

第 9 章 模型设定与数据问题

如果模型设定(model specification)不当, 比如解释变量选择不当、测量误差、函数形式不妥等, 则会出现“设定误差”(specification error), 即模型本身的设定所带来的误差。

数据本身也可能存在问题, 比如多重共线性、对回归结果影响很大的极端数据等。

9.1 遗 漏 变 量

由于某些数据难以获得, 遗漏变量现象几乎难以避免。

假设真实的模型(true model)为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \quad (9.1)$$

其中，解释变量 x_1, x_2 与扰动项 ε 不相关。

而实际估计的模型(estimated model)为

$$y = \alpha + \beta x_1 + u \quad (9.2)$$

遗漏变量(omitted variable) x_2 被归入扰动项 $u = \gamma x_2 + \varepsilon$ 。

遗漏变量是否一定导致不一致的估计？

考虑两种情形：

(1) 遗漏变量 x_2 与解释变量 x_1 不相关，即 $\text{Cov}(x_1, x_2) = 0$ 。

扰动项 $u = \gamma x_2 + \varepsilon$ 与解释变量 x_1 不相关，因为

$$\text{Cov}(x_1, u) = \text{Cov}(x_1, \gamma x_2 + \varepsilon) = \gamma \text{Cov}(x_1, x_2) + \text{Cov}(x_1, \varepsilon) = 0 + 0 = 0$$

(9.3)

虽然存在遗漏变量，但 OLS 依然可一致地估计回归系数。

由于遗漏变量 x_2 被归入扰动项中，可能增大扰动项的方差，影响 OLS 估计的精确度。

(2) 遗漏变量 x_2 与解释变量 x_1 相关，即 $\text{Cov}(x_1, x_2) \neq 0$ 。

根据大样本理论，OLS 估计不一致，称为“遗漏变量偏差” (omitted variable bias)。

这种偏差在实践中较常见，成为某些实证研究的致命伤。

比如，研究教育投资回报时，个人能力因无法观测而遗漏，但能力与教育年限正相关。

存在遗漏变量本身并不要紧；关键在于，遗漏变量不能与方程的解释变量相关。

解决遗漏变量偏差的方法主要有：

- (i) 加入尽可能多的控制变量(control variable);
- (ii) 随机实验与自然实验;
- (iii) 工具变量法(第 10 章);
- (iv) 使用面板数据(第 12 章);

第(i)种方法“加入尽可能多的控制变量”着眼于直接解决遗漏变量问题。

首先从理论出发，列出所有可能对被解释变量有影响的变量，然后尽可能去收集数据。

如果有些相关变量无法获得，则需从理论上说明，遗漏变量不会与解释变量相关，或相关性很弱。

例 李宏彬等(2012)通过就业调查数据，研究“官二代”大学毕业生的起薪是否高于非官二代。

由于可能存在遗漏变量，该文包括了尽可能多的控制变量，比如年龄、性别、城镇户口、父母收入、父母学历、高考成绩、大学成绩、文理科、党员、学生会干部、兼职实习经历、拥有技术等级证书等。

解决遗漏变量偏差的第(ii)种方法为随机实验或自然实验。

物理学常使用“控制实验”(controlled experiment)来研究 x 对 y 的因果关系，即给定影响 y 的所有其他因素，单独让 x 变化，然后观察 y 如何变化。

控制实验的方法在其他学科未必可行。

比如，医学对新药 x 疗效的实验。

由于参加实验者的体质与生活方式不同，不可能完全控制所有其他因素，故无法进行严格的控制实验。

统计学之父费舍尔(Ronald Fischer)提出了随机(控制)实验(randomized controlled experiment)的概念。

考虑回归模型：

$$y = \alpha + \beta x + \varepsilon \quad (9.4)$$

x 完全随机地决定(比如，通过抛硬币或电脑随机数)，故 x 与任何其他变量都相互独立。

由于 x 独立于 ε ，故 $\text{Cov}(x, \varepsilon) = 0$ ，无论遗漏多少解释变量，OLS 都一致。

如果 x 是取值为 0 或 1 的虚拟变量，可将样本数据分为两组。

称 “ $x=1$ ” 的组为 “实验组” 或 “处理组” (treatment group)，比如医学实验中吃药的那组。

称 “ $x=0$ ” 的组为 “控制组” (control group) 或 “对照组”，比如

医学实验中不吃药的那组。

随机实验的核心思想是，将实验人群(或个体)随机分为两组，即实验组与控制组，则这两组除了 x 不同外，在所有其他方面都没有系统差别。

例 农学中将地块随机分成三组(很难找到土壤条件完全一样的地块)，分别给予不同的施肥量，然后考察施肥的效果。

例 班级规模是否影响学习成绩？由于遗漏变量的存在，观测数据很难回答此问题。比如，规模较小的班级可能位于好学区，师资好，家庭也富有。

美国田纳西州进行了为期四年的随机实验(Project STAR，即 Student-Teacher Achievement Ratio)。

将幼儿园至小学三年级的学生随机分为三组。

第一组为普通班，每班 22-25 名学生；

第二组为小班，每班 13-17 名学生；

第三组也为小班，但配备一名教学助理(teacher's aide)。

教师也随机分到这三类班级。

结果发现，班级规模对学习成绩的影响在统计上显著，但在经济上并不显著(即此效应本身比较小，普通班与小班的成绩差距类似于男生与女生的成绩差距)。

随机实验说服力强，但通常成本高。

另一实验方法为“自然实验”(natural experiment)或“准实验”(quasi experiment),即由于某些并非为了实验目的而发生的外部突发事件,使得当事人仿佛被随机分在实验组或控制组。

例 最低工资对就业的影响。提高法定最低工资(minimum wage)在多大程度上会影响对低技能工人的需求?

Card and Krueger (1994)考虑一个自然实验。

1992 年,新泽西州通过法律将最低工资从每小时\$4.25 提高到\$5.05,但相邻的宾夕法尼亚州最低工资保持不变。

这两个州的雇主仿佛被随机分配到实验组(新泽西州)与控制组

(宾夕法尼亚州)。

Card and Kruger (1994)收集两个州的快餐店在实施新法前后雇佣人数的数据；结果发现，提高最低工资对低技能工人的就业几乎没有影响。

例(一个失败的例子) 京杭大运河流经省份的人均 GDP 平均而言高于其他省份。这是否可以归功于京杭大运河对区域经济增长的促进作用？

隋炀帝确定京杭大运河的位置时，他是在地图上随机画了一条线吗？

由于影响被解释变量的因素往往很多，而局限于数据的可得性(availability)，故在任何实证研究中几乎总存在遗漏变量。

一篇专业水准的实证论文几乎总要说明，如何在存在遗漏变量的情况下避免遗漏变量偏差。

9.2 无 关 变 量

与遗漏变量相反的情形是，在回归方程中加入了与被解释变量无关的变量。假设真实的模型为

$$y = \alpha + \beta x_1 + \varepsilon \quad (9.5)$$

其中， $\text{Cov}(x_1, \varepsilon) = 0$ 。而实际估计的模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + (\varepsilon - \gamma x_2) \quad (9.6)$$

由于真实参数 $\gamma = 0$ (x_2 对 y 无影响)，故可将模型写为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \quad (9.7)$$

由于 x_2 与 y 无关，根据“无关变量”的定义， x_2 也与 y 的扰动项 ε 无关，即 $\text{Cov}(x_2, \varepsilon) = 0$ 。

扰动项 ε 与所有解释变量均无关，故 OLS 一致，即 $\text{plim}_{n \rightarrow \infty} \hat{\beta} = \beta$ ，
 $\text{plim}_{n \rightarrow \infty} \hat{\gamma} = \gamma = 0$ 。

引入无关变量后，受到无关变量的干扰，估计量 $\hat{\beta}$ 的方差一般会增大。

对于解释变量的选择最好遵循经济理论的指导。

9.3 建模策略：“由小到大”还是“由大到小”

“由小到大” (specific to general) 的建模方式首先从最简单的小模型开始，逐渐增加解释变量。

比如，先将被解释变量 y 对关键解释变量 x 回归，然后再加入其他控制变量 z 。

但小模型很可能存在遗漏变量偏差，系数估计不一致， t 检验、 F 检验都失效，很难确定如何取舍变量。

“由大到小” (general to specific) 的建模方式从尽可能大的模型开始，收集所有可能的解释变量，再逐步剔除不显著的解释变量(可依次剔除最不显著，即 p 值最大的变量)。

虽然冒着包含无关变量的危险，但危害性没有遗漏变量严重。

但在实际操作上，很难找到与被解释变量相关的所有解释变量。

实证研究中，常采用以上两种策略的折中方案。

9.4 解释变量个数的选择

好的经济理论使用简洁的模型来很好地描述复杂的经济现实。

但这两个目标常常矛盾。

加入过多的解释变量可提高模型的解释力(比如增大拟合优度 R^2)，但也牺牲了模型的简洁性(parsimony)。

需在模型的解释力与简洁性之间找到最佳平衡。

在时间序列模型里，常需选择解释变量滞后的期数(比如，确定自回归模型的阶数)。

更一般地，要确定解释变量的个数。

可供选择的权衡标准如下。

(1) 校正可决系数 \bar{R}^2 ：选择解释变量的个数 K 以最大化 \bar{R}^2 。

(2) “赤池信息准则”(Akaike Information Criterion, AIC)：选择解释变量的个数 K ，使得目标函数最小化：

$$\min_K \text{AIC} \equiv \ln \left(\frac{\text{SSR}}{n} \right) + \frac{2}{n} K \quad (9.8)$$

其中，SSR 为残差平方和 $\sum_{i=1}^n e_i^2$ 。

第一项为对模型拟合度的奖励(减少残差平方和 SSR)。

第二项为对解释变量过多的惩罚(为解释变量个数 K 的增函数)。

当 K 上升时，第一项下降而第二项上升。

(3) “贝叶斯信息准则” (Bayesian Information Criterion, BIC)或“施瓦茨信息准则” (Schwarz Information Criterion, SIC 或 SBIC): 选择解释变量的个数 K , 使得目标函数最小化:

$$\min_K \text{BIC} \equiv \ln \left(\frac{\text{SSR}}{n} \right) + \frac{\ln n}{n} K \quad (9.9)$$

BIC 准则与 AIC 准则仅第二项有差别。

一般来说, $\ln n > 2$, 故 BIC 准则对于解释变量过多的惩罚比 AIC 准则更为严厉。

BIC 准则更强调模型的简洁性。

在时间序列模型中, 常用信息准则确定滞后阶数。

考虑 p 阶自回归模型,

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t \quad (9.10)$$

滞后阶数 p 可通过信息准则来确定。

根据 BIC 准则计算的 \hat{p} 是真实滞后阶数 p 的一致估计量。

但根据 AIC 计算的 \hat{p} 却不一致，即使在大样本中也可能高估 p 。

由于现实样本通常有限，而 BIC 准则可能导致模型过小(对解释变量过多的惩罚太严厉)，故 AIC 准则依然很常用。

在 Stata 中作完回归后，计算信息准则的命令为

```
estat ic
```

其中，“ic”表示 information criterion(信息准则)。

(4) “由大到小的序贯 t 规则” (general-to-specific sequential t rule)。

这种方法常用于时间序列模型，比如 $AR(p)$ 。

首先，设最大滞后期 p_{\max} ，令 $\hat{p} = p_{\max}$ 进行估计，并对最后一阶系数的显著性进行 t 检验。

如果接受该系数为 0，则令 $\hat{p} = p_{\max} - 1$ ，重新进行估计，再对(新的)最后一阶系数的显著性进行 t 检验，如果显著，则停止；否则，令 $\hat{p} = p_{\max} - 2$ ；以此类推。

以数据集 icecream.dta 为例，考虑应引入气温(temp)的几阶滞后项。

首先，使用信息准则。

```
. use icecream.dta,clear  
  
. quietly reg consumption temp price income  
  
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	30	39.57876	58.61944	4	-109.2389	-103.6341

加入气温的一阶滞后项(L.temp), 重新估计。

```
. qui reg consumption temp L.temp price income  
  
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	29	37.85248	63.41576	5	-116.8315	-109.995

增加解释变量 L.temp 后, AIC 与 BIC 都下降了。

引入气温的二阶滞后项(L2.temp):

```
. qui reg consumption temp L.temp L2.temp price  
income
```

```
. estat ic
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	28	36.08382	61.12451	6	-110.249	-102.2558

加入气温的二阶滞后项后，AIC 与 BIC 比仅包括气温的滞后项上升了。

仅包含气温的滞后项可达到 AIC 与 BIC 的最小值。

其次，使用序贯 t 规则，并假设 $p_{\max} = 2$ ：

```
. reg consumption temp L.temp L2.temp price
income
```

Source	SS	df	MS	Number of obs = 28		
Model	.103722201	5	.02074444	F(5, 22) =	21.92	
Residual	.020822754	22	.000946489	Prob > F =	0.0000	
Total	.124544954	27	.004612776	R-squared =	0.8328	
				Adj R-squared =	0.7948	
				Root MSE =	.03077	
consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
temp						
--.	.0047858	.0013502	3.54	0.002	.0019856	.007586
L1.	-.0010836	.0022905	-0.47	0.641	-.0058338	.0036666
L2.	-.0008022	.0013414	-0.60	0.556	-.0035841	.0019797
price	-.7326035	.7214324	-1.02	0.321	-2.228763	.7635558
income	.0026704	.0011308	2.36	0.027	.0003252	.0050156
_cons	.1883478	.23949	0.79	0.440	-.3083241	.6850196

L2.temp 的系数高度不显著(p 值为 0.556)。

令 $\hat{p} = p_{\max} - 1 = 1$ (去掉 L2.temp), 重新估计。

```
. reg consumption temp L.temp price income
```

Source	SS	df	MS	Number of obs = 29		
Model	.103387183	4	.025846796	F(4, 24) =	28.98	
Residual	.021406049	24	.000891919	Prob > F =	0.0000	
Total	.124793232	28	.004456901	R-squared =	0.8285	
				Adj R-squared =	0.7999	
				Root MSE =	.02987	
consumption	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
temp						
--.	.0053321	.0006704	7.95	0.000	.0039484	.0067158
L1.	-.0022039	.0007307	-3.02	0.006	-.0037119	-.0006959
price	-.8383021	.6880205	-1.22	0.235	-2.258307	.5817025
income	.0028673	.0010533	2.72	0.012	.0006934	.0050413
_cons	.1894822	.2323169	0.82	0.423	-.2899963	.6689607

L.temp 的系数在 1%水平上显著(p 值为 0.006), 故最终选择 $\hat{p}=1$; 此结果与信息准则的结果相同。

9.5 对函数形式的检验

很多经济关系是非线性的。多元线性回归可看作是非线性关系的一阶线性近似。

如果存在非线性项, 但被遗漏, 则会导致遗漏变量偏差, 这是模型设定误差的一种形式。

假设真实模型为

$$y = \alpha + \beta x + (\gamma x^2 + \varepsilon) \quad (9.11)$$

其中, $\text{Cov}(x, \varepsilon) = 0$, 而平方项 γx^2 被遗漏。

解释变量与扰动项相关：

$$\text{Cov}(x, \gamma x^2 + \varepsilon) = \gamma \text{Cov}(x, x^2) + \text{Cov}(x, \varepsilon) = \gamma \text{Cov}(x, x^2) \neq 0 \quad (9.12)$$

遗漏高次项会导致遗漏变量偏差。

“Ramsey’s RESET 检验”(Regression Equation Specification Error Test)(Ramsey, 1969)的基本思想：如果怀疑遗漏非线性项，就把非线性项引入方程，检验其系数是否显著。

假设线性回归模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \quad (9.13)$$

记此回归的拟合值为

$$\hat{y} = \hat{\alpha} + \hat{\beta}x_1 + \hat{\gamma}x_2 \quad (9.14)$$

既然 \hat{y} 是解释变量的线性组合, \hat{y}^2 就包含了解释变量二次项(含平方项与交叉项)的信息, \hat{y}^3 就包含了解释变量三次项的信息, 以此类推。

考虑辅助回归:

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 \hat{y}^2 + \delta_3 \hat{y}^3 + \delta_4 \hat{y}^4 + \varepsilon \quad (9.15)$$

对 $H_0: \delta_2 = \delta_3 = \delta_4 = 0$ 作 F 检验。

如果拒绝 H_0 , 说明模型中应有高次项;

如果接受 H_0 , 则可使用线性模型。

RESET 检验的缺点是，在拒绝 H_0 的情况下，并不提供具体遗漏哪些高次项的信息。

也可直接将解释变量 x_1 与 x_2 的高次项放入辅助回归中，比如

$$y = \alpha + \beta x_1 + \gamma x_2 + \delta_2 x_1^2 + \delta_3 x_2^2 + \delta_4 x_1 x_2 + \varepsilon \quad (9.16)$$

然后检验 $H_0 : \delta_2 = \delta_3 = \delta_4 = 0$ 。

关于如何确定回归方程的函数形式，最好从经济理论出发。

在缺乏理论指导的情况下，可先从线性模型出发，然后进行 RESET 检验，看是否应加入非线性项。

在 Stata 中作完回归，进行 RESET 检验的命令为

```
estat ovtest,rhs
```

其中，“ovtest”表示 omitted variable test，因为遗漏高次项的后果类似于遗漏解释变量。

选择项“rhs”表示使用解释变量的幂为非线性项，即方程(9.16)；默认使用 $\hat{y}^2, \hat{y}^3, \hat{y}^4$ 为非线性项，即方程(9.15)。

以数据集 grilic.dta 为例。

首先，并进行线性 OLS 回归。

```
. use grilic.dta,clear
```

```
. qui reg lnw s expr tenure smsa rns
```

使用拟合值的高次项进行 RESET 检验。

```
. estat ovtest
```

```
Ramsey RESET test using powers of the fitted values of lnw
Ho:  model has no omitted variables
      F(3, 749) =      1.51
      Prob > F =      0.2114
```

p 值为 0.2114，可接受原假设，未发现遗漏高次项。

直接使用解释变量的高次项进行 RESET 检验。

```
. estat ovtest,rhs
```



```
Ramsey RESET test using powers of the independent variables
Ho: model has no omitted variables
      F(9, 743) =      2.03
      Prob > F =      0.0336
```

可在 5% 水平上拒绝原假设，认为遗漏了高阶非线性项。

工资对数与工龄(expr)的关系可能存在非线性。

引入工龄的平方项，记为 expr2，再进行回归。

```
. gen expr2=expr^2
```

```
. reg lnw s expr expr2 tenure smsa rns
```

Source	SS	df	MS	Number of obs = 758		
Model	49.7985818	6	8.29976364	F(6, 751) = 69.65		
Residual	89.487568	751	.11915788	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.3575		
				Adj R-squared = 0.3524		
				Root MSE = .34519		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.1029839	.0058299	17.66	0.000	.0915391	.1144287
expr	.0018351	.0157708	0.12	0.907	-.0291249	.0327951
expr2	.0051819	.0020645	2.51	0.012	.001129	.0092348
tenure	.037085	.0077374	4.79	0.000	.0218954	.0522746
smsa	.1419045	.0279979	5.07	0.000	.086941	.1968679
rns	-.0843448	.0286965	-2.94	0.003	-.1406797	-.0280098
_cons	4.119327	.0850277	48.45	0.000	3.952407	4.286247

工龄平方(expr2)在 1%显著为正，但工龄本身(expr)变得很不显著；因为二者存在多重共线性(参见下一节)。

使用解释变量的高次项进行 RESET 检验:

```
. estat ovtest,rhs
```

```
(note:  expr2 dropped because of collinearity)
(note:  expr2^2 dropped because of collinearity)

Ramsey RESET test using powers of the independent variables
      Ho:  model has no omitted variables
           F(11, 741) =      1.73
           Prob > F =      0.0626
```

可以在 5% 水平上，接受“无遗漏变量”的原假设。

9.6 多重共线性

如果在解释变量中，某一解释变量可由其他解释变量线性表出，则存在“严格多重共线性”(strict multicollinearity)。

此时数据矩阵 \mathbf{X} 不满列秩, $(\mathbf{X}'\mathbf{X})^{-1}$ 不存在, 无法定义 OLS 估计量 $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ 。

比如, 解释变量 x_2 正好是解释变量 x_3 的两倍, 则无法区分 x_2 与 x_3 对被解释变量 y 的影响。

严格多重共线性是对数据的最低要求, 现实中较少出现。

在实践中更为常见的为近似(非严格)的多重共线性, 简称“多重共线性”(multicollinearity)或“共线性”。

多重共线性的主要表现是, 如果将第 k 个解释变量 x_k 对其余解释变量 $\{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_K\}$ 进行回归, 所得可决系数(记为 R_k^2)较高。

在多重共线性的情况下，OLS 仍是 BLUE，因为高斯-马尔可夫定理并未排除多重共线性的情形。

但 BLUE 只保证 OLS 估计量在所有线性无偏估计量中相对而言方差最小，并不意味着 OLS 估计量的方差在绝对意义上小。

如果存在严格多重共线性，则矩阵 $(X'X)$ 不可逆。

在多重共线性的情况下，矩阵 $(X'X)$ 变得“几乎不可逆”， $(X'X)^{-1}$ 变得很“大”，致使方差 $\text{Var}(\hat{\beta} | X) = \sigma^2 (X'X)^{-1}$ 增大，系数估计变得不准确。

X 中元素轻微变化就会引起 $(X'X)^{-1}$ 很大变化，导致 OLS 估计值 $\hat{\beta}$ 发生很大变化。

多重共线性的通常症状是，虽然整个回归方程的 R^2 较大、 F 检验也很显著，但单个系数的 t 检验却不显著。

另一症状是，增减解释变量使得系数估计值发生较大变化(比如，加入的解释变量与已有解释变量构成多重共线性)。

如果两个(或多个)解释变量之间高度相关，则不易区分它们对被解释变量的单独影响力。

在严格多重共线性的极端情况下，一个变量刚好是其他变量的倍数，则完全无法区分。

R_k^2 越高，解释变量 x_k 与其他解释变量的共线性越严重，则 x_k 的系数估计量 $\hat{\beta}_k$ 的方差越大。可以证明：

$$\text{Var}(\hat{\beta}_k | \mathbf{X}) = \frac{\sigma^2}{(1 - R_k^2)S_k} \quad (9.17)$$

其中， $\sigma^2 \equiv \text{Var}(\varepsilon)$ 为扰动项的方差。

$S_k \equiv \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2$ 为 x_k 的离差平方和，反映 x_k 的变动幅度。

如果 x_k 变动很少，很难准确地估计 x_k 对 y 的作用。

在极端情况下， x_k 完全不变(为常数)， $S_k = 0$ ，则完全无法估计 $\hat{\beta}_k$ (x_k 与常数项构成严格多重共线性)。

方差 $\text{Var}(\hat{\beta}_k | \mathbf{X})$ 与 $(1 - R_k^2)$ 成反比。

定义解释变量 x_k 的“方差膨胀因子” (Variance Inflation Factor, VIF) 为

$$\text{VIF}_k \equiv \frac{1}{1 - R_k^2} \quad (9.18)$$

可将方差写为

$$\text{Var}(\hat{\beta}_k | \mathbf{X}) = \text{VIF}_k \cdot \frac{\sigma^2}{S_k} \quad (9.19)$$

方差膨胀因子 VIF_k 越大, x_k 的多重共线性问题越严重, 其方差 $\text{Var}(\hat{\beta}_k | \mathbf{X})$ 将变得越大。

对于 K 个解释变量 $\{x_1, \dots, x_K\}$ ，可计算相应的方差膨胀因子 $\{\text{VIF}_1, \dots, \text{VIF}_K\}$ 。

判断是否存在多重共线性的经验规则是， $\{\text{VIF}_1, \dots, \text{VIF}_K\}$ 的最大值不应超过 10。

求解 “ $10 = \frac{1}{1 - R_k^2}$ ” 可知，相应的 R_k^2 不应超过 0.9。

解释变量 x_k 的多重共线性越严重， R_k^2 越接近于 1，方差膨胀因子 VIF_k 将急剧上升。

在 Stata 中画函数(9.18)考察 VIF_k 对 R_k^2 的依赖性:

```
. twoway function VIF=1/(1-x),xtitle(R2)
xline(0.9,lp(dash)) yline(10,lp(dash))
xlabel(0.1(0.1)1) ylabel(10 100 200 300)
```

其中，选择项 “xtitle(R2)” 指示横轴的标题为 R2；
“xline(0.9,lp(dash))” 与 “yline(10,lp(dash))” 分别表示在横轴 0.9 与纵轴 10 的位置画一条虚线；
“xlabel(0.1(0.1)1)” 表示在横轴上，从 0.1 至 1，每隔 0.1 的位置给出标签；“ylabel(10 100 200 300)” 表示在纵轴上 10、100、200 与 300 的位置给出标签，参见图 9.1。

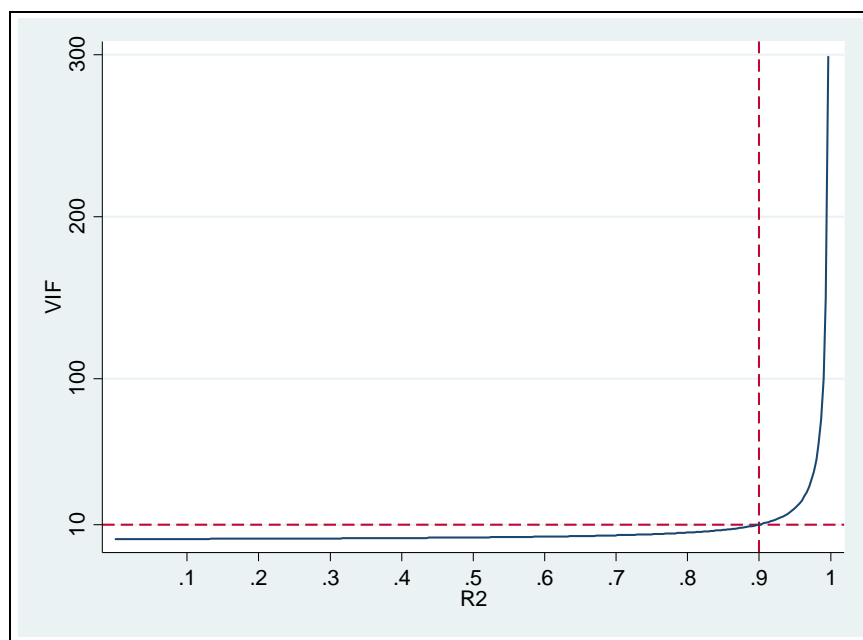


图 9.1 VIF_k 与 R_k^2 的关系

如发现多重共线性，可采取以下处理方法。

(1) 如不关心具体的回归系数，只关心整个方程的预测能力，可不必理会多重共线性(假设整个方程显著)。多重共线性的主要后果

是使得对单个变量的贡献估计不准，但所有变量的整体效应仍可较准确地估计。

(2) 如关心具体的回归系数，但多重共线性并不影响所关心变量的显著性，也可不必理会。

在方差膨胀的情况下，系数依然显著；如没有多重共线性，只会更显著。

(3) 如多重共线性影响所关心变量的显著性，应设法进行处理。

比如，增大样本容量，剔除导致严重共线性的变量，将变量标准化(详见下文)，或对模型设定进行修改。

解释变量之间的相关性普遍存在，在一定程度上也是允许的。

处理多重共线性的最常见方法是“无为而治” (do nothing)。

在 Stata 中作完回归后，可使用如下命令计算 VIF。

```
estat vif
```

以数据集 `grilic.dta` 为例。

首先，考察线性回归的 VIF。

```
. use grilic.dta,clear
```

```
. qui reg lnw s expr tenure iq smsa rns
```

```
. estat vif
```

Variable	VIF	1/VIF
expr	1.12	0.893267
s	1.07	0.930295
tenure	1.06	0.944083
smsa	1.04	0.964256
rns	1.03	0.970508
Mean VIF	1.06	

最大的 VIF 为 1.12，远小于 10，不必担心多重共线性。

如在模型中引入解释变量的平方项，则易引起多重共线性，因为 x 与 x^2 较相关。

在回归中加入教育年限(s)的平方项，记为 s2，再进行多重共线性检验。

```
. gen s2=s^2
. reg lnw s s2 expr tenure smsa rns
```

Source	SS	df	MS	Number of obs = 758		
Model	49.1549871	6	8.19249785	F(6, 751) = 68.26		
Residual	90.1311627	751	.120014864	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.3529		
				Adj R-squared = 0.3477		
				Root MSE = .34643		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.0339768	.0729216	0.47	0.641	-.1091776	.1771312
s2	.0024636	.0026079	0.94	0.345	-.002656	.0075832
expr	.0375502	.0063558	5.91	0.000	.0250729	.0500275
tenure	.0356461	.007743	4.60	0.000	.0204456	.0508466
smsa	.1390585	.0280915	4.95	0.000	.0839113	.1942058
rns	-.0864204	.0289057	-2.99	0.003	-.143166	-.0296747
_cons	4.57118	.5021415	9.10	0.000	3.585412	5.556948

教育年限(s)与其平方项(s2)都很不显著。二者之间可能存在多重共线性。

计算各变量的 VIF 值。

```
. estat vif
```

Variable	VIF	1/VIF
s	167.07	0.005986
s2	166.30	0.006013
expr	1.13	0.885254
tenure	1.06	0.944065
rns	1.04	0.963378
smsa	1.04	0.963750
Mean VIF	56.27	

变量 s 与 s2 的 VIF 分别达到 167.07 与 166.30，远大于 10，存在多重共线性。

将 s2 对 s 进行回归。

```
. reg s2 s
```


Source	SS	df	MS	Number of obs = 758		
				F(1, 756) = .		
Model	2916802.81	1	2916802.81	Prob > F = 0.0000		
Residual	17941.0733	756	23.7315785	R-squared = 0.9939		
				Adj R-squared = 0.9939		
Total	2934743.89	757	3876.8083	Root MSE = 4.8715		
s2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	27.81281	.0793331	350.58	0.000	27.65707	27.96854
_cons	-188.1622	1.078081	-174.53	0.000	-190.2785	-186.0458

R^2 高达 0.9939，说明 s 与 s2 所包含信息基本相同，导致严重的多重共线性。

如回归方程中包含解释变量的多项式(比如， $\beta x + \gamma x^2$)，通常导致多重共线性。

可能的解决方法是将变量标准化，即减去均值，除以标准差：

$$\tilde{x} \equiv \frac{x - \bar{x}}{s_x} \quad (9.20)$$

\bar{x} 为变量 x 的样本均值， s_x 为样本标准差， \tilde{x} 为标准化之后的变量；然后，以 \tilde{x} 及其平方 \tilde{x}^2 作为解释变量。

继续上例，先计算变量 s 的均值与标准差，将其标准化，并记标准化变量为 sd ：

```
. sum s
```

Variable	Obs	Mean	Std. Dev.	Min	Max
s	758	13.40501	2.231828	9	18

```
. gen sd=(s-r(mean))/r(sd)
```

```
. gen sd2=sd^2
```

```
. reg lnw sd sd2 expr tenure smsa rns
```

Source	SS	df	MS	Number of obs = 758		
Model	49.154987	6	8.19249783	F(6, 751) = 68.26		
Residual	90.1311629	751	.120014864	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.3529		
				Adj R-squared = 0.3477		
				Root MSE = .34643		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sd	.2232421	.014444	15.46	0.000	.1948866	.2515975
sd2	.0122714	.0129899	0.94	0.345	-.0132294	.0377723
expr	.0375502	.0063558	5.91	0.000	.0250729	.0500275
tenure	.0356461	.007743	4.60	0.000	.0204456	.0508466
smsa	.1390585	.0280915	4.95	0.000	.0839113	.1942058
rns	-.0864204	.0289057	-2.99	0.003	-.143166	-.0296747
_cons	5.469338	.0319719	171.07	0.000	5.406574	5.532103

标准化的线性项(sd)在 1%水平上显著为正，而标准化的平方项(sd2)不显著；多重共线性似乎有所缓解。

计算方差膨胀因子。

```
. estat vif
```

Variable	VIF	1/VIF
sd	1.32	0.759911
sd2	1.23	0.811990
expr	1.13	0.885254
tenure	1.06	0.944065
rns	1.04	0.963378
smsa	1.04	0.963750
Mean VIF	1.14	

VIF 的最大值仅为 1.32，基本不存在多重共线性。

将 sd2 对 sd 进行回归：

```
. reg sd2 sd
```

Source	SS	df	MS	Number of obs = 758		
Model	152.821515	1	152.821515	F(1, 756) = 159.77		
Residual	723.111439	756	.956496612	Prob > F = 0.0000		
				R-squared = 0.1745		
				Adj R-squared = 0.1734		
Total	875.932954	757	1.1571109	Root MSE = .97801		
sd2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sd	.4493082	.0355462	12.64	0.000	.3795271	.5190892
_cons	.9986808	.0355228	28.11	0.000	.9289457	1.068416

R^2 仅为 0.1745。

由于 sd2 在上面的回归中不显著，去掉 sd2，再次回归：

```
. reg lnw sd expr tenure smsa rns
```

Source	SS	df	MS	Number of obs = 758		
Model	49.0478812	5	9.80957624	F(5, 752) = 81.75		
Residual	90.2382686	752	.119997698	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.3521		
				Adj R-squared = 0.3478		
				Root MSE = .34641		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sd	.2290816	.0130535	17.55	0.000	.2034559	.2547073
expr	.0381189	.0063268	6.02	0.000	.0256986	.0505392
tenure	.0356146	.0077424	4.60	0.000	.0204153	.0508138
smsa	.1396666	.0280821	4.97	0.000	.0845379	.1947954
rns	-.0840797	.0287973	-2.92	0.004	-.1406124	-.0275471
_cons	5.479606	.0300656	182.26	0.000	5.420584	5.538629

sd 的回归系数为 0.2291，似乎偏高。

但 sd 为标准化的变量，故 sd 变化一单位，等于 s 变化一个标准差，即 2.231828 年。

以此推算 s 的系数，即教育投资的年回报率应为

```
. dis .2290816/2.231828
```

```
.10264304
```

再次对比未将变量 s 标准化的回归：

```
. reg lnw s expr tenure smsa rns
```

Source	SS	df	MS	Number of obs = 758		
Model	49.0478814	5	9.80957628	F(5, 752) = 81.75		
Residual	90.2382684	752	.119997697	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.3521		
				Adj R-squared = 0.3478		
				Root MSE = .34641		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.102643	.0058488	17.55	0.000	.0911611	.114125
expr	.0381189	.0063268	6.02	0.000	.0256986	.0505392
tenure	.0356146	.0077424	4.60	0.000	.0204153	.0508138
smsa	.1396666	.0280821	4.97	0.000	.0845379	.1947954
rns	-.0840797	.0287973	-2.92	0.004	-.1406124	-.0275471
_cons	4.103675	.085097	48.22	0.000	3.936619	4.270731

是否将变量 s 标准化, 对于回归结果(回归系数、标准误)没有任何实质性影响。

9.7 极端数据

如果样本数据中的少数观测值离大多数观测值很远，可能对OLS的回归系数产生很大影响。

这些数据称为“极端观测值”(outliers)或“高影响力数据”(influential data)，参见图9.2。

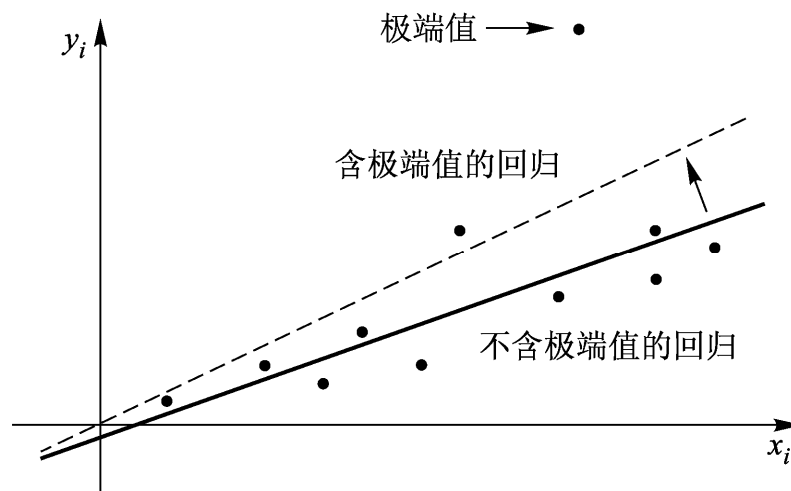


图 9.2 极端观测值对回归系数的影响

以数据集 nerlove.dta 为例。

首先，进行回归；然后人为构造极端值，再回归，并比较结果。

```
. use nerlove.dta,clear
. reg lntc lnq lnpl lnpl lnpl lnpl
```

Source	SS	df	MS	Number of obs = 145		
Model	269.524728	4	67.3811819	F(4, 140) = 437.90		
Residual	21.5420958	140	.153872113	Prob > F = 0.0000		
Total	291.066823	144	2.02129738	R-squared = 0.9260		
				Adj R-squared = 0.9239		
				Root MSE = .39227		
lntc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnq	.7209135	.0174337	41.35	0.000	.6864462	.7553808
lnpl	.4559645	.299802	1.52	0.131	-.1367602	1.048689
lnpk	-.2151476	.3398295	-0.63	0.528	-.8870089	.4567136
lnpf	.4258137	.1003218	4.24	0.000	.2274721	.6241554
_cons	-3.566513	1.779383	-2.00	0.047	-7.084448	-.0485779

将第一个观测值的产量对数(lnq)乘以 100，再次进行回归。

```
. replace lnq=lnq*100 if _n==1
(1 real change made)
```

其中，“_n”表示第 n 个观测值，故“_n==1”表示第 1 个观测值。

```
. reg lntc lnq lnpl lnpl lnpl lnpl
```

Source	SS	df	MS	Number of obs = 145		
Model	7.4424142	4	1.86060355	F(4, 140) = 0.92		
Residual	283.624409	140	2.02588864	Prob > F = 0.4551		
Total	291.066823	144	2.02129738	R-squared = 0.0256		
				Adj R-squared = -0.0023		
				Root MSE = 1.4233		
lntc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnq	.0156026	.0218326	0.71	0.476	-.0275616	.0587668
lnpl	1.214014	1.093264	1.11	0.269	-.9474289	3.375456
lnpk	-1.096614	1.233079	-0.89	0.375	-3.534477	1.341248
lnpf	-.2427032	.3641193	-0.67	0.506	-.9625867	.4771803
_cons	7.230075	6.388544	1.13	0.260	-5.400419	19.86057

人为制造极端值后，回归系数的估计值变化很大，所有系数都变得不显著， R^2 也从 0.926 降为 0.0256 (\bar{R}^2 变为负数)。

将此人造极端值去掉，再对比回归结果。

```
. reg lntc lnq lnpl lnpl lnpl lnpl if _n>1
```

Source	SS	df	MS	Number of obs = 144		
Model	251.560166	4	62.8900416	F(4, 139) =	406.10	
Residual	21.5261194	139	.154864168	Prob > F =	0.0000	
				R-squared =	0.9212	
				Adj R-squared =	0.9189	
Total	273.086286	143	1.90969431	Root MSE =	.39353	

lntc	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lnq	.7225462	.0182135	39.67	0.000	.6865348	.7585576
lnpl	.4445846	.3028466	1.47	0.144	-.1541969	1.043366
lnpk	-.2219834	.3415869	-0.65	0.517	-.8973614	.4533947
lnpf	.4311295	.1019964	4.23	0.000	.2294645	.6327944
_cons	-3.55227	1.78566	-1.99	0.049	-7.082837	-.0217022

去掉极端值后的回归结果又“恢复正常”。

如何发现极端数据？

对于一元回归，可通过画 (x, y) 的散点图来考察，但对多元回归行不通。

某个观测值的影响力可通过去掉此观测值对回归系数的影响来衡量。

记 $\hat{\beta}$ 为全样本的 OLS 估计值，

而 $\hat{\beta}^{(i)}$ 为去掉第 i 个观测值后的 OLS 估计值。

关心 $(\hat{\beta} - \hat{\beta}^{(i)})$ 的变化幅度及如何决定。

定义第 i 个观测数据对回归系数的“影响力”或“杠杆作用”(leverage)为

$$\text{lev}_i \equiv \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \quad (9.21)$$

其中, $\mathbf{x}_i \equiv (1 \ x_{i2} \ \cdots \ x_{iK})'$ 包含个体 i 的全部解释变量, 而 $\mathbf{X} = (\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n)'$ 为数据矩阵。

lev_i 与 $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})$ 存在如下关系:

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)} = \left(\frac{1}{1 - \text{lev}_i} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i e_i \quad (9.22)$$

lev_i 越大, 则 $(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})$ 的变化越大。

可以证明，所有观测数据的影响力 lev_i 满足：

(i) $0 \leq \text{lev}_i \leq 1, (i = 1, \dots, n)$;

(ii) $\sum_{i=1}^n \text{lev}_i = K$ (解释变量个数)。

因此，影响力 lev_i 的平均值为 (K/n) 。

如果某些数据的 lev_i 比平均值 (K/n) 高很多，则可能对回归系数有很大影响。

在 Stata 作完回归后，计算影响力 lev_i 的命令为

```
predict lev,leverage
```

此命令将计算所有观测数据的影响力，并记为变量 lev (可自行命名)。

回到数据集 nerlove.dta。

```
. use nerlove.dta,clear  
  
. qui reg lntc lnq lnpl lnpl lnpl lnpl  
  
. predict lev,leverage
```



```
. sum lev
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lev	145	.0344828	.0202164	.009924	.1177335

```
. dis r(max)/r(mean)
```

```
3.4142728
```

lev 的最大值是其平均值的 3.41 倍，似乎不大。

下面看 lev 最大的三个数值：

```
. gsort -lev
```

此命令将观测值按变量 lev 的降序排列。

如使用命令 “sort lev”，只能按升序排列。

下面看 lev 取值最大的三个数据。

```
. list lev in 1/3
```

	lev
1.	.1177335
2.	.1001472
3.	.0983759

再次人为制造极端数据，将第一个观测值的产量对数(lnq)乘以100，然后计算 lev。

```
. replace lnq=lnq*100 if _n==1  
. qui reg lntc lnq lnpl lnpg lnprf  
. predict lev1,lev
```

```
. sum lev1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lev1	145	.0344828	.0807897	.0083048	.9801415

```
. dis r(max)/r(mean)
```

```
28.424102
```

lev 的最大值是其平均值的 28.42 倍，故存在极端观测值。

如何处理极端数据？

首先，应检查是否因数据输入有误导致极端观测值。

其次，对极端观测值的个体进行背景调查，看是否由与研究课题无关的特殊现象所致，必要时可删除极端数据。

最后，比较稳健的做法是同时汇报“全样本”(full sample)与删除极端数据后的“子样本”(subsample)的回归结果，让读者自己做判断。

9.8 虚 拟 变 量

如使用“定性数据”(qualitative data)或“分类数据”(categorical data)，需要引入“虚拟变量”，即取值为 0 或 1 的变量。

比如，性别分男女，可定义

$$D = \begin{cases} 1, & \text{男} \\ 0, & \text{女} \end{cases} \quad (9.23)$$

对于全球的五大洲，则需要四个虚拟变量，即

$$D_1 = \begin{cases} 1, & \text{亚洲} \\ 0, & \text{其他} \end{cases}, D_2 = \begin{cases} 1, & \text{美洲} \\ 0, & \text{其他} \end{cases}, D_3 = \begin{cases} 1, & \text{欧洲} \\ 0, & \text{其他} \end{cases}, D_4 = \begin{cases} 1, & \text{非洲} \\ 0, & \text{其他} \end{cases} \quad (9.24)$$

如果 $D_1 = D_2 = D_3 = D_4 = 0$ ，则表明为大洋洲。

在有常数项的模型中，如果定性指标共分 M 类，则最多只能在方程中放入 $(M - 1)$ 个虚拟变量。

如在方程中包含 M 个虚拟变量，会产生严格多重共线性，因为如果将这 M 个虚拟变量在数据矩阵 \mathbf{X} 中对应的列向量相加，就会得到与常数项完全相同的向量，即 $(1 \cdots 1)'$ （因为 M 类中必居其一）。

称为“虚拟变量陷阱” (dummy variable trap)。

如模型中没有常数项，可放入 M 个虚拟变量。

例 假设样本中只有四位个体，分别属于三类。其中，前两位个体属于第一类，第三位个体属于第二类，而第四位个体属于第三类。相应地，定义三个虚拟变量 D_1, D_2, D_3 ，则其取值分别为：

$$D_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad (9.25)$$

考虑将这三个虚拟变量同时放入回归方程，并包含常数项：

$$y = \alpha + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \quad (9.26)$$

这三个虚拟变量之和正好就是常数项，因为

$$D_1 + D_2 + D_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \quad (9.27)$$

故方程(9.26)存在严格多重共线性，即虚拟变量陷阱。

解决方法之一是去掉一个虚拟变量：

$$y = \alpha + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \quad (9.28)$$

解决方法之二是去掉常数项：

$$y = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \varepsilon \quad (9.29)$$

在模型中引入虚拟变量，会带来什么影响呢？

考虑有关中国经济的时间序列模型：

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad (t = 1950, \dots, 2000) \quad (9.30)$$

经济结构可能在 1978 年后有变化，引入虚拟变量：

$$D_t = \begin{cases} 1, & \text{若 } t \geq 1978 \\ 0, & \text{其他} \end{cases} \quad (9.31)$$

根据引入虚拟变量的方式，考虑两种情况。

(1) 仅仅引入虚拟变量 D_t 本身，回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \varepsilon_t \quad (9.32)$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + \beta x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases} \quad (9.33)$$

仅引入虚拟变量相当于在不同时期给予不同截距项，见图 9.3。

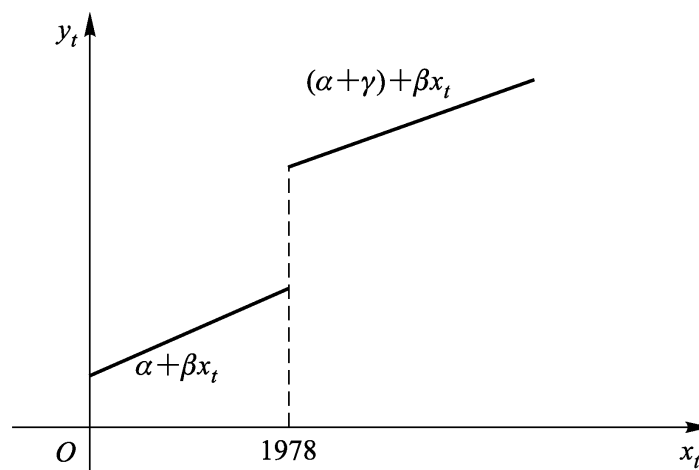


图 9.3 仅引入虚拟变量的效果

(2) 引入虚拟变量 D_t ，以及虚拟变量与解释变量的“互动项”(interaction term) $D_t x_t$ ，回归方程为

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \quad (9.34)$$

该模型等价于

$$y_t = \begin{cases} \alpha + \beta x_t + \varepsilon_t, & \text{若 } t < 1978 \\ (\alpha + \gamma) + (\beta + \delta)x_t + \varepsilon_t, & \text{若 } t \geq 1978 \end{cases} \quad (9.35)$$

引入虚拟变量及其互动项，相当于在不同时期使用不同的截距项与斜率，参见图 9.4。

如果仅仅引入互动项，则仅改变斜率(比较少见)。

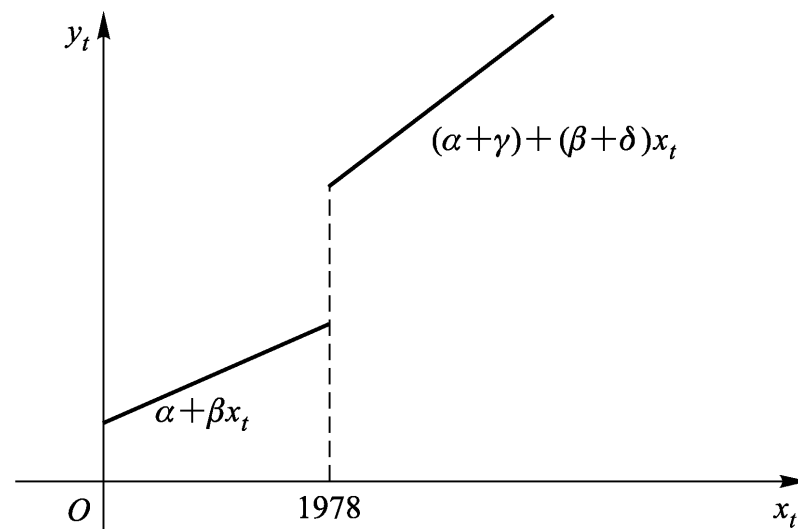


图 9.4 引入虚拟变量及其互动项的效果

在 Stata 中，假设时间变量为 year，可使用如下命令生成上文的虚拟变量：

```
gen d=(year>=1978)
```

其中，“()”表示对括弧内的表达式“`year>=1978`”进行逻辑判断。如果此表达式为真，则取值为 1；反之，取值为 0。

假设有 30 个省的名字储存于变量 `province`，希望为每个省设立一个虚拟变量，分别记为“`prov1, prov2, ..., prov30`”，可使用如下命令：

```
tabulate province, generate(prov)
```

其中，“tabulate”表示将变量按其取值列表；

选择项“generate(prov)”表示根据此变量的不同取值生成以“`prov`”开头的虚拟变量。

由此生成的虚拟变量，将按照变量 `province` 的字母顺序而排序。

在回归时可使用变量的简略写法：

```
reg x1 x2 x3 prov2-prov30
```

为了避免虚拟变量陷阱，略去第一个省的虚拟变量 prov1。

相应的回归模型为

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sum_{i=2}^{30} \delta_i \text{prov}_i + \varepsilon_i \quad (9.36)$$

第一个省的截距项为 α ，第二个省的截距项为 $\alpha + \delta_2$ ，第三个省的截距项为 $\alpha + \delta_3$ ，以此类推。

9.9 经济结构变动的检验

对于时间序列而言，模型系数的稳定性(stability)很重要。

如存在“结构变动”(structural break)，但未加考虑，也是模型设定误差。

考虑结构变动日期已知的情形。

检验中国经济是否在 1978 年发生结构变动。

定义第 1 个时期为 $1950 \leq t < 1978$ ，第 2 个时期为 $1978 \leq t \leq 2010$ 。

两个时期对应的回归方程分别为

$$y_t = \alpha_1 + \beta_1 x_t + \varepsilon_t \quad (1950 \leq t < 1978) \quad (9.37)$$

$$y_t = \alpha_2 + \beta_2 x_t + \varepsilon_t \quad (1978 \leq t \leq 2010) \quad (9.38)$$

原假设为，经济结构在这两个时期内没有变化，即“ $H_0: \alpha_1 = \alpha_2, \beta_1 = \beta_2$ ”，共有两个约束。

如有 K 个解释变量(包含常数项)，则 H_0 共有 K 个约束。

在无约束的情况下，对两个时期，分别进行回归。

在有约束的情况下，可将模型合并为

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad (1950 \leq t \leq 2010) \quad (9.39)$$

其中, $\alpha = \alpha_1 = \alpha_2$, $\beta = \beta_1 = \beta_2$ 。

可将所有样本数据合在一起回归, 即方程(9.39)。

传统的“邹检验”(Chow, 1960)通过作三个回归来检验“无结构变动”的原假设。

(1) 回归整个样本, $1950 \leq t \leq 2010$, 得到残差平方和, 记为 SSR^* 。

(2) 回归第 1 部分子样本, $1950 \leq t < 1978$, 得到残差平方和 SSR_1 。

(3) 回归第 2 部分子样本, $1978 \leq t \leq 2010$, 得到残差平方和 SSR_2 。

将整个样本一起回归为“有约束 OLS”，其残差平方和为 SSR^* 。

将样本一分为二，分别进行回归为“无约束 OLS”，其残差平方和为

$$SSR = SSR_1 + SSR_2 \quad (9.40)$$

$SSR^* \geq SSR = SSR_1 + SSR_2$ ，因为有约束 OLS 的拟合优度比无约束 OLS 更差。

如 H_0 成立(无结构变动)，则 $(SSR^* - SSR_1 - SSR_2)$ 应较小；施加约束后，不应使残差平方和上升很多。

如 $(SSR^* - SSR_1 - SSR_2)$ 很大，则倾向于认为 H_0 不成立，存在结构变动。

根据第 5 章, 在对 m 个线性约束进行联合检验时, 似然比检验原理的 F 统计量为

$$F = \frac{(\text{SSR}^* - \text{SSR})/m}{\text{SSR}/(n - K)} \sim F(m, n - K) \quad (9.41)$$

其中, SSR 为无约束的残差平方和, SSR^* 为有约束的残差平方和, n 为样本容量, 而 K 为无约束回归的参数个数。

回到检验结构变动的情形, 如果有 K 个解释变量(含常数项), 则共有 K 个约束条件, 而无约束回归的参数个数为 $2K$ 。

检验结构变动的 F 统计量:

$$F = \frac{(\text{SSR}^* - \text{SSR}_1 - \text{SSR}_2)/K}{(\text{SSR}_1 + \text{SSR}_2)/(n - 2K)} \sim F(K, n - 2K) \quad (9.42)$$

其中， n 为样本容量， K 为有约束回归的参数个数(含常数项)。

对于一元回归的例子， $K = 2$ 。

检验结构变动的另一方法是引入虚拟变量，并检验所有虚拟变量以及其与解释变量交叉项的系数的联合显著性。

对于 $K = 2$ 的情形，可进行如下回归：

$$y_t = \alpha + \beta x_t + \gamma D_t + \delta D_t x_t + \varepsilon_t \quad (9.43)$$

检验联合假设“ $H_0 : \gamma = \delta = 0$ ”。

所得 F 统计量与邹检验完全相同，故虚拟变量法与邹检验等价。

与邹检验相比，虚拟变量法的优点包括：

(1) 只需生成虚拟变量即可检验，十分方便；

(2) 邹检验在“球形扰动项”(同方差、无自相关)的假设下得到，并不适用于异方差或自相关的情形。

在异方差或自相关的情况下，仍可使用虚拟变量法，只要在估计方程(9.43)时，使用异方差自相关稳健的 HAC 标准误即可。

(3) 如发现结构变动，邹检验不提供究竟是截距项还是斜率变动的信息，虚拟变量法可提供这些信息。

以数据集 `consumption.dta` 为例，考察中国的消费函数是否在 1992 年发生了结构变化。

先看中国 1978—2013 年“居民人均消费”(c)与“人均国内总产值”(y)的年度(year)时间趋势图(如图 9.5)，以当年价格计。

```
. use consumption.dta,clear  
  
. twoway connect c y year,msymbol(circle)  
msymbol(triangle)
```

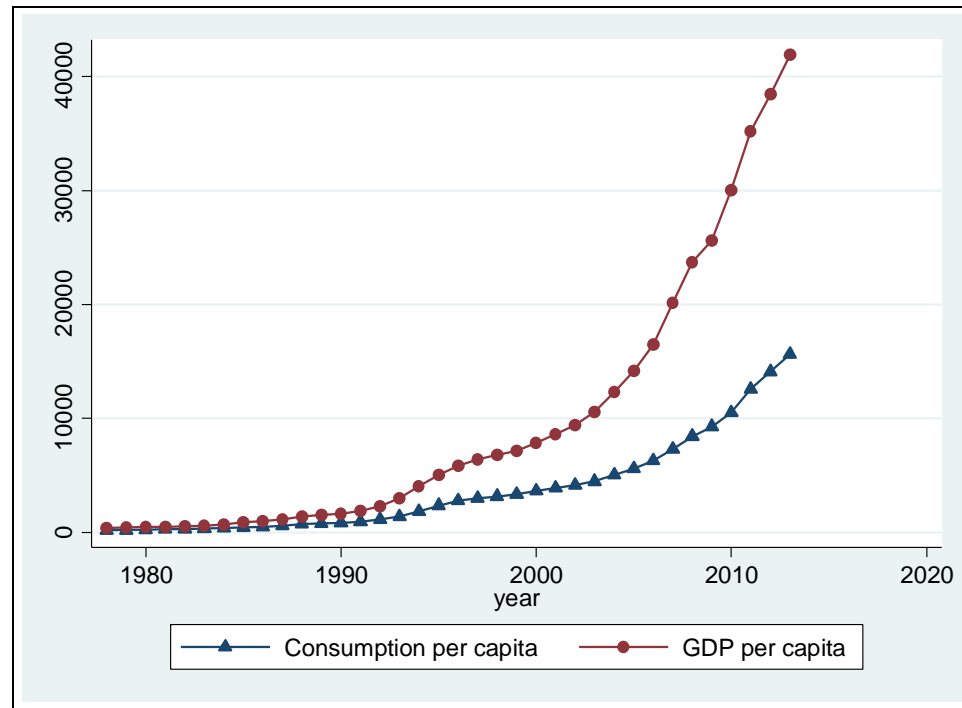


图 9.5 居民人均消费与人均国内总产值时间趋势

二者走势有较强相关性。但图右边有空白，不够美观。

将命令略加改进，并在 1992 年处画一条垂直线(见图 9.6):

```
. twoway connect c y year,msymbol(circle)  
msymbol(triangle) xlabel(1980(10)2010)  
xline(1992)
```

选择项 “`xlabel(1980(10)2010)`” 指示在横轴 (即 X 轴)1980-2010 年之间, 每隔 10 年做个标注(label);

选择项 “`xline(1992)`” 表示在横轴 1992 年的位置画一条直线。

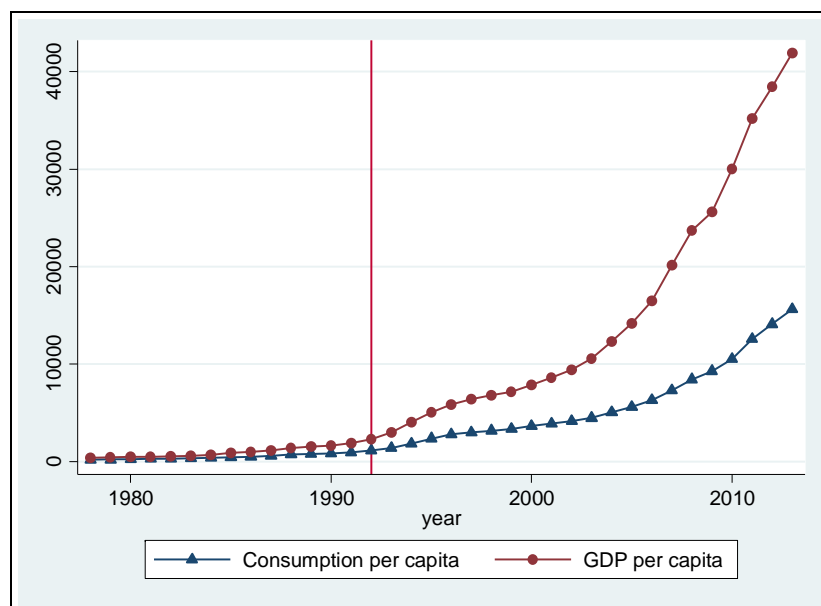


图 9.6 改进的趋势图

考察简单的消费函数：

$$c_t = \alpha + \beta y_t + \varepsilon_t$$

首先，使用传统的邹检验来检验消费函数是否在 1992 年发生结构变动。

分别对整个样本、1992 年之前及之后的子样本进行回归，获得其残差平方和：

```
. reg c y
```

Source	SS	df	MS	Number of obs = 36		
Model	617812224	1	617812224	F(1, 34) = 7139.56		
Residual	2942143.12	34	86533.6213	Prob > F = 0.0000		
Total	620754367	35	17735839.1	R-squared = 0.9953		
				Adj R-squared = 0.9951		
				Root MSE = 294.17		
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y	.3572642	.0042282	84.50	0.000	.3486715	.3658569
_cons	339.0701	63.83305	5.31	0.000	209.3458	468.7945

```
. scalar ssr=e(rss)
```

其中，“scalar”表示标量，将此回归的残差平方和(e(rss))记为标量 ssr。

对 1992 年之前的子样本进行回归。

```
. reg c y if year<1992
```

Source	SS	df	MS	Number of obs = 14		
Model	829125.648	1	829125.648	F(1, 12) = 4344.64		
Residual	2290.06599	12	190.838833	Prob > F = 0.0000		
Total	831415.714	13	63955.0549	R-squared = 0.9972		
				Adj R-squared = 0.9970		
				Root MSE = 13.814		
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y	.4996452	.0075803	65.91	0.000	.4831292	.5161612
_cons	12.89123	7.908212	1.63	0.129	-4.339283	30.12174

```
. scalar ssr1=e(rss)
```

此命令将 1992 年之前的子样本回归的残差平方和记为 ssr1。

对 1992 年及之后的子样本进行回归。

```
. reg c y if year>=1992
```

Source	SS	df	MS	Number of obs = 22		
Model	366038781	1	366038781	F(1, 20) = 4889.80		
Residual	1497151	20	74857.5501	Prob > F = 0.0000		
Total	367535932	21	17501711	R-squared = 0.9959		
				Adj R-squared = 0.9957		
				Root MSE = 273.6		
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y	.3444589	.004926	69.93	0.000	.3341836	.3547343
_cons	658.1088	95.04293	6.92	0.000	459.8527	856.3648

```
. scalar ssr2=e(rss)
```

此命令将 1992 年之后的子样本回归的残差平方和记为 ssr2。

由于 $n = 36$, $K = 2$, $n - 2K = 32$, 可计算 F 统计量如下:

```
. di ((ssr-ssr1-ssr2)/2)/((ssr1+ssr2)/32)  
15.394558
```

故 F 统计量等于 15.39。

其次, 使用虚拟变量法进行结构变动的检验。

生成虚拟变量 d (对于 1992 年及以后, $d=1$; 反之, $d=0$); 以及虚拟变量 d 与人均收入 y 的互动项 yd :

```
. gen d=(year>1991)  
. gen yd=y*d
```

引入 d 与 yd，进行全样本 OLS 回归：

```
. reg c y d yd
```

Source	SS	df	MS	Number of obs = 36		
Model	619254926	3	206418309	F(3, 32) = 4405.23		
Residual	1499441.07	32	46857.5333	Prob > F = 0.0000		
				R-squared = 0.9976		
				Adj R-squared = 0.9974		
Total	620754367	35	17735839.1	Root MSE = 216.47		
c	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
y	.4996452	.1187794	4.21	0.000	.2576994	.741591
d	645.2175	144.9484	4.45	0.000	349.9673	940.4678
yd	-.1551863	.1188434	-1.31	0.201	-.3972623	.0868897
_cons	12.89123	123.9181	0.10	0.918	-239.5216	265.3041

检验 d 与 yd 的联合显著性：

```
. test d yd
```

```
( 1)  d = 0
( 2)  yd = 0

      F(  2,    32) =   15.39
      Prob > F =    0.0000
```

虚拟变量法所得 F 统计量为 15.39，与邹检验完全相同。

p 值为 0.0000，可在 1% 水平上拒绝“无结构变动”的原假设。

上述检验仅在球形扰动项(同方差、无自相关)的情况下才成立。

下面进行异方差与自相关的检验。

```
. qui reg c y
. estat imtest,white
```

```
White's test for Ho: homoskedasticity
    against Ha: unrestricted heteroskedasticity

    chi2(2)      =      6.31
    Prob > chi2   =      0.0427
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	6.31	2	0.0427
Skewness	4.11	1	0.0425
Kurtosis	4.76	1	0.0291
Total	15.19	4	0.0043

可在 5% 的水平上拒绝“同方差”的原假设。

为进行自相关的 BG 检验，首先设定变量 `year` 为时间变量。

```
. tsset year
```

```
time variable:  year, 1978 to 2013
               delta:  1 unit
```



```
. estat bgodfrey
```

Breusch-Godfrey LM test for autocorrelation			
lags(p)	chi2	df	Prob > chi2
1	28.109	1	0.0000
H0: no serial correlation			

可在 1%水平上强烈拒绝“无自相关”的原假设。

故此模型的扰动项存在异方差与自相关。

应使用异方差自相关稳健的标准误，通过虚拟变量法检验结构变动。

首先，计算 HAC 标准误的截断参数。

```
. dis 36^(1/4)
```

2.4494897

应将截断参数设为 3。进行 Newey-West 回归。

```
. newey c y d yd, lag(3)
```

Regression with Newey-West standard errors				Number of obs	=	36
maximum lag: 3				F(3, 32)	=	2455.09
				Prob > F	=	0.0000
c	Newey-West		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
y	.4996452	.0099228	50.35	0.000	.4794332	.5198573
d	645.2175	139.943	4.61	0.000	360.1629	930.2721
yd	-.1551863	.013774	-11.27	0.000	-.1832431	-.1271295
_cons	12.89123	10.16563	1.27	0.214	-7.815475	33.59793

检验虚拟变量 d 及其互动项 yd 的联合显著性。

```
. test d yd
```

```
( 1)  d = 0  
( 2)  yd = 0  
  
      F(  2,    32) =    73.05  
      Prob > F =    0.0000
```

p 值为 0.0000，可在 1% 水平上拒绝“无结构变动”的原假设，认为中国的消费函数在 1992 年发生了结构变动。

9.10 缺失数据与线性插值

有时会出现某些时期数据缺失(missing data)的情形,尤其是历史较久远的数据。

缺失的观测值在 Stata 中以 “.” 来表示。

在运行 Stata 命令时(比如 `reg`), 会自动将缺失的观测值从样本中去掉, 导致样本容量损失。

在数据缺失不严重的情况下, 为保持样本容量, 可采用“线性插值”(linear interpolation)的方法补上缺失数据。

已知 x_{t-1} 与 x_{t+1} , 但缺失 x_t 的数据, 则 x_t 对时间 t 的线性插值为

$$\hat{x}_t = \frac{x_{t-1} + x_{t+1}}{2} \quad (9.44)$$

更一般地，假设与 x (通常为时间) 对应的 y 缺失，而最临近的两个点分别为 (x_0, y_0) 与 (x_1, y_1) ，且 $x_0 < x < x_1$ ，则 y 对 x 的线性插值 \hat{y} 满足(参见图 9.7):

$$\frac{\hat{y} - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \quad (9.45)$$

整理可得

$$\hat{y} = \frac{y_1 - y_0}{x_1 - x_0} (x - x_0) + y_0 \quad (9.46)$$

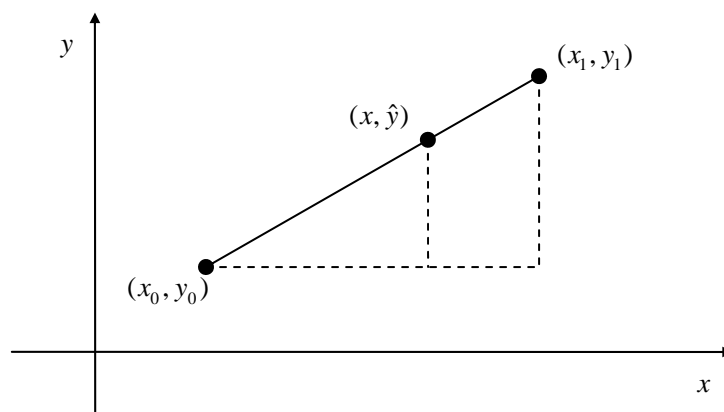


图 9.7 线性插值示意图

线性插值的基本假设是变量以线性速度均匀变化。

如变量 y 有指数增长趋势(比如 **GDP**)，应先取对数，再用 $\ln y$ 进行线性插值，以避免偏差。

如需要以原变量 y 进行回归，可将线性插值的对数值 $\widehat{\ln y}$ 取反对数(antilog)，即计算 $\exp(\widehat{\ln y})$ 。

线性插值的 Stata 命令为

```
ipolate y x, gen(newvar)
```

其中，“ipolate”表示 interpolate，即将变量 y 对变量 x 进行线性插值，并将插值的结果记为新变量 newvar。

以数据集 consumption.dta 为例。

```
. use consumption.dta, clear
```

假设 1980 年、1990 年、2000 年及 2010 年的人均 GDP 数据缺失。

首先，生成缺失这些年份数据的人均 GDP 变量，记为 y1。

```
. gen y1=y  
  
. replace y1=. if year==1980 | year==1990 |  
year==2000 | year==2010  
(4 real changes made, 4 to missing)
```

直接用 y1 对 year 进行线性插值，将结果记为 y2。

```
. ipolate y1 year,gen(y2)
```

由于人均 GDP 有指数增长趋势，故更好的做法是，先对 y1 取对数，进行线性插值，再取反对数，将结果记为 y3。

```
. gen lny1=log(y1)  
(4 missing values generated)
```



```
. ipolate lny1 year, gen(lny3)
```

```
. gen y3=exp(lny3)
```

对比这两种方法的效果。

```
. list year y y2 y3 if year==1980 | year==1990  
| year==2000 | year==2010
```

	year	y	y2	y3
3.	1980	463.25	455.705	454.2445
13.	1990	1644	1705.88	1695.613
23.	2000	7857.68	7890.105	7856.112
33.	2010	30015.1	30402.659	30022.13

直接插值的结果 y2 倾向于高估真实值 y，整体估计效果不如先取对数再插值的结果 y3 (1980 年的结果是例外)。

9.11 变量单位的选择

在选择变量单位时，应尽量避免变量间的数量级差别过于悬殊，以免出现计算机运算的较大误差。

比如，通货膨胀率通常小于 1，而如果模型中有 **GDP** 这个变量，则 **GDP** 应该使用亿或万亿作为单位。

否则，变量 **GDP** 的取值将是通货膨胀率的很多倍，即数据矩阵 **X** 中某列的数值是另一列的很多倍，可能使计算机在对 $(\mathbf{X}'\mathbf{X})^{-1}$ 进行数值计算时出现较大误差。

电脑的存储空间有限，实际上只能作近似计算，即精确到小数点后若干位。