估计量为 $(\hat{\boldsymbol{\beta}}_1(\hat{\gamma}), \hat{\boldsymbol{\beta}}_2(\hat{\gamma}), \hat{\gamma})$ 。

在一定的条件下,Hansen(2000)导出了 $\hat{\gamma}$ 的大样本渐近分布,在此基础上构造 $\hat{\gamma}$ 的置信区间,并对原假设“ $H_0: \gamma = \gamma_0$ ”进行似然比检验(参见下文)。

类似地,可以考虑包含“多个门限值”(multiple thresholds)的门限回归。比如,对于门限变量 $q_i$ ,假设两个门限值为 $\gamma_1 < \gamma_2$ ,则门限回归模型为

$$y_i = \boldsymbol{\beta}'_1 \mathbf{x}_i \cdot \mathbf{1}(q_i \leq \gamma_1) + \boldsymbol{\beta}'_2 \mathbf{x}_i \cdot \mathbf{1}(\gamma_1 < q_i \leq \gamma_2) + \boldsymbol{\beta}'_3 \mathbf{x}_i \cdot \mathbf{1}(q_i > \gamma_2) + \varepsilon_i \quad (25.15)$$

## 25.4 面板数据的门限回归

对于面板数据  $\{y_{it}, \mathbf{x}_{it}, q_{it}; 1 \leq i \leq n, 1 \leq t \leq T\}$ , 其中  $i$  表示个体,  $t$  表示时间, Hansen(1999) 考虑了如下的固定效应(fixed effects)门限回归模型:

$$\begin{cases} y_{it} = \mu_i + \boldsymbol{\beta}'_1 \mathbf{x}_{it} + \varepsilon_{it}, & \text{若 } q_{it} \leq \gamma \\ y_{it} = \mu_i + \boldsymbol{\beta}'_2 \mathbf{x}_{it} + \varepsilon_{it}, & \text{若 } q_{it} > \gamma \end{cases} \quad (25.16)$$

其中,  $q_{it}$  为门限变量(可以是解释变量  $\mathbf{x}_{it}$  的一部分),  $\gamma$  为待估计的门限值, 扰动项  $\varepsilon_{it}$  为独立同分布的。假设解释变量  $\mathbf{x}_{it}$  为外生变量, 与扰动项  $\varepsilon_{it}$  不相关。因此,  $\mathbf{x}_{it}$  不包含被解释变量  $y_{it}$  的滞后值, 不是动态面板(dynamic panel)。个体截距项  $\mu_i$  的存在表明, 这是固定效应模型。使用示性函数  $\mathbf{1}(\cdot)$ , 可以将模型更简洁地表示为

$$y_{it} = \mu_i + \boldsymbol{\beta}'_1 \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} \leq \gamma) + \boldsymbol{\beta}'_2 \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} > \gamma) + \varepsilon_{it} \quad (25.17)$$

假设  $n$  较大,  $T$  较小(短面板), 故大样本的渐近理论基于“ $n \rightarrow \infty$ ”而展开。定义  $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ ,

$\mathbf{x}_{it}(\gamma) = \begin{pmatrix} \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} \leq \gamma) \\ \mathbf{x}_{it} \cdot \mathbf{1}(q_{it} > \gamma) \end{pmatrix}$ , 则方程(25.17)可进一步简化为

$$y_{it} = \mu_i + \boldsymbol{\beta}' \mathbf{x}_{it}(\gamma) + \varepsilon_{it} \quad (25.18)$$

对于第  $i$  位个体, 将方程(25.18)两边对时间求平均可得

$$\bar{y}_i = \mu_i + \boldsymbol{\beta}' \bar{\mathbf{x}}_i(\gamma) + \bar{\varepsilon}_i \quad (25.19)$$

其中,  $\bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^T y_{it}$ ,  $\bar{\mathbf{x}}_i(\gamma) \equiv \frac{1}{T} \sum_{t=1}^T \mathbf{x}_{it}(\gamma)$ ,  $\bar{\varepsilon}_i \equiv \frac{1}{T} \sum_{t=1}^T \varepsilon_{it}$ 。将方程(25.18)减去方程(25.19), 可得模型的离差形式:

$$y_{it} - \bar{y}_i = \boldsymbol{\beta}' [\mathbf{x}_{it}(\gamma) - \bar{\mathbf{x}}_i(\gamma)] + (\varepsilon_{it} - \bar{\varepsilon}_i) \quad (25.20)$$

记  $y_{it}^* \equiv y_{it} - \bar{y}_i$ ,  $\mathbf{x}_{it}^*(\gamma) \equiv \mathbf{x}_{it}(\gamma) - \bar{\mathbf{x}}_i(\gamma)$ ,  $\varepsilon_{it}^* \equiv \varepsilon_{it} - \bar{\varepsilon}_i$ , 则可得

$$y_{it}^* = \boldsymbol{\beta}' \mathbf{x}_{it}^*(\gamma) + \varepsilon_{it}^* \quad (25.21)$$

仍然使用两步法进行估计。首先, 给定  $\gamma$  的取值, 用 OLS 对方程(25.21)进行一致估计(组内估计量), 得到估计系数  $\hat{\boldsymbol{\beta}}(\gamma)$  以及残差平方和  $SSR(\gamma)$ 。其次, 对于  $\gamma \in \{q_{it}; 1 \leq i \leq n, 1 \leq t \leq T\}$  ( $\gamma$  最多有  $nT$  个可能取值), 选择  $\hat{\gamma}$ , 使得  $SSR(\hat{\gamma})$  最小。最后得到估计系数  $\hat{\boldsymbol{\beta}}(\hat{\gamma})$ 。如果不希望某个子样本中的观测值过少, 则可以限制  $\gamma$  的取值, 比如不考虑  $\{q_{it}\}$  中最大 5% 或最小 5% 的取值。

对于是否存在“门限效应”(threshold effect), 可以检验以下原假设:

$$H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \quad (25.22)$$

如果此原假设成立, 则不存在门限效应。此时, 模型简化为

$$y_{it} = \mu_i + \boldsymbol{\beta}'_1 \mathbf{x}_{it} + \varepsilon_{it} \quad (25.23)$$

对于这个标准的固定效应面板模型, 可以将其转化为离差形式, 然后用 OLS 来估计(组内估计量)。记在“ $H_0: \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ ”约束下所得到的残差平方和为  $SSR^*$ , 以区别于无约束的残差平方和  $SSR(\hat{\gamma})$ 。显然,  $SSR^* \geq SSR(\hat{\gamma})$ 。如果  $[SSR^* - SSR(\hat{\gamma})]$  越大, 加上约束条件后使得  $SSR$  增大

越多，则越应该倾向于拒绝“ $H_0: \beta_1 = \beta_2$ ”。

Hansen(1999)提出使用以下似然比检验(LR)统计量：

$$LR \equiv [SSR^* - SSR(\hat{\gamma})]/\hat{\sigma}^2 \quad (25.24)$$

其中， $\hat{\sigma}^2 \equiv \frac{SSR(\hat{\gamma})}{n(T-1)}$  为对扰动项方差的一致估计。然而，如果原假设“ $H_0: \beta_1 = \beta_2$ ”成立，则不存在门限效应，也就无所谓门限值  $\gamma$  等于多少。因此，在  $H_0$  成立的情况下，无论  $\gamma$  取什么值，对模型都没有影响，故参数  $\gamma$  不可识别。因此，检验统计量 LR 的渐近分布并非标准的  $\chi^2$  分布，而依赖于样本矩(sample moments)，无法将其临界值列表，但可以用自助法(bootstrap)来得到其临界值。

如果拒绝“ $H_0: \beta_1 = \beta_2$ ”，则认为存在门限效应，可以进一步对门限值进行检验，即检验“ $H_0: \gamma = \gamma_0$ ”。定义似然比检验统计量为

$$LR(\gamma) \equiv [SSR(\gamma) - SSR(\hat{\gamma})]/\hat{\sigma}^2 \quad (25.25)$$

可以证明，在“ $H_0: \gamma = \gamma_0$ ”成立的情况下， $LR(\gamma)$  的渐近分布虽然仍然是非标准的，但其累积分布函数为  $(1 - e^{-x/2})^2$ ，可以直接算出其临界值。由此，可以利用统计量  $LR(\gamma)$  来计算  $\gamma$  的置信区间。

类似地，可以考虑多门限值的面板回归模型。以两个门限值为例：

$$y_{it} = \mu_i + \beta_1' x_{it} \cdot \mathbf{1}(q_{it} \leq \gamma_1) + \beta_2' x_{it} \cdot \mathbf{1}(\gamma_1 < q_{it} \leq \gamma_2) + \beta_3' x_{it} \cdot \mathbf{1}(q_{it} > \gamma_2) + \varepsilon_{it} \quad (25.26)$$

其中，门限值  $\gamma_1 < \gamma_2$ 。同样地，可以将这个模型转换为离差形式，并仍用两步法进行估计。首先，给定  $(\gamma_1, \gamma_2)$ ，使用 OLS 估计离差模型，得到残差平方和  $SSR(\gamma_1, \gamma_2)$ 。其次，选择  $(\gamma_1, \gamma_2)$  使得  $SSR(\gamma_1, \gamma_2)$  最小化。

## 25.5 门限回归的计算机操作

目前，尚无门限回归的官方 Stata 程序。门限回归的发明者 Bruce Hansen 在其个人网站<sup>①</sup>提供了进行门限回归的 Matlab 与 Gauss 程序。

### 习题

**25.1** 使用数据集 mpyr.dta，估计货币流通速度的对数( $\log v$ )与名义利率( $r$ )的以下非线性回归模型：

$$\log v_i = \beta_1 + \beta_2 r_i^{\beta_3} + \varepsilon_i \quad (25.27)$$

将其结果与线性回归比较。其中，非线性参数  $\beta_3$  显著吗？

① 网址为 <http://www.ssc.wisc.edu/~bhansen/>，或搜索“Bruce Hansen economics”。

# 第26章 分位数回归

## 26.1 为什么需要分位数回归

在迄今为止的回归模型中,我们着重考察解释变量  $\mathbf{x}$  对被解释变量  $y$  的条件期望  $E(y|\mathbf{x})$  的影响,实际上均值回归。但我们真正关心的是  $\mathbf{x}$  对整个条件分布  $y|\mathbf{x}$  的影响,而条件期望  $E(y|\mathbf{x})$  只是刻画条件分布  $y|\mathbf{x}$  集中趋势的一个指标而已。如果条件分布  $y|\mathbf{x}$  不是对称分布 (symmetric distribution),则条件期望  $E(y|\mathbf{x})$  很难反映整个条件分布的全貌。如果能够估计出条件分布  $y|\mathbf{x}$  的若干重要的条件分位数 (conditional quantiles),比如中位数 (median)、 $1/4$  分位数 (lower quartile)、 $3/4$  分位数 (upper quartile),就能对条件分布  $y|\mathbf{x}$  有更全面的认识。另一方面,使用 OLS 的古典“均值回归”,由于最小化的目标函数为残差平方和 ( $\sum_{i=1}^n e_i^2$ ),故容易受极端值 (outliers) 的影响。

为此,Koenker and Bassett(1978)提出“分位数回归”(Quantile Regression,简记 QR),使用残差绝对值的加权平均(比如,  $\sum_{i=1}^n |e_i|$ )作为最小化的目标函数,故不易受极端值影响,较为稳健。更重要的是,分位数回归还能提供关于条件分布  $y|\mathbf{x}$  的全面信息。下面首先回顾有关总体分位数与样本分位数的概念。

## 26.2 总体分位数

假设  $Y$  为连续型随机变量,其累积分布函数为  $F_y(\cdot)$ ,则  $Y$  的“总体  $q$  分位数”(population  $q^{\text{th}}$  quantile,  $0 < q < 1$ ),记为  $y_q$ ,满足以下定义式:

$$q = P(Y \leq y_q) = F_y(y_q) \quad (26.1)$$

即总体  $q$  分位数  $y_q$  正好将总体分布分为两部分,其中小于或等于  $y_q$  的概率为  $q$ ,而大于  $y_q$  的概率为  $(1 - q)$ 。如果  $q = 1/2$ ,则为中位数,正好将总体分为两个相等的部分,一半在中位数之上,而另一半在中位数之下。如果  $F_y(\cdot)$  严格单调递增,则有

$$y_q = F_y^{-1}(q) \quad (26.2)$$

其中,  $F_y^{-1}(\cdot)$  为  $F_y(\cdot)$  的逆函数,参见图 26.1。

以上讨论的是单一变量的分位数。对于回归模型而言,记条件分布  $y|\mathbf{x}$  的累积分布函数为  $F_{y|\mathbf{x}}(\cdot)$ 。条件分布  $y|\mathbf{x}$  的总体  $q$  分位数,记为  $y_q$ ,满足以下定义式:

$$q = F_{y|\mathbf{x}}(y_q) \quad (26.3)$$

假设  $F_{y|\mathbf{x}}(\cdot)$  严格单调递增,则有

$$y_q = F_{y|\mathbf{x}}^{-1}(q) \quad (26.4)$$

由于条件累积分布函数  $F_{y|x}(\cdot)$  依赖于  $x$ , 故条件分布  $y|x$  的总体  $q$  分位数  $y_q$  也依赖于  $x$ , 可以明确地写为  $y_q(x)$ , 称为“条件分位数函数”(conditional quantile function)。也就是说, 条件分位数  $y_q(x)$  是解释变量  $x$  的函数。更进一步, 对于线性回归模型而言, 如果扰动项满足同方差的假定, 或扰动项异方差的形式为乘积形式, 则  $y_q(x)$  是  $x$  的线性函数。证明如下。考虑以下模型:

$$y = x'\beta + u$$

$$u = x'\alpha + \varepsilon$$

$$\varepsilon \sim \text{iid}(0, \sigma^2)$$

其中, 不失一般性, 假设  $x'\alpha > 0$ <sup>①</sup>。如果  $x'\alpha$  为常数, 则扰动项  $u$  为同方差; 反之, 则为乘积形式的异方差。根据定义, 条件分位数函数  $y_q(x)$  满足

(条件分位数的定义)

$$\begin{aligned} q &= P\{y \leq y_q(x)\} \\ &= P\{x'\beta + u \leq y_q(x)\} \quad (\text{代入 } y = x'\beta + u) \\ &= P\{u \leq y_q(x) - x'\beta\} \quad (\text{移项}) \\ &= P\{x'\alpha + \varepsilon \leq y_q(x) - x'\beta\} \quad (\text{代入 } u = x'\alpha + \varepsilon) \\ &= P\left\{\frac{\varepsilon \leq y_q(x) - x'\beta}{x'\alpha}\right\} \quad (\text{两边同除以 } x'\alpha > 0) \\ &= F_\varepsilon\left(\frac{y_q(x) - x'\beta}{x'\alpha}\right) \quad (\text{累积分布函数的定义}) \end{aligned}$$

其中,  $F_\varepsilon(\cdot)$  为  $\varepsilon$  的累积分布函数。因此,

$$\frac{y_q(x) - x'\beta}{x'\alpha} = F_\varepsilon^{-1}(q) \quad (26.6)$$

$$y_q(x) = x'\beta + x'\alpha F_\varepsilon^{-1}(q) = x'[\beta + \alpha F_\varepsilon^{-1}(q)] \quad (26.7)$$

从方程(26.7)可以看出,  $y_q(x)$  是  $x$  的线性函数。在同方差的情况下,  $x'\alpha$  为常数, 则所有条件分位数函数  $\{y_q(x), 0 < q < 1\}$  的“斜率”都等于  $\beta$ , 只有截距项  $x'\alpha F_\varepsilon^{-1}(q)$  依赖于  $q$ 。在一般情况下, 条件分位数函数的“斜率”也依赖于  $q$ , 通常记为  $\beta_q$ 。

由于上面这个结果, 在下文中, 我们假设条件分位数函数是解释变量  $x$  的线性函数。

## 26.3 样本分位数

对于随机变量  $Y$ , 如果总体的  $q$  分位数  $y_q$  未知, 则可以使用样本  $q$  分位数  $\hat{y}_q$  来估计  $y_q$ 。通常的做法是, 首先将样本数据  $\{y_1, y_2, \dots, y_n\}$  按照从小到大的顺序排列为  $\{y_{(1)}, y_{(2)}, \dots, y_{(n)}\}$ , 则  $\hat{y}_q$  等于第  $[nq]$  个最小观测值, 其中  $n$  为样本容量,  $[nq]$  表示大于或等于  $nq$  并离  $nq$  最近的正整数。

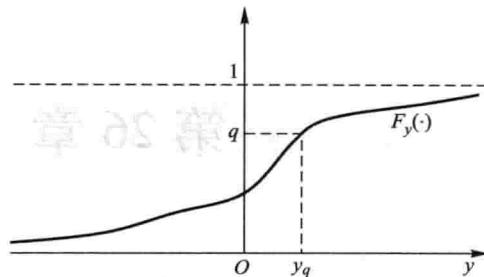


图 26.1 总体  $q$  分位数与累积分布函数

① 如果  $x'\alpha < 0$ , 则可以把  $x'\alpha + \varepsilon$  写为  $(-x'\alpha) + (-\varepsilon)$ 。

数。比如,  $n = 97$ ,  $q = 0.25$ , 则  $[nq] = [97 \times 0.25] = [24.25] = 25$ 。

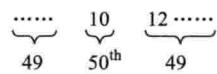
但是, 上述样本分位数的计算方法不容易推广到回归模型的情形。一种更方便的等价方法是, 将样本分位数看成某个最小化问题的解。事实上, 样本均值也可以看成是最小化残差平方和问题的最优解, 即

$$\min_{\mu} \sum_{i=1}^n (y_i - \mu)^2 \Rightarrow \mu = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (26.8)$$

类似地, 样本中位数可以视为是“最小化残差绝对值之和”问题的最优解, 即

$$\min_{\mu} \sum_{i=1}^n |y_i - \mu| \Rightarrow \mu = \text{median}\{y_1, y_2, \dots, y_n\} \quad (26.9)$$

为什么求解这个最小化问题会得到样本中位数呢? 因为只要上式中  $\mu$  的取值偏离中位数, 就会使得残差绝对值之和上升。直观来看, 考虑一个样本容量为 99 的



样本, 假设其样本中位数(即第 50 个最小观测值)为 10, 参见图 26.2。假设第 51 个最小观测值为 12。如果在上述最小化问题中, 让  $\mu = 12$  而不是 10, 则对于前 50 个观测值而言, 其对应的残差绝对值  $|y_i - \mu|$  都将

增加 2; 而对于后 49 个观测值而言, 其对应的残差绝对值  $|y_i - \mu|$  都将减少 2。故总变动为  $(50 \times 2) - (49 \times 2) = 2$ , 故第 51 个最小观测值不如第 50 个最小观测值(中位数)更能使目标函数最小化。根据同样的道理, 第 49 个最小观测值也不如第 50 个最小观测值。由此可知(严格证明见下文), 第 50 个最小观测值(中位数)是优解。

图 26.2

**命题** 可以将样本  $q$  分位数视为以下最小化残差绝对值的加权平均问题的最优解:

$$\min_{\mu} \sum_{i:y_i \geq \mu}^n q |y_i - \mu| + \sum_{i:y_i < \mu}^n (1-q) |y_i - \mu| \Rightarrow \mu = \hat{y}_q \quad (26.10)$$

**例** 如果  $q = 1/4$ , 则满足 “ $y_i \geq \mu$ ” 条件的观测值只得到  $1/4$  的权重, 而满足 “ $y_i < \mu$ ” 条件的其余观测值则得到  $3/4$  的权重。直观来看, 因为估计的是  $1/4$  分位数(位于总体的底部), 故较大的观测值得到的权重较小, 而较小的观测值得到的权重较大。

**证明:** 将目标函数(26.10)中的绝对值去掉可得

$$\min_{\mu} \sum_{i:y_i \geq \mu}^n q(y_i - \mu) + \sum_{i:y_i < \mu}^n (1-q)(\mu - y_i) \quad (26.11)$$

对  $\mu$  求一阶导数可得

$$\sum_{i:y_i \geq \mu}^n q(-1) + \sum_{i:y_i < \mu}^n (1-q) = 0 \quad (26.12)$$

假设  $y_{(k)} < \mu \leq y_{(k+1)}$ , 其中  $y_{(k)}$  为第  $k$  个最小观测值, 则共有  $k$  个观测值满足 “ $y_i < \mu$ ”,  $(n-k)$  个观测值满足 “ $y_i \geq \mu$ ”, 故

$$-(n-k)q + k(1-q) = 0 \quad (26.13)$$

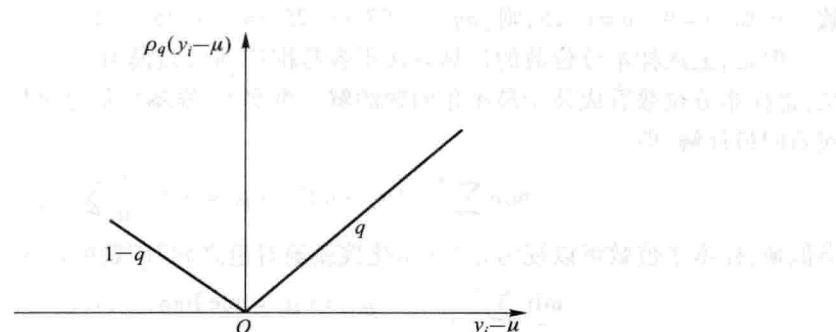
经整理可得

$$k = nq \quad (26.14)$$

当然,  $k$  还必须是整数。由此可知, 最优解  $\mu = y_{(\lfloor nq \rfloor)} = \hat{y}_q$ , 即样本分位数。为了证明最小化问题的二阶条件也满足, 只要说明目标函数为凸函数即可。为此, 定义函数  $\rho_q(\cdot)$  为

$$\rho_q(y_i - \mu) \equiv \begin{cases} q |y_i - \mu|, & \text{若 } y_i \geq \mu \\ (1-q) |y_i - \mu|, & \text{若 } y_i < \mu \end{cases} \quad (26.15)$$

函数  $\rho_q(y_i - \mu)$  的形状如图 26.3, 故也称为“倾斜的绝对值函数”(tilted absolute value function)或“打钩函数”(check function)。从图形易知,  $\rho_q(\cdot)$  为凸函数。而目标函数(26.11)可以写为  $\sum_{i=1}^n \rho_q(y_i - \mu)$ , 即  $n$  个凸函数之和, 故仍然是凸函数。



## 26.4 分位数回归的估计方法

下面将单变量情形下对样本分位数的估计方法推广到线性回归模型。假设条件分布  $y|x$  的总体  $q$  分位数  $y_q(x)$  是  $x$  的线性函数, 即

$$y_q(x_i) = x'_i \boldsymbol{\beta}_q \quad (26.16)$$

其中,  $\boldsymbol{\beta}_q$  被称为“ $q$  分位数回归系数”, 其估计量  $\hat{\boldsymbol{\beta}}_q$  可以由以下最小化问题来定义:

$$\min_{\boldsymbol{\beta}_q} \sum_{i:y_i \geq x'_i \boldsymbol{\beta}_q}^n q |y_i - x'_i \boldsymbol{\beta}_q| + \sum_{i:y_i < x'_i \boldsymbol{\beta}_q}^n (1-q) |y_i - x'_i \boldsymbol{\beta}_q| \quad (26.17)$$

如果  $q=1/2$ , 则为“中位数回归”(median regression)。此时, 目标函数简化为

$$\min_{\boldsymbol{\beta}_q} \sum_{i=1}^n |y_i - x'_i \boldsymbol{\beta}_q| \quad (26.18)$$

故中位数回归也被称为“最小绝对离差估计量”(Least Absolute Deviation Estimator, 简记 LAD)。显然, 它比均值回归(OLS)更不易受到极端值的影响, 故更加稳健。

由于分位数回归的目标函数带有绝对值, 不可微分, 故通常使用线性规划(linear programming)<sup>①</sup>的方法来计算  $\hat{\boldsymbol{\beta}}_q$ 。可以证明, 样本分位数回归系数  $\hat{\boldsymbol{\beta}}_q$  是总体分位数回归系数  $\boldsymbol{\beta}_q$  的一致估计量, 而且  $\hat{\boldsymbol{\beta}}_q$  服从渐近正态分布, 即

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_q - \boldsymbol{\beta}_q) \xrightarrow{d} N(\mathbf{0}, \text{Avar}(\hat{\boldsymbol{\beta}}_q)) \quad (26.19)$$

其中, 渐近方差  $\text{Avar}(\hat{\boldsymbol{\beta}}_q) = \mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1}$  (形式上仍为夹心估计量),  $\mathbf{A} = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f_{u_q}(0|x_i) \mathbf{x}_i \mathbf{x}_i'$ ,

$\mathbf{B} = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n q(1-q) \mathbf{x}_i \mathbf{x}_i'$ , 而  $f_{u_q}(0|x_i)$  是扰动项  $u_q \equiv y - x' \boldsymbol{\beta}_q$  的条件密度函数在  $u_q = 0$  处的取值。因此, 要计算  $\hat{\boldsymbol{\beta}}_q$  的协方差矩阵  $\text{Avar}(\hat{\boldsymbol{\beta}}_q)$ , 首先要估计  $f_{u_q}(0|x_i)$ 。这是 Stata 的默认方法。Stata 也提供自助法作为计算  $\text{Avar}(\hat{\boldsymbol{\beta}}_q)$  的另一方法。

对于  $q$  分位数回归, 可使用准  $R^2$  度量其拟合优度, 其定义为:

① 即目标函数与约束条件均为线性函数的最优化问题。

$$1 - \frac{\sum_{i:y_i \geq \hat{y}_q}^n q |y_i - \mathbf{x}'_i \hat{\beta}_q| + \sum_{i:y_i < \hat{y}_q}^n (1-q) |y_i - \mathbf{x}'_i \hat{\beta}_q|}{\sum_{i:y_i \geq \hat{y}_q}^n q |y_i - \hat{y}_q| + \sum_{i:y_i < \hat{y}_q}^n (1-q) |y_i - \hat{y}_q|} \quad (26.20)$$

其中,  $\hat{y}_q$  为样本  $q$  分位数, 上式第二项的分子为  $q$  分位数回归目标函数的最小值 (sum of weighted deviations about estimated quantiles), 而分母为“sum of weighted deviations about raw quantiles”。

例 Buchinsky(1994) 使用分位数回归研究 1963—1987 年间美国工资结构的变化。

例 张车伟、薛欣欣(2008) 使用分位数回归分析国有与非国有部门工资差异的决定因素。

## 26.5 分位数回归的 Stata 命令及实例

Stata 提供了以下有关分位数回归的命令。

(1) 只作一个分位数回归, 使用默认的协方差矩阵计算方法<sup>①</sup>

`qreg y x1 x2 x3` (默认为中位数回归)

`qreg y x1 x2 x3, q(#)` (#分位数回归)

(2) 只作一个分位数回归, 使用自助法计算协方差矩阵

`set seed #` (指定产生随机数的种子, 以便每次均能得到同样的结果)

`bsqreg y x1 x2 x3, reps(#)` q(#)

(3) 同时作多个分位数回归 (simultaneous quantile regressions), 使用自助法计算协方差矩阵

`sqreg y x1 x2 x3, q(.25 .5 .75) reps(#)` (同时计算 0.25, 0.5 与 0.75 分位数回归, 自助法重复#次)

`test [q25 = q50 = q75]: x1` (检验这三个分位数回归 x1 的系数是否相等)

(4) 将不同分位数回归的系数及其置信区间进行画图比较

首先, 下载非官方命令 `grqreg` (表示 graph quantile regression):

`net install grqreg.pkg` (下载安装命令 `grqreg`)

然后进行如下操作:

`set seed #`

`bsqreg y x1 x2 x3, reps(#)` q(.5)

(为了得到自助标准误而先作中位数回归)

`grqreg, cons ci ols olscl`

其中, 选择项“`cons`”表示对常数项也进行比较, 选择项“`ci`”表示包括估计系数的 95% 置信区间, 选择项“`ols`”表示提供 OLS 估计系数作为参照系, 选择项“`olscl`”表示提供 OLS 估计系数的 95% 置信区间。

下面以数据集 `grilic.dta` 为例(参见第 10 章)。作为参照系, 首先进行 OLS 回归<sup>②</sup>:

```
. use grilic.dta, clear  
. reg lw s iq expr tenure rns smsa,
```

① 也可使用非官方命令 `qreg2`, 下载方法为“`ssc install qreg2`”。该命令提供异方差稳健的标准误。

② 此处复制了第 10 章的结果, 且不考虑内生解释变量的问题。

Linear regression						Number of obs = 758
						F( 6, 751) = 71.89
						Prob > F = 0.0000
						R-squared = 0.3600
						Root MSE = .34454
lw	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
s	.0927874	.0069763	13.30	0.000	.0790921	.1064826
iq	.0032792	.0011321	2.90	0.004	.0010567	.0055016
expr	.0393443	.0066603	5.91	0.000	.0262692	.0524193
tenure	.034209	.0078957	4.33	0.000	.0187088	.0497092
rns	-.0745325	.0299772	-2.49	0.013	-.1333815	-.0156834
smsa	.1367369	.0277712	4.92	0.000	.0822186	.1912553
_cons	3.895172	.1159286	33.60	0.000	3.667589	4.122754

下面进行中位数回归：

```
. qreg lw s iq expr tenure rns smsa, nolog
```

Median regression						Number of obs = 758
Raw sum of deviations	261.36	(about 5.684)				
Min sum of deviations	203.7577					
					Pseudo R2	= 0.2204
lw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.1014672	.0085368	11.89	0.000	.0847084	.1182261
iq	.0045716	.0013901	3.29	0.001	.0018426	.0073005
expr	.0359681	.0080899	4.45	0.000	.0200865	.0518497
tenure	.0425991	.0098489	4.33	0.000	.0232645	.0619338
rns	-.0362275	.0369023	-0.98	0.327	-.1086715	.0362164
smsa	.1318406	.0358176	3.68	0.000	.0615262	.202155
_cons	3.629197	.1400148	25.92	0.000	3.354331	3.904064

从上表可知,增加一年教育(s)能够使工资的中位数增加 10.1%,略大于对工资平均数的影响(OLS 系数估计值为 9.3%)。

下面使用自助法来计算分位数回归的标准误。为便于复制结果,指定随机数的种子:

```
. set seed 10101
. bsqreg lw s iq expr tenure rns smsa, reps(400) q(.5)
```

(fitting base model)						
(bootstrapping .....						
> .....						
> .....						
> .....						)
Median regression, bootstrap(400) SEs					Number of obs = 758	
Raw sum of deviations	261.36	(about 5.684)				
Min sum of deviations	203.7577				Pseudo R2	= 0.2204
lw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.1014672	.0094649	10.72	0.000	.0828863	.1200481
iq	.0045716	.001477	3.10	0.002	.0016719	.0074712
expr	.0359681	.0093081	3.86	0.000	.0176951	.0542411
tenure	.0425991	.0096584	4.41	0.000	.0236384	.0615599
rns	-.0362275	.0472525	-0.77	0.444	-.1289901	.0565351
smsa	.1318406	.0368775	3.58	0.000	.0594454	.2042359
_cons	3.629197	.1736001	20.91	0.000	3.288398	3.969997

对比自助标准误与 Stata 的默认标准误可知,二者相差不大。

也可以同时估计多个分位数回归,比如,1/10,5/10 与 9/10 分位数:

```
. set seed 10101
```

```
. sqreg lw s iq expr tenure rns smsa, reps(400) q(.1 .5 .9) nodots
```

其中,选择项“nodots”表示不显示自助抽样过程中的点。

Simultaneous quantile regression						Number of obs = 758
bootstrap(400) SEs						.10 Pseudo R2 = 0.1580
						.50 Pseudo R2 = 0.2204
						.90 Pseudo R2 = 0.2230
		Bootstrap				
		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
q10	lw	.0761644	.0139931	5.44	0.000	.0486941 .1036347
	s	.0052221	.0019361	2.70	0.007	.0014213 .009023
	iq	.0336056	.0108269	3.10	0.002	.0123511 .0548601
	expr	.0335323	.0182892	1.83	0.067	-.0023717 .0694363
	tenure	-.0730452	.0527144	-1.39	0.166	-.1765304 .03044
	rns	.1328733	.0482216	2.76	0.006	.0382082 .2275385
	smsa	3.493085	.2372351	14.72	0.000	3.027363 3.958808
q50	lw	.1014672	.0094601	10.73	0.000	.0828958 .1200387
	s	.0045716	.0013566	3.37	0.001	.0019084 .0072348
	iq	.0359681	.0103368	3.48	0.001	.0156757 .0562605
	expr	.0425991	.0100634	4.23	0.000	.0228434 .0623549
	tenure	-.0362275	.0456606	-0.79	0.428	-.1258652 .0534101
	rns	.1318406	.0365317	3.61	0.000	.0601242 .2035571
	smsa	3.629197	.1687068	21.51	0.000	3.298005 3.96039
q90	lw	.0825558	.0115389	7.15	0.000	.0599035 .1052082
	s	.0042117	.001989	2.12	0.035	.0003071 .0081163
	iq	.0484154	.01051	4.61	0.000	.027783 .0690478
	expr	.0256796	.008211	3.13	0.002	.0095603 .0417989
	tenure	-.072669	.0525121	-1.38	0.167	-.175757 .0304191
	rns	.1281147	.0520661	2.46	0.014	.0259023 .2303271
	smsa	4.36716	.165961	26.31	0.000	4.041358 4.692963

进一步,可以检验在以上三个分位数回归中,教育年限(s)的系数是否相等:

```
. test [q10]=[q50]=[q90]: s
```

```
( 1) [q10]s - [q50]s = 0  
( 2) [q10]s - [q90]s = 0
```

```
F( 2, 751) = 2.32  
Prob > F = 0.0989
```

结果表明,可以在 10% 的显著性水平上认为,以上分位数回归系数不完全相等。

为了便于比较,下面把 OLS 与“1/10,5/10,9/10 分位数”的系数估计值及标准误列表:

```
. qui reg lw s iq expr tenure rns smsa
```

```
. est sto OLS
```

```
. qui qreg lw s iq expr tenure rns smsa, q(.1)
```

```
. est sto QR_10
```

```
. qui qreg lw s iq expr tenure rns smsa, q(.5)
```

```
. est sto QR_50
. qui qreg lw s iq expr tenure rns smsa, q(.9)
. est sto QR_90
. esttab OLS QR_10 QR_50 QR_90, se mtitles star(* 0.1 ** 0.05 *** 0.01)
```

	(1) OLS	(2) QR_10	(3) QR_50	(4) QR_90
s	0.0928*** (0.00667)	0.0762*** (0.0154)	0.101*** (0.00854)	0.0826*** (0.0117)
iq	0.00328*** (0.00108)	0.00522** (0.00237)	0.00457*** (0.00139)	0.00421** (0.00206)
expr	0.0393*** (0.00631)	0.0336*** (0.0113)	0.0360*** (0.00809)	0.0484*** (0.0135)
tenure	0.0342*** (0.00771)	0.0335** (0.0144)	0.0426*** (0.00985)	0.0257 (0.0171)
rns	-0.0745*** (0.0288)	-0.0730 (0.0602)	-0.0362 (0.0369)	-0.0727 (0.0520)
smsa	0.137*** (0.0279)	0.133** (0.0576)	0.132*** (0.0358)	0.128** (0.0505)
_cons	3.895*** (0.109)	3.493*** (0.241)	3.629*** (0.140)	4.367*** (0.226)
N	758	758	758	758
Standard errors in parentheses				
* p<0.1, ** p<0.05, *** p<0.01				

以教育投资的回报率(s的系数)为例。上表显示,随着分位数的增加(1/10→5/10→9/10),教育年限(s)的分位数回归系数呈现先升后降的趋势(7.6%→10.1%→8.3%)。这表明,教育年限对工资的条件分布的两端之影响小于对其中间部分的影响。也就是说,增加教育年限对于低工资者与高工资者的影响都比较小,而最大受益者为中间阶层。

另一方面,估计系数的标准误则呈现先降后升的趋势(0.015→0.009→0.012)。这说明,对于条件分布两端的分位数回归系数的估计较不准确。

进一步,把分位数回归系数随着分位数的变化情形更直观地图示,结果如图 26.4。

```
. qui bsqreg lw s iq expr tenure rns smsa, q(.50) reps(400)
. grqreg, cons ci ols olscl
```

以图 26.4 第一行第二列的小图为例。该图显示,随着分位数的变化,教育年限(s)的分位数回归系数(即教育回报率)的变化。此图的基本形状印证了在前面表格中,教育年限的分位数回归系数先升后降的格局。另外,上图还显示,在条件分布的两端,95% 的置信区间通常变得更宽了(因为系数估计值的标准误变大了)。

目前,分位数回归仍是一个活跃的前沿领域,特别是分位数回归与各种类型的数据相结合,

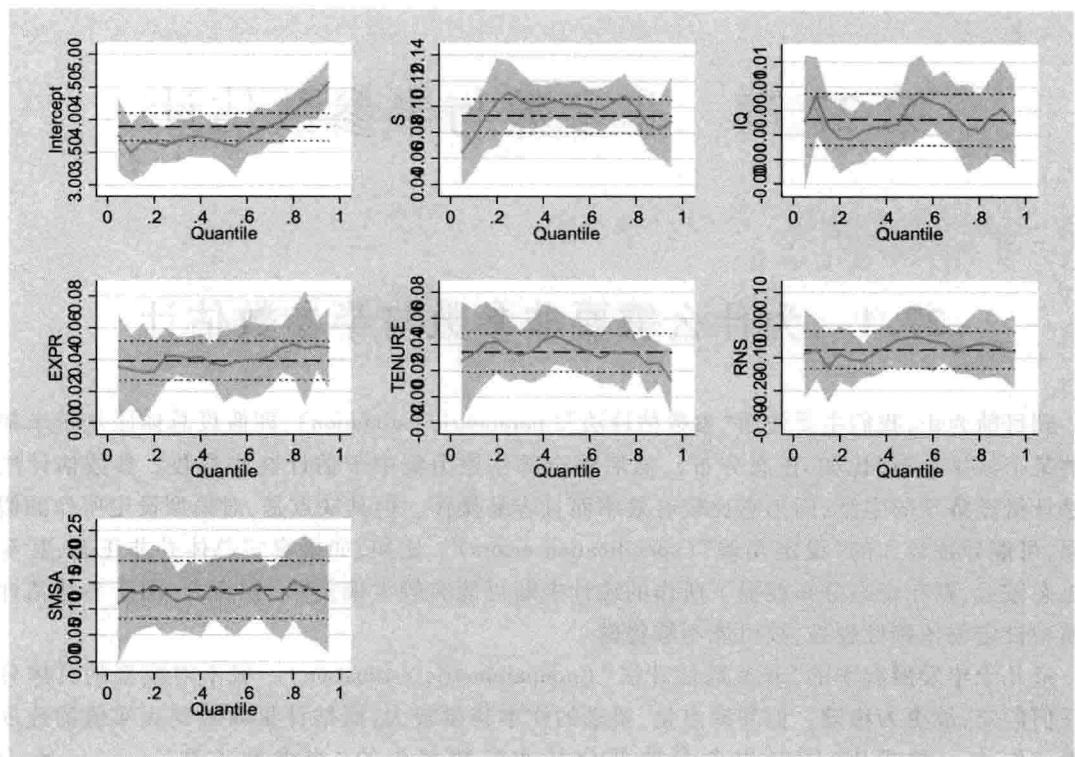


图 26.4 分位数回归系数的变化

比如面板分位数回归(Koenker, 2004)、单位根分位数回归(Koenker and Xiao, 2004)、分位数自回归(Koenker and Xiao, 2006)等。

## 习 题

26.1 参照本章实例, 使用数据集 nerlove.dta 对成本函数进行分位数回归(参见第4章)。

26.2 参照本章实例, 使用数据集 hprice2a.dta 对房价的决定因素进行分位数回归(参见第7章)。

# 第 27 章 非参数与半参数估计

## 27.1 为什么需要非参数与半参数估计

到目前为止,我们主要使用“参数估计法”(parametric estimation),即假设总体服从带未知参数的某个具体分布(比如,正态分布),然后将全部注意力集中于估计这些参数。参数估计法依然是计量经济学的主流,因为它比较有效率而且容易操作。但其缺点是,对模型设定所作的假定较强,可能导致较大的“设定误差”(specification errors)。比如,如果真实总体并非正态,甚至偏离正态较远,则在正态分布前提下所作的统计推断可能有较大偏差<sup>①</sup>。换言之,由于参数估计法对模型设定的依赖性较强,故可能不够稳健。

近几十年发展起来的“非参数估计法”(nonparametric estimation)一般不对模型的具体分布作任何假定,故更为稳健。但其缺点是,要求的样本容量较大,而估计量收敛到真实值的速度也较慢。作为一种折中,同时包含参数部分与非参数部分的“半参数方法”(semiparametric estimation)应运而生,它降低了对样本容量的要求,又具有一定的稳健性。总之,非参数及半参数方法与传统的参数估计法是互补关系,当后者不太适用时,则可以考虑前者。

## 27.2 对密度函数的非参数估计

在统计学中,经常需要根据样本数据来推断总体的分布,即密度函数。如果采用参数估计法,则要先对总体分布的具体形式进行假定。比如,假设总体服从正态分布  $N(\mu, \sigma^2)$ ,然后估计参数  $(\mu, \sigma^2)$ ,就可以得到对密度函数的估计。如上所述,如果真实总体与正态分布相去甚远,则根据参数估计法作出的统计推断可能有较大偏差。

在估计密度函数时,如果不假设总体分布的具体形式,则为非参数方法。最原始的非参数方法是画直方图(histogram),即将数据的取值范围等分为若干组(bin),然后计算数据落入每一组的频率,以此画图,作为对密度函数的估计。用直方图来估计密度函数的缺点是,即使随机变量是连续的,直方图也始终是不连续的阶梯函数。

为了得到对密度函数的光滑估计,Rosenblatt(1956)提出“核密度估计法”(kernel density estimation)。为此,首先考察直方图的数学本质,然后再推广到核密度估计。

假设要估计连续型随机变量  $x$  在  $x_0$  处的概率密度  $f(x_0)$ 。由于概率密度  $f(x_0)$  是累积分布函数  $F(x)$  在  $x_0$  处的导数,根据微积分中导数的定义:

<sup>①</sup> 参见 DiNardo and Tobias(2001)的例子。

$$\begin{aligned} f(x_0) &= \lim_{h \rightarrow 0} \frac{F(x_0 + h) - F(x_0 - h)}{2h} \\ &= \lim_{h \rightarrow 0} \frac{P(x_0 - h < x < x_0 + h)}{2h} \end{aligned} \quad (27.1)$$

其中, 区间  $(x_0 - h, x_0 + h)$  为在  $x_0$  附近的小邻域。对于样本  $\{x_1, x_2, \dots, x_n\}$ , 可以用数据落入区间  $(x_0 - h, x_0 + h)$  的频率来估计概率  $P(x_0 - h < x < x_0 + h)$ , 得到以下直方图估计量:

$$\begin{aligned} \hat{f}_{\text{HIST}}(x_0) &= \frac{\sum_{i=1}^n \mathbf{1}(x_0 - h < x_i < x_0 + h)/n}{2h} \\ &= \frac{1}{nh} \sum_{i=1}^n \frac{1}{2} \cdot \mathbf{1}\left\{\left|\frac{x_i - x_0}{h}\right| < 1\right\} \end{aligned} \quad (27.2)$$

其中,  $\mathbf{1}(\cdot)$  为示性函数, 即当括弧中的表达式为真时, 取值为 1; 反之, 则取值为 0。从方程(27.2)的第二个表达式可以清楚地看出, 直方图估计量  $\hat{f}_{\text{HIST}}(x_0)$  对于所有在区间  $(x_0 - h, x_0 + h)$  内的观测值都给予相同权重, 而对于在此区间以外的观测值则权重为 0。这个区间的半径  $h$  定义了“在  $x_0$  附近邻域的大小”(size of neighborhood around  $x_0$ ), 被称为“带宽”(bandwidth)。这个区间的直径  $2h$  则被称为“窗宽”(window width)。

直方图之所以得到不光滑的密度估计, 根本原因是由于它使用了示性函数作为“权重函数”(weighting function), 以及它的各组之间不允许交叠。为了得到光滑的密度估计, 核密度估计法使用更一般的权重函数, 并允许各组之间交叠。核密度估计量为

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(x_i - x_0)/h] \quad (27.3)$$

其中, 函数  $K(\cdot)$  称为“核函数”(kernel function), 本质上就是权重函数。带宽  $h$  越大, 在  $x_0$  附近邻域越大, 则估计的密度函数  $\hat{f}(x)$  越光滑<sup>①</sup>, 故称带宽  $h$  为“光滑参数”(smoothing parameter)。一般假设核函数  $K(z)$  满足以下性质:

- (i)  $K(z)$  连续且关于原点对称(偶函数);
- (ii)  $\int_{-\infty}^{+\infty} K(z) dz = 1$ ,  $\int_{-\infty}^{+\infty} zK(z) dz = 0$ ,  $\int_{-\infty}^{+\infty} |K(z)| dz < +\infty$ ;
- (iii) 或者①存在  $z_0 > 0$ , 使得当  $|z| > z_0$  时,  $K(z) = 0$ ; 或者②当  $|z| \rightarrow +\infty$  时,  $|z|K(z) \rightarrow 0$ ;
- (iv)  $\int_{-\infty}^{+\infty} z^2 K(z) dz = \gamma$ , 其中  $\gamma$  为常数。

条件(ii)要求核函数的曲线下面积(积分)为 1(将核函数标准化), 并满足一些有界条件(boundedness conditions)。条件(iii)①比条件(iii)②更强, 在实践中, 常常采用条件(iii)①, 即如果超出某个邻域范围  $[-z_0, z_0]$ , 则权重变为 0。不失一般性, 常将邻域  $[-z_0, z_0]$  标准化为  $[-1, 1]$ 。条件(iv)也是一个有界条件(在某些证明中使用)。

比较常见的核函数参见表 27.1。其中, 均匀核也用于直方图, 只是在用均匀核进行核密度估计时并不固定分组, 而在每个点上进行估计。最为流行的核函数为二次核(也称 Epanechnikov 核, 参见图 27.1)与高斯核。除高斯核满足条件(iii)②外, 表 27.1 所列其他核函数均满足条件(iii)①。这些核函数的共同特点是, 离原点越近, 则核函数取值越大, 并在原点处达到最大值;

<sup>①</sup> 比如, 在直方图的情形下, 如果只分一组, 则估计的密度函数将是无穷光滑的均匀分布。然而, 如果带宽  $h$  很大, 即只分很少的几个组, 则直方图将很难反映真实分布的概貌(即过度光滑)。

这意味着,越近的点给予的权重越大。

表 27.1 常用的核函数

核函数名称	核函数的数学形式	$\delta$
均匀核 (uniform or rectangular)	$\frac{1}{2} \cdot \mathbf{1}( z  < 1)$	1.351 0
三角核 (triangular or Bartlett)	$(1 -  z ) \cdot \mathbf{1}( z  < 1)$	—
伊番科尼可夫核 (Epanechnikov) <sup>①</sup> 或二次核 (quadratic)	$\frac{3}{4} (1 - z^2) \cdot \mathbf{1}( z  < 1)$	1.718 8
四次核 (quartic) 或双权核 (biweight)	$\frac{15}{16} (1 - z^2)^2 \cdot \mathbf{1}( z  < 1)$	2.036 2
三权核 (Triweight)	$\frac{35}{32} (1 - z^2)^3 \cdot \mathbf{1}( z  < 1)$	2.312 2
三三核 (Tricubic)	$\frac{70}{81} (1 -  z ^3)^3 \cdot \mathbf{1}( z  < 1)$	—
高斯核 (Gaussian or Normal)	$\frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$	0.776 4

注:其中  $\delta$  为用来计算“Silverman 嵌入估计”的常数,参见下文。

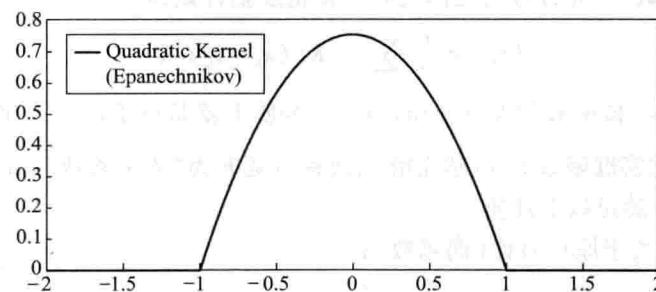


图 27.1 二次核 (Epanechnikov 核)

给定核函数  $K(\cdot)$  与带宽  $h$ ,就可以估计核密度  $\hat{f}(x_0)$ 。在 Stata 中,默认设置为在等距离的  $\min(n, 50)$  个点来计算  $\hat{f}(x_0)$ ,然后连成光滑的密度函数。当然,也可以人为指定估计的点数。

### 27.3 核密度估计的性质

由于核密度估计使用了在  $x_0$  附近的点  $x$  来估计  $\hat{f}(x_0)$ ,而一般地,如果  $x \neq x_0$ ,则  $f(x) \neq f(x_0)$ ,故核密度估计通常是有偏的。通过二阶泰勒近似,可以证明(见附录),其偏差为

$$\text{Bias}(x_0) \equiv E[\hat{f}(x_0)] - f(x_0) \approx \frac{1}{2} h^2 f''(x_0) \int_{-\infty}^{+\infty} z^2 K(z) dz \quad (27.4)$$

① Stata 记此函数为“epan2”,而默认的核函数“epan”与此略有不同。

即偏差与  $h^2$  成正比, 为  $h^2$  的同阶无穷小, 记为  $O(h^2)$ <sup>①</sup>。直观来看, 带宽  $h$  越大, 则将使用离  $x_0$  更远的点在估计  $f(x_0)$ , 导致偏差增大。更进一步, 由于偏差大致与  $h^2$  成正比, 故随着  $h$  增大, 偏差将以平方的速度迅速上升。如果当样本容量  $n \rightarrow \infty$  时, 让带宽  $h \rightarrow 0$  (正如在画直方图时, 样本容量越大, 则可以把每组分得越细), 则偏差将在大样本中消失。另一方面, 密度函数的二阶导数  $f''(x_0)$  越大, 即在  $x_0$  处的曲率越大, 则  $x_0$  附近的函数值波动越大, 导致  $x_0$  附近观测点所包含的信息量下降, 也会引起偏差增大。最后, 偏差还取决于核函数  $K(z)$ 。

类似地, 可以证明核密度估计的方差为(参见附录)

$$\text{Var}[\hat{f}(x_0)] = \frac{1}{nh} f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz + o(1/nh) \quad (27.5)$$

其中,  $o(1/nh)$  表示是  $(1/nh)$  的高阶无穷小<sup>②</sup>。因此,  $\text{Var}[\hat{f}(x_0)] = O(1/nh)$ , 即是  $(1/nh)$  的同阶无穷小。直观来看, 样本容量  $n$  越大, 则方差越小; 另一方面, 带宽  $h$  越大, 由于使用了更多观测点来估计  $f(x_0)$ , 故方差越小。如果当样本容量  $n \rightarrow \infty$  时, 让  $nh \rightarrow \infty$  (虽然  $h \rightarrow 0$ , 但  $h$  趋于 0 的速度比样本容量  $n \rightarrow \infty$  的速度更慢, 故二者的乘积仍然趋于无穷), 则此方差将在大样本中消失。

### 核密度估计的一致性

如果当  $n \rightarrow \infty$  时, 让带宽  $h \rightarrow 0$  且  $nh \rightarrow \infty$ , 则偏差  $\text{Bias}(x_0)$  与方差  $\text{Var}[\hat{f}(x_0)]$  在大样本下都趋于 0, 故根据均方收敛可知,  $\hat{f}(x_0)$  是  $f(x_0)$  的一致估计量。

### 核密度估计的渐近正态性

如果核函数  $K(z)$  的条件(iv)得到满足, 则通过中心极限定理可以证明,  $\hat{f}(x_0)$  服从渐近正态分布:

$$\sqrt{nh} [\hat{f}(x_0) - f(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N\left(0, f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz\right) \quad (27.6)$$

利用估计量  $\hat{f}(x_0)$  的渐近正态性, 可以进行区间估计。由此可见, 核密度估计量的收敛速度为  $\sqrt{nh}$ 。由于最优带宽  $h^*$  与  $n^{-0.2}$  成正比(参见下节), 故  $\sqrt{nh} = \sqrt{n \cdot n^{-0.2}} = \sqrt{n^{0.8}} = n^{0.4} < n^{0.5} = \sqrt{n}$ , 这意味着非参数估计量的收敛速度  $n^{0.4}$  慢于参数估计量的通常收敛速度  $n^{0.5}$ 。

## 27.4 最优带宽

如果带宽  $h$  越大, 则  $x_0$  附近的邻域越大, 故偏差也越大(偏差与  $h^2$  成正比); 另一方面, 带宽  $h$  越大, 则  $\hat{f}(x_0)$  越光滑, 即方差  $\text{Var}[\hat{f}(x_0)]$  越小, 参见方差的表达式(27.5)。在选择“最优带宽”(optimal bandwidth)  $h^*$  时, 通常希望最小化均方误差(MSE), 即估计量方差与偏差平方之和:

$$\min_h \text{MSE}[\hat{f}(x_0)] = [\text{Bias}(x_0)]^2 + \text{Var}[\hat{f}(x_0)] \quad (27.7)$$

由于  $\text{Bias}(x_0) = O(h^2)$ , 故  $[\text{Bias}(x_0)]^2 = O(h^4)$ , 而  $\text{Var}[\hat{f}(x_0)] = O(1/nh)$ , 故此最小化问题可以大致写为

① 记函数  $a(h)$  为  $O(h^k)$ , 如果当  $h \rightarrow 0$  时,  $a(h)/h^k$  的极限是一个有限的数。

② 记函数  $a(h)$  为  $o(h^k)$ , 如果当  $h \rightarrow 0$  时,  $a(h)/h^k$  的极限为 0。

$$\min_h \text{MSE}[\hat{f}(x_0)] = k_1 h^4 + (k_2/nh) \quad (27.8)$$

其中,  $k_1, k_2$  为常数。对  $h$  求导, 可得一阶条件为

$$4k_1 h^3 + k_2 \frac{1}{n} (-1/h^2) = 0 \quad (27.9)$$

$$h = (4k_1/k_2)^{-0.2} n^{-0.2} \quad (27.10)$$

由此可知, 最优带宽为  $h^* = O(n^{-0.2})$ 。显然, 随着  $n$  增大,  $n^{-0.2} = 1/\sqrt[5]{n}$  的下降速度远慢于  $n^{-1} = 1/n$ , 参见图 27.2。故当  $n \rightarrow \infty$  时, 最优带宽  $h^* \rightarrow 0$ , 而且  $nh^* = n \cdot O(n^{-0.2}) = O(n^{0.8}) \rightarrow \infty$ 。因此, 选择最优带宽  $h^*$ , 就能保证核密度估计的一致性。

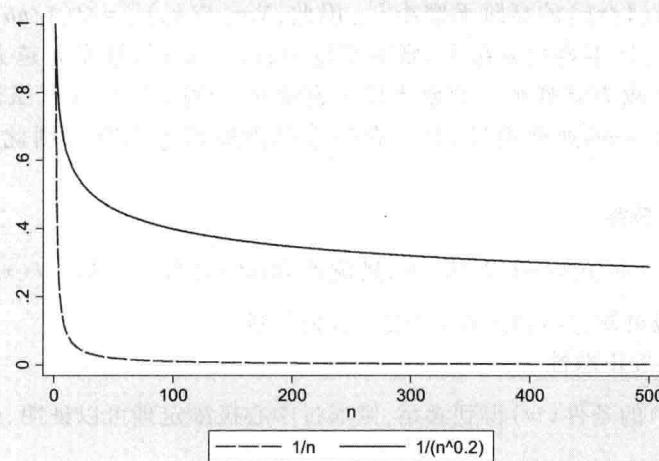


图 27.2 对比  $n^{-0.2}$  与  $n^{-1}$  的下降速度

然而, 均方误差  $\text{MSE}[\hat{f}(x_0)]$  依然取决于  $x_0$ 。如果希望得到对于  $x_0$  所有可能取值的均方误差的整体度量, 则可以最小化如下的“积分均方误差”(Integrated Mean Squared Error, 简记 IMSE):

$$\min_h \text{IMSE} \equiv \int_{-\infty}^{+\infty} \text{MSE}[\hat{f}(x_0)] dx_0 \quad (27.11)$$

求解这个最小化问题, Silverman(1986) 证明最优带宽为(参见附录)

$$h^* = \delta \left[ \int_{-\infty}^{+\infty} f''(x_0)^2 dx_0 \right]^{-0.2} n^{-0.2} \quad (27.12)$$

其中, 常数  $\delta \equiv \left[ \int_{-\infty}^{+\infty} K(z)^2 dz / \left( \int_{-\infty}^{+\infty} z^2 K(z) dz \right)^2 \right]^{0.2}$  仅依赖于核函数。从方程(27.12)可以看出, 最优带宽  $h^*$  还取决于密度函数的曲率( $f''(x_0)$ )。当密度函数波动较大(highly variable)时, 如果选择较大的邻域, 将带来更大的偏差, 故最优带宽  $h^*$  较小。

由于  $\delta$  依赖于核函数, 故最优带宽  $h^*$  也依赖于核函数。可以证明, 如果对于不同的核函数分别使用相应的最优带宽, 则积分均方误差  $\text{IMSE}(h^*)$  差别不大。能使  $\text{IMSE}(h^*)$  最小化的核函数为“伊番科尼可夫核”(Epanechnikov or quadratic), 也是 Stata 默认的核函数, 尽管它只有微弱的优势。总之, 对于最优带宽的选择远比核函数的选择更重要。使用不同核函数得到的密度估计一般非常接近。

但是, 最优带宽  $h^*$  仍然依赖于待估密度函数的二阶导数  $f''(x_0)$ 。如果样本来自正态总体,

则可以计算出,  $\int_{-\infty}^{+\infty} f''(x_0)^2 dx_0 = 3/(8\sqrt{\pi}\sigma^5) = 0.2116/\sigma^5$ , 故

$$h^* = 1.3643\delta n^{-0.2}s \quad (27.13)$$

其中,  $s$  为样本标准差。为了防止样本标准差受极端值的影响, 常使用“Silverman 嵌入估计”(Silverman's plug-in estimate) :

$$h^* = 1.3643\delta n^{-0.2} \min(s, iqr/1.349) \quad (27.14)$$

其中, “ $iqr$ ”为样本四分位距(sample interquartile range), 即样本  $3/4$  分位数与  $1/4$  分位数之间的距离。使用  $\min(s, iqr/1.349)$  是为了防止存在极端值而导致  $s$  被高估。使用嵌入估计通常能得到很好的结果, 但为了保险起见, 可以比较两倍嵌入估计与一半嵌入估计的效果。在实践中, 也常使用“眼球法”(eyeball method), 即用肉眼对带宽进行判断, 是否使用当前带宽导致密度函数的图形“过度光滑”(oversmoothed) 或“不够光滑”(undersmoothed), 然后微调到合适的带宽。

## 27.5 多元密度函数的核估计

对于  $k$  维随机变量  $x$ , 可以类似地进行“多元密度函数的核估计”(multivariate kernel density estimator) :

$$\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(x_i - x_0)/h] \quad (27.15)$$

其中,  $K(\cdot)$  是  $k$  维核函数, 即权重函数。 $K(\cdot)$  通常为一维核函数的乘积, 也可以使用多维正态的密度函数。多元密度函数核估计的性质与一元的情形相似。但最优带宽为  $h^* = O(n^{-1/(k+4)})$  (大于一元情形下的最优带宽), 而  $\hat{f}(x_0)$  收敛到真实值的速度也更慢。

在多维情况下, 容易出现“数据稀疏”(sparseness of data) 的问题, 即在  $x_0$  附近的观测点很少。另外, 即便是二元密度函数, 也要画成三维的立体图; 更高维则无法画图。

估计多维密度函数的用途之一是为了估计条件密度函数(conditional density function)。由于条件密度  $f(y|x) = f(x,y)/f(x)$ , 故可以用  $\hat{f}(y|x) = \hat{f}(x,y)/\hat{f}(x)$  作为条件密度的估计量, 其中  $\hat{f}(x,y)$  与  $\hat{f}(x)$  分别为二维与一维的密度函数核估计。

## 27.6 非参数核回归

计量经济学的首要方法为回归。考虑以下非参数一元回归模型:

$$\begin{aligned} y_i &= m(x_i) + \varepsilon_i \\ \varepsilon_i &\sim \text{iid}(0, \sigma_\varepsilon^2) \end{aligned} \quad (27.16)$$

该模型的困难(与优点)在于,  $m(\cdot)$  是未知函数(连函数形式也未知)。也正因为如此, 该模型可以更好地反映真实的回归关系。非参数回归的思想是, 对于每一个  $i(i=1, \dots, n)$ , 分别估计  $m(x_i)$ , 从而得到对回归函数  $m(x)$  的估计。在某种意义上, 我们不寻求  $m(x)$  的解析解, 而是寻找其数值解。

考虑最简单的情形。假设对于  $x$  的某个特定取值, 比如  $x_0$ , 都有若干个  $y$  的观测值, 比如  $n_0$

个。显然,可以把这  $n_0$  个  $y$  观测值的平均值作为  $m(x_0)$  的估计量。但在现实数据中,  $n_0$  可能很小(对于连续变量,很可能  $n_0$  仅为 1),导致估计量的方差过大。解决数据稀疏问题的一种方法是,对  $x_0$  附近邻域中的观测值也进行加权平均,即“局部加权平均估计量”(local weighted average estimator) :

$$\hat{m}(x_0) = \sum_{i=1}^n w_{i0,h} y_i \quad (27.17)$$

其中,权重  $w_{i0,h}$  是  $(x_i, x_0, h)$  的函数,即  $w_{i0,h} = w(x_i, x_0, h)$ ,且满足  $\sum_{i=1}^n w_{i0,h} = 1$ 。 $x_i$  是  $x_0$  附近的点,而  $h$  仍然是带宽或光滑参数。Nadaraya(1964)与 Watson(1964)使用核函数来定义以下权重,得到“核回归估计量”(kernel regression estimator) :

$$w_{i0,h} = \frac{K[(x_i - x_0)/h]}{\sum_{i=1}^n K[(x_i - x_0)/h]} \quad (27.18)$$

因此,核回归估计量可以写为

$$\hat{m}(x_0) = \frac{\sum_{i=1}^n K[(x_i - x_0)/h] y_i}{\sum_{i=1}^n K[(x_i - x_0)/h]} \quad (27.19)$$

由于使用了  $x_0$  附近邻域中观测值的信息,核回归估计量  $\hat{m}(x_0)$  也是有偏的。可以证明(参见附录),在大样本下其偏差为

$$\text{Bias}(x_0) \equiv E[\hat{m}(x_0)] - m(x_0) = h^2 \left[ m'(x_0) \frac{f'(x_0)}{f(x_0)} + \frac{1}{2} m''(x_0) \right] \int_{-\infty}^{+\infty} z^2 K(z) dz \quad (27.20)$$

因此,  $\text{Bias}(x_0) = O(h^2)$ 。而核回归估计的方差为

$$\text{Var}[\hat{m}(x_0)] = \frac{1}{nh} \frac{\sigma_s^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz + o(1/nh) \quad (27.21)$$

故  $\text{Var}[\hat{m}(x_0)] = O(1/nh)$ 。因此,与核密度估计的情形类似,如果当  $n \rightarrow \infty$  时,让带宽  $h \rightarrow 0$ ,而且  $nh \rightarrow \infty$ ,则根据均方收敛,核回归估计是一致的。可以证明(参见附录),如果  $\{x_1, x_2, \dots, x_n\}$  为独立同分布的,则核回归估计量  $\hat{m}(x_0)$  服从渐近正态分布:

$$\sqrt{nh} [\hat{m}(x_0) - m(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N \left( 0, \frac{\sigma_s^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz \right) \quad (27.22)$$

由于  $\text{Bias}(x_0) = O(h^2)$  且  $\text{Var}[\hat{m}(x_0)] = O(1/nh)$ ,最小化 IMSE 的结果显示,最优带宽为  $h^* = O(n^{-0.2})$ 。然而,由于最优带宽  $h^*$  取决于待估计回归函数的导数(因为  $m'(x_0), m''(x_0)$  出现在偏差的表达式(27.20)中),而要估计  $m'(x_0), m''(x_0)$  又需要指定最优带宽  $h^*$ ,故在实践上常使用以下“交叉核实”(Cross Validation,简记 CV)的方法来确定最优带宽  $h^*$ 。其思想是,在估计  $\hat{m}(x_i)$  时,不使用  $y_i$  的信息,看其余观测值预测  $y_i$  的能力有多强;而这个能力又取决于带宽  $h$ 。故选择带宽  $h$ ,使得此预测能力最强,即最小化以下目标函数:

$$\min_h \text{CV}(h) \equiv \sum_{i=1}^n [y_i - \hat{m}_{-1}(x_i)]^2 \pi(x_i) \quad (27.23)$$

其中,  $\hat{m}_{-1}(x_i) \equiv \frac{\sum_{j \neq i} w_{ji,h} y_j}{\sum_{j \neq i} w_{ji,h}}$  是对  $m(x_i)$  的“去掉一个观测值”估计量(leave-one-out estimate),即

$j = 1, \dots, n$ ,但  $j \neq i$ 。 $\pi(x_i)$  是权重函数(weighting function),主要是为了给边界附近的端点(end points)更小的权重,以避免端点可能对估计量带来较大的扭曲。比如,可以不考虑  $x_i$  的 5% 分位数以下与 95% 分位数以上的观测值,即对这些观测值令  $\pi(x_i) = 0$ ,而对其余观测值令  $\pi(x_i) = 1$ 。

之所以要去掉自身第  $i$  个观测值是因为,如果将它保留,则总可以选择足够小的带宽  $h$  使得对于任何  $i$ ,都有  $\hat{m}(x_i) = y_i$ <sup>①</sup>,故  $CV(h) = 0$  得到最小化。但这个过小的带宽  $h$  并不最优。 $CV(h)$  也被称为“预测误差的估计值”(estimated prediction error),因为它衡量的是用  $\hat{m}_{-i}(x_i)$  来预测  $y_i$  会带来多大的预测误差。可以证明,最小化  $CV(h)$  与最小化 IMSE 是渐近等价的。在实践上,交叉核实并非决定最优带宽的完美方法,故仍经常辅之以不严格的眼球法。另外,可以证明,能使  $IMSE(h^*)$  最小化的核函数为“伊番科尼可夫核”(Epanechnikov 或 quadratic kernel),尽管它只有微弱的优势。

## 27.7 多元核回归

更一般地,对于  $k$  维解释向量  $x$ ,考虑如下非参数多元回归模型:

$$y_i = m(x_i) + \varepsilon_i = m(x_{1i}, x_{2i}, \dots, x_{ki}) + \varepsilon_i \quad (27.24)$$

其中,  $m(\cdot)$  是未知的多元函数。在  $x_0$  处的核回归估计量可以写为

$$\hat{m}(x_0) = \frac{\sum_{i=1}^n K[(x_i - x_0)/h] y_i}{\sum_{i=1}^n K[(x_i - x_0)/h]} \quad (27.25)$$

其中,  $K(\cdot)$  为  $k$  维核函数。多元核回归估计量的性质与一元核回归相似。但最优带宽为  $h^* = O(n^{-1/(k+4)})$  (大于一元情形下的最优带宽),而  $\hat{m}(x_0)$  收敛到真实值的速度也更慢。多元回归的解释变量越多,则收敛的速度越慢,对样本容量的要求也就越大。这种“维度的诅咒”(curse of dimensionality)大大限制了多元非参数回归的应用。一种解决方法是,使用半参数估计(semiparametric estimation)来降低模型中非参数部分的维度,参见下文。

## 27.8 $k$ 近邻回归

事实上,核回归估计是局部加权平均估计量(local weighted average estimator)的一个特例,因为它使用的是一个特别的权重。选择权重的另一方式是,对于最靠近  $x_0$  的  $k$  个  $x_i$  的观测值都给予相同的权重,而对其余观测值则给予权重 0,即只对最靠近  $x_0$  的  $k$  个观测值进行简单算术平均(不加权)。记  $N_k(x_0)$  为最靠近  $x_0$  的  $k$  个  $x_i$  观测值的集合(包括  $x_0$  自身),则  $k$  近邻估计量( $k$ -nearest neighbor estimator)可以定义为

$$\hat{m}_{KNN}(x_0) = \frac{1}{k} \sum_{i=1}^n \mathbf{1}\{x_i \in N_k(x_0)\} \cdot y_i \quad (27.26)$$

这个估计量可以看成是使用“均匀核”(uniform kernel)的核估计,但其带宽却是可变的,而且可以不对称(左边的带宽不一定等于右边的带宽)。“对称化”(symmetrized)的  $k$  近邻估计量则对小于  $x_0$  的  $(k-1)/2$  观测值与大于  $x_0$  的  $(k-1)/2$  观测值进行简单算术平均。

在某种意义上, $k$  近邻估计量相当于移动平均(moving average)。由于  $k$  近邻估计量使用的是简单算术平均,而不是核回归估计所使用的加权平均,故前者可能不如后者光滑(正如直方图

<sup>①</sup> 可以选择足够小的带宽,使得  $x_i$  邻域中的观测值只剩下  $y_i$ ,故  $\hat{m}(x_i) = y_i$ 。

不如核密度估计光滑)。而且,越靠近端点,则可用于进行移动平均的样本点就越少,估计量也就越不准确。这种边界问题(boundary problem)可通过下面的“局部线性回归”来缓解。

## 27.9 局部线性回归

前面介绍的核回归估计量,实际上是一个“局部常数估计”(local constant estimator),因为它假定在  $x_0$  附近的某个邻域里,  $m(x)$  均等于一个常数。局部线性回归则假定  $m(x)$  在  $x_0$  附近的某个邻域里为线性函数,即在该邻域里,  $m(x) = a_0 + b_0(x - x_0)$ , 然后使用加权最小二乘法(WLS)来估计这个线性函数(因为不同观测点到  $x_0$  的距离不同,所包含的信息量也因此不同),即最小化以下目标函数:

$$\min_{[a_0, b_0]} \sum_{i=1}^n K[(x_i - x_0)/h] [y_i - a_0 - b_0(x_i - x_0)]^2 \quad (27.27)$$

其中,  $K(\cdot)$  为核函数,即权重函数。显然,离  $x_0$  越近,则权重越大(除非使用均匀核,则权重都一样,等价于标准 OLS 回归)。因此,在  $x_0$  附近的小邻域里,  $\hat{m}(x) = \hat{a}_0 + \hat{b}_0(x - x_0)$ 。当  $x$  正好等于  $x_0$  时,  $\hat{m}(x_0) = \hat{a}_0$ , 而一阶导数  $\hat{m}'(x_0) = \hat{b}_0$ 。这种方法称为“局部线性回归”(local linear regression),由 Fan(1992)首倡,故也称为“范回归”(Fan regression)。Fan(1992)表明,局部线性回归不仅能较好地解决“边界问题”,而且比核回归更有效率且适用于更多数据类型。直观来看,如果带宽足够小,则在此小邻域内,一般的函数都可以很好地用线性函数来近似,故局部线性回归具有较好的性质。

更一般地,可以假定  $m(x)$  在  $x_0$  附近的某个邻域为  $p$  级多项式。“局部  $p$  级多项式估计量”(local polynomial estimator of degree  $p$ ) 最小化以下目标函数:

$$\min_{[a_0, b_0]} \sum_{i=1}^n K[(x_i - x_0)/h] \left[ y_i - a_{0,0} - a_{0,1}(x_i - x_0) - \cdots - \frac{a_{0,p}}{p!}(x_i - x_0)^p \right]^2 \quad (27.28)$$

局部线性或局部多项式估计量有助于缓解前面提及的在两端估计不准确的边界问题。一个常用的局部回归估计量为 Cleveland(1979)所提出的“局部加权散点光滑估计量”(Locally weighted scatterplot smoothing,简记 Lowess),这是局部多项式估计量的一个变种(variant)或升级版。该估计量使用“三三核”(tricubic kernel),同时使用可变带宽  $h_{0,k}$ (由  $x_0$  到其最近的  $k$  个观测值的距离所决定<sup>①</sup>),以及对较大的残差 [ $y_i - \hat{m}(x_i)$ ] 给予较小的权重。Lowess 估计的优点是,它使用了可变带宽(依数据的稠密程度而定),对于极端值更稳健,而且缓解了在两端估计不准的边界问题。

## 27.10 非参数估计的 Stata 命令及实例

Stata 提供了如下有关非参数估计的命令。

### 1. 直方图与核密度估计

`hist y, width(#) start(#)` (画  $y$  的直方图,指定组宽与起始点)

<sup>①</sup> 在 Stata 中,默认  $k = 0.8n$ ,其中  $n$  为样本容量。

```
kdensity y,bwidth(#) n(#) (对 y 进行核密度估计,指定带宽与估计点数)
kdensity y,norm (同时画一个正态分布,作为对比)
```

有关直方图与核密度估计的 Stata 实例,参见第 4 章。

## 2. 核密度回归

```
ssc install kernreg1 (下载安装命令 kernreg1)
```

```
kernreg1 y x,bwidth(#) kercode(#) npoint(#) gen(mhvar gridvar)
```

其中,选择项“`bwidth(#)`”用于指定带宽,默认值为最优带宽;“`kercode(#)`”为必选项,用于指定核函数,其中 1 = Uniform,2 = Triangle,3 = Epanechnikov,4 = Quartic(Biweight),5 = Triweight,6 = Gaussian,7 = Cosinus;“`npoint(#)`”为必选项,用于指定将解释变量“`x`”的取值范围分为多少个等距离的点(称为“网格点”,grid point),并在这些点上进行核密度回归;选择项“`gen(mhvar gridvar)`”产生两个变量,其中“`gridvar`”为解释变量“`x`”的网格点,而“`mhvar`”为对应的核密度回归值。另外,该程序将最优带宽记为“全局宏”(global macro)“`S_1`”(字母“`s`”为大写),故可以用“`$S_1`”来调用它。

下面以数据集 `mpyr.dta` 为例(参见第 21 章),将 `logv`(货币流通速度的对数)对 `r`(名义利率)进行核密度回归,结果如图 27.3:

```
. use mpyr.dta,clear
. kernreg1 logv r,k(3) np(100) gen(logv_kern r_grid)
```

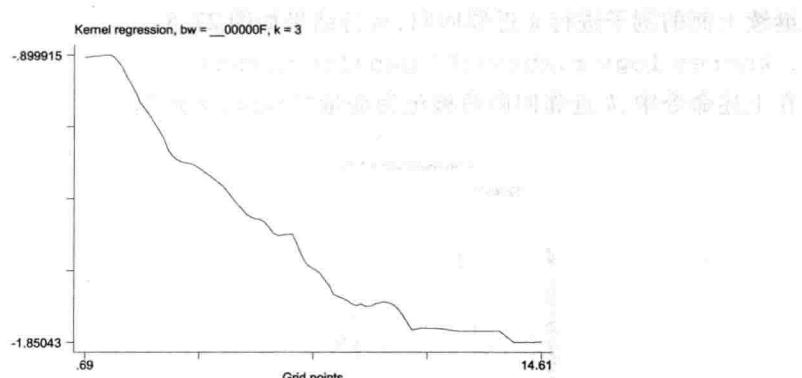


图 27.3 核密度回归

此命令使用了伊番科尼可夫核( $k=3$ ),在 100 个等距离的网格点进行估计,并将解释变量 `r`(利率)的网格点记为变量“`r_grid`”,对应的核密度回归值记为“`logv_kern`”。

更直观地,可将散点图、线性回归线及核密度回归线画在一张图上(如图 27.4):

```
. graph twoway (scatter logv r)(lfit logv r,lpattern("-"))(line logv_kern r_grid)
```

由于该命令的默认带宽为最优带宽,故可以利用该程序计算最优带宽。

```
. dis $S_1
```

```
. 90306211
```

这表明,最优带宽为 0.90。

## 3. $k$ 近邻回归

$k$  近邻回归可通过非官方命令 `knnreg` 来实现:

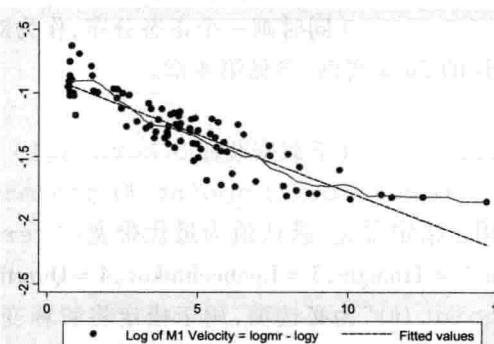


图 27.4 散点图、线性回归线与核密度回归

```
net install.snp10.pkg (下载安装命令 knnreg)
```

该命令的格式为

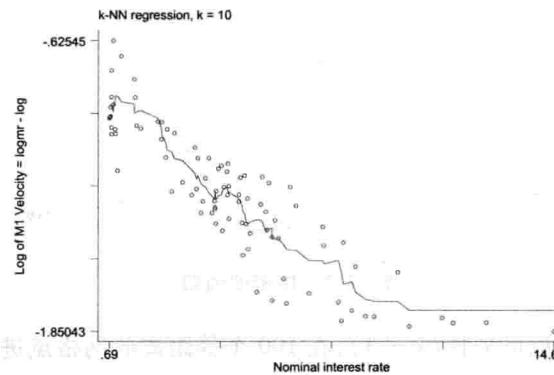
```
knnreg y x, knum(#) gen(mkvar)
```

其中,选择项“`knum(#)`”为必选项,用于指定  $k$  近邻回归中  $k$  的取值;选择项“`gen(mkvar)`”将  $k$  近邻回归值赋予变量“`mkvar`”。

继续上面的例子进行  $k$  近邻回归,运行结果如图 27.5:

```
. knnreg logv r, knum(10) gen(logv_knn)
```

在上述命令中,  $k$  近邻回归值被记为变量“`logv_knn`”。

图 27.5  $k$  近邻回归

#### 4. 局部多项式回归

```
lpoly y x, bwidth(#) kernel(kernel) degree(#) gen(gridvar newvar)
```

其中,选择项“`bwidth(#)`”用于指定带宽,默认值为“rule-of-thumb (ROT) bandwidth estimator”;“`kernel(kernel)`”用于指定核函数,默认值为“`kernel(epanechnikov)`”;选择项“`degree(#)`”用于指定多项式的级数,默认值为“`degree(0)`”,即只有常数项(局部均值平滑, local-mean smoothing);选择项“`gen(gridvar newvar)`”将产生两个变量,其中“`gridvar`”为解释变量“`x`”的网格点,而“`newvar`”为对应的局部多项式回归值。

继续上面的例子进行局部多项式回归,结果如图 27.6:

```
. lpoly logv r,b($S_1) degree(1) gen(r_grid_l logv_lpoly)
```

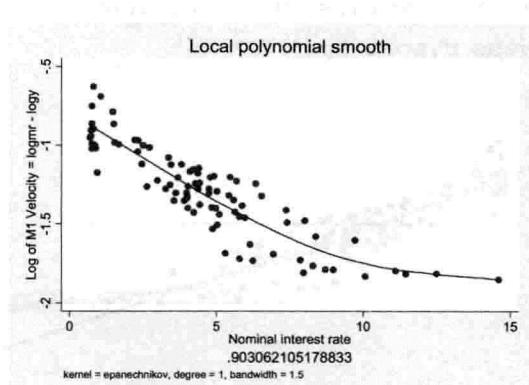


图 27.6 局部多项式回归

其中,选择项“`b($S_1)`”表示使用前面核密度回归的最优带宽(0.90);选择项“`degree(1)`”表明,这是一个局部线性回归;选择项“`gen(r_grid_l logv_lpoly)`”将解释变量 `r`(利率)的网格点记为“`r_grid_l`”,而把对应的局部回归值记为“`logv_lpoly`”。

### 5. Lowess 回归

```
lowess y x,bwidth(#) gen(newvar)
```

其中,选择项“`bwidth(#)`”用于指定带宽,默认值为0.8;选择项“`gen(newvar)`”将被解释变量的光滑值(smoothed values)赋予变量“`newvar`”。

继续以数据集 mpyr.dta 为例,结果如图 27.7:

```
. lowess logv r,b($S_1) gen(logv_lowess)
```

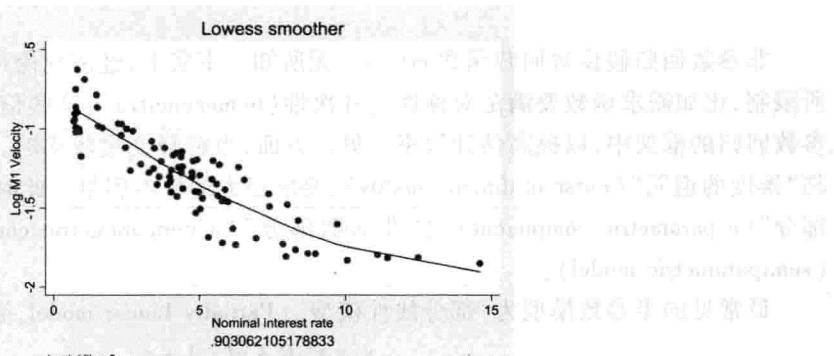


图 27.7 Lowess 回归

其中,选择项“`b($S_1)`”表示使用前面核密度回归的最优带宽(0.90);选择项“`gen(logv_lowess)`”把被解释变量的光滑值记为“`logv_lowess`”。

最后,将以上所有各图合并在一起(如图 27.8):

```
. graph twoway (scatter logv r)(lfit logv r,lp("-"))(line logv_knn r,lp("-"))(line logv_kern r_grid)(line logv_lpoly r_grid_l,lp("_."))(line logv_lowess r,lp("."))
```

```
graph twoway (scatter logv r)(lfit logv r,lp("-"))(line logv_knn r,sort lp("-"))(line logv_kern r_grid,sort)(line logv_lpoly r_grid_l,sort lp("-."))(line logv_lowess r,sort lp("."))
```

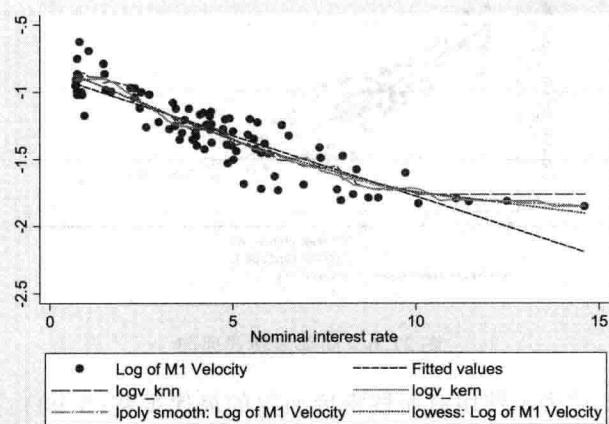


图 27.8 各种非参数回归的比较

从图 27.8 可知,以上各种非参数回归的结果很接近,但与线性回归有所不同。非参数估计对数据的拟合程度明显优于线性回归(尤其在右端)。总之,非参数回归以牺牲部分光滑度为代价(线性回归的光滑度为无穷),可以更好地拟合数据。

## 27.11 半参数估计

非参数回归假设对回归函数  $m(x)$  一无所知。事实上,经济理论可能对  $m(x)$  的具体形式有所限制,比如需求函数要满足对称性与齐次性(homogeneity)。这些信息应该而且可以被纳入非参数回归的框架中,以提高估计效率。另一方面,当解释变量较多时,完全的非参数方法可能面临“维度的诅咒”(curse of dimensionality),要求很大的样本容量。此时,可以使用同时包含“参数部分”(a parametric component)与“非参数部分”(a nonparametric component)的“半参数模型”(semiparametric model)。

最常见的半参数模型为“部分线性模型”(Partially Linear model,简记 PL),即

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + g(\mathbf{z}_i) + \varepsilon_i \quad (27.29)$$

其中,参数部分  $\mathbf{x}'_i \boldsymbol{\beta}$  为线性函数,而非参数部分  $g(\mathbf{z}_i)$  为未知函数(连函数形式也不知道),并假设扰动项  $\varepsilon_i$  均值独立于  $\mathbf{x}_i, \mathbf{z}_i$ ,即  $E(\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i) = 0$ 。Robinson(1988)提出了以下“罗宾逊差分估计量”(Robinson difference estimator)。给定  $\mathbf{z}_i$ ,对方程(27.29)两边取条件期望可得

$$E(y_i | \mathbf{z}_i) = E(\mathbf{x}'_i \boldsymbol{\beta} + g(\mathbf{z}_i) + \underbrace{E(\varepsilon_i | \mathbf{z}_i)}_{=0}) \quad (27.30)$$

其中,根据迭代期望定律,

$$E(\varepsilon_i | \mathbf{z}_i) = E_{\mathbf{x}_i} E[(\varepsilon_i | \mathbf{z}_i) | \mathbf{x}_i] = E_{\mathbf{x}_i} \underbrace{E(\varepsilon_i | \mathbf{x}_i, \mathbf{z}_i)}_{=0} = 0 \quad (27.31)$$

将方程(27.29)减去方程(27.30)可得

$$y_i - E(y_i | z_i) = [x_i - E(x_i | z_i)]' \beta + \varepsilon_i \quad (27.32)$$

在这个“差分方程”中,未知函数  $g(z_i)$  被消去了,而条件期望  $E(y_i | z_i)$  与  $E(x_i | z_i)$  则可以用非参数方法来估计(比如,核回归)。假设  $\hat{E}(y_i | z_i)$  与  $\hat{E}(x_i | z_i)$  分别为对  $E(y_i | z_i)$  与  $E(x_i | z_i)$  的非参数估计,则可对以下线性方程进行最小二乘估计:

$$y_i - \hat{E}(y_i | z_i) = [x_i - \hat{E}(x_i | z_i)]' \beta + u_i \quad (27.33)$$

记此估计量为  $\hat{\beta}_{PL}$ 。由于使用了条件期望的估计量来替代条件期望本身,故扰动项不再是  $\varepsilon_i$ , 记新扰动项为  $u_i$ 。假设  $\varepsilon_i$  为 iid( $0, \sigma^2$ ), 根据大样本 OLS 理论, 可以证明,

$$\sqrt{n}(\hat{\beta}_{PL} - \beta) \xrightarrow{d} N\left[0, \sigma^2 \left( \operatorname{plim}_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n w_i w_i' \right)^{-1} \right] \quad (27.34)$$

其中,  $w_i = x_i - E(x_i | z_i)$ 。在上式中, 只要用  $\hat{E}(x_i | z_i)$  来替代  $E(x_i | z_i)$  就可以估计渐近协方差矩阵  $\text{Avar}(\hat{\beta}_{PL})$ 。如果存在异方差, 则可以使用标准的夹心估计量来得到稳健标准误。

最后, 可以得到对  $g(z_i)$  的非参数估计, 即

$$\hat{g}(z_i) = \hat{E}(y_i | z_i) - \hat{E}(x_i | z_i)' \hat{\beta} \quad (27.35)$$

对于部分线性模型使用罗宾逊差分估计量, 可通过非官方命令 `semipar` 来实现:

```
.ssc install semipar (下载安装命令 semipar)
```

该命令的格式为

```
semipar y x1 x2 x3, nonpar(z) robust cluster(varname) kernel(kernel) gen(varname) ci partial(varname)
```

其中, 必选项“`nonpar(z)`”用来指定半参模型中非参部分, 选择项“`robust`”表示使用异方差稳健标准误, 选择项“`cluster(varname)`”表示使用聚类稳健标准误。选择项“`gen(varname)`”用来记录  $\hat{E}(y_i | z_i)$  (nonparametric fit of the dependent variable); 选择项“`ci`”表示在图上给出此 nonparametric fit 的置信区间; 选择项“`partial(varname)`”用来记录  $[y_i - \hat{E}(y_i | z_i)]$  (dependent variable partialled out from the nonparametric fit)。选择项“`kernel(kernel)`”用来指定用于核回归的核函数, 默认为高斯核, 即“`kernel(gaussian)`”, 更多备选核函数参见“`help semipar`”。

下面以第 4 章的数据集 `nerlove.dta` 为例。假设我们不确定原材料价格( $\ln PF$ )对总成本( $\ln TC$ )的作用函数形式, 故考虑以下部分线性模型:

$$\ln TC_i = \beta_1 \ln Q_i + \beta_2 \ln PL_i + \beta_3 \ln PK_i + g(\ln PF_i) + \varepsilon_i \quad (27.36)$$

其中,  $g(\cdot)$  为未知函数。

作为参照, 首先进行 OLS 回归。

```
.use nerlove.dta, clear
.reg lntc lnq lnpl lnpk lnpf, r
.est sto reg
```

Linear regression						
Number of obs = 145						
F( 4, 140) = 177.19						
Prob > F = 0.0000						
R-squared = 0.9260						
Root MSE = .39227						
lntc		Robust				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnq	.7209135	.0325376	22.16	0.000	.656585	.785242
lnpl	.4559645	.260326	1.75	0.082	-.0587139	.9706429
lnpk	-.2151476	.3233711	-0.67	0.507	-.8544698	.4241745
lnpf	.4258137	.0740741	5.75	0.000	.2793653	.5722622
_cons	-3.566513	1.718304	-2.08	0.040	-6.963693	-.1693331

其次, 使用罗宾逊差分估计量进行半参估计。

```
.semipar lntc lnq lnpl lnpk,nonpar(lnpf) robust xtitle(lnPF) ytitle(lnTC) ci  
.est sto semipar
```

Number of obs = 145						
R-squared = 0.9273						
Adj R-squared = 0.9258						
Root MSE = 0.3749						
lntc		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnq	.7237013	.030765	23.52	0.000	.6628846	.784518
lnpl	.3398899	.2931278	1.16	0.248	-.2395684	.9193481
lnpk	-.3549753	.3252955	-1.09	0.277	-.9980231	.2880725

该命令还会输出被解释变量 lnTC 对非参变量 lnPF 的核回归图(参见图 27.9)。

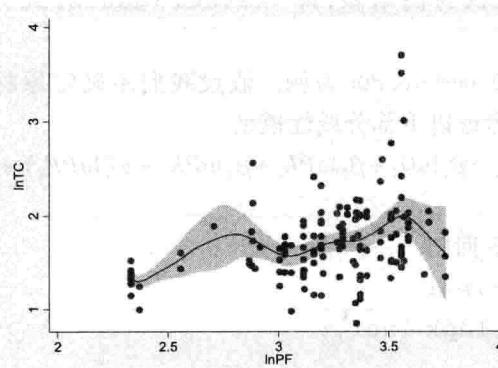


图 27.9 lnTC 对 lnPF 的核回归图

下面,对比这两种回归方法的结果。

```
.esttab reg semipar, mtitles r2 se star(* 0.1 ** 0.05 *** 0.01)
```

	(1) reg	(2) semipar
lnq	0.721*** (0.0325)	0.724*** (0.0308)
lnpl	0.456* (0.260)	0.340 (0.293)
lnpk	-0.215 (0.323)	-0.355 (0.325)
lnpf	0.426*** (0.0741)	
_cons	-3.567** (1.718)	
N	145	145
R-sq	0.926	0.927

Standard errors in parentheses  
\* p<0.1, \*\* p<0.05, \*\*\* p<0.01

从上表可知,半参估计结果与线性回归差别不大。当然,本例纯粹为演示而设,在其他情况下,二者可能有较大差别。

## 习题

### 27.1 使用数据集 mpyr.dta,进行以下估计。

- (1) 画变量 logv 的直方图;
- (2) 画变量 logv 的核密度图,使用眼球法选择适当的带宽;
- (3) 作为对照,在 logv 的核密度图上加上正态分布密度图。

### 27.2 参照本章的实例,使用数据集 consumption\_china.dta 对中国的消费函数进行以下非参数估计。

- (1) 核密度回归;
- (2) k 近邻回归;
- (3) 局部多项式回归(一阶);
- (4) Lowess 回归。

## 附录

### A27.1 证明核密度估计的偏差

证明:假设样本  $|x_1, x_2, \dots, x_n|$  为独立同分布的。核密度估计量的公式为  $\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(x_i - x_0)/h]$ , 其期望值为

$$\begin{aligned} E[\hat{f}(x_0)] &= E\left[\frac{1}{h} K((x - x_0)/h)\right] \quad (\text{对 } x \text{ 求期望}) \\ &= \int_{-\infty}^{+\infty} \frac{1}{h} K((x - x_0)/h) f(x) dx \quad (\text{期望的定义}) \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} K(z)f(x_0 + hz) dz \quad (\text{积分变换 } z \equiv (x - x_0)/h, x = x_0 + hz) \\
&\approx \int_{-\infty}^{+\infty} K(z) \left[ f(x_0) + f'(x_0)hz + \frac{1}{2}f''(x_0)(hz)^2 \right] dz \quad (\text{二阶泰勒近似}) \\
&= f(x_0) \underbrace{\int_{-\infty}^{+\infty} K(z) dz}_{=1} + hf'(x_0) \underbrace{\int_{-\infty}^{+\infty} zK(z) dz}_{=0} + \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz \quad (\text{核函数的性质}) \\
&= f(x_0) + \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz
\end{aligned}$$

因此,  $\text{Bias}(x_0) \equiv E[\hat{f}(x_0)] - f(x_0) \approx \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz$  (移项)。

### A27.2 证明核密度估计的方差公式

证明: 假设  $\{y_i\}$  为独立同分布的, 则  $\text{Var}(\bar{y}) = \frac{1}{n}\text{Var}(y) = \frac{1}{n}\{E(y^2) - [E(y)]^2\}$ , 故

$$\text{Var}[\hat{f}(x_0)] = \frac{1}{n}E\left\{\frac{1}{h}K((x - x_0)/h)\right\}^2 - \frac{1}{n}\left\{E\left[\frac{1}{h}K((x - x_0)/h)\right]\right\}^2 \quad (27.37)$$

其中, 上式右边的第一项等于

$$\begin{aligned}
E\left\{\frac{1}{h}K((x - x_0)/h)\right\}^2 &= \int_{-\infty}^{+\infty} \frac{1}{h^2}[K((x - x_0)/h)]^2 f(x) dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{h}K(z)^2 f(x_0 + hz) dz \quad (\text{积分变换 } z \equiv (x - x_0)/h) \\
&\approx \int_{-\infty}^{+\infty} \frac{1}{h}K(z)^2 [f(x_0) + f'(x_0)hz] dz \quad (\text{一阶泰勒近似}) \\
&= \frac{1}{h}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz + f'(x_0) \int_{-\infty}^{+\infty} zK(z)^2 dz \quad (\text{分成两项})
\end{aligned}$$

根据附录 A27.1,  $E\left[\frac{1}{h}K((x - x_0)/h)\right] \approx f(x_0) + \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz$ , 故方程 (27.37) 第二项中的  $E\left[\frac{1}{h}K((x - x_0)/h)\right]^2 = \left[f(x_0) + \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz\right]^2$ 。综合以上结果,

$$\begin{aligned}
\text{Var}[\hat{f}(x_0)] &\approx \frac{1}{nh}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz + \frac{1}{n}f'(x_0) \int_{-\infty}^{+\infty} zK(z)^2 dz - \\
&\quad \frac{1}{n}\left[f(x_0) + \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz\right]^2 \quad (27.38)
\end{aligned}$$

当  $n \rightarrow \infty$  且  $h \rightarrow 0$  时, 上式的性质由第一项  $\frac{1}{nh}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz$  所决定 (其余两项为高阶无穷小), 故

$$\text{Var}[\hat{f}(x_0)] = \frac{1}{nh}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz + o(1/nh)。$$

### A27.3 证明最优带宽

证明: 由于  $\text{IMSE} = \int_{-\infty}^{+\infty} \text{MSE}[\hat{f}(x_0)] dx_0 = \int_{-\infty}^{+\infty} \{[\text{Bias}(x_0)]^2 + \text{Var}[\hat{f}(x_0)]\} dx_0$ ,

而  $\text{Bias}(x_0) = \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz$ ,  $\text{Var}[\hat{f}(x_0)] \approx \frac{1}{nh}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz$ , 故

$$\begin{aligned}
\text{IMSE} &= \int_{-\infty}^{+\infty} \left( \left[ \frac{1}{2}h^2f''(x_0) \int_{-\infty}^{+\infty} z^2K(z) dz \right]^2 \right) dx_0 + \int_{-\infty}^{+\infty} \left[ \frac{1}{nh}f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \right] dx_0 \\
&= \frac{1}{4}h^4 \int_{-\infty}^{+\infty} \left( [f''(x_0)]^2 \left[ \int_{-\infty}^{+\infty} z^2K(z) dz \right]^2 \right) dx_0 + \frac{1}{nh} \int_{-\infty}^{+\infty} \left[ f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \right] dx_0 \\
&= \frac{1}{4}h^4 \left[ \int_{-\infty}^{+\infty} z^2K(z) dz \right]^2 \int_{-\infty}^{+\infty} [f''(x_0)]^2 dx_0 + \frac{1}{nh} \left[ \int_{-\infty}^{+\infty} K(z)^2 dz \right] \underbrace{\left[ \int_{-\infty}^{+\infty} f(x_0) dx_0 \right]}_{=1} \\
&= \frac{1}{4}h^4 \left[ \int_{-\infty}^{+\infty} z^2K(z) dz \right]^2 \int_{-\infty}^{+\infty} [f''(x_0)]^2 dx_0 + \frac{1}{nh} \int_{-\infty}^{+\infty} K(z)^2 dz
\end{aligned}$$

将上式对  $h$  求导，并令其导数为 0，

$$\frac{\partial \text{IMSE}}{\partial h} = h^3 \left[ \int_{-\infty}^{+\infty} z^2 K(z) dz \right]^2 \int_{-\infty}^{+\infty} [f''(x_0)]^2 dx_0 + \frac{1}{n} \int_{-\infty}^{+\infty} K(z)^2 dz \left( -\frac{1}{h^2} \right) = 0 \quad (27.39)$$

整理可得最优带宽为  $h^* = \delta \left[ \int_{-\infty}^{+\infty} f''(x_0)^2 dx_0 \right]^{-0.2} n^{-0.2}$ , 其中常数

$$\delta \equiv \left[ \int_{-\infty}^{+\infty} K(z)^2 dz / \left( \int_{-\infty}^{+\infty} z^2 K(z) dz \right)^2 \right]^{0.2} \quad (27.40)$$

#### A27.4 证明核回归估计量 $\hat{m}(x_0)$ 服从渐近正态分布

$$\sqrt{nh} [\hat{m}(x_0) - m(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N\left(0, \frac{\sigma_e^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz\right) \quad (27.41)$$

证明：由于核回归估计量是一个加权平均，即  $\hat{m}(x_0) = \sum_{i=1}^n w_{i0,h} y_i$ ，故

$$\hat{m}(x_0) - m(x_0) = \sum_{i=1}^n w_{i0,h} y_i - \sum_{i=1}^n w_{i0,h} m(x_0) \quad (\text{权重之和为 } 1)$$

$$= \sum_{i=1}^n w_{i0,h} [y_i - m(x_0)] \quad (\text{两项合并})$$

$$= \sum_{i=1}^n w_{i0,h} [m(x_i) - m(x_0) + \varepsilon_i] \quad (\text{代入模型的假定 } y_i = m(x_i) + \varepsilon_i)$$

由于权重  $w_{i0,h} = \frac{K[(x_i - x_0)/h]}{\sum_{i=1}^n K[(x_i - x_0)/h]}$ ，而核密度估计  $\hat{f}(x_0) = \frac{1}{nh} \sum_{i=1}^n K[(x_i - x_0)/h]$ ，故权重可以写为

$$w_{i0,h} = \frac{\frac{1}{nh} K[(x_i - x_0)/h]}{\hat{f}(x_0)}。因此，用  $\sqrt{nh}$  来标准化变量可得$$

$$\begin{aligned} \sqrt{nh} [\hat{m}(x_0) - m(x_0)] &= \sqrt{nh} \sum_{i=1}^n \frac{\frac{1}{nh} K[(x_i - x_0)/h]}{\hat{f}(x_0)} [m(x_i) - m(x_0) + \varepsilon_i] \quad (\text{代入 } w_{i0,h}) \\ &= \frac{1}{\sqrt{nh}} \frac{\sum_{i=1}^n K[(x_i - x_0)/h] [m(x_i) - m(x_0) + \varepsilon_i]}{\hat{f}(x_0)} \quad (\text{合并整理}) \end{aligned}$$

我们已经知道  $\hat{f}(x_0)$  是  $f(x_0)$  的一致估计量，即  $\hat{f}(x_0) \xrightarrow{P} f(x_0)$ 。故只需要考虑上式中分子的渐近分布。将上式中的分子分成两个部分：

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] [m(x_i) - m(x_0)] + \frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] \varepsilon_i \quad (27.42)$$

对于第一部分，在大数定律成立的情况下，它将收敛到其期望值，因为

$$\begin{aligned} &\mathbb{E} \left\{ \frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] [m(x_i) - m(x_0)] \right\} \\ &= \frac{\sqrt{n}}{\sqrt{h}} \mathbb{E} [K((x - x_0)/h) [m(x) - m(x_0)]] \quad (\{x_1, x_2, \dots, x_n\} \text{ 为独立同分布的，省略下标 } i) \\ &= \frac{\sqrt{n}}{\sqrt{h}} \int_{-\infty}^{+\infty} K((x - x_0)/h) [m(x) - m(x_0)] f(x) dx \quad (\text{期望的定义}) \\ &= \sqrt{nh} \int_{-\infty}^{+\infty} K(z) [m(x_0 + hz) - m(x_0)] f(x_0 + hz) dz \quad (\text{积分变换 } z \equiv \frac{x - x_0}{h}) \\ &\approx \sqrt{nh} \int_{-\infty}^{+\infty} K(z) \left[ hzm'(x_0) + \frac{1}{2} h^2 z^2 m''(x_0) \right] [f(x_0) + f'(x_0) hz] dz \quad (\text{泰勒展开}) \end{aligned}$$

将上式的被积函数展开成四项。根据核函数的性质，其中第一项为

$$\sqrt{nh} \int_{-\infty}^{+\infty} K(z) hzm'(x_0) f(x_0) dz = h \sqrt{nh} m'(x_0) f(x_0) \int_{-\infty}^{+\infty} z K(z) dz = 0 \quad (27.43)$$

而第四项为

$$\frac{1}{2} \sqrt{nh} \int_{-\infty}^{+\infty} K(z) h^3 z^3 m''(x_0) f'(x_0) dz = \frac{1}{2} h^3 \sqrt{nh} m''(x_0) f'(x_0) \int_{-\infty}^{+\infty} z^3 K(z) dz \quad (27.44)$$

显然,第四项是第二项与第三项的高阶无穷小,故可以忽略。因此,只要考虑第二项与第三项即可:

$$\begin{aligned} & E \left\{ \frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] [m(x_i) - m(x_0)] \right\} \\ & \approx \sqrt{nh} \int_{-\infty}^{+\infty} K(z) \left[ h^2 z^2 m'(x_0) f'(x_0) + \frac{1}{2} h^2 z^2 m''(x_0) f(x_0) \right] dz \quad (\text{泰勒展开}) \\ & = \sqrt{nh} h^2 \left[ m'(x_0) f'(x_0) + \frac{1}{2} h m''(x_0) f(x_0) \right] \int_{-\infty}^{+\infty} z^2 K(z) dz \quad (\text{合并整理}) \\ & = \sqrt{nh} f(x_0) \text{Bias}(x_0) \end{aligned}$$

其中,  $\text{Bias}(x_0) = h^2 \left[ m'(x_0) \frac{f'(x_0)}{f(x_0)} + \frac{1}{2} h m''(x_0) \right] \int_{-\infty}^{+\infty} z^2 K(z) dz$

对于第二部分  $\frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] \varepsilon_i$ , 显然其中每一项的期望值均为 0, 而每一项的方差为

$$\begin{aligned} \text{Var} \{ K[(x_i - x_0)/h] \varepsilon_i \} &= E \{ K^2[(x_i - x_0)/h] \varepsilon_i^2 \} \quad (\text{公式 } \text{Var}(y) = E(y^2) - [E(y)]^2) \\ &= E_x E_{\varepsilon} \{ K^2[(x_i - x_0)/h] \varepsilon_i^2 | x \} \quad (\text{迭代期望公式}) \\ &= E_x \{ K^2[(x_i - x_0)/h] \text{Var}(\varepsilon_i^2 | x) \} \quad (\text{对 } \varepsilon \text{ 求期望}) \\ &= \int_{-\infty}^{+\infty} K^2[(x_i - x_0)/h] \text{Var}(\varepsilon_i^2 | x) f(x) dx \quad (\text{对 } x \text{ 求期望}) \\ &= h \int_{-\infty}^{+\infty} K(z)^2 \text{Var}(\varepsilon_i^2 | x_0 + hz) f(x_0 + hz) dz \quad (\text{积分变换 } z = \frac{x - x_0}{h}) \\ &= h \text{Var}(\varepsilon_i^2 | x_0) f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \quad (\text{大样本下 } h \rightarrow 0) \\ &= h \sigma_{\varepsilon}^2 f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \quad (\text{根据假定, } \varepsilon_i \sim \text{iid}(0, \sigma_{\varepsilon}^2)) \end{aligned}$$

因此,  $\text{Var} \left\{ \frac{1}{\sqrt{nh}} \sum_{i=1}^n K[(x_i - x_0)/h] \varepsilon_i \right\} = \frac{1}{nh} \sum_{i=1}^n \text{Var} \{ K[(x_i - x_0)/h] \varepsilon_i \}$

$$\begin{aligned} &= \frac{1}{nh} \sum_{i=1}^n \text{Var} \{ K[(x_i - x_0)/h] \varepsilon_i \} \\ &= \frac{1}{nh} n \cdot h \sigma_{\varepsilon}^2 f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \\ &= \sigma_{\varepsilon}^2 f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \end{aligned}$$

综合第一部分与第二部分可知,

$$\sqrt{nh} [\hat{m}(x_0) - m(x_0)] \xrightarrow{d} N \left( \sqrt{nh} f(x_0) \text{Bias}(x_0), \sigma_{\varepsilon}^2 f(x_0) \int_{-\infty}^{+\infty} K(z)^2 dz \right) \quad (27.45)$$

因此,

$$\sqrt{nh} [\hat{m}(x_0) - m(x_0)] \xrightarrow{d} N \left( \sqrt{nh} \text{Bias}(x_0), \frac{\sigma_{\varepsilon}^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz \right) \quad (27.46)$$

故

$$\sqrt{nh} [\hat{m}(x_0) - m(x_0) - \text{Bias}(x_0)] \xrightarrow{d} N \left( 0, \frac{\sigma_{\varepsilon}^2}{f(x_0)} \int_{-\infty}^{+\infty} K(z)^2 dz \right) \quad (27.47)$$

# 第28章 处理效应

## 28.1 处理效应与选择难题

在经济学中,我们常希望评估某项目或政策实施后的效应,比如政府推出的就业培训项目(job training program)。此类研究被称为“项目效应评估”(program evaluation)<sup>①</sup>,而项目效应也常被称为“处理效应”(treatment effect)<sup>②</sup>。项目参与者的全体构成“实验组”或“处理组”(treatment group, or the treated),而未参与项目者则构成“控制组”(control group)或“对照组”(comparison group)。有关处理效应的专著包括Morgan and Winship(2007), Guo and Fraser(2009), Rosenbaum(2010)以及Imbens and Wooldridge(2009)的文献综述。

考虑就业培训的处理效应评估。一个自然(天真)的做法是直接对比实验组与控制组的未来收入或就业状况。如果这样做,常会发现参加就业培训者的未来收入比未参加者更低。难道就业培训反而有害?值得注意的是,是否参加培训是参加者自我选择(self selection)的结果,岗位好收入高的人群并不需要参加就业培训,而就业培训的参加者多为失业、低收入者,甚至刑满释放犯。由于实验组与对照组成员的初始条件不完全相同,故存在“选择偏差”(selection bias)。另外,即使实验组的未来收入低于对照组,我们真正感兴趣的问题是,实验组的未来收入是否会比这些人如果未参加培训项目的(假想)未来收入更高<sup>③</sup>。

为此,Rubin(1974)提出了以下“反事实框架”(a counterfactual framework),称为“鲁宾因果模型”(Rubin Causal Model,简记RCM)。以虚拟变量 $D_i = \{0, 1\}$ 表示个体*i*是否参与此项目,即1为参与,而0为未参与。通常称 $D_i$ 为“处理变量”(treatment variable),反映个体*i*是否得到了“处理”(treatment)。记其未来收入或其他感兴趣的结果(outcome of interest)为 $y_i$ 。我们想知道 $D_i$ 是否对 $y_i$ 有因果作用。对于个体*i*,其未来收入 $y_i$ 可能有两种状态,取决于是否参加此项目,即

$$y_i = \begin{cases} y_{1i} & \text{若 } D_i = 1 \\ y_{0i} & \text{若 } D_i = 0 \end{cases} \quad (28.1)$$

其中, $y_{0i}$ 表示个体*i*未参加项目的未来收入,而 $y_{1i}$ 表示个体*i*参加项目的未来收入。我们想知道 $(y_{1i} - y_{0i})$ ,即个体*i*参加该项目的因果效应。如果个体*i*参加了项目,则可观测到 $y_{1i}$ ,但看不到 $y_{0i}$ ;除非让此人“穿越”到过去,改写历史而选择不参加此项目,才能观测到 $y_{0i}$ 。反之,如果个体*i*未参加项目,则可观测到 $y_{0i}$ ,但看不到 $y_{1i}$ 。总之,由于个体只能处于一种状态(要么参加项目,要

① “项目评估”在汉语中主要指投资项目的可行性分析,而“项目效应评估”(program evaluation)的含义完全不同。

② 直译为“治疗效应”,因为它最早盛行于医学领域对于药物疗效的评估。也译为“处置效应”。

③ 类似地,1993年诺贝尔经济学奖得主罗伯特·福格尔(Robert Fogel)在使用“历史计量学”(Cliometrics)评估铁路对美国经济发展的贡献时,将1890年的美国经济与“假想”(hypothetical)的没有铁路的1890年美国经济进行了对比。

么不参加),故只能观测到  $y_{0i}$  或  $y_{1i}$ ,而无法同时观测到  $y_{0i}$  与  $y_{1i}$ 。这实际上是一种“数据缺失”(missing data)问题。

表达式(28.1)将  $y_i$  写为分段函数。更简洁地,可将  $y_i$  写为

$$y_i = (1 - D_i)y_{0i} + D_i y_{1i} = y_{0i} + \underbrace{(y_{1i} - y_{0i})}_{\text{处理效应}} D_i \quad (28.2)$$

其中,  $(y_{1i} - y_{0i})$  为个体  $i$  参加项目的因果效应或处理效应(treatment effect)。显然,不同个体的处理效应可能不同,故应将  $(y_{0i}, y_{1i}, D_i)$  视为来自三维随机向量  $(y_0, y_1, D)$  总体的一个随机抽样(random draw)。假设样本为 iid, 即对于任何  $i \neq j$ ,  $(y_{0i}, y_{1i}, D_i)$  的概率分布与  $(y_{0j}, y_{1j}, D_j)$  相同, 而且二者相互独立。这意味着不存在溢出效应, 即个体  $i$  是否参加项目, 不影响任何其他个体。此假定被称为“个体处理效应稳定假设”(Stable Unit Treatment Value Assumption, 简记 SUTVA)。SUTVA 假定排除了个体间的社会互动(social interactions)或一般均衡效应(general equilibrium effects)。由于处理效应  $(y_{1i} - y_{0i})$  为随机变量, 故我们关心其期望值, 即“平均处理效应”(Average Treatment Effect, 简记 ATE), 也称为“平均因果效应”(Average Causal Effect, 简记 ACE):

$$\text{ATE} \equiv E(y_{1i} - y_{0i}) \quad (28.3)$$

ATE 表示从总体中随机抽取某个体的期望处理效应,无论该个体是否参与项目。有些学者批评此定义过于宽泛,因为总体中的某些个体可能根本无资格参加项目。比如,在估计培训项目的平均处理效应时,我们并不希望将百万富翁也包括在内。但 ATE 的定义依然十分有用,因为总可以通过重新定义总体而将无资格参加项目者排除在外,比如,限定培训前收入(pretraining income)在某临界水平之下。

另一常用概念为仅考虑项目实际参加者的平均处理效应,称为“参与者平均处理效应”(Average Treatment Effect on the Treated, 简记 ATT 或 ATET)或“参与者处理效应”(Treatment Effect on the Treated, 简记 TOT), 即

$$\text{ATT} \equiv E(y_{1i} - y_{0i} | D_i = 1) \quad (28.4)$$

对于政策制定者而言,ATT 可能更为重要,因为它衡量的是项目参与者的毛收益(比如,可对比此毛收益与项目成本,进行成本收益分析)。在特殊情况下,ATE 与 ATT 可能相等,但一般不等。由于不能同时观测  $y_{0i}$  与  $y_{1i}$ ,应如何估计 ATE 或 ATT? 如果简单地比较项目参与者与未参与者的未来收入,则会导致选择偏差,因为

$$\begin{aligned} \underbrace{E(y_{1i} | D_i = 1) - E(y_{0i} | D_i = 0)}_{\text{参与者与未参与者的平均差异}} &= \underbrace{E(y_{1i} | D_i = 1) - E(y_{0i} | D_i = 1)}_{\text{ATT}} \\ &\quad + \underbrace{E(y_{0i} | D_i = 1) - E(y_{0i} | D_i = 0)}_{\text{选择偏差}} \end{aligned} \quad (28.5)$$

在上式中,先减去  $E(y_{0i} | D_i = 1)$ , 然后再加上它。上式将项目参与者与未参与者的平均收入之差分解为两部分,其中第一项为 ATT,而第二项为参与者的平均  $y_{0i}$  与未参与者的平均  $y_{0i}$  之差(即这两类人如果都未参与项目的收入差距),即选择偏差。由于低收入者通常更倾向于选择参加培训项目,故选择偏差一般为负,导致实验组与控制组的收入之差[即  $E(y_{1i} | D_i = 1) - E(y_{0i} | D_i = 0)$ ]低估参与者平均处理效应(ATT)。如果选择偏差的绝对值足够大,则可能导致  $E(y_{1i} | D_i = 1) - E(y_{0i} | D_i = 0) < 0$ , 出现参加培训者的收入反而低于未参加者的情形。

类似地,可以定义“非参与者平均处理效应”(Average Treatment Effect on the Untreated, 简记 ATU)为

$$\text{ATU} \equiv E(y_{1i} - y_{0i} | D_i = 0) \quad (28.6)$$

由于个体通常会根据其参加项目的预期收益  $E(y_{1i} - y_{0i})$  而自我选择是否参加项目, 导致对平均处理效应的估计带来困难, 这被称为“选择难题”(the selection problem)。

需要注意的是, 在估计处理效应时面临的选择难题, 与第 14 章的样本选择问题(Heckman, 1979)有所不同, 后者考虑的是所获样本是否为总体的代表性样本。具体来说, 样本选择问题通常不考虑某项目或政策的效应, 故个体间的差异并不在于是否得到处理, 而在于是否能进入样本(即被解释变量  $y_i$  是否可观测), 通常  $D_i = 1$  意味着  $y_i$  可观测, 而  $D_i = 0$  则意味着  $y_i$  不可观测。而在处理效应模型中, 无论  $D_i = 1$  或 0, 结果变量  $y_i$  均可观测。

## 28.2 通过随机分组解决选择难题

选择难题并非不可克服。解决方法之一是通过随机分组(random assignment)<sup>①</sup>, 使得个体  $i$  的  $D_i$ (即是否参加项目)通过抛硬币或电脑随机数而决定, 则  $D_i$  独立于  $(y_{0i}, y_{1i})$ 。此时, ATE = ATT, 因为  $E(y_{1i} - y_{0i} | D_i = 1) = E(y_{1i} - y_{0i})$  [由于  $(y_{1i} - y_{0i})$  独立于  $D_i$ , 故条件期望等于无条件期望]。对 ATE 的估计也非常简单, 只要比较实验组与控制组的平均收入即可, 因为

$$E(y_{1i} | D_i = 1) - E(y_{0i} | D_i = 0) = E(y_{1i}) - E(y_{0i}) = \text{ATE} = \text{ATT} \quad (28.7)$$

上式能够成立, 也是因为  $D_i$  独立于  $(y_{0i}, y_{1i})$ 。因此, 在随机分组的情况下, 只需要计算样本中实验组与控制组的平均收入之差, 即可一致地估计平均处理效应。此估计量即第 18 章介绍的“差额估计量”(differences estimator), 也称为“difference-in-means”, 服从渐近正态分布。

事实上, 上述结果在更弱的均值独立(mean independence)条件下也成立, 即只要  $y_{0i}, y_{1i}$  都均值独立于  $D_i$ , 则表达式(28.7)依然成立。进一步, 如果我们只关心 ATT, 则只需要  $y_{0i}$  均值独立于  $D_i$  即可, 不需要对  $y_{1i}$  与  $D_i$  的关系进行任何限制。这是因为, 如果  $y_{0i}$  均值独立于  $D_i$ , 即  $E(y_{0i} | D_i) = E(y_{0i})$ , 则公式(28.5)中的选择偏差为 0, 因为  $E(y_{0i} | D_i = 1) - E(y_{0i} | D_i = 0) = E(y_{0i}) - E(y_{0i}) = 0$ 。由于随机分组可以很好地解决选择偏差, 故随机实验或准实验的研究方法在经济学各领域日益盛行。

然而, 随机分组并非在所有情况下都可行(可能成本太高)。如果只有观测数据(observational data), 则很可能不满足“ $y_{0i}$  均值独立于  $D_i$ ”的假设, 因为该假设意味着个体  $i$  是否参加项目( $D_i$ )与其不参加项目的收入( $y_{0i}$ )不相关。因此, 在大多数情况下, 需要使用以下两类方法。第一类方法假设个体依可测变量选择是否参加项目(参见本章第 3~7 节), 而第二类方法假设个体依不可测变量选择(参见本章第 8 节)。

## 28.3 依可测变量选择

通常, 除了  $(y_i, D_i)$  之外, 还可以观测到个体  $i$  的一些特征, 比如年龄、性别、培训前收入, 记为向量  $\mathbf{x}_i$ , 也称为“协变量”(covariates)。这样, 总体可由  $(y_0, y_1, D, \mathbf{x})$  来表示。如果个体  $i$  对  $D_i$

<sup>①</sup> 有关随机实验, 详见第 18 章。

的选择完全取决于可观测的  $\mathbf{x}_i$ , 称为“依可测变量选择”(selection on observables)<sup>①</sup>, 则可以找到估计处理效应的合适方法(即使没有合适的工具变量)。如果个体对  $D_i$  的选择完全取决于  $\mathbf{x}_i$ , 则在给定  $\mathbf{x}_i$  的情况下, 潜在结果  $(y_{0i}, y_{1i})$  将独立于  $D_i$ , 这就是 Rosenbaum and Rubin(1983) 所引入的“可忽略性”假设:

**假定 28.1 可忽略性(Ignorability)**。给定  $\mathbf{x}_i$ , 则  $(y_{0i}, y_{1i})$  独立于  $D_i$ , 记为  $(y_{0i}, y_{1i}) \perp D_i | \mathbf{x}_i$ , 其中“ $\perp$ ”表示相互独立。

“可忽略性”的含义是, 给定  $\mathbf{x}_i$ , 则  $(y_{0i}, y_{1i})$  对于  $D_i$  的影响可以忽略, 故名。可忽略性也称为“无混淆性”(unconfoundedness)<sup>②</sup>, “条件独立假定”(Conditional Independence Assumption, 简记 CIA), 或“依可测变量选择”(selection on observables)。假定 28.1 意味着, 给定  $\mathbf{x}_i$ , 则  $(y_{0i}, y_{1i})$  在处理组与控制组的分布完全一样, 即

$$F(y_{0i}, y_{1i} | \mathbf{x}_i, D_i = 1) = F(y_{0i}, y_{1i} | \mathbf{x}_i, D_i = 0) \quad (28.8)$$

其中,  $F(\cdot)$  表示分布函数。在很多情况下, 只需要更弱的均值独立即可:

**假定 28.2 均值可忽略性(Ignorability in Mean)**。 $E(y_{0i} | \mathbf{x}_i, D_i) = E(y_{0i} | \mathbf{x}_i)$ , 而且  $E(y_{1i} | \mathbf{x}_i, D_i) = E(y_{1i} | \mathbf{x}_i)$ 。这意味着, 在给定  $\mathbf{x}_i$  的情况下,  $y_{0i}$  与  $y_{1i}$  都均值独立于  $D_i$ 。

如果假定 28.1 或 28.2 成立, 则原则上可将  $\mathbf{x}_i$  直接作为控制变量引入以下回归方程, 以解决遗漏变量问题:

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \gamma D_i + \varepsilon_i \quad (28.9)$$

然而, 我们并不清楚  $\mathbf{x}_i$  是否应以线性形式进入上述方程。如果遗漏了非线性项, 则依然可能存在遗漏变量偏差。解决方法之一为基于鲁宾反事实框架的匹配估计量, 参见下一节。从回归方程(28.9)也可以看出, 可忽略性假定是一个很强的假定; 它意味着回归方程已包括了所有相关变量, 故不存在任何与解释变量相关的遗漏变量。另一方面, 如果  $\mathbf{x}_i$  中已包含较丰富的协变量(a rich set of covariates), 则可认为可忽略性假定基本得到满足, 遗漏变量偏差较小。

## 28.4 匹配估计量的思想

假设个体  $i$  属于处理组, 匹配估计量的基本思路是, 找到属于控制组的某个体  $j$ , 使得个体  $j$  与个体  $i$  的可测变量取值尽可能相似(匹配), 即  $\mathbf{x}_i \approx \mathbf{x}_j$ 。基于可忽略性假设, 则个体  $i$  与个体  $j$  进入处理组的概率相近, 具有可比性; 故可将  $y_j$  作为  $y_{0i}$  的估计量, 即  $\hat{y}_{0i} = y_j$ 。因此, 可将  $(y_i - \hat{y}_{0i}) = y_i - y_j$  作为对个体  $i$  处理效应的度量。对处理组中的每位个体都如此进行匹配; 类似地, 对控制组每位个体也进行匹配, 然后对每位个体的处理效应进行平均, 即可得到“匹配估计量”(matching estimators)。

由于匹配的具体方法不同, 故存在不同的匹配估计量。这里有两个技术细节。首先, 是否放回。如果不放回(no replacement), 则每次都将匹配成功的个体( $i, j$ )从样本中去掉, 不再参与其余匹配; 如果有放回(with replacement), 则依然将匹配成功个体留在样本中, 参与其余匹配(这导

<sup>①</sup> 这并不要求  $D_i$  是  $\mathbf{x}_i$  的确定性函数。仍然可允许除  $\mathbf{x}_i$  以外的随机因素对  $D_i$  产生影响, 只要这些随机因素独立于  $(\mathbf{x}_i, y_{0i}, y_{1i})$ 。

<sup>②</sup> “unconfoundedness”的含义是, 只要把  $\mathbf{x}_i$  包括在回归方程中, 就能完全解决遗漏变量偏差, 从而避免各变量间作用关系的混淆(confounding)。

致一位个体可能与多位不同组个体匹配)。其次,是否允许并列(ties),比如控制组个体  $j$  与  $k$  的可测变量都与处理组个体  $i$  一样接近。如果允许并列,则将  $y_j$  与  $y_k$  的平均值作为  $\hat{y}_{0i}$  的估计量,即  $\hat{y}_{0i} = (y_j + y_k)/2$ 。如果不允许并列,则计算机程序将根据数据排序选择个体  $j$  或  $k$ ;此时,匹配结果可能与数据排序有关,故一般建议先将样本随机排序<sup>①</sup>,再进行匹配。

以上为一对一(one-to-one)匹配,也可以进行一对多匹配,比如一对四匹配,即针对每位个体寻找四位不同组的最近个体进行匹配。一般来说,匹配估计量存在偏差(bias)<sup>②</sup>,除非在“精确匹配”(exact matching)的情况下,即对于所有匹配都有  $x_i = x_j$ 。更常见的为“非精确匹配”(inexact matching),即只能保证  $x_i \approx x_j$ 。在非精确匹配的情况下,如果进行一对一匹配,则偏差较小,但方差较大;而进行一对多匹配可降低方差(因为使用了更多信息),但代价是偏差增大(因为使用了更远的信息)。Abadie et al(2004)建议进行一对四匹配,在一般情况下可最小化均方误差(MSE)。

下面以一个具体例子来说明匹配的过程<sup>③</sup>。假设样本容量为 7,其中包括 3 位控制组个体与 4 位处理组个体。同时假设  $x_i$  仅包含一个变量  $x_i$ 。具体数据参见表 28.1。下面,进行有放回的一对一匹配,且允许并列。

表 28.1 匹配估计量的简单例子

$i$	$D_i$	$x_i$	$y_i$	匹配结果	$\hat{y}_{0i}$	$\hat{y}_{1i}$
1	0	2	7	{5}	7	8
2	0	4	8	{4,6}	8	7.5
3	0	5	6	{4,6}	6	7.5
4	1	3	9	{1,2}	7.5	9
5	1	2	8	{1}	7	8
6	1	3	6	{1,2}	7.5	6
7	1	1	5	{1}	7	5

首先,考虑个体 1 的匹配。由于个体 1 属于控制组( $D_1 = 0$ ),故在处理组(即个体 4—7)中寻找最佳匹配。由于  $x_1 = 2$ ,而  $x_5 = 2$ ,故个体 1 的匹配为个体 5,记为 {5}。因此,  $\hat{y}_{01} = y_1 = 7$ ,而  $\hat{y}_{11} = y_5 = 8$ 。

其次,考虑个体 2 的匹配。由于  $x_2 = 4$ ,在处理组中没有完全相同的匹配,而最近匹配为  $x_4 = x_6 = 3$ ,故个体 2 的匹配结果为 {4,6}。因此,  $\hat{y}_{02} = y_2 = 8$ ,而  $\hat{y}_{12} = (y_4 + y_6)/2 = 7.5$ 。类似地,个体 3 的匹配结果也是 {4,6}。

再次,考虑个体 4 的匹配。由于个体 4 属于处理组( $D_4 = 1$ ),故在控制组(即个体 1—3)中寻找最佳匹配。由于  $x_4 = 3$ ,在控制组中没有完全相同的匹配,而最近匹配为  $x_1 = 2$  与  $x_2 = 4$ ,故个体 4 的匹配结果为 {1,2}。因此,  $\hat{y}_{04} = (y_1 + y_2)/2 = 7.5$ ,而  $\hat{y}_{14} = y_4 = 9$ 。类似地,可以得到个体 5—7 的匹配结果,参见表 28.1。

① 比如,产生服从均匀分布的一组随机数,然后根据此组随机数进行样本排序。为了保证结果的可重复性,在产生随机数之前,应确定随机数的“种子”(seed)。

② 正如非参数估计一般存在偏差,因为它使用了附近邻域的信息,参见第 27 章。

③ 此例来自 Abadie et al(2004)。

最后,在Stata中分别计算ATE(考虑整个样本的匹配结果),ATT(只考虑参加者的匹配结果)以及ATU(只考虑未参加者的匹配结果)。由于此例子很简单,故可手工计算。

```
. dis "ATE = " ((8 - 7) + (7.5 - 8) + (7.5 - 6) + (9 - 7.5) + (8 - 7) + (6 - 7.5) + (5 - 7)) / 7
ATE = .14285714
dis "ATT = " ((9 - 7.5) + (8 - 7) + (6 - 7.5) + (5 - 7)) / 4
ATT = -.25
dis "ATU = " ((8 - 7) + (7.5 - 8) + (7.5 - 6)) / 3
ATU = .66666667
```

## 28.5 倾向得分匹配

更一般地, $x_i$ 可能包括多个变量,比如 $x_i$ 为 $K$ 维向量。此时,如果直接使用 $x_i$ 进行匹配,则意味着要在高维度空间进行匹配,可能遇到数据稀疏的问题,即很难找到与 $x_i$ 相近的 $x_j$ 与之匹配。为此,一般使用某函数 $f(x_i)$ ,将 $K$ 维向量 $x_i$ 的信息压缩到一维,进而根据 $f(x_i)$ 进行匹配。方法之一为使用向量范数(vector norm),即在向量空间(vector space)定义的距离函数。考虑 $x_i$ 与 $x_j$ 之间的相似度或距离,定义“马氏距离”(Mahalanobis distance)为

$$d(i,j) = (x_i - x_j)' \hat{\Sigma}_x^{-1} (x_i - x_j) \quad (28.10)$$

其中,二次型矩阵 $\hat{\Sigma}_x^{-1}$ 为 $x$ 的样本协方差矩阵之逆矩阵,它的作用相当于“权重矩阵”(weighting matrix)<sup>①</sup>。使用马氏距离进行匹配,被称为“马氏匹配”(Mahalanobis matching)。有时,也使用主对角元素为各变量方差的对角矩阵之逆矩阵作为权重矩阵。通过协变量的某个距离函数进行匹配,统称为“协变量匹配”(covariate matching)。

马氏匹配的缺点是,如果 $x$ 包括的变量较多或样本容量不够大,则不容易找到好的匹配,比如,尽管个体 $j$ 与个体 $i$ 的(相对)马氏距离最近,但绝对距离可能依然很远。为此,统计学家Rosenbaum and Rubin(1983)提出使用“倾向得分”(propensity score,简记p-score)来度量距离。

**定义** 个体 $i$ 的倾向得分为,在给定 $x_i$ 的情况下,个体 $i$ 进入处理组的条件概率,即 $p(x_i) = P(D_i = 1 | x = x_i)$ ,或简记 $p(x)$ (省略下标 $i$ )。

在使用样本数据估计 $p(x)$ 时,可使用参数估计(比如,probit或logit)或非参数估计(参见第27章),而最流行的方法为logit。使用倾向得分来度量个体之间距离的好处在于,它不仅是一维变量,而且取值介于[0,1]之间。比如,即使 $x_i$ 与 $x_j$ 距离很远,但仍可能 $p(x_i) \approx p(x_j)$ 。

使用倾向得分作为距离函数进行匹配,称为“倾向得分匹配”(Propensity Score Matching,简记PSM)。PSM的理论依据在于,如果可忽略性假定成立,则只需在给定 $p(x)$ 的情况下, $(y_{0i}, y_{1i})$ 就独立于 $D_i$ 。

**命题(倾向得分定理)**  $(y_0, y_1) \perp D | x \Rightarrow (y_0, y_1) \perp D | p(x)$

**证明:**由于 $D$ 为虚拟变量,故只需证明 $P[D = 1 | y_0, y_1, p(x)]$ 与 $y_0, y_1$ 无关即可。

<sup>①</sup> 权重矩阵的作用相当于GMM估计量目标函数中的二次型,参见第10章。

$$\begin{aligned}
 & P[D = 1 | y_0, y_1, p(\mathbf{x})] \\
 &= E[D | y_0, y_1, p(\mathbf{x})] \\
 &= E_{y_0, y_1, \mathbf{x}}[E(D | y_0, y_1, \mathbf{x}) | y_0, y_1, p(\mathbf{x})] \quad (\text{迭代期望定律}) \\
 &= E_{y_0, y_1, \mathbf{x}}[E(D | \mathbf{x}) | y_0, y_1, p(\mathbf{x})] \quad (\text{可忽略性假定}) \\
 &= E_{y_0, y_1, \mathbf{x}}[p(\mathbf{x}) | y_0, y_1, p(\mathbf{x})] \\
 &= p(\mathbf{x})
 \end{aligned} \tag{28.11}$$

此证明的倒数第二行使用了以下关系式：

$$E(D | \mathbf{x}) = 1 \cdot P(D = 1 | \mathbf{x}) + 0 \cdot P(D = 0 | \mathbf{x}) = p(\mathbf{x}) \tag{28.12}$$

当然,为了能够进行匹配,需要在  $\mathbf{x}$  的每个可能取值上都同时存在处理组与控制组的个体。这就是下面的“重叠假定”(overlap assumption)或“匹配假定”(matching assumption)。

**假定 28.3 重叠假定。**对于  $\mathbf{x}$  的任何可能取值,都有  $0 < p(\mathbf{x}) < 1$ 。

此假定意味着处理组与控制组这两个子样本存在重叠,故名“重叠假定”;另外,它又是进行匹配的前提,故也称“匹配假定”。它保证了处理组与控制组的倾向得分取值范围有相同的部分(common support),参见图 28.1。如果假定 28.3 不成立,则意味着可能存在某些  $\mathbf{x}$ ,使得  $p(\mathbf{x}) = 1$ ,即这些个体都属于处理组,无法找到与之匹配的控制组个体;另一方面,也可能存在某些  $\mathbf{x}$ ,使得  $p(\mathbf{x}) = 0$ ,即这些个体都属于控制组,无法找到与其匹配的处理组个体。

在进行匹配时,为了提高匹配质量,通常仅保留倾向得分重叠部分的个体(尽管这样做会损失样本容量)。具体来说,如果某处理组个体的倾向得分高于控制组倾向得分的最大值或低于控制组倾向得分的最小值,则去掉该处理组个体(在使用 Stata 命令 psmatch2 时,可用选择项 common 来实现)。如果倾向得分的共同取值范围太小,则会导致偏差。

通过倾向得分匹配计算平均处理效应的一般步骤如下。

(1) 选择协变量  $\mathbf{x}_i$ 。尽量将可能影响  $(y_{0i}, y_{1i})$  与  $D_i$  的相关变量包括进来,以保证可忽略性假设得到满足。如果协变量  $\mathbf{x}_i$  选择不当或太少,导致可忽略性假设不满足,将引起偏差。

(2) 估计倾向得分,一般使用 logit 回归。Rosenbaum and Rubin(1985)建议使用形式灵活的 logit 模型,比如包括  $\mathbf{x}_i$  的高次项与互动项。

(3) 进行倾向得分匹配。如果倾向得分估计得较准确,则应使得  $\mathbf{x}_i$  在匹配后的处理组与控制组之间分布较均匀,比如,匹配后的处理组均值  $\bar{x}_{\text{treat}}$  与控制组均值  $\bar{x}_{\text{control}}$  较接近;这个过程在统计学上称为“数据平衡”(data balancing)。但  $\bar{x}_{\text{treat}}$  与  $\bar{x}_{\text{control}}$  的差距显然与计量单位有关,故一般针对  $\mathbf{x}$  的每个分量  $x$  考察如下“标准化差距”(standardized differences)或“标准化偏差”(standardized bias):

$$\frac{|\bar{x}_{\text{treat}} - \bar{x}_{\text{control}}|}{\sqrt{(s_{x,\text{treat}}^2 + s_{x,\text{control}}^2)/2}} \tag{28.13}$$

其中,  $s_{x,\text{treat}}^2$  与  $s_{x,\text{control}}^2$  分别为处理组与控制组变量  $x$  的样本方差。一般要求此标准化差距不超过

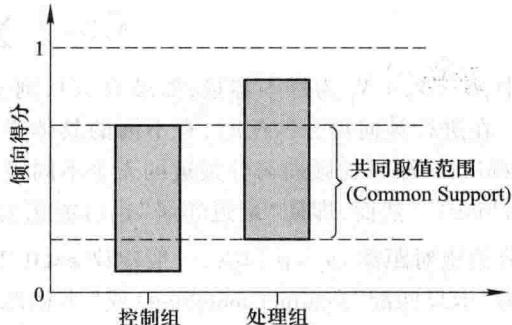


图 28.1 倾向得分的共同取值范围

10%；如果超过，则应回到第(2)步、甚至第(1)步，重新估计倾向得分；或者改变具体的匹配方法（参见下文）。

(4) 根据匹配后样本(matched sample)计算平均处理效应。参加者平均处理效应(ATT)估计量的一般表达式为

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i:D_i=1} (y_i - \hat{y}_{0i}) \quad (28.14)$$

其中， $N_1 = \sum_i D_i$  为处理组个体数，而  $\sum_{i:D_i=1}$  表示仅对处理组个体进行加总。类似地，也可为控制组的每位个体  $j$  寻找处理组的相应匹配。未参加者平均处理效应(ATU)估计量的一般表达式为

$$\widehat{ATU} = \frac{1}{N_0} \sum_{j:D_j=0} (\hat{y}_{1j} - y_j) \quad (28.15)$$

其中， $N_0 = \sum_j (1 - D_j)$  为控制组个体数，而  $\sum_{j:D_j=0}$  表示仅对控制组个体进行加总。整个样本(包括参加者与未参加者)的平均处理效应(ATE)估计量的一般表达式为

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_{1i} - \hat{y}_{0i}) \quad (28.16)$$

其中， $N = N_0 + N_1$  为样本容量；如果  $D_i = 1$ ，则  $\hat{y}_{1i} = y_i$ ；如果  $D_i = 0$ ，则  $\hat{y}_{0i} = y_i$ 。

在进行倾向得分匹配时，有不同的具体方法。方法之一为“ $k$  近邻匹配”( $k$ -nearest neighbor matching)，即寻找倾向得分最近的  $k$  个不同组个体。如果  $k = 1$ ，则为“一对一匹配”(one-to-one matching)。然而，即使“最近邻居”也可能相去甚远，从而失去可比性。为此，方法之二限制倾向得分的绝对距离  $|p_i - p_j| \leq \varepsilon$ ，一般建议  $\varepsilon \leq 0.25\hat{\sigma}_{pscore}$ ，其中  $\hat{\sigma}_{pscore}$  为倾向得分的样本标准差；这被称为“卡尺匹配”(caliper matching)或“半径匹配”(radius matching)。方法之三为“卡尺内最近邻匹配”(nearest-neighbor matching within caliper)，即在给定的卡尺  $\varepsilon$  范围寻找最近匹配，此法较为流行。

以上三种方法本质上都是近邻匹配法，其匹配结果为最近的部分个体，然后进行简单算术平均。另一类匹配方式为整体匹配法，每位个体的匹配结果为不同组的全部个体(但通常去掉在 common support 之外的个体)，只是根据个体距离不同给予不同的权重(近者权重大，远者权重小，超出一定范围权重可为 0)。比如，在使用表达式(28.14)估计 ATT 时， $\hat{y}_{0i}$  的估计量为

$$\hat{y}_{0i} = \sum_{j:D_j=0} w(i, j) y_j \quad (28.17)$$

其中， $w(i, j)$  为适用于配对  $(i, j)$  的权重。如果使用核函数来计算权重  $w(i, j)$ ，则为方法之四“核匹配”(kernel matching)(Heckman et al., 1997, 1998)，其权重表达式为

$$w(i, j) = \frac{K[(x_j - x_i)/h]}{\sum_{k:D_k=0} K[(x_k - x_i)/h]} \quad (28.18)$$

其中， $h$  为指定带宽(bandwidth)， $K(\cdot)$  为核函数。将表达式(28.18)代入方程(28.17)，可将  $\hat{y}_{0i}$  视为“核回归估计量”(kernel regression estimator)，参见第 27 章。如果不进行核回归，而使用局部线性回归来估计  $w(i, j)$ ，则为方法之五“局部线性回归匹配”(local linear regression matching)。方法之六使用更为光滑的“三次样条”(cubic spline)来估计  $w(i, j)$ ，称为“样条匹配”(spline matching)。

在实际进行匹配时，究竟应使用以上哪种具体方法或参数(比如， $k$  近邻匹配的  $k$  取值，是否放回，如何处理并列)，目前文献中尚无明确指南。一般认为，不存在适用于一切情形的绝对好方法，只能根据具体数据来选择匹配方法。比如，如果控制组个体并不多( $N_0$  较小)，则应进行有

放回的匹配。又比如,如果存在较多具有可比性的控制组个体,则可考虑一对多匹配或核匹配,以提高匹配效率。在实践中,一般建议尝试不同的匹配方法,然后比较其结果(类似于敏感度分析);如果不同方法的结果相似,则说明结果是稳健的,不依赖于具体方法;反之,如果存在较大差异,则应考察造成此差异的原因。

在一定意义上,匹配估计量可视为一种再抽样方法(resampling)。因此,在方法论上,PSM 试图通过匹配再抽样的方法使得观测数据尽可能地接近随机实验数据,其思想可以追溯到费舍尔(Ronald Fischer)提出的随机实验设计。然而,尽管 PSM 可能在很大程度上减少观测数据的偏差,但它本身也有如下局限性:

- (1) PSM 通常要求比较大的样本容量以得到高质量的匹配。
- (2) PSM 要求处理组与控制组的倾向得分有较大的共同取值范围(common support);否则,将丢失较多观测值,导致剩下的样本不具有代表性。
- (3) PSM 只控制了可测变量的影响,如果存在依不可测变量选择(selection on unobservable),仍会带来“隐性偏差”(hidden bias)。

## 28.6 倾向得分匹配的 Stata 实例

倾向得分匹配可通过下载非官方命令 psmatch2 来实现:

```
ssc install psmatch2, replace
```

其中,选择项“replace”表示以该命令的最新版本替代计算机中可能已有的旧版命令。由于该命令仍在不断更新中,故建议使用选择项“replace”。

该命令的一般格式为

```
psmatch2 D x1 x2 x3, outcome(y) logit ties ate common odds pscore  
(varname) quietly
```

其中,“D”为处理变量(treatment variable),“x1 x2 x3”为协变量,“outcome(y)”用来指定变量“y”为结果变量(outcome variable)。选择项“logit”表示使用 logit 来估计倾向得分,默认方法为 probit。选择项“ties”表示包括所有倾向得分相同的并列个体,默认按照数据排序选择其中一位个体。选择项“ate”表示同时汇报 ATE,ATU 与 ATT,默认仅汇报 ATT。选择项“common”表示仅对共同取值范围(common support)内个体进行匹配,默认对所有个体进行匹配。选择项“odds”表示使用几率比(odds ratio,即  $p/(1-p)$ )进行匹配;默认使用倾向得分  $p$  进行匹配。选择项“pscore(varname)”用来指定某变量作为倾向得分,默认通过“x1 x2 x3”来估计倾向得分。选择项“quietly”表示不汇报对倾向得分的估计过程。

针对不同的匹配方法,命令 psmatch2 提供了一系列选择项。

- (1) psmatch2 D x1 x2 x3, outcome(y) neighbor(k) noreplacement

选择项“neighbor(k)”表示进行  $k$  近邻匹配( $k$  为正整数);默认  $k=1$ ,即一对一匹配。选择项“noreplacement”表示无放回匹配,默认认为有放回;该选项只能用于一对一匹配。

- (2) psmatch2 D x1 x2 x3, outcome(y) radius caliper(real)

选择项“radius”表示进行半径匹配(也称卡尺匹配),其中“caliper(real)”用来指定卡尺  $\epsilon$ ,必须为正实数。

- (3) psmatch2 D x1 x2 x3, outcome(y) neighbor(k) caliper(real)

选择项“neighbor (k)”与“caliper (real)”表示进行卡尺内的  $k$  近邻匹配。

(4) psmatch2 D x1 x2 x3, outcome(y) kernel kerneltype (type) bwidth (real)

选择项“kernel”表示进行核匹配, 其中“kerneltype (type)”用来指定核函数, 默认使用二次核(epan kernel), “bwidth (real)”用来指定带宽, 默认带宽为 0.06。

(5) psmatch2 D x1 x2 x3, outcome(y) llr kerneltype (type) bwidth (real)

选择项“llr”表示进行局部线性回归匹配, 其中“kerneltype (type)”用来指定核函数, 默认使用三三核(tricubic kernel), “bwidth (real)”用来指定带宽, 默认带宽为 0.8。

(6) psmatch2 D x1 x2 x3, outcome(y) spline

选择项“spline”表示进行样条匹配。

(7) psmatch2 D x1 x2 x3, outcome(y) mahal(varlist) ai(m)

选择项“mahal(varlist)”表示进行马氏匹配, 并指定用于计算马氏距离的协变量。选择项“ai(m)”表示使用由 Abadie and Imbens(2006)提出的异方差稳健标准误, 该选项仅适用于使用马氏距离的  $k$  近邻匹配; 其中  $m$  须为正整数, 表示用于计算稳健标准误的近邻个数(一般可让  $m = k$ )。使用此命令进行马氏匹配时, 无法使用选择项“ties”或“common”。

命令 psmatch2 还带有以下两个“估计后命令”(post-estimation commands), 分别用来检验匹配后数据是否平衡, 以及画图显示倾向得分的共同取值范围。

pstest x1 x2 x3, both graph

此命令将显示变量“x1 x2 x3”在匹配后是否平衡, 选择项“both”表示同时显示匹配前的数据平衡情况, 默认仅显示匹配后情形。选择项“graph”表示图示各变量匹配前后的平衡情况。

psgraph, bin(#)

此命令将画直方图, 显示倾向得分的共同取值范围, 选择项“bin(#)”用来指定直方图的分组数, 默认为 20 组(处理组与控制组各分为 10 组)。

下面以数据集 ldw\_exper.dta 为例进行演示。该数据集由 Dehejia and Wahba(1999)构建, 为 Abadie et al(2004)所使用, 原始数据来自 Lalonde(1986)。该数据集包括以下变量: 结果变量 re78(1978 年实际收入), 处理变量 t(是否参加就业培训), 协变量 age(年龄), educ(教育年限), black(是否黑人), hisp(是否拉丁裔), married(是否已婚), re74(1974 年实际收入), re75(1975 年实际收入), u74(1974 年是否失业), 以及 u75(1975 年是否失业)。

. use ldw\_exper.dta, clear

作为参照, 首先进行一元回归。

. reg re78 t, r

Linear regression						Number of obs = 445
						F( 1, 443) = 7.15
						Prob > F = 0.0078
						R-squared = 0.0178
						Root MSE = 6.5795
		Robust				[95% Conf. Interval]
re78		Coef.	Std. Err.	t	P> t	
t		1.794343	.6708247	2.67	0.008	.475949 3.112737
_cons		4.554802	.3402038	13.39	0.000	3.886188 5.223416

上表显示,在未控制任何协变量的情况下,平均处理效应为 1.794,即参加就业培训平均能使 1978 年实际收入提高 1794 美元(变量 re78 的单位为千美元),且在 1% 水平上显著。由于可能存在选择偏差,此结果并不可信。而且, $R^2$ 很低,仅为 0.0178(即是否参加就业培训仅能解释 1978 年实际收入 1.78% 的变动)。

下面,直接引入协变量,进行更为可信的多元回归。

```
. reg re78 t age educ black hisp married re74 re75 u74 u75 ,r
```

Linear regression		Number of obs = 445 F( 10, 434) = 2.53 Prob > F = 0.0057 R-squared = 0.0582 Root MSE = 6.5093				
re78		Robust				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
t	1.672042	.6617972	2.53	0.012	.3713161	2.972768
age	.0536677	.040388	1.33	0.185	-.0257127	.133048
educ	.4029471	.1610925	2.50	0.013	.0863287	.7195655
black	-2.039466	1.038581	-1.96	0.050	-4.080739	.0018068
hisp	.4246486	1.427471	0.30	0.766	-2.380968	3.230265
married	-.1466618	.8640396	-0.17	0.865	-1.844884	1.551561
re74	.1235727	.127147	0.97	0.332	-.1263278	.3734731
re75	.0194585	.14063	0.14	0.890	-.2569421	.2958591
u74	1.380999	1.554643	0.89	0.375	-1.674566	4.436564
u75	-1.071817	1.408301	-0.76	0.447	-3.839755	1.696121
_cons	.2214288	2.824293	0.08	0.938	-5.329565	5.772422

上表显示,加入协变量后,平均处理效应降为 1.672(变化不大),且显著性水平接近 1%( $p$  值为 1.2%)。在协变量中,除了 educ(教育年限)与 black(是否黑人)在 5% 水平上显著外,其余协变量均不显著。

下面进行倾向得分匹配。为此,先将数据随机排序。

```
. set seed 10101
. gen ranorder = runiform()
. sort ranorder
```

下面进行一对一匹配。由于样本容量并不十分大,进行有放回匹配,且允许并列(如果进行无放回匹配,将损失约 1/5 样本)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75 , outcome(re78) n(1) ate ties logit common
```

Logistic regression				Number of obs	=	445
				LR chi2(9)	=	11.70
				Prob > chi2	=	0.2308
				Pseudo R2	=	0.0194
<b>Log likelihood = -296.25026</b>						
t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
age	.0142619	.0142116	1.00	0.316	-.0135923	.0421162
educ	.0499776	.0564116	0.89	0.376	-.060587	.1605423
black	-.347664	.3606532	-0.96	0.335	-1.054531	.3592032
hisp	-.9284851	.50661	-1.83	0.067	-1.921422	.0644521
married	.1760431	.2748817	0.64	0.522	-.3627151	.7148012
re74	-.0339278	.0292559	-1.16	0.246	-.0912683	.0234127
re75	.01221	.0471351	0.26	0.796	-.0801731	.1045932
u74	-.1516037	.3716369	-0.41	0.683	-.8799987	.5767913
u75	-.3719486	.317728	-1.17	0.242	-.9946841	.2507869
_cons	-.4736308	.8244205	-0.57	0.566	-2.089465	1.142204
There are observations with identical propensity score values. The sort order of the data could affect your results. Make sure that the sort order is random before calling psmatch2.						
Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.99436488	1.4105933	.839875971	1.68
	ATU	4.52683013	6.15618973	1.6293596	.	.
	ATE			1.53668776	.	.
Note: S.E. does not take into account that the propensity score is estimated.						
psmatch2: Treatment assignment	psmatch2: Common support					
	Off suppo	On suppor	Total			
Untreated	11	249	260			
Treated	2	183	185			
Total	13	432	445			

上表上部显示了 logit 回归的结果。上表中部显示, ATT 估计值为 1.411, 对应的 t 值为 1.68, 小于 1.96 的临界值, 故不显著。ATE 与 ATU 的估计值与 ATT 类似, 但不汇报标准误。其中, “Unmatched” 报汇匹配前样本估计的结果, 与前面一元回归的结果完全一样。上表下部汇报观测值是否在共同取值范围中。在总共 445 个观测值中, 控制组 (Untreated) 共有 11 个不在共同取值范围中 (off support), 处理组 (Treated) 共有 2 个不在共同取值范围中 (off support), 其余 432 个观测值均在共同取值范围中 (on support)。

上表中部的 Note 显示, 所汇报的标准误并未考虑倾向得分为估计所得的事实 (即假设倾向得分为真实值, 然后推导标准误)<sup>①</sup>; 此标准误的另一假设为同方差, 也可能不成立。为此, 可以考虑使用自助法来得到标准误 (参见第 19 章), 尽管自助标准误也未必正确 (Abadie and Imbens, 2008)。

```
. set seed 10101
. bootstrap r(att)r(atu)r(ate), reps(500): psmatch2 t age educ black
hisp married re74 re75 u74 u75,outcome(re78)n(1)ate ties logit common
```

① 这个理论问题正在得到解决, 参见 Abadie and Imbens(2012)。

Bootstrap results		Number of obs	=	445
		Replications	=	500
command: psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78) n(1) ate ties logit common				
<u>_bs_1:</u> r(att)				
<u>_bs_2:</u> r(atu)				
<u>_bs_3:</u> r(ate)				
	Observed Coef.	Bootstrap Std. Err.	z	P> z  [95% Conf. Interval]
<u>_bs_1</u>	1.410593	.9303588	1.52	0.129 -.4128764 3.234063
<u>_bs_2</u>	1.62936	.991854	1.64	0.100 -.3146385 3.573358
<u>_bs_3</u>	1.536688	.8059366	1.91	0.057 -.042919 3.116295

上表显示,ATT 的自助标准误为 0.93。上表还汇报了前面不曾给出的 ATU 与 ATE 标准误。根据这些自助标准误可知,ATE 与 ATU 在 10% 水平上显著(ATE 的  $p$  值为 0.057, 接近 5%), 而 ATT 并不显著。

下面, 使用命令 ptest 来考察此匹配结果是否较好地平衡了数据。

```
. quietly psmatch2 t age educ black hisp married re74 re75 u74 u75,  
outcome(re78) n(1) ate ties logit common  
. ptest age educ black hisp married re74 re75 u74 u75, both graph
```

Variable	Unmatched Matched	Mean		%reduct	t-test	
		Treated	Control		%bias	bias
age	Unmatched	25.816	25.054	10.7		
	Matched	25.781	25.383	5.6	47.7	
educ	Unmatched	10.346	10.088	14.1		
	Matched	10.322	10.415	-5.1	63.9	
black	Unmatched	.84324	.82692	4.4		
	Matched	.85246	.86339	-2.9	33.0	
hisp	Unmatched	.05946	.10769	-17.5		
	Matched	.06011	.04372	5.9	66.0	
married	Unmatched	.18919	.15385	9.4		
	Matched	.18579	.19126	-1.4	84.5	
re74	Unmatched	2.0956	2.107	-0.2		
	Matched	2.0672	1.9222	2.7	-1166.6	
re75	Unmatched	1.5321	1.2669	8.4		
	Matched	1.5299	1.6446	-3.6	56.7	
u74	Unmatched	.70811	.75	-9.4		
	Matched	.71038	.75956	-11.1	-17.4	
u75	Unmatched	.6	.68462	-17.7		
	Matched	.60656	.63388	-5.7	67.7	

上表显示,匹配后(Matched)大多数变量的标准化偏差(% bias)小于10%,只是变量u74的偏差为11.1%,似乎可以接受;而且大多数t检验的结果不拒绝处理组与控制组无系统差异的原假设(re75与u75为例外)。对比匹配前(Unmatched)的结果,大多数变量的标准化偏差均大幅缩小,但变量re74与u74的偏差反而有所增加。

Summary of the distribution of the abs(bias)					
BEFORE MATCHING					
Percentiles	Smallest				
1%	.2159919	.2159919			
5%	.2159919	4.388661			
10%	.2159919	8.386325	Obs	9	
25%	8.386325	9.36407	Sum of Wgt.	9	
50%	9.414047		Mean	10.19509	
		Largest	Std. Dev.	5.726406	
75%	14.12198	10.72771			
90%	17.68094	14.12198	Variance	32.79173	
95%	17.68094	17.45611	Skewness	-.2354408	
99%	17.68094	17.68094	Kurtosis	2.252342	
AFTER MATCHING					
Percentiles	Smallest				
1%	1.447804	1.447804			
5%	1.447804	2.735669			
10%	1.447804	2.93891	Obs	9	
25%	2.93891	3.62947	Sum of Wgt.	9	
50%	5.094981		Mean	4.906021	
		Largest	Std. Dev.	2.787843	
75%	5.709197	5.613233			
90%	11.05192	5.709197	Variance	7.772069	
95%	11.05192	5.933004	Skewness	1.057563	
99%	11.05192	11.05192	Kurtosis	3.751284	
Sample	Pseudo R2	LR chi2	p>chi2	MeanBias	MedBias
Raw	0.019	11.75	0.227	10.2	9.4
Matched	0.018	10.65	0.301	4.9	5.1

上表主要显示了匹配前后偏差绝对值的分布特征。各变量标准化偏差的匹配前后变化,参见图28.2。从图28.2可以直观地看出,大多数变量的标准化偏差在匹配后缩小了。

下面,画条形图来显示倾向得分的共同取值范围。

. psgraph

此命令的输出结果参见图28.3。

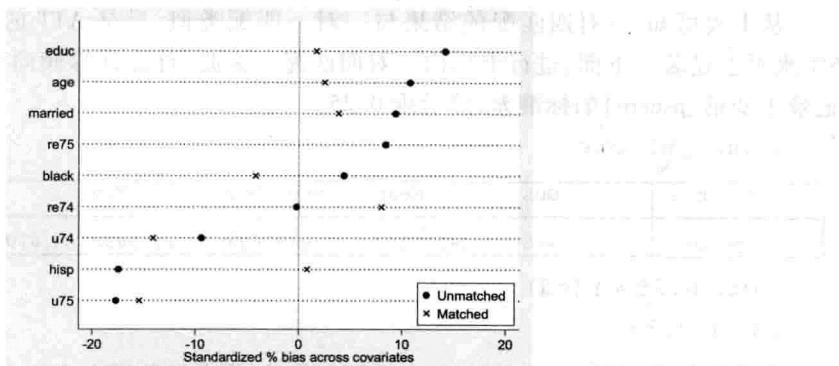


图 28.2 各变量的标准化偏差图示

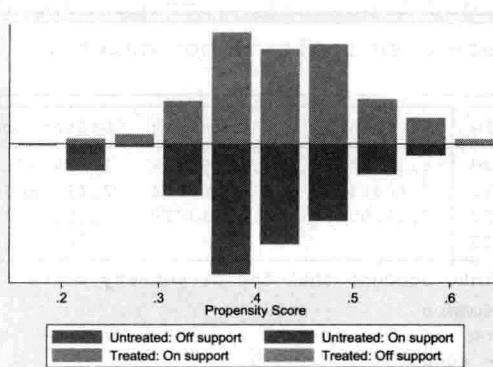


图 28.3 倾向得分的共同取值范围

从图 28.3 可以直观地看出, 大多数观测值均在共同取值范围内(on support), 故在进行倾向得分匹配时仅会损失少量样本。

下面, 进行  $k$  近邻匹配, 并令  $k=4$ 。为节省空间, 使用选择项“quietly”略去对倾向得分估计结果的汇报。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome  
(re78) n(4) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.32359171	2.08136646	.727464025	2.86
	ATU	4.52683013	5.80969564	1.28286551	.	.
	ATE			1.62111939	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support		
	Off suppo	On suppor	Total
Untreated	11	249	260
Treated	2	183	185
Total	13	432	445

从上表可知,一对四匹配的结果与一对一匹配类似,只是 ATT 的估计值有较大差异,且在 5% 水平上显著。下面,进行卡尺内一对四匹配。为此,首先计算倾向得分(由 Stata 自动生成并记录于变量\_pscore)的标准差,然后乘 0.25。

```
. sum _pscore
```

Variable	Obs	Mean	Std. Dev.	Min	Max
_pscore	445	.4157303	.0791695	.1556892	.6102826

```
. dis 0.25 * r(sd)
```

```
. 01979237
```

由此可知, $0.25\hat{\sigma}_{pscore} \approx 0.02$ 。为了保守起见,将卡尺范围定为 0.01,这意味着对倾向得分相差 1% 的观测值进行一对四匹配。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome  
(re78) n(4) cal(0.01) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.29263811	2.11232006	.729511502	2.90
	ATU	4.51395698	5.70213019	1.18817321	.	.
	ATE			1.58423615	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support			Total	S.E.	T-stat
	Off	suppo	On suppor			
Untreated	16	244		260		
Treated	2	183		185		
Total	18	427		445		

上表显示,卡尺内一对四匹配的结果与简单的一对四匹配比较接近,这说明大多数一对四匹配均发生在卡尺 0.01 的范围内,不存在太远的“近邻”。下面,进行半径(卡尺)匹配。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome  
(re78) radius cal(0.01) ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.54598746	1.85897071	.704721737	2.64
	ATU	4.51395698	6.14730708	1.63335009	.	.
	ATE			1.73004464	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support			Total	S.E.	T-stat
	Off	suppo	On suppor			
Untreated	16	244		260		
Treated	2	183		185		
Total	18	427		445		

此匹配结果依然类似。下面,进行核匹配(使用默认的核函数与带宽)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome  
(re78) kernel ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.58467178	1.8202864	.685942087	2.65
	ATU	4.52683013	6.13018211	1.60335198	.	.
	ATE			1.69524781	.	.

Note: S.E. does not take into account that the propensity score is estimated.

psmatch2: Treatment assignment	psmatch2: Common support			Total
	Off suppo	On suppor		
Untreated	11	249		260
Treated	2	183		185
Total	13	432		445

结果依然类似。然后,进行局部线性回归匹配(使用默认的核函数与带宽)。

```
. psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome  
(re78) llr ate ties logit common quietly
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.40495818	4.51340325	1.89155492	.	.
	ATU	4.52683013	6.27518361	1.74835348	.	.
	ATE			1.8090152	.	.

Note: Sample S.E.

psmatch2: Treatment assignment	psmatch2: Common support			Total
	Off suppo	On suppor		
Untreated	11	249		260
Treated	2	183		185
Total	13	432		445

上表未汇报 ATT 的标准误;为此,使用自助法以得到自助标准误(费时较长)。

```
. set seed 10101  
. bootstrap r(att) r(atu) r(ate), reps(500): psmatch2 t age educ black  
hisp married re74 re75 u74 u75, outcome (re78) llr ate ties logit  
common quietly
```

Bootstrap results			Number of obs	=	445
			Replications	=	500
command: psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78) llr ate ties logit common quietly					
<u>_bs_1:</u> r(att) <u>_bs_2:</u> r(atu) <u>_bs_3:</u> r(ate)					
	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]
_bs_1	1.891555	.7049953	2.68	0.007	.5097895 3.27332
_bs_2	1.748353	.7959967	2.20	0.028	.1882286 3.308478
_bs_3	1.809015	.6743811	2.68	0.007	.4872525 3.130778

根据上表的自助标准误,对平均处理效应的三种度量均至少在5%水平上显著。

下面,进行样条匹配(同样使用自助法计算标准误)。为此,需要先安装一个非官方命令spline。

```
. net install.snp7_1.pkg (下载安装命令 spline)①
.set seed 10101
.bootstrap r(att) r(atu) r(ate), reps(500): psmatch2 t age educ black
hisp married re74 re75 u74 u75, outcome(re78) spline ate ties logit
common quietly
```

Bootstrap results			Number of obs = 445					
			Replications = 500					
command: psmatch2 t age educ black hisp married re74 re75 u74 u75, outcome(re78)								
spline ate ties logit common quietly								
	Observed Coef.	Bootstrap Std. Err.	z	P> z	Normal-based [95% Conf. Interval]			
_bs_1	1.749681	.6590988	2.65	0.008	.457871 3.041491			
_bs_2	1.576453	.7084574	2.23	0.026	.187902 2.965004			
_bs_3	1.649834	.6532842	2.53	0.012	.3694209 2.930248			

估计结果仍然类似。总之,以上各种倾向得分匹配的结果表明,参加就业培训的平均处理效应为正,不仅在经济上显著(平均可使1978年实际收入增加近两千元),而且在统计上显著(大多数在5%水平上显著)。

下面,进行马氏匹配,并使用 Adabie and Imbens(2006)提供的异方差稳健标准误。

```
. psmatch2 t,outcome(re78) mahal(age educ black hisp married re74 re75
u74 u75) n(4) ai(4) ate
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
re78	Unmatched	6.34914538	4.55480228	1.79434311	.632853552	2.84
	ATT	6.34914538	4.42361842	1.92552697	.707891623	2.72
	ATU	4.55480228	6.50818167	1.95337939	.902160108	2.17
	ATE			1.94180029	.785347215	2.47

Note: Sample S.E.

psmatch2: Treatment assignment	psmatch2:		Total
	Common support	On suppor	
Untreated	260	260	
Treated	185	185	
Total	445	445	

<sup>①</sup> 或输入命令“findit.snp7\_1”寻找下载地址。

上表显示,无论是平均处理效应的估计值还是显著性,马氏匹配的结果与倾向得分匹配类似;这也说明了以上结果的稳健性。

## 28.7 偏差校正匹配估计量

由于在倾向得分匹配第一阶段估计倾向得分时存在不确定性(可使用 `probit`, `logit` 或非参数估计;即使确定用 `logit`,模型的具体设定仍取决于研究者),Abadie and Imbens(2002, 2004, 2006, 2011)又重新回到更简单的马氏距离,进行有放回且允许并列(ties)的  $k$  近邻匹配。这样,研究者在计算匹配估计量时,只需做少量的主观决定。

更重要的是,由于非精确匹配(inexact matching)一般存在偏差,Abadie and Imbens 提出了偏差校正的方法,通过回归的方法来估计偏差,然后得到“偏差校正匹配估计量”(bias-corrected matching estimator)。另外,Abadie and Imbens 还通过在处理组或控制组内部进行二次匹配,来得到在异方差条件下也成立的稳健标准误。

偏差校正匹配估计量可通过非官方命令 `nnmatch` 来实现,其下载方法为

```
ssc install nnmatch,replace      (下载安装命令 nnmatch)
```

该命令的基本句式为

```
nnmatch y D x1 x2 x3,metric(maha) tc(att) tc(atc) m(k) robust(#) biasadj  
(bias |varlist) pop
```

其中,选择项“`metric(maha)`”表示使用马氏距离,即权重矩阵为样本协方差矩阵的逆矩阵;默认权重矩阵是主对角线元素为各变量样本方差的对角矩阵之逆矩阵(称为 inverse variance)。选择项“`tc(att)`”表示估计 ATT,选择项“`tc(atc)`”表示估计 ATU(此处 c 表示 control,是 untreated 的同义词),默认值为 `tc(ate)`,即估计 ATE。选择项“`m(k)`”表示进行  $k$  近邻匹配,默认值为  $k=1$ 。选择项“`robust(#)`”表示计算异方差稳健的标准误,其中#须为正整数,表示用于计算稳健标准误的近邻个数(一般可让# =  $k$ )。选择项“`biasadj(bias)`”表示根据原来的协变量进行偏差校正,默认不进行偏差校正;也可以通过选择项“`biasadj(varlist)`”来指定用于偏差校正的变量名单。选择项“`pop`”表示估计“总体平均处理效应”(Population Average Treatment Effects,简记 PATE),其估计值与一般的“样本平均处理效应”(Sample Average Treatment Effects,简记 SATE)相同,只是标准误略微不同;默认估计 SATE。

下面以数据集 `ldw_exper.dta` 为例。首先,通过一对四匹配来估计 ATT,不做偏差校正,但使用异方差稳健标准误。

```
. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att)  
m(4)robust(4)
```

. nnmatch re78 t age educ black hisp married re74 re75 u74 u75, tc(att) m(4) robust(4)					
Matching estimator: Average Treatment Effect for the Treated					
Weighting matrix: inverse variance					
Number of obs = 445					
Number of matches (m) = 4					
Number of matches, robust std. err. (h) = 4					
re78	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
SATT	1.994622	.7526339	2.65	0.008	.5194864 3.469757

Matching variables: age educ black hisp married re74 re75 u74 u75

上表显示,权重矩阵为默认的“inverse variance”,即主对角线元素为各变量样本方差的对角矩阵之逆矩阵。ATT 的估计值为 1.995,且在 1% 水平上显著。

其次,重复以上命令,但进行偏差校正。

. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att)  
m(4) robust(4) bias(bias)

Matching estimator: Average Treatment Effect for the Treated					
Weighting matrix: inverse variance					
Number of obs = 445					
Number of matches (m) = 4					
Number of matches, robust std. err. (h) = 4					
re78	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
SATT	1.838424	.7526339	2.44	0.015	.363289 3.31356

Matching variables: age educ black hisp married re74 re75 u74 u75

Bias-adj variables: age educ black hisp married re74 re75 u74 u75

上表显示,经过偏差校正后,ATT 的估计值减少为 1.838,且仅在 5% 水平上显著( $p$  值为 1.5%)。最后,以样本协方差矩阵的逆矩阵为权重矩阵,使用马氏距离进行匹配。

. nnmatch re78 t age educ black hisp married re74 re75 u74 u75,tc(att)  
m(4) robust(4) bias(bias) metric(maha)

Matching estimator: Average Treatment Effect for the Treated					
Weighting matrix: Mahalanobis					
Number of obs = 445					
Number of matches (m) = 4					
Number of matches, robust std. err. (h) = 4					
re78	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
SATT	1.796419	.7489237	2.40	0.016	.3285557 3.264283

Matching variables: age educ black hisp married re74 re75 u74 u75

Bias-adj variables: age educ black hisp married re74 re75 u74 u75

由上表可知,无论使用哪种加权矩阵,对估计结果影响不大。

## 28.8 双重差分倾向得分匹配

以上介绍的各种匹配估计量均依赖于可忽略性假定,即依可测变量选择;故不适用于依不可测变量选择的情形。对于观测数据,如果怀疑存在依不可测变量选择,则大致有以下几种处理方法。

(1) 尽量使用更多的相关可测变量,以满足可忽略性假定,然后使用匹配估计量。显然,如果 $\mathbf{x}_i$ 中包括的变量太少,则不太可能满足可忽略性假定。

(2) 如果影响处理变量 $D_i$ 的不可观测变量不随时间而变,而且有面板数据,则可使用“双重差分倾向得分匹配估计量”(differences-in-differences PSM estimator)。

(3) 使用断点回归法,特别是模糊断点回归,参见本章第9~12节。

(4) 使用工具变量法,比如,Imbens and Angrist(1994)提出的局部平均处理效应IV估计量,以及本章第13节的处理效应模型。工具变量法的最大局限在于,通常很难找到有效的工具变量,因为此工具变量既要影响个体选择参与项目的决定 $D_i$ ,但又必须与是否参加项目的潜在结果( $y_{1i}, y_{0i}$ )无关。

(5) 根据依可测变量选择的影响来估计依不可测变量选择的影响,参见 Altonji et al(2005)。

本节主要关注第(2)种方法,即双重差分PSM,由 Heckman et al(1997, 1998)所提出。假设有两期面板数据,记实验前的时期为 $t'$ ,实验后的时期为 $t$ 。在时期 $t'$ ,实验还未发生,故所有个体(无论处理组还是控制组)的潜在结果均可记为 $y_{0t'}$ 。在时期 $t$ ,实验已经发生,故可能有两种潜在结果,分别记为 $y_{1t}$ (如果参与实验)与 $y_{0t}$ (如果未参与实验)。

双重差分PSM成立的前提为以下均值可忽略性假定:

$$E(y_{0t} - y_{0t'} | \mathbf{x}, D=1) = E(y_{0t} - y_{0t'} | \mathbf{x}, D=0) \quad (28.19)$$

如果假定(28.19)成立,则可一致地估计ATT:

$$\widehat{ATT} = \frac{1}{N_1} \sum_{i: i \in I_1 \cap S_p} [(y_{1ti} - y_{0ti}) - \sum_{j: j \in I_0 \cap S_p} w(i, j)(y_{0tj} - y_{0t'i})] \quad (28.20)$$

其中, $S_p$ 为共同取值范围的集合(common support), $I_1 = \{i : D_i = 1\}$ (处理组的集合), $I_0 = \{i : D_i = 0\}$ (控制组的集合), $N_1$ 为集合 $I_1 \cap S_p$ 所包含的处理组个体数,而 $w(i, j)$ 为对应于配对 $(i, j)$ 的权重,可通过核匹配或局部线性回归匹配的方法来确定(参见本章第5节)。表达式(28.20)的括弧内第一项 $(y_{1ti} - y_{0ti})$ 为处理组个体*i*实验前后的变化,而第二项中的 $(y_{0tj} - y_{0t'i})$ 则为控制组个体*j*的前后变化。概括起来,双重差分PSM法的步骤如下。

第一步,根据处理变量 $D_i$ 与协变量 $\mathbf{x}_i$ 估计倾向得分。

第二步,对于处理组的每位个体*i*,确定与其匹配的全部控制组个体(即确定集合 $S_p$ )。

第三步,对于处理组的每位个体*i*,计算其结果变量的前后变化 $(y_{1ti} - y_{0ti})$ 。

第四步,对于处理组的每位个体*i*,计算与其匹配的全部控制组个体的前后变化 $(y_{0tj} - y_{0t'i})$ ,其中 $j \in I_0 \cap S_p$ 。

第五步,针对 $(y_{1ti} - y_{0ti})$ 与 $(y_{0tj} - y_{0t'i})$ ,根据公式(28.20)进行倾向得分核匹配或局部线性回归匹配,即得到 $\widehat{ATT}$ 。

双重差分PSM法的优点在于它可以控制不可观测(unobservable)但不随时间变化(time invariant)的组间差异,比如处理组与控制组分别来自两个不同的区域,或处理组与控制组使用了不同的调查问卷。

在Stata中操作双重差分PSM法,既可手工进行(手工计算每位个体的前后变化,然后使用

命令 `psmatch2`), 也可使用非官方命令 `diff` 来自动进行(该命令可用于估计一般的双重差分估计量, 参见第 18 章)。

该命令的下载方法为

```
ssc install diff,replace
```

使用命令 `diff` 进行双重差分 PSM 估计的基本句式为：

其中，“`outcome_var`”为结果变量，必选项“`treat(varname)`”用来指定处理变量，必选项“`period(varname)`”用来指定实验期虚拟变量（实验期 = 1，非实验期 = 0）。必选项“`id(varname)`”用来指定个体 ID（这是进行匹配的前提）。必选项“`kernel`”表示进行基于倾向得分的核匹配（命令 `diff` 不提供其他匹配方法），选择项“`ktype(kernel)`”用于指定核函数，默认为二次核。必选项“`cov(varlist)`”用来指定用于估计倾向得分的协变量，选择项“`report`”表示汇报对倾向得分的估计结果。选择项“`logit`”表示使用 Logit 估计倾向得分，默认为 Probit。选择项“`support`”表示仅使用共同取值范围（common support）内的观测值进行匹配。选择项“`test`”表示检验在倾向得分匹配后，各变量在实验组与控制组的分布是否平衡（balancing test）。

下面以数据集 `cardkrueger1994.dta` 为例(参见第 18 章)。

```
. use cardkrueger1994,clear  
. diff fte,t(treated) p(t)kernel^id(id) logit cov(bk kfc roys) report  
support
```

KERNEL PROPENSITY SCORE DIFFERENCE-IN-DIFFERENCES  
 ESTIMATION ON THE COMMON SUPPORT

Report - Propensity score estimation with the logit command:

Iteration 0: log likelihood = -198.21978  
 Iteration 1: log likelihood = -196.77862  
 Iteration 2: log likelihood = -196.7636  
 Iteration 3: log likelihood = -196.7636

Logistic regression

	Number of obs	=	404
	LR chi2(3)	=	2.91
	Prob > chi2	=	0.4053
	Pseudo R2	=	0.0073

Log likelihood = -196.7636

treated	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
bk	.3108387	.3561643	0.87	0.383	-.3872306 1.008908
kfc	.6814511	.4335455	1.57	0.116	-.1682824 1.531185
roys	.520356	.4011747	1.30	0.195	-.265932 1.306644
_cons	1.05315	.2998708	3.51	0.000	.465414 1.640886

ATENTION: \_pscore is estimated at the baseline

Number of observations in the DIFF-IN-DIFF: 795

Baseline	Follow-up
Control: 78	154
Treated: 326	641
404	391

R-square: 0.02844

DIFFERENCE IN DIFFERENCES ESTIMATION

Outcome Variable	BASE LINE			FOLLOW UP			DIFF-IN-DIFF
	Control	Treated	Diff(BL)	Control	Treated	Diff(FU)	
fte	21.656	17.065	-4.591	18.914	17.499	-1.414	3.177
Std. Error	0.573	1.096	1.237	0.578	1.115	1.256	1.763
t	37.77	17.47	-3.71	16.91	17.17	-2.06	1.80
P> t	0.000	0.000	0.000***	0.000	0.000	0.260	0.072*

\* Means and Standard Errors are estimated by linear regression

\*\*Inference: \*\*\* p<0.01; \*\* p<0.05; \* p<0.1

上表上部的 Logit 回归结果表明,各品牌虚拟变量对于处理变量的解释力很弱(准  $R^2$  仅为 0.007),均不显著。这暗示此数据集并不适合进行倾向得分匹配。故本例纯粹出于演示的目的。上表下部最后一列显示,平均处理效应的系数估计值为 3.177,但仅在 10% 水平上显著。下面检验进行倾向得分匹配后,是否使得各变量在处理组与控制组的分布变得平衡。

```
. diff fte,t(treated) p(t) kernel id(id) logit cov(bk kfc roys) support
test
```

TWO-SAMPLE T TEST TEST ON THE COMMON SUPPORT					
Number of observations (baseline): 404					
	Baseline	Follow-up			
Control:	78	-	78		
Treated:	326	-	326		
	404				
t-test at period = 0:					
Weighted Variable(s)	Mean Control	Mean Treated	Diff.	t	Pr( T > t )
fte	21.656	17.065	-4.591	3.22	0.0014***
bk	0.618	0.408	-0.210	3.55	0.0004***
kfc	0.104	0.209	0.104	2.60	0.0097***
roys	0.183	0.252	0.068	1.42	0.1570

\*\*\* p<0.01; \*\* p<0.05; \* p<0.1  
Attention: option kernel weighs variables in cov(varlist)  
Means and t-test are estimated by linear regression

从上表结果可知,进行匹配后,有两个协变量(bk 与 kfc)的均值在处理组与控制组之间依然存在显著差异。这再次验证此数据集并不适用双重差分 PSM 法。一个显然的原因是,数据集中的协变量太少了。对于这个数据集而言,使用一般的双重差分法即可。

## 28.9 断点回归的思想

依可测变量选择的一种特殊情形是,有时处理变量  $D_i$  完全由某连续变量  $x_i$  是否超过某断点(cutoff point)所决定。据以进行分组的变量  $x_i$  称为“分组变量”(assignment variable, forcing variable 或 running variable)。比如,考察上大学对工资收入的影响,并假设上大学与否( $D_i$ )完全取决于由高考成绩  $x_i$  是否超过 500 分:

$$D_i = \begin{cases} 1 & \text{若 } x_i \geq 500 \\ 0 & \text{若 } x_i < 500 \end{cases} \quad (28.21)$$

记不上大学与上大学的两种潜在结果分别为  $(y_{0i}, y_{1i})$ 。由于  $D_i$  是  $x_i$  的确定性函数,故在给定  $x_i$  的情况下,可将  $D_i$  视为常数,不可能与任何变量有关系,因此  $D_i$  独立于  $(y_{0i}, y_{1i})$ ,满足可忽略性假定。但此时,并不能使用倾向得分匹配法,因为重叠假定完全不满足,对于所有处理组成员,都有  $x_i \geq 500$ ;而所有控制组成员都有  $x_i < 500$ ,二者完全没有交集!因此,匹配估计量此路不通,需另辟蹊径。

显然,处理变量  $D_i$  为  $x_i$  的函数,记为  $D(x_i)$ 。由于函数  $D(x_i)$  在  $x = 500$  处存在一个断点(discontinuity),这提供了估计  $D_i$  对  $y_i$  因果效应的机会。对于高考成绩为 498,499,500,或 501

的考生,可以认为他们在各方面(包括可观测变量与不可观测变量)都没有系统差异。他们高考成绩的细微差异只是由于“上帝之手”随机抽样的结果(考试成绩本身含随机因素)<sup>①</sup>,导致成绩为500或501的考生上大学(进入处理组),而成绩为498或499的考生落榜(进入控制组)。因此,由于制度原因,仿佛对高考成绩在小邻域 $[500 - \varepsilon, 500 + \varepsilon]$ 之间的考生进行了随机分组,故可视为准实验(quasi experiment)。由于存在随机分组,故可一致地估计在 $x = 500$ 附近的局部平均处理效应(Local Average Treatment Effect,简记 LATE)<sup>②</sup>,即

$$\begin{aligned} \text{LATE} &\equiv E(y_{1i} - y_{0i} | x = 500) \\ &= E(y_{1i} | x = 500) - E(y_{0i} | x = 500) \\ &= \lim_{x \downarrow 500} E(y_{1i} | x) - \lim_{x \uparrow 500} E(y_{0i} | x) \end{aligned} \quad (28.22)$$

其中,  $\lim_{x \downarrow 500}$  与  $\lim_{x \uparrow 500}$  分别表示从500的右侧与左侧取极限(即右极限与左极限)。在上式最后一步推导中,假设条件期望函数  $E(y_{1i} | x)$  与  $E(y_{0i} | x)$  为连续函数<sup>③</sup>,故其极限值等于函数取值。

更一般地,断点可以是某常数  $c$ ,而分组规则为

$$D_i = \begin{cases} 1 & \text{若 } x_i \geq c \\ 0 & \text{若 } x_i < c \end{cases} \quad (28.23)$$

假设在实验前(pretreatment),结果变量  $y_i$  与  $x_i$  之间存在如下线性关系:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \dots, n) \quad (28.24)$$

不失一般性,假设  $D_i = \mathbf{1}(x_i \geq c)$  的处理效应为正,则  $y_i$  与  $x_i$  之间的线性关系在  $x = c$  处就存在一个向上跳跃(jump)的断点,参见图 28.4。由于在  $x = c$  附近,个体在各方面均无系统差别,故造成条件期望函数  $E(y_i | x)$  在此跳跃的唯一原因只可能是  $D_i$  的处理效应。基于此逻辑,可将此跳跃视为在  $x = c$  处  $D_i$  对  $y_i$  的因果效应。

我们知道,在方程中引入虚拟变量的效果就是在不同的子样本中产生不同的截距项(参见第9章)。

因此,为了估计此跳跃,可将方程(28.24)改写为:

$$y_i = \alpha + \beta(x_i - c) + \delta D_i + \gamma(x_i - c)D_i + \varepsilon_i \quad (i = 1, \dots, n) \quad (28.25)$$

在上式中,变量  $(x_i - c)$  为  $x_i$  的标准化,使得  $(x_i - c)$  的断点为0。引入互动项  $\gamma(x_i - c)D_i$  是为了允许在断点两侧的回归线斜率可以不同<sup>④</sup>。对方程(28.25)进行 OLS 回归,所得  $\hat{\delta}$  就是在  $x = c$  处局部平均处理效应(LATE)的估计量。由于此回归存在一个断点,故称为“断点回归”(Regression Discontinuity,简记 RD)或“断点回归设计”(Regression Discontinuity Design,简记 RDD)。

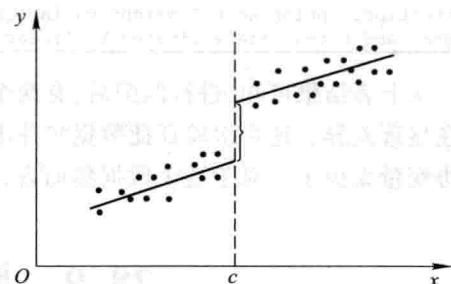


图 28.4 断点回归示意图

<sup>①</sup> 如果考生能够完全控制分组变量  $x$  的取值(比如通过自身努力),则断点回归将失效。但在一般情况下,考生无法精确地控制成绩(imprecise control over the assignment variable)。因此,在断点附近的考生,成绩大于或小于断点的概率大约都是二分之一,形成局部的随机分组(local randomization)。

<sup>②</sup> 此 LATE 不同于 Imbens 和 Angrist(1994) 基于工具变量所定义的 LATE。

<sup>③</sup> 严格来说,只需要在  $x = c$  这一点连续即可。

<sup>④</sup> 方程(28.25)等价于使用断点两侧的数据分别进行回归,然后计算两侧截距项之差。如果断点两侧的回归线斜率不同,但未包括互动项,即相当于强迫两侧的斜率相同。这会导致断点右(左)侧的观测值影响对左(右)侧截距项的估计(我们不希望有此影响),从而引起偏差。

RDD)。需要注意的是,在有互动项的情况下,如果在方程(28.25)使用 $x_i$ 而非标准化变量( $x_i - c$ ),则 $\hat{\delta}$ 虽然度量断点两侧回归线的截距之差,但并不等于这两条回归线在 $x = c$ 处的跳跃距离<sup>①</sup>。

由于在断点附近仿佛存在随机分组,故一般认为断点回归是内部有效性(internal validity)比较强的一种准实验。在某种意义上,断点回归可视为“局部随机实验”(local randomized experiment);而且,可通过考察协变量在断点两侧的分布是否有显著差异来检验此随机性。另一方面,断点回归仅推断在断点处的因果关系,并不一定能推广到其他样本值,故外部有效性(external validity)受局限。

断点回归由Thistlewaite and Campbell(1960)首次使用,但直到1990年代末才引起经济学家的重视。Hahn et al(2001)提供了断点回归的计量经济学理论基础。目前,断点回归在教育经济学、劳动经济学、健康经济学、政治经济学(Political Economy)以及区域经济学等领域的应用仍方兴未艾。参见Imbens and Lemieux(2008),Van Der Klaauw(2008)以及Lee and Lemieux(2010)的文献综述。

**例** Thistlewaite and Campbell(1960)使用断点回归研究奖学金对于未来学业成就的影响。由于奖学金由学习成绩决定,故成绩刚好达到获奖标准与差一点达到的学生具有可比性。

**例** Angrist and Lavy(1999)在研究班级规模对成绩的影响时,利用以色列教育系统的一项制度进行断点回归;该制度限定班级规模的上限为40名学生,一旦超过40名学生(比如41名学生),则该班级将被一分为二。参见习题。

## 28.10 精确断点回归

断点回归可分为两种类型。一种类型是上节介绍的“精确断点回归”(Sharp Regression Discontinuity,简记SRD),其特征是在断点 $x = c$ 处,个体得到处理的概率从0跳跃为1。另一种类型为“模糊断点回归”(Fuzzy Regression Discontinuity,简记FRD),其特征是在断点 $x = c$ 处,个体得到处理的概率从 $a$ 跳跃为 $b$ ,其中 $0 < a < b < 1$ 。本节关注精确断点回归,下节介绍模糊断点回归。

使用方程(28.25)来估计精确断点回归,存在两个问题。首先,如果回归函数包含高次项,比如二次项 $(x - c)^2$ ,则会导致遗漏变量偏差。其次,既然断点回归是局部的随机实验,则原则上只应使用断点附近的观测值,但方程(28.25)却使用了整个样本。为了解决这两个问题,可在方程(28.25)中引入高次项(比如二次项),并限定 $x$ 的取值范围为 $(c - h, c + h)$ :

$$\begin{aligned} y_i = & \alpha + \beta_1(x_i - c) + \delta D_i + \gamma_1(x_i - c) D_i \\ & + \beta_2(x_i - c)^2 + \gamma_2(x_i - c)^2 D_i + \varepsilon_i \quad (c - h < x < c + h) \end{aligned} \quad (28.26)$$

其中, $\hat{\delta}$ 为对LATE的估计量,并可使用稳健标准误来控制可能存在的异方差。但上式并未确定 $h$ 的取值,而且仍然依赖于具体的函数形式。为此,研究者开始转向非参数回归。与前面的参数回归相比,非参数回归的优点在于不依赖于具体的函数形式,而且可以通过最小化均方误差

<sup>①</sup> 存在互动项意味着,断点两侧的回归线斜率不同,并非平行线,故在断点处的跳跃距离并不等于二者的截距项之差。

(MSE)来选择最优带宽  $h$ 。直观来看,  $h$  越小, 则偏差(bias)越小, 但离  $x = c$  很近的点可能很少, 导致方差变大; 反之,  $h$  越大, 则方差越小, 但由于包含了离  $x = c$  较远的点导致偏差变大。

最简单的非参数方法就是比较  $y$  在两个区间  $(c - h, c)$  与  $[c, c + h)$  的均值。但这种方法缺乏效率, 且要求在这两个区间有较多观测值。另一种非参数方法为核回归(kernel regression), 即以核函数计算权重, 对带宽  $h$  范围内的观测值进行加权平均(参见第27章)。但核回归的边界性质并不理想, 而我们关心的恰恰是回归函数在端点的取值。为此, 一般推荐使用局部线性回归(local linear regression), 即最小化如下目标函数:

$$\min_{[\alpha, \beta, \delta, y]} \sum_{i=1}^n K[(x_i - c)/h][y_i - \alpha - \beta(x_i - c) - \delta D_i - \gamma(x_i - c)D_i]^2 \quad (28.27)$$

其中,  $K(\cdot)$  为核函数。局部线性回归的实质是, 在一个小邻域  $(c - h, c + h)$  内进行加权最小二乘法估计, 此权重由核函数来计算, 离  $c$  越近的点权重越大。针对断点回归, 较常用的核函数为三角核(triangular kernel)与矩形核(rectangular kernel, 即均匀核)。如果使用矩形核, 则为标准 OLS 回归, 等价于上文的参数回归。此估计量也称为“局部沃尔德估计量”(local Wald estimator)。

下面考察最优带宽的选择。记  $m_1(x) \equiv E(y_1 | x)$ ,  $m_0(x) \equiv E(y_0 | x)$ , 则  $\delta = m_1(c) - m_0(c)$ ,  $\hat{\delta} = \hat{m}_1(c) - \hat{m}_0(c)$ 。Imbens and Kalyanaraman(2009) 提出通过最小化两个回归函数在断点处的均方误差来选择最优带宽:

$$\min_h E\{[\hat{m}_1(c) - m_1(c)]^2 + [\hat{m}_0(c) - m_0(c)]^2\} \quad (28.28)$$

另外, 也可在方程(28.26)或(28.27)加入影响结果变量  $y_i$  的其他协变量  $w_i$ , 可通过 Stata 命令 rd 的选择项“cov(varlist)”来实现。由于断点回归可视为局部随机实验, 故是否包括协变量  $w_i$  并不影响断点回归估计量的一致性。加入协变量  $w_i$  的好处在于, 如果这些协变量对于被解释变量  $y_i$  有解释力, 则可以减少扰动项方差, 使得估计更为准确。然而, 如果所加入的协变量为内生变量, 与扰动项相关, 则反而会干扰对 LATE 的估计。

另外, 如果协变量  $w_i$  在  $x = c$  处的条件密度函数也存在跳跃, 则不宜将  $\hat{\delta}$  全部归功于该项目的处理效应。事实上, 断点回归的一个隐含假设是, 协变量  $w_i$  的条件密度在  $x = c$  处连续。为了检验此假设, 可将  $w_i$  中每个变量作为被解释变量, 进行断点回归, 考察其分布是否在  $x = c$  处有跳跃; 这可通过 Stata 命令 rd 的选择项“x(varlist)”来实现。

在进行断点回归时, 还应注意可能存在“内生分组”(endogenous sorting): 如果个体事先知道分组规则, 并可通过自身努力而完全控制分组变量(complete manipulation), 则可自行选择进入处理组或控制组, 导致在断点附近的内生分组而非随机分组, 引起断点回归失效。另一方面, 如果个体事先不清楚分组规则, 或只能部分地控制分组变量(partial manipulation), 则一般不存在此担忧。对于内在分组的可能性, 可从理论上讨论, 也可根据数据进行检验。假设存在内生分组, 则个体将自行选择进入断点两侧, 导致在断点两侧的分布不均匀, 即分组变量  $x$  的密度函数  $f(x)$  在断点  $x = c$  处不连续, 出现左极限不等于右极限的情形。为此, McCrary(2008) 提出检验以下原假设:

$$H_0: \theta \equiv \ln \lim_{x \downarrow c} f(x) - \ln \lim_{x \uparrow c} f(x) = \ln f^+ - \ln f^- = 0 \quad (28.29)$$

通过计算  $\hat{\theta}$  及其标准误, 即可检验密度函数  $f(x)$  是否  $x = c$  处连续。另外, 根据同样的逻辑, 内生分组也可能使得协变量  $w_i$  在  $x = c$  两侧分布不均匀; 故检验协变量  $w_i$  的条件密度在  $x = c$  处是否连续也有帮助。

由于断点回归在操作上存在不同选择, 故在实践中, 一般建议同时汇报以下各种情形, 以保证结果的稳健性。

- (1) 分别汇报三角核与矩形核的局部线性回归结果(后者等价于线性参数回归);
- (2) 分别汇报使用不同带宽的结果(比如,最优带宽及其二分之一或两倍带宽);
- (3) 分别汇报包含协变量与不包含协变量的情形。
- (4) 进行模型设定检验,包括检验分组变量与协变量的条件密度是否在断点处连续。

## 28.11 模糊断点回归

模糊断点回归的特征是,在断点  $x = c$  处,个体得到处理的概率从  $a$  跳跃为  $b$ ,其中  $0 < a < b < 1$ ,参见图 28.5。这意味着,即使  $x > c$ ,也不一定得到处理,只不过得到处理的概率在  $x = c$  处有一个不连续的跳跃。显然,所谓“模糊断点回归”,其断点并不模糊(断点很明确地在  $x = c$  处),只不过分组变量  $x$  跨过断点  $c$  的后果并非泾渭分明,只是得到处理的概率存在跳跃。在某种意义上,精确断点回归可视为模糊断点回归的特例或极限情形。

回到高考录取上大学的例子。事实上,高考成绩上线并不能完全保证上大学,因为能否上大学还取决于填报的志愿,甚至有些上线考生放弃上大学的机会;另一方面,即使成绩未上线,但也可能因某种特长而得到加分,从而得到上大学的机会。这表明,分数线并不完全决定上大学。然而,上大学的概率确实在分数线的位置上有一个不连续的跳跃。

在模糊断点的情况下,处理变量  $D$  并不完全由分组变量  $x$  所决定。一般来说,影响处理变量  $x$  的其他因素也会影响结果变量  $y$ ,导致在方程(28.25)或(28.27)中处理变量  $D$  与扰动项  $\varepsilon$  相关,故 OLS 估计量不一致。比如,虽然成绩上线却因志愿不妥而落榜者多有较深实力,而这种不可观测的实力可以影响结果变量  $y$ 。

为了在模糊断点的情况下识别平均处理效应,需要引入以下条件独立假定。

**假定 28.4** 给定  $x$ ,则  $(y_1 - y_0)$  独立于  $D$ ,即  $(y_{1i} - y_{0i}) \perp D_i | x_i$ 。

此假定意味着,在给定分组变量  $x$  的情况下, $D$  可以与  $y_0$  相关,但不能与参加项目的收益  $(y_{1i} - y_{0i})$  相关。由于  $y = y_0 + D(y_1 - y_0)$ ,故

$$\begin{aligned} E(y|x) &= E(y_0|x) + E[D(y_1 - y_0)|x] \\ &= E(y_0|x) + E(D|x) \cdot E[(y_1 - y_0)|x] \end{aligned} \quad (28.30)$$

其中,  $E[(y_1 - y_0)|x]$  是我们想要估计的平均处理效应,而  $E(D|x)$  为倾向得分。在上式的第二步使用了条件独立假定。对上式两边从  $c$  的右边取极限可得

$$\lim_{x \downarrow c} E(y|x) = \lim_{x \downarrow c} E(y_0|x) + \lim_{x \downarrow c} E(D|x) \cdot \lim_{x \downarrow c} E[(y_1 - y_0)|x] \quad (28.31)$$

同理,对上式两边从  $c$  的左边取极限可得

$$\lim_{x \uparrow c} E(y|x) = \lim_{x \uparrow c} E(y_0|x) + \lim_{x \uparrow c} E(D|x) \cdot \lim_{x \uparrow c} E[(y_1 - y_0)|x] \quad (28.32)$$

假设函数  $E(D|x)$ , $E(y_0|x)$  与  $E(y_1|x)$  在  $x = c$  处连续,则其左极限等于右极限,也等于其函数值,故  $\lim_{x \downarrow c} E(y_0|x) = \lim_{x \uparrow c} E(y_0|x)$ ,而且  $\lim_{x \downarrow c} E[(y_1 - y_0)|x] = \lim_{x \uparrow c} E[(y_1 - y_0)|x] =$

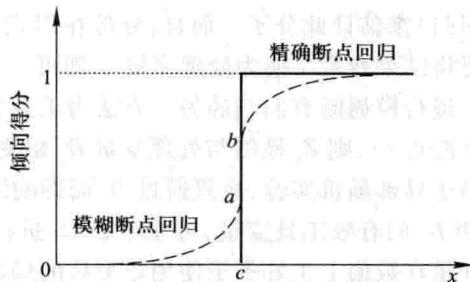


图 28.5 精确断点回归与模糊断点回归

$E[(y_1 - y_0) | x = c]$ 。因此,将方程(28.31)减去(28.32)可得

$$\lim_{x \downarrow c} E(y|x) - \lim_{x \uparrow c} E(y|x) = [\lim_{x \downarrow c} E(D|x) - \lim_{x \uparrow c} E(D|x)] \cdot E[(y_1 - y_0) | x = c] \quad (28.33)$$

根据模糊断点回归的定义可知,  $\lim_{x \downarrow c} E(D|x) - \lim_{x \uparrow c} E(D|x) = b - a \neq 0$ , 故可将其作为分母:

$$LATE \equiv E[(y_1 - y_0) | x = c] = \frac{\lim_{x \downarrow c} E(y|x) - \lim_{x \uparrow c} E(y|x)}{\lim_{x \downarrow c} E(D|x) - \lim_{x \uparrow c} E(D|x)} \quad (28.34)$$

显然,上式的分子就是精确断点回归的 LATE,而分母为得到处理的概率(即倾向得分)在断点  $c$  处的跳跃( $b - a$ )。表达式(28.34)是精确断点回归表达式(28.22)的推广,因为在精确断点的情况下, $b - a = 1$ ,将其代入(28.34)式即可得到(28.22)式。

由于表达式(28.34)的分子就是精确断点回归的 LATE,故可用精确断点回归(比如,局部线性回归)来估计此分子。而且,分母在形式上与分子完全一样,故也可用精确断点回归来估计,只要将结果变量  $y$  换为处理变量  $D$  即可。

进行模糊断点回归的另一方法为工具变量法。定义  $Z_i = \mathbf{1}(x_i \geq c)$ (即分组变量是否大于或等于断点  $c$ ),则  $Z_i$  显然与处理变量  $D_i$  相关,满足相关性。另一方面,  $Z_i = \mathbf{1}(x_i \geq c)$  在断点  $c$  附近相当于局部随机实验,故只通过  $D_i$  而影响结果变量  $y_i$ ,与扰动项  $\varepsilon_i$  不相关,满足外生性。因此,  $Z_i$  为  $D_i$  的有效工具变量,可使用 2SLS 进行估计。可以证明,如果使用相同的带宽  $h$ ,则此 2SLS 估计量在数值上正好等于使用矩形核的局部线性回归估计量。

以上介绍的断点回归均假设在断点附近仿佛存在局部随机分组。但如果分组变量为年龄(时间)或地理区域,则这种解释一般行不通,被 Lee and Lemieux(2010)称为“非随机断点设计”(Nonrandomized discontinuity design)。比如,以年龄 65 岁为分界线,年满 65 岁即可获得退休金。此时,分组变量为时间,是个确定性过程,个体无法控制。此时,须考虑以下三种可能性。首先,年满 65 岁是否使得个体有资格参加其他项目,从而通过其他渠道影响结果变量。其次,虽然年满 65 岁即可获得退休金,但退休金的效应可能需要几年后才能体现(即可能存在动态效应)。最后,由于个体可以预见 65 岁以后将得到退休金,故可能在 65 岁之前就调整其经济行为。对于这些可能性,都应进行具体分析,才能得到令人信服的结论。

另一种非随机断点设计使用地理区域作为分组变量,以某种区域分界线作为断点,进行“地理断点回归”(geographic RD)。比如,Black(1999)通过比较在学区分界线两侧的房价来测算居民对高质量小学教育的支付意愿(willingness to pay)。由于个体一般可以选择住在学区分界线的哪一侧,故很难视为局部随机分组。此时,需要说明的是,在分界线两侧,除了处理变量不同外(一侧的学生去一所学校,而另一侧的学生去另一所学校),在其他方面均几乎没有差别。为了保证分界线两侧的可比性,Black(1999)剔除了分界线为主要街道或高速公路的部分分界线(主要街道或高速公路两侧的社区可能有较大差别,尽管距离很近)。

**例(强制矿工制度的长期经济影响)** 在 1573—1812 年期间,西班牙殖民者在秘鲁与玻利维亚实行了一种称为“mining mita”的强制矿工征用制度。该制度规定,在离矿山较近的一定区域内,每个土著社区须提供其成年男性人口的七分之一作为强制矿工。为了研究此制度的长期经济影响,Dell(2010)使用断点回归来比较此区域分界线两侧的当代家庭消费与儿童发育不良比例。为了保证分界线两侧具有可比性,该研究剔除了分界线一侧为平原而另一侧为安第斯山脉的部分分界线。

## 28.12 断点回归的 Stata 实例

断点回归可通过非官方 Stata 命令 `rd` 来实现：

```
ssc install rd,replace (下载安装命令 rd)
```

该命令使用局部线性回归来估计断点回归模型，其基本句式为

```
rd y D x, z0(real) strineq mbw(numlist) graph bdep oxline kernel  
(rectangle) cov(varlist) x(varlist)
```

其中，“y”为结果变量，“D”为处理变量，而“x”为分组变量。选择项“`z0(real)`”用来指定断点位置，默认值为“`z0(0)`”，即断点为原点。如果省略处理变量 `D`，则默认为精确断点回归，并根据分组变量 `x` 来计算处理变量，即如果 `x` 大于或等于断点 `z0`，则 `D` 取值为 1；反之，`D` 取值为 0。选择项“`strineq`”表示根据严格不等式 (strong inequality) 来计算处理变量，即如果 `x` 大于断点 `z0`，则 `D` 取值为 1；反之，`D` 取值为 0。

选择项“`mbw(numlist)`”用来指定最优带宽的倍数，默认值为“`mbw(50 100 200)`”，即根据最优带宽的 0.5, 1 与 2 倍进行局部线性回归，其中 100 对应于根据 Imbens 和 Kalyanaraman (2009) 计算的最优带宽。选择项“`graph`”表示根据所选的每一带宽，画出其局部线性回归图。选择项“`bdep`”表示通过画图来考察断点回归估计量对带宽的依赖性 (bandwidth dependence)，而选择项“`oxline`”表示在此图的默认带宽 (即最优带宽) 上画一条直线，以便识别。选择项“`kernel(rectangle)`”表示使用矩形核 (即均匀核)，默认使用三角核。选择项“`cov(varlist)`”用来指定加入局部线性回归的协变量。选择项“`x(varlist)`”表示检验这些协变量是否在断点处有跳跃 (估计跳跃值及其显著性)。

下面以命令 `rd` 自带的数据集 `votex.dta` 为例演示断点回归的操作。该数据集用于考察美国国会选区如果有一名民主党众议员对该选区联邦支出 (federal expenditure within a Congressional district) 的影响。传统上，民主党倾向于大政府，故一个选区如果有民主党众议员，则该议员可能为该选区争取更多的联邦支出。然而，直接对二者进行回归可能存在遗漏变量问题或双向因果关系。为此，使用该民主党候选人的得票比例作为分组变量，以 0.5 作为断点 (在两党政治中，得票比例大于或等于 0.5 则当选，反之落选)，进行断点回归。数据集 `votex.dta` 的主要变量包括：结果变量 `lne` (选区联邦开支的对数)，分组变量 `d` (民主党候选人得票比例减去 0.5)，处理变量 `win` (民主党候选人当选)，以及一系列协变量。

```
. use votex.dta,clear
```

先来看一下数据集中的变量。

```
. d
```

obs:	349	102nd Congress		
vars:	19	4 Feb 2013 15:33		
size:	37,692			
<hr/>				
variable	storage name	display type	value format	variable label
fips	byte	%8.0g	fips	State code
district	byte	%8.0g		Congr district
d	double	%10.0g		Dem vote share minus .5
win	byte	%9.0g		Dem Won Race
lne	float	%9.0g		Log fed expenditure in district
i	byte	%9.0g		Incumbent
votingpop	long	%12.0g		Voting Age Population
votpop	double	%10.0g		Voting Age Population Share
populatn	long	%12.0g		Population
black	double	%12.0g		Black Population Share
blucllr	double	%12.0g		Blue-collar Population Share
farmer	double	%12.0g		Farmer Population Share
fedwrkr	double	%12.0g		Fed Worker Population Share
forborn	double	%12.0g		Foreign Born Population Share
manuf	double	%12.0g		Manufactur Population Share
unemployd	double	%12.0g		Unemp Population Share
union	float	%9.0g		Unionized Population Share
urban	double	%12.0g		Urban Population Share
veterans	double	%12.0g		Veteran Population Share
<hr/>				
Sorted by: fips district				

首先, 使用最优带宽以及默认的三角核进行精确断点回归, 并画图(参见图 28.6)。

. rd lne d,gr mbw(100)

Two variables specified; treatment is assumed to jump from zero to one at Z=0.						
Assignment variable Z is d						
Treatment variable X_T unspecified						
Outcome variable y is lne						
Command used for graph: lpoly; Kernel used: triangle (default)						
Bandwidth: .29287776; loc Wald Estimate: -.07739553						
Estimating for bandwidth .29287775925349						
<hr/>						
lne	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889

从上表可知, 局部沃尔德估计值(local Wald estimate)为负, 且很不显著; 说明拥有民主党众议员的选区并不能吸引更多的联邦开支。图 28.6 也显示, 条件期望函数  $E(lne|d)$  只在断点  $d = 0$  处稍微向下跳跃。

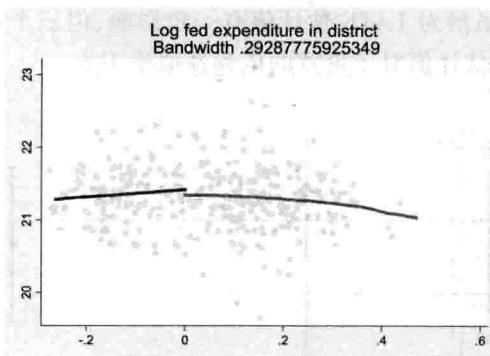


图 28.6 使用最优带宽与三角核的断点回归图

其次,加入协变量宽重复上述估计,但省略画图。

```
. rd lne d,mbw(100) cov(i votpop black blucllr farmer fedwrkr forborn  
manuf unemployd union urban veterans)
```

```
Two variables specified; treatment is  
assumed to jump from zero to one at Z=0.
```

```
Assignment variable Z is d  
Treatment variable X_T unspecified  
Outcome variable y is lne
```

```
Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
lwald	.0543733	.0921634	0.59	0.555	-.1262636 .2350102

上表显示,LATE 估计值变为正,但依然很不显著。

再次,去掉协变量,但同时估计三种带宽,并画出估计值对带宽的依赖性(参见图 28.7)。

```
. rd lne d,gr bdep oxline
```

```
Two variables specified; treatment is  
assumed to jump from zero to one at Z=0.
```

```
Assignment variable Z is d  
Treatment variable X_T unspecified  
Outcome variable y is lne
```

```
Command used for graph: lpoly; Kernel used: triangle (default)  
Bandwidth: .29287776; loc Wald Estimate: -.07739553
```

```
Bandwidth: .14643888; loc Wald Estimate: -.09491495
```

```
Bandwidth: .58575552; loc Wald Estimate: -.0543086
```

```
Estimating for bandwidth .29287775925349
```

```
Estimating for bandwidth .146438879626745
```

```
Estimating for bandwidth .58575551850698
```

lne	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438 .1295889
lwald50	-.0949149	.1454442	-0.65	0.514	-.3799804 .1901505
lwald200	-.0543086	.0911788	-0.60	0.551	-.2330157 .1243985

从上表可知,改变带宽虽然对 LATE 估计值有一定影响,但三个估计值均为负,且依然不显著。从图 28.7 也可以看出,估计值对于带宽的依赖性似乎不大。

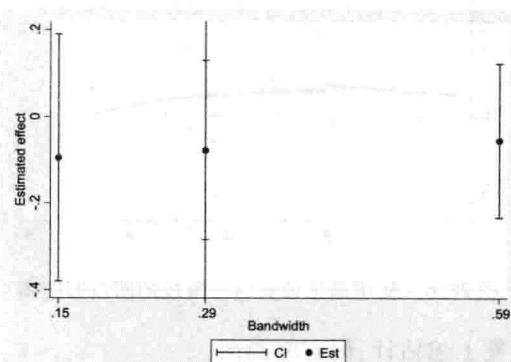


图 28.7 断点回归估计值对带宽的依赖性

进行断点回归后,还需要对其设定进行检验。下面,检验协变量在断点处的条件密度是否存在跳跃。

```
. rd lne d,mbw(100) x(i votpop black blucllr farmer fedwrkr forborn  
manuf unemployd union urban veterans)
```

```
Two variables specified; treatment is  
assumed to jump from zero to one at Z=0.
```

```
Assignment variable Z is d  
Treatment variable X_T unspecified  
Outcome variable y is lne
```

```
Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
i	-.0044941	.1208008	-0.04	0.970	-.2412592 .2322711
votpop	-.0082128	.0062347	-1.32	0.188	-.0204326 .0040071
black	-.0036113	.020048	-0.18	0.857	-.0429046 .0356821
blucllr	.0026193	.0057316	0.46	0.648	-.0086144 .013853
farmer	-.0078737	.0037566	-2.10	0.036	-.0152366 -.0005109
fedwrkr	.0001617	.0037584	0.04	0.966	-.0072046 .0075281
forborn	-.015235	.0120682	-1.26	0.207	-.0388882 .0084183
manuf	.0147223	.0100352	1.47	0.142	-.0049463 .0343908
unemployd	-.0007393	.0019069	-0.39	0.698	-.0044769 .0029982
union	-2.25e-06	3.66e-06	-0.61	0.540	-9.43e-06 4.94e-06
urban	.0370978	.0559882	0.66	0.508	-.072637 .1468326
veterans	.0015796	.0036205	0.44	0.663	-.0055164 .0086756
lwald	-.0773955	.1056062	-0.73	0.464	-.28438 .1295889

上表显示,除了变量 farmer(农民占人口比例)外,所有协变量的条件密度函数在断点处都是连续的。下面,使用 McCrary (2008) 的方法检验分组变量的密度函数是否在断点处不连续。McCrary 检验分两步进行。第一步,将分组变量在断点  $c$  两侧尽量等距离细分,画很不光滑的直方图(a very undersmoothed histogram),记组距(bin size)为  $b$ ,记每组的中心位置为变量  $X_j =$

$\left\{ \cdots, c - \frac{3b}{2}, c - \frac{b}{2}, c + \frac{b}{2}, c + \frac{3b}{2}, \cdots \right\}$ 。然后计算每组的标准化频率 (normalized cellsize), 即频数除以  $nb$  ( $n$  为样本容量), 记为  $Y_j$ 。第二步, 使用三角核, 将  $Y_j$  对  $X_j$  进行局部线性回归(参见第 27 章); 针对分组变量的取值  $r_0 = \{\cdots, c - 2b, c - b, c + b, c + 2b, \cdots\}$ , 可得密度函数估计值  $\hat{f}(r_0)$  (记为 fhat) 及标准误  $SE[\hat{f}(r_0)]$  (记为 se\_fhat)。通过方程 (28.29) 可计算  $\hat{\theta} = \widehat{\ln f^+} - \widehat{\ln f^-}$  (称为 “log difference in height”), 然后检验密度函数是否在  $c$  处连续。

McCrory 检验可通过非官方 Stata 命令 DCdensity 来实现(其中, DC 表示 Discontinuity), 下载地址为 <http://emlab.berkeley.edu/~jmccrary/DCdensity/>。将该网页的“DCdensity.ado”文件下载到文件夹“\ado\plus”即可(在 Stata 中输入命令 sysdir 即可显示此文件夹的位置)。

命令 DCdensity 的基本句式为

```
. DCdensity assign_var, breakpoint (#) generate (Xj Yj r0 fhat se_fhat)
graphname(filename)
```

其中, “assign\_var”为分组变量, 必选项“breakpoint (#)”用来指定断点位置, 必选项“generate (Xj Yj r0 fhat se\_fhat)”用来指定输出变量名, 而选择项“graphname (filename)”用来指定密度函数图的文件名。

下面, 将该命令应用于此数据集的分组变量 d, 画图结果参见图 28.8。

```
. DCdensity d, breakpoint (0) generate (Xj Yj r0 fhat se_fhat) graphname
(rd.eps)
```

```
Using default bin size calculation, bin size = .017174107
Using default bandwidth calculation, bandwidth = .10514868

Discontinuity estimate (log difference in height): -.429396753
                                                (.444361558)

Performing LLR smoothing.
45 iterations will be performed
....
Exporting graph as rd.eps
(note: file rd.eps not found)
(file rd.eps written in EPS format)
```

从上表可知,  $\hat{\theta} = -0.43$ , 而标准误为 0.44, 故可接受密度函数在  $c$  处连续的原假设。从图 28.8 也可以看出, 断点两侧密度函数估计值的置信区间有很大部分重叠, 故断点两侧的密度函

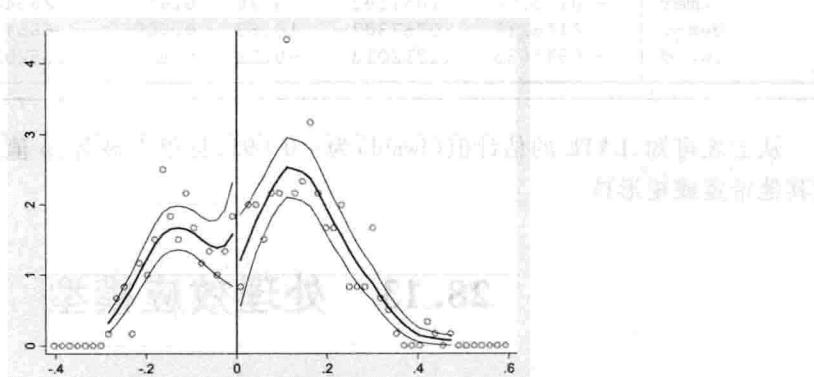


图 28.8 分组变量密度函数在断点处的连续性

数不存在显著差异。

以上为精确断点回归。纯粹为了演示模糊断点回归，下面随机生成一个新的处理变量 ranwin。变量 ranwin 不由分组变量 d 完全决定，但与原来的处理变量 win 高度相关（win 完全由 d 决定）。

```
. set seed 10101
. gen ranwin = cond(uniform() < .1, 1 - win, win)
```

其中，“uniform()”表示生成一个在 [0, 1] 区间服从均匀分布的随机变量，而命令“cond(uniform() < .1, 1 - win, win)”表示，如果此随机变量小于 0.1，则变量 ranwin = 1 - win；否则，变量 ranwin = win。因此，ranwin 也是虚拟变量，且与 win 高度相关。

下面，看一下 ranwin 与 win 的分布情况。

```
. tab ranwin win
```

ranwin	Dem Won Race		Total
	0	1	
0	118	20	138
1	13	198	211
Total	131	218	349

从上表可知，在大多数情况下，ranwin 与 win 的取值相同。下面使用最优带宽与默认的三角核进行模糊断点回归。

```
. rd lne ranwin d,mbw(100)
```

```
Three variables specified; jump in treatment
at Z=0 will be estimated. Local Wald Estimate
is the ratio of jump in outcome to jump in treatment.
```

```
Assignment variable Z is d
Treatment variable X_T is ranwin
Outcome variable y is lne
```

```
Estimating for bandwidth .29287775925349
```

lne	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
numer	-.0773955	.1051192	-0.74	0.462	-.2834254 .1286343
denom	.8158211	.0767307	10.63	0.000	.6654317 .9662105
lwald	-.0948683	.1312013	-0.72	0.470	-.3520182 .1622816

从上表可知，LATE 的估计值(lwald)为 -0.095，且很不显著(p 值为 0.47)。读者可自行尝试其他带宽或矩形核。

## 28.13 处理效应模型

解决依不可测变量选择问题的另一方法是遵循 Heckman (1979) 样本选择模型的传统，直接对处理变量  $D_i$  进行结构建模。为此，Maddala (1983) 提出了以下“处理效应模型”(treatment

effects model) :

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \gamma D_i + \varepsilon_i \quad (28.35)$$

假设处理变量由以下“处理方程”(treatment equation)所决定:

$$D_i = \mathbf{1}(\mathbf{z}'_i \boldsymbol{\delta} + u_i) \quad (28.36)$$

其中,  $\mathbf{1}(\cdot)$  为示性函数(indicator function)。 $\mathbf{z}_i$  可以与  $\mathbf{x}_i$  有重叠的变量, 但  $\mathbf{z}_i$  中至少有一个变量, 比如  $z_{ii}$ , 不在  $\mathbf{x}_i$  中。进一步, 假设  $\text{Cov}(z_{ii}, \varepsilon_i) = 0$ , 即虽然  $z_{ii}$  影响个体是否参与项目  $D_i$ , 但并不直接影响结果变量  $y_i$  (只通过  $D_i$  间接影响  $y_i$ ); 故可将  $z_{ii}$  视为  $D_i$  的工具变量。假设扰动项  $(\varepsilon_i, u_i)$  服从二维正态分布:

$$\begin{pmatrix} \varepsilon_i \\ u_i \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho \sigma_\varepsilon \\ \rho \sigma_\varepsilon & 1 \end{pmatrix} \right] \quad (28.37)$$

其中,  $\rho$  为  $(\varepsilon_i, u_i)$  的相关系数, 而  $u_i$  的方差被标准化为 1 (因为  $u_i$  是 Probit 模型的扰动项, 参见第 11 章)。我们允许  $\rho \neq 0$ , 这正是模型内生性的来源。反之, 如果  $\rho = 0$ , 则不存在内生性, 可直接用 OLS 得到对方程(28.35)的一致估计。对于参加者而言,  $y_i$  的条件期望为

$$\begin{aligned} E(y_i | D_i = 1, \mathbf{x}_i, \mathbf{z}_i) &= \mathbf{x}'_i \boldsymbol{\beta} + \gamma + E(\varepsilon_i | D_i = 1, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \gamma + E(\varepsilon_i | \mathbf{z}'_i \boldsymbol{\delta} + u_i > 0, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \gamma + E(\varepsilon_i | u_i > -\mathbf{z}'_i \boldsymbol{\delta}, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + \gamma + \rho \sigma_\varepsilon \lambda(-\mathbf{z}'_i \boldsymbol{\delta}) \end{aligned} \quad (28.38)$$

其中,  $\lambda(\cdot)$  为反米尔斯函数, 即  $\lambda(c) = \frac{\phi(c)}{1 - \Phi(c)}$ 。在上式推导的最后一步, 用到了偶然断尾的条件期望公式(参见第 14 章)。类似地, 未参加者的条件期望为

$$\begin{aligned} E(y_i | D_i = 0, \mathbf{x}_i, \mathbf{z}_i) &= \mathbf{x}'_i \boldsymbol{\beta} + E(\varepsilon_i | D_i = 0, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + E(\varepsilon_i | \mathbf{z}'_i \boldsymbol{\delta} + u_i \leq 0, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} + E(\varepsilon_i | u_i \leq -\mathbf{z}'_i \boldsymbol{\delta}, \mathbf{x}_i, \mathbf{z}_i) \\ &= \mathbf{x}'_i \boldsymbol{\beta} - \rho \sigma_\varepsilon \lambda(\mathbf{z}'_i \boldsymbol{\delta}) \end{aligned} \quad (28.39)$$

将方程(28.38)减去方程(28.39), 可得参加者与未参加者的条件期望之差:

$$E(y_i | D_i = 1, \mathbf{x}_i, \mathbf{z}_i) - E(y_i | D_i = 0, \mathbf{x}_i, \mathbf{z}_i) = \gamma + \rho \sigma_\varepsilon [\lambda(-\mathbf{z}'_i \boldsymbol{\delta}) + \lambda(\mathbf{z}'_i \boldsymbol{\delta})] \quad (28.40)$$

显然, 如果直接比较处理组与控制组的平均收益  $y_i$ , 将遗漏上式右边第二项  $\rho \sigma_\varepsilon [\lambda(-\mathbf{z}'_i \boldsymbol{\delta}) + \lambda(\mathbf{z}'_i \boldsymbol{\delta})]$ , 导致不一致的估计(除非  $\rho = 0$ )。为了将处理组与控制组放在一起进行回归, 定义个体  $i$  的风险(hazard)为<sup>①</sup>

$$\lambda_i = \begin{cases} \lambda(-\mathbf{z}'_i \boldsymbol{\delta}) & \text{若 } D_i = 1 \\ -\lambda(\mathbf{z}'_i \boldsymbol{\delta}) & \text{若 } D_i = 0 \end{cases} \quad (28.41)$$

这样, 可以将方程(28.38)与(28.39)合并为一个方程(对于参加者与未参加者都适用):

$$E(y_i | \mathbf{x}_i, \mathbf{z}_i) = \mathbf{x}'_i \boldsymbol{\beta} + \gamma D_i + \rho \sigma_\varepsilon \lambda_i \quad (28.42)$$

为此, 可进行类似于 Heckit 的两步法估计。

第一步: 用 Probit 估计方程  $P(D_i = 1 | \mathbf{z}_i) = \Phi(\mathbf{z}'_i \boldsymbol{\delta})$ , 得到估计值  $\hat{\boldsymbol{\delta}}$ , 计算  $\hat{\lambda}_i$ 。

第二步: 用 OLS 回归  $y_i \xrightarrow{\text{OLS}} \mathbf{x}_i, D_i, \hat{\lambda}_i$ , 得到估计值  $\hat{\boldsymbol{\beta}}, \hat{\gamma}, \hat{\rho \sigma_\varepsilon}$ 。

<sup>①</sup> “风险函数”也称为反米尔斯函数。

在 Stata 中,称  $\lambda_i$  的系数估计值  $\widehat{\rho\sigma_\varepsilon}$  为 lambda。两步法的优点是计算方便;其缺点在于,第一步的估计误差被带入第二步中,导致效率损失。更有效率的做法是,使用最大似然估计法 (MLE),同时估计所有模型参数。需要注意的是,上述处理效应模型依赖于对结构方程的正确设定,如果模型设定有误, $z_{ii}$  不是有效工具变量(比如  $z_{ii}$  与  $\varepsilon_i$  相关),或扰动项不服从正态分布,都会导致不一致的估计。

处理效应模型的 Stata 命令为

```
treatreg y x1 x2 x3, treat(D = z1 z2 z3) twostep first
```

其中,“y”为结果变量,“x1 x2 x3”为直接影响 y 的自变量。“treat(D = z1 z2 z3)”表示处理方程,其中“D”为处理变量,“z1 z2 z3”为影响 D 的变量。选择项“twostep”表示使用两步法,默认使用 MLE;选择项“first”表示汇报第一阶段 Probit 回归的结果。

下面以数据集“labor.dta”为例来说明处理效应模型的 Stata 操作<sup>①</sup>。被解释变量为 ww(妻子工资),wa(妻子年龄),cit(是否住在大城市),以及 we(妻子受教育年限);其他变量包括 wmed(妻子母亲的受教育年限),以及 wfed(妻子父亲的受教育年限)。我们希望知道妻子是否上大学对其工资的影响。

```
. use labor.dta, clear
```

假设受教育年限超过 12 年即为受过大学教育,生成以下虚拟变量:

```
. gen wc = (we > 12)
```

首先,作为参照,进行 OLS 回归。

```
. reg ww wa cit wc, r
```

Linear regression							Number of obs = 250
							F( 3, 246) = 3.29
							Prob > F = 0.0212
							R-squared = 0.0555
							Root MSE = 2.54
ww	Robust						[95% Conf. Interval]
	Coef.	Std. Err.	t	P> t			
wa	-.0104985	.0204216	-0.51	0.608	-.050722 .029725		
cit	.1278922	.3225258	0.40	0.692	-.507372 .7631564		
wc	1.332192	.4321052	3.08	0.002	.4810939 2.183289		
_cons	2.278337	.8543208	2.67	0.008	.5956206 3.961054		

由上表可知,变量 wc(妻子是否上大学)的系数为 1.33,且在 1% 水平上显著。然而,妻子是否上大学显然为内生虚拟变量。一种可能的情形是,(不可观测的)妻子能力越高,则越可能上大学,也越可能获得高工资,故 OLS 可能高估了上大学的作用。

下面,假设 wmed(妻子母亲的受教育年限)与 wfed(妻子父亲的受教育年限)都是影响 wc(妻子是否上大学)的外生因素,进行两步法估计。

```
. treatreg ww wa cit, treat(wc = wmed wfed) two first
```

<sup>①</sup> 此数据集由 Stata 提供,截取自 Berndt(1996)。

Probit regression		Number of obs = 250			
		LR chi2(2) = 51.93			
		Prob > chi2 = 0.0000			
Log likelihood = -121.31755		Pseudo R2 = 0.1763			
<hr/>					
wc	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
wmed	.1198888	.0319862	3.75	0.000	.0571971 .1825806
wfed	.0960764	.0290583	3.31	0.001	.0391233 .1530295
cons	-2.631496	.3308389	-7.95	0.000	-3.279928 -1.983063
<hr/>					
Treatment-effects model -- two-step estimates		Number of obs = 250			
		Wald chi2(3) = 3.67			
		Prob > chi2 = 0.2998			
<hr/>					
ww	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
wa	-.0111623	.020152	-0.55	0.580	-.0506594 .0283348
cit	.1276102	.33619	0.38	0.704	-.53131 .7865305
wc	1.257995	.8007428	1.57	0.116	-.3114319 2.827422
cons	2.327482	.9610271	2.42	0.015	.4439031 4.21106
<hr/>					
wc	wmed	.1198888	3.75	0.000	.0571971 .1825806
	wfed	.0960764	3.31	0.001	.0391233 .1530295
	cons	-2.631496	.3308389	-7.95	0.000 -3.279928 -1.983063
<hr/>					
hazard					
lambda	.0548738	.5283928	0.10	0.917	-.9807571 1.090505
<hr/>					
	rho	0.02178			
	sigma	2.5198211			
<hr/>					

上表上部显示,变量 wmed 与 wfed 在第一阶段 Probit 回归中均有显著的正作用。上表下部显示,尽管变量 wc 的系数估计值仍为 1.26(仅略小于 OLS 估计值),却不显著( $p$  值为 11.6%)。下面,进行更有效率的 MLE 估计。

```
. treatreg ww wa cit,treat(wc = wmed wfed) nolog
```

Treatment-effects model -- MLE				Number of obs = 250		
				Wald chi2(3) = 4.11		
				Prob > chi2 = 0.2501		
<hr/>						
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
ww	wa	-.0110424	.0199652	-0.55	0.580	-.0501735 .0280887
	cit	.127636	.3361938	0.38	0.704	-.5312917 .7865638
	wc	1.271327	.7412951	1.72	0.086	-,1815842 2.724239
	_cons	2.318638	.9397573	2.47	0.014	.4767478 4.160529
wc	wmed	.1198055	.0320056	3.74	0.000	.0570757 .1825352
	wfed	.0961886	.0290868	3.31	0.001	.0391795 .1531977
	_cons	-2.631876	.3309128	-7.95	0.000	-3.280453 -1.983299
	/athrho	.0178668	.1899898	0.09	0.925	-.3545063 .3902399
	/lnsigma	.9241584	.0447455	20.65	0.000	.8364588 1.011858
	rho	.0178649	.1899291			-.3403659 .371567
	sigma	2.519747	.1127473			2.308179 2.750707
	lambda	.0450149	.4786442			-.8931105 .9831404
<hr/>						
LR test of indep. eqns. (rho = 0): chi2(1) = 0.01 Prob > chi2 = 0.9251						

上表显示,MLE 的估计结果与两步法类似。然而,上表底部的似然比检验结果却没有拒绝原假设“ $H_0: \rho = 0$ ”( $p$  值高达 0.93)。从表面上看,这意味着不存在内生性,可直接进行 OLS 估计。然而,也可能存在其他模型设定误差,比如忽略了妻子是否进行劳动力市场的内生选择(对于 40% 的观测值,ww = 0,这意味着妻子无工作),或者忽略了高次项或互动项。

## 习 题

**28.1** 使用数据集 jtrain3.dta 估计参加就业培训(train)对 1978 年实际收入(re78)的处理效应。该数据集中的变量解释类似于本章正文使用的数据集 ldw\_exper.dta。

- (1) 把 re78 对 train 进行一元回归。
- (2) 加入控制变量进行多元回归。
- (3) 进行一对一的倾向得分匹配估计。
- (4) 进行  $k$  近邻倾向得分匹配,令  $k = 4$ 。
- (5) 进行卡尺内一对四倾向得分匹配。
- (6) 进行倾向得分核匹配。
- (7) 进行倾向得分线性回归匹配。
- (8) 进行马氏匹配。
- (9) 计算偏差校正的匹配估计量。

**28.2** 使用数据集 angrist.dta<sup>①</sup>,利用断点回归估计班级规模对阅读成绩的影响。详见 Angrist and Lavy (1999)。

① 该数据集的下载地址为 [http://www.ats.ucla.edu/stat/stata/examples/methods\\_matter/chapter9/angrist.dta](http://www.ats.ucla.edu/stat/stata/examples/methods_matter/chapter9/angrist.dta)。

# 第 29 章 空间计量经济学

## 29.1 地理学第一定律

许多经济数据都涉及一定的空间位置。比如,研究全国各省的国内生产总值、投资、贸易、研发等数据。本书此前各章很少关注各省经济之间的互动,通常假设各省的变量相互独立。但常识告诉我们,各省经济有着广泛的联系,而且距离越近的省份联系越密切。根据 Tobler(1970),“所有事物都与其他事物相关联,但较近的事物比较远的事物更关联”(Everything is related to everything else, but near things are more related than distant things)。这被称为“地理学第一定律”(First Law of Geography)。

事实上,各省之间的距离信息并不难获得,比如是否相邻,直线距离或运输距离;只是此前一直未利用这些信息。将各省的变量数据,再加上各省的位置信息,即可得到“空间数据”(spatial data 或 areal data)。所谓空间数据,就是在原来的横截面或面板数据上,加上横截面单位的位置信息(或相互距离)。研究如何处理空间数据的计量经济学分支,称为“空间计量经济学”(spatial econometrics)。空间计量经济学的最大特色在于充分考虑横截面单位之间的空间依赖性(spatial dependence)。更一般地,空间效应(spatial effects)包括空间依赖性与“空间异质性”(spatial heterogeneity)。由于标准的计量经济学也考虑横截面单位之间的异质性(比如异方差),故空间计量经济学的关注重点为空间依赖性。

空间计量经济学诞生于 20 世纪 70 年代。近年来,空间计量经济学蓬勃发展并进入主流,可归功于两方面。首先,由于 GIS(地理信息系统)的发展,空间数据或包含地理信息的数据(georeferenced data)日益增多。其次,在经济理论方面,人们越来越关注经济行为人之间的互动,而不仅仅停留于代表性厂商或个人。比如,在考察同伴效应(peer effect),相邻效应(neighborhood effect),溢出效应(spillover effect)或网络效应(network effect)时,都需要明确地考虑空间因素。克鲁格曼(Paul Krugman)也因为倡导新贸易理论与新经济地理(new economic geography)而获得 2008 年诺贝尔经济学奖(Krugman, 1991)。

有关空间计量经济学的专著包括 Cliff and Ord(1973, 1981), Anselin(1988), Arbia(2006), LeSage and Pace(2009), 以及沈体雁等(2010)的中文教材。Anselin(2010)提供了较新的文献综述。

## 29.2 空间权重矩阵

进行空间计量分析的前提是度量区域之间的空间距离。记来自  $n$  个区域的空间数据为  $\{x_i\}_{i=1}^n$ , 下标  $i$  表示区域  $i$ 。记区域  $i$  与区域  $j$  之间的距离为  $w_{ij}$ , 则可定义“空间权重矩阵”

(spatial weighting matrix)如下：

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix} \quad (29.1)$$

其中,主对角线上元素  $w_{11} = \cdots = w_{nn} = 0$  (同一区域的距离为0)。显然,空间权重矩阵  $W$  为对称矩阵。最常用的 距离函数为“相邻”(contiguity),即如果区域  $i$  与区域  $j$  有共同的边界,则  $w_{ij} = 1$ ;反之,则  $w_{ij} = 0$ 。比照(国际)象棋中不同棋子的行走路线,相邻关系可分为以下几种,参见图 29.1。

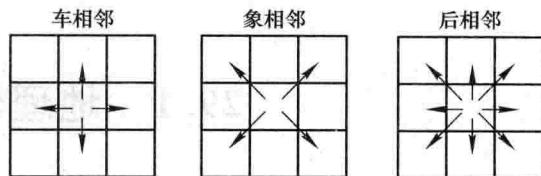


图 29.1 常用相邻关系

(1) 车相邻(rook contiguity):两个相邻的区域有共同的边。

(2) 象相邻(bishop contiguity):两个相邻的区域有共同的顶点,但没有共同的边。

(3) 后相邻(queen contiguity):两个相邻的区域有共同的边或顶点。

在实践中,为了区分“边”与“点”,常须设定一个最小距离,在此距离以下为点,而在此距离以上为边。究竟使用车、象或后相邻,取决于具体情况。比如,区域  $i$  与区域  $j$  仅在一点相交(象相邻),但有一条主要高速公路通过此点连接两区域,则不宜使用车相邻。

举一个简单例子,假设有如下四个区域,其变量取值分别为  $x = (x_1 \ x_2 \ x_3 \ x_4)'$ ,参见图 29.2。

图 29.2 假想的四个区域

针对图 29.2 中的四个区域,其空间权重矩阵为:

$$W = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \quad (29.2)$$

矩阵(29.2)的第一行表示,区域 1 与其余三个区域均相邻;第二行表示,区域 2 与区域 1、区域 3 相邻,但不与区域 4 相邻;以此类推。空间权重矩阵考虑的是一阶邻居,还可以考虑二阶邻居,即邻居的邻居,可用矩阵  $W^2$  来表示。需要注意的是,矩阵  $W^2$  的主对角线上元素一般不再为 0,这意味着邻居的邻居也包括自己。在实践中,有时对空间权重矩阵进行“行标准化”(row standardization),即将矩阵中的每个元素(记为  $\tilde{w}_{ij}$ )除以其所在行元素之和,以保证每行元素之和为 1:

$$\tilde{w}_{ij} \equiv \frac{\tilde{w}_{ij}}{\sum_j \tilde{w}_{ij}} \quad (29.3)$$

当然,如果区域  $i$  为孤岛,与其他区域均不相邻,则上式分母为 0,并不适用。此时,可将分母改为  $\max(1, \sum_j \tilde{w}_{ij})$ <sup>①</sup>。不包含孤岛的行标准化矩阵也称为“行随机矩阵”(row-stochastic matrix),因为它的所有元素均介于 0 与 1 之间,且每行元素之和为 1,在形式上与离散型概率分

① 样本中如果包含孤岛,可能导致程序计算困难,故在实践中一般去掉孤岛。比如,在研究美国各州的空间数据时,常去掉夏威夷州与阿拉斯加州。

布一样。将(29.2)式的空间权重矩阵行标准化可得(仍记为  $\mathbf{W}$ )：

$$\mathbf{W} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix} \quad (29.4)$$

行标准化的好处在于,如果将行标准化矩阵  $\mathbf{W}$  乘  $\mathbf{x}$ ,则可得到每个区域邻居的平均值。比如,在上例中:

$$\mathbf{Wx} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} (x_2 + x_3 + x_4)/3 \\ (x_1 + x_3)/2 \\ (x_1 + x_2 + x_4)/3 \\ (x_1 + x_3)/2 \end{pmatrix} \quad (29.5)$$

比如,区域 1 的邻居为区域 2,3 和 4,而上式右边第一行元素正好为  $(x_2 + x_3 + x_4)/3$ ,即区域 1 邻居的平均值;以此类推。比照时间序列中时间滞后(time lag)的概念,  $\mathbf{Wx}$  也被称为  $\mathbf{x}$  的“空间滞后”(spatial lag),即  $\mathbf{x}$  邻居的平均取值。

显然,行标准化之后的空间权重矩阵一般不再是对称矩阵,参见矩阵(29.4),这是它的缺陷之一。另外,由于每行元素之和均为 1,这意味着区域  $i$  所受其邻居的影响之和一定等于区域  $j$  所受其邻居的影响之和(任意  $i \neq j$ );此假定可能过强,这是行标准化的另一局限。

定义相邻关系的另一方法基于区域间的距离。记区域  $i$  与区域  $j$  的距离为  $d_{ij}$ ,可定义空间权重如下:

$$w_{ij} = \begin{cases} 1 & \text{若 } d_{ij} < d \\ 0 & \text{若 } d_{ij} \geq d \end{cases} \quad (29.6)$$

其中,  $d$  为事先给定的距离临界值。另外,也可以不用相邻关系,而直接以距离之倒数(inverse distance)作为空间权重:

$$w_{ij} = \frac{1}{d_{ij}} \quad (29.7)$$

在上式中,距离  $d_{ij}$  既可以是地理距离,比如直线距离或大圆距离(great circle distance);也可以是基于运输成本或旅行时间的经济距离;甚至社交网络中的距离。

**例** 林光平等(2005)使用基于地理相邻关系的简单权重矩阵  $\mathbf{W}$  来研究我国 28 个省市在 1978—2002 年期间实际人均 GDP 的收敛情况。但相邻地区经济上的相互关系并不完全相同。例如,河北省虽然在地理上与北京、天津、山西、内蒙古、山东、河南相邻,但很明显河北省与北京、天津的经济密切程度高于其他各省。为此,林光平等(2005)使用地区间人均 GDP 的差额作为测度地区间“经济距离”的指标,并引入经济空间权重矩阵  $\mathbf{W}^* = \mathbf{W} \times \mathbf{E}$ ,其中矩阵  $\mathbf{E}$  的主对角线元素均为 0,而非主对角线的  $(i,j)$  元素为  $E_{ij} = \frac{1}{|\bar{Y}_i - \bar{Y}_j|}$ ,  $\bar{Y}_i$  为地区  $i$  在样本期间的人均实际 GDP 平均值。结果发现,将经济距离引入空间权重矩阵能更好地拟合我国地区经济的发展状况。

## 29.3 空间自相关

在确定是否使用空间计量方法时,首先要考察数据是否存在空间依赖性。如果不存在,则使用标准的计量方法即可;如果存在,则可使用空间计量方法。比照时间序列(time series),空间数据有时也称为“空间序列”(spatial series)。这是因为,时间序列可视为在时间轴上分布的随机过程,而空间数据(序列)则为在空间分布的随机过程。时间序列的一个重要特性是可能存在自相关,特别是一阶自相关。而对于空间序列,自相关的情形则更为复杂;因为时间序列只可能在一个方向上相关(过去影响现在,但现在无法影响过去),而空间序列则可以在多个方向上相关,而且可以互相影响( $x_i$ 影响 $x_j$ ,而 $x_j$ 也影响 $x_i$ )。

“空间自相关”(spatial autocorrelation)可理解为位置相近的区域具有相似的变量取值。如果高值与高值聚集在一起,低值与低值聚集在一起,则为“正空间自相关”(positive spatial autocorrelation);反之,如果高值与低值相邻,则为“负空间自相关”(negative spatial autocorrelation);后者较少见。如果高值与低值完全随机地分布,则不存在空间自相关。

考虑空间序列 $\{x_i\}_{i=1}^n$ 。基于空间自相关的复杂性,文献中提出了一系列度量空间自相关的方法,其中最为流行的是“莫兰指数 $I$ ”(Moran's  $I$ )(Moran, 1950):

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{S^2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \quad (29.8)$$

其中, $S^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ 为样本方差, $w_{ij}$ 为空间权重矩阵的 $(i,j)$ 元素(用来度量区域 $i$ 与区域 $j$ 之间的距离),而 $\sum_{i=1}^n \sum_{j=1}^n w_{ij}$ 为所有空间权重之和。如果空间权重矩阵为行标准化,则 $\sum_{i=1}^n \sum_{j=1}^n w_{ij} = n$ 。此时,莫兰指数 $I$ 可写为:

$$I = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (29.9)$$

莫兰指数 $I$ 的取值一般介于-1到1之间,大于0表示正自相关,即高值与高值相邻、低值与低值相邻;小于0表示负自相关,即高值与低值相邻。一般来说,正自相关比负自相关更为常见。如果莫兰指数 $I$ 接近于0,则表明空间分布是随机的,不存在空间自相关。莫兰指数 $I$ 可视为观测值与其空间滞后(spatial lag)的相关系数。如果将观测值与其空间滞后画成散点图,称为“莫兰散点图”(Moran scatterplot),则莫兰指数 $I$ 就是该散点图回归线的斜率。

为了进行严格检验,须导出莫兰指数 $I$ 的渐近分布。考虑原假设“ $H_0: \text{Cov}(x_i, x_j) = 0, \forall i \neq j$ ”(即不存在空间自相关)。在此原假设下,可以证明莫兰指数 $I$ 的期望值为

$$E(I) = \frac{-1}{n-1} \quad (29.10)$$

莫兰指数 $I$ 的方差表达式更为复杂,记为 $\text{Var}(I)$ 。可以证明,标准化的莫兰指数 $I$ 服从渐近标准正态分布:

$$I^* \equiv \frac{I - E(I)}{\sqrt{\text{Var}(I)}} \xrightarrow{d} N(0, 1) \quad (29.11)$$

因此,可使用标准正态的临界值进行检验。在使用莫兰指数  $I$  检验空间自相关时,须注意两个问题。问题之一,莫兰指数  $I$  取决于空间矩阵  $W$ ,如果空间矩阵设定不正确,则可能导致错误的结果。问题之二,莫兰指数  $I$  的核心成分为  $(x_i - \bar{x})(x_j - \bar{x})$ ,其隐含假设是  $\{x_i\}_{i=1}^n$  的期望值为常数 (constant mean),不存在任何趋势 (trend)。如果存在趋势,则可能导致检验结果出现偏差。为了解决问题一,须仔细选择合适的空间矩阵,或使用不同的空间矩阵以考察结果的稳健性。为了解决问题二,可引入协变量,通过回归的方法去掉趋势,然后对残差项进行莫兰指数  $I$  检验。

以上的莫兰指数  $I$  也被称为“全局莫兰指数  $I'$ ”(global Moran's  $I$ ),因为它考察的是整个空间序列  $\{x_i\}_{i=1}^n$  的空间集聚情况。如果想知道某区域  $i$  附近的空间集聚情况,则可使用“局部莫兰指数  $I''$ ”(local Moran's  $I$ ):

$$I_i = \frac{(x_i - \bar{x})}{S^2} \sum_{j=1}^n w_{ij}(x_j - \bar{x}) \quad (29.12)$$

局部莫兰指数  $I$  的含义与全局莫兰指数  $I$  相似。正的  $I_i$  表示区域  $i$  的高(低)值被周围的高(低)值所包围;负的  $I_i$  则表示区域  $i$  的高(低)值被周围的低(高)值所包围。莫兰指数  $I$  并非唯一的空间自相关指标,另一常用指标为“吉尔里指数  $C$ ”(Geary's  $C$ ) (Geary, 1954),也称为“吉尔里相邻比率”(Geary's Contiguity Ratio):

$$C = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - x_j)^2}{2 \left( \sum_{i=1}^n \sum_{j=1}^n w_{ij} \right) \left[ \sum_{i=1}^n (x_i - \bar{x})^2 \right]} \quad (29.13)$$

与莫兰指数  $I$  不同,吉尔里指数  $C$  的核心成分为  $(x_i - x_j)^2$ 。吉尔里指数  $C$  的取值一般介于 0 到 2 之间(2 不是严格上界),大于 1 表示负相关,等于 1 表示不相关,而小于 1 表示正相关。因此,吉尔里指数  $C$  与莫兰指数  $I$  呈反向变动;一般认为,前者比后者对于局部空间自相关更为敏感。在不存在空间自相关的原假设下,可以证明吉尔里指数  $C$  的期望值为 1,而方差的表达式较复杂,记为  $\text{Var}(C)$ 。可以证明,标准化的吉尔里指数  $C$  服从渐近标准正态分布:

$$C^* = \frac{C - 1}{\sqrt{\text{Var}(C)}} \xrightarrow{d} N(0, 1) \quad (29.14)$$

因此,可使用标准化的吉尔里指数  $C^*$  检验空间自相关。然而,莫兰指数  $I$  与吉尔里指数  $C$  的共同缺点在于,即无法分别“热点”(hot spot)与“冷点”(cold spot)区域。所谓热点区域,即高值与高值聚集的区域;而冷点区域则是低值与低值聚集的区域。热点区域与冷点区域都表现为正自相关。为此,Getis and Ord(1992)提出了以下“Getis-Ord 指数  $G$ ”:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} x_i x_j}{\sum_{i=1}^n \sum_{j=1, j \neq i}^n x_i x_j} \quad (29.15)$$

其中,  $x_i > 0, \forall i$ ; 而  $w_{ij}$  来自非标准化的对称空间权重矩阵,且所有元素均为 0 或 1。显然,如果样本中高值聚集在一起,则  $G$  较大;如果低值聚集在一起,则  $G$  较小。在无空间自相关的原假设下,可以证明,  $E(G) = \frac{\sum_{i=1}^n \sum_{j=1, j \neq i}^n w_{ij}}{n(n-1)}$ 。如果  $G$  值大于此期望值,则表示存在热点区域;如果  $G$  值小于此期望值,则表示存在冷点区域。类似地,标准化的  $G$  也服从渐近标准正态分布:

$$G^* = \frac{G - E(G)}{\sqrt{\text{Var}(G)}} \xrightarrow{d} N(0, 1) \quad (29.16)$$

如果  $G^* > 1.96$ ,则可在 5% 水平上拒绝无空间自相关的原假设,认为存在空间正自相关,且

存在热点区域。反之,如果  $G^* < -1.96$ ,则可在 5% 水平上拒绝无空间自相关的原假设,认为存在空间正自相关,且存在冷点区域。如果要考察某区域  $i$  是否为热点或冷点,则可使用“局部 Getis-Ord 指数  $G^*$ ”:

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j} \quad (29.17)$$

需要指出的是,以上各种空间自相关指标仅提供是否存在空间效应的初步检验,深入检验有赖于建立正式的空间计量模型<sup>①</sup>。有关空间自相关的指标计算与假设检验可通过下载非官方 Stata 命令来实现(或通过命令“findit spatreg”查找并下载):

```
. net install sg162.pkg
```

此命令将下载安装以 `spat` 开头的系列命令:`spatwmat`(用于定义空间权重矩阵);`spatgsa`(进行全局空间自相关检验,“`gsa`”表示“global spatial autocorrelation”);`spatlsa`(进行局部空间自相关检验,“`lsa`”表示“local spatial autocorrelation”);`spatcorr`(考察空间自相关指标对距离临界值  $d$  的依赖性,参见方程(29.6));`spatdiag`(针对 OLS 回归结果,诊断是否存在空间效应);以及`spatreg`(估计空间滞后与空间误差模型,参见下两节)。

下面以数据集 `columbusdata.dta` 与 `columbuswm.dta` 为例。这两个数据集来自 Anselin(1988),前者包含美国俄亥俄州哥伦布市(Columbus, Ohio)49个社区的社区编号(`id`)、犯罪率(`crime`)、房价(`hoval`)与家庭收入(`income`)的数据,而后者为这49个社区基于相邻关系的空间权重矩阵;这些社区的分布参见图 29.3。本研究考察房价与家庭收入对犯罪率的作用。显然,各个社区的犯罪率是相关的,故应进行空间计量分析。

运行 `spat` 系列命令(除 `spatcorr` 外)的前提是通过命令 `spatwmat` 来定义空间权重矩阵。为此,将数据集 `columbuswm.dta` 与 `columbusdata.dta` 置于 Stata 的当前目录下。如果不知道当前目录在哪里,可通过输入命令 `pwd`(表示“path of working directory”)来显示当前目录;然后再输入以下命令:

```
. spatwmat using columbuswm.dta, name(W)
```

其中,必选项“`name(W)`”表示将根据数据集 `columbuswm.dta` 生成的空间权重矩阵命令为 `W`。更多选择项参见“`help spatwmat`”。

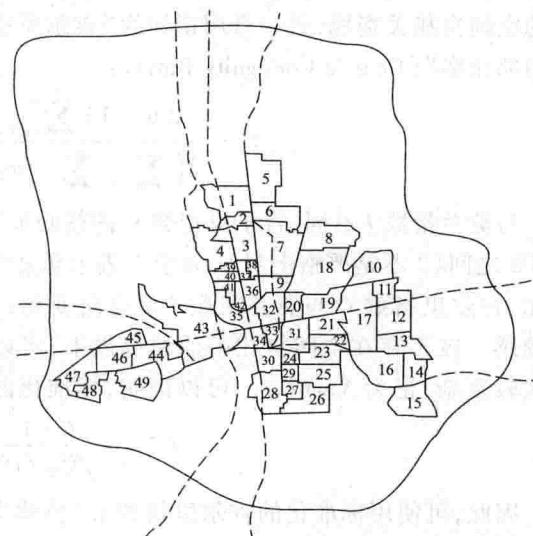


图 29.3 美国俄亥俄州哥伦布市的 49 个社区

```
The following matrix has been created:
```

```
1. Imported binary weights matrix W
Dimension: 49x49
```

<sup>①</sup> 空间相关系数与空间计量模型的关系正如简单相关系数与多元回归的关系,即简单相关系数仅提供初步证据,而更深入的分析则有赖于多元回归模型。

上表显示,已生成  $49 \times 49$  的空间权重矩阵  $W$ ,其中元素均为 0 或 1 (binary)。如果想看此矩阵,可输入命令“matrix list W”(为节省空间,运行结果从略)。

下面,计算被解释变量 crime 的全局自相关指标及相应检验。

```
. use columbusdata.dta, clear
. spatgsa crime, weights(W) moran geary go twotail
```

其中,必选项“weights(W)”指定空间权重矩阵为  $W$ 。选择项“moran geary go”分别表示计算莫兰指数  $I$ ,吉尔里指数  $C$  以及 Getis-Ord 指数  $G$ ;这三个选择项至少须选一。选择项“twotail”表示进行双边检验,默认为单边检验(即认为只可能存在正空间自相关)。

<b>Measures of global spatial autocorrelation</b>					
<b>Weights matrix</b>					
Name: W					
Type: Imported (binary)					
Row-standardized: No					
<b>Moran's I</b>					
Variables	I	E(I)	sd(I)	z	p-value*
crime	0.521	-0.021	0.087	6.212	0.000
<b>Geary's c</b>					
Variables	c	E(c)	sd(c)	z	p-value*
crime	0.584	1.000	0.109	-3.835	0.000
<b>Getis &amp; Ord's G</b>					
Variables	G	E(G)	sd(G)	z	p-value*
crime	0.126	0.099	0.006	4.714	0.000

\*2-tail test

从上表可知,这三个全局空间自相关指标均强烈拒绝“无空间自相关”的原假设,即认为存在空间自相关。下面计算局部空间自相关指标(为节省空间,仅计算局部莫兰指数  $I$ )。

```
. spatlsa crime, w(W) moran twotail
```

命令 spatlsa 的句型及选择项与 spatgsa 类似。

**Measures of local spatial autocorrelation**

Weights matrix

Name: W

Type: Imported (binary)

Row-standardized: No

Moran's Ii (Residential burglaries &amp; vehicle thefts pe)

Location	Ii	E(Ii)	sd(Ii)	z	p-value*
1	1.586	-0.063	1.674	0.985	0.325
2	0.019	-0.083	1.912	0.054	0.957
3	0.460	-0.125	2.289	0.255	0.798
4	-7.442	-0.083	1.912	-3.850	0.000
5	1.474	-0.042	1.381	1.097	0.273
6	0.375	-0.083	1.912	0.240	0.810
7	2.429	-0.167	2.581	1.006	0.315
8	-0.363	-0.042	1.381	-0.233	0.816
9	4.409	-0.125	2.289	1.981	0.048
10	0.161	-0.083	1.912	0.128	0.898
11	-0.324	-0.063	1.674	-0.156	0.876
12	2.703	-0.063	1.674	1.652	0.099
13	4.959	-0.083	1.912	2.638	0.008
14	2.495	-0.063	1.674	1.528	0.126
15	0.935	-0.042	1.381	0.707	0.479
16	4.846	-0.104	2.113	2.342	0.019
17	3.414	-0.208	2.815	1.287	0.198
18	0.195	-0.146	2.443	0.140	0.889
19	-0.024	-0.125	2.289	0.044	0.965
20	1.180	-0.104	2.113	0.607	0.544
21	-0.496	-0.083	1.912	-0.216	0.829
22	0.214	-0.083	1.912	0.156	0.876
23	-0.210	-0.146	2.443	-0.026	0.979
24	3.465	-0.125	2.289	1.569	0.117
25	-0.103	-0.104	2.113	0.000	1.000
26	1.090	-0.063	1.674	0.689	0.491
27	0.814	-0.083	1.912	0.470	0.639
28	0.272	-0.083	1.912	0.186	0.853
29	0.035	-0.125	2.289	0.070	0.944
30	2.539	-0.125	2.289	1.164	0.244
31	8.793	-0.188	2.704	3.321	0.001
32	11.771	-0.146	2.443	4.877	0.000
33	10.351	-0.125	2.289	4.577	0.000
34	8.276	-0.104	2.113	3.965	0.000
35	1.600	-0.146	2.443	0.714	0.475
36	9.910	-0.167	2.581	3.904	0.000
37	4.529	-0.146	2.443	1.913	0.056
38	7.290	-0.104	2.113	3.498	0.000
39	0.501	-0.083	1.912	0.306	0.760
40	3.692	-0.125	2.289	1.667	0.095
41	2.400	-0.063	1.674	1.471	0.141
42	2.763	-0.104	2.113	1.357	0.175
43	0.465	-0.063	1.674	0.315	0.753
44	2.114	-0.083	1.912	1.149	0.250
45	2.035	-0.042	1.381	1.503	0.133
46	6.216	-0.104	2.113	2.991	0.003
47	2.304	-0.042	1.381	1.698	0.089
48	2.499	-0.042	1.381	1.840	0.066
49	2.173	-0.042	1.381	1.604	0.109

\*2-tail test

上表分别列出了 49 个社区的莫兰指数  $I$  及检验结果。对于某些社区,可以强烈拒绝“无空间自相关”的原假设,这与全局空间自相关的检验结果相一致。