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)。

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

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

其中，解释变量x_1, x_2与扰动项ε不相关，并省略了表示个体的下标 i。

实际估计的模型(estimated model)为

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

对比以上两个方程可知，**遗漏变量**(omitted variable)x_2被归入新扰动项$u=\gamma x_2+\varepsilon$中了。

遗漏变量是否一定导致不一致的估计？考虑以下两种情形：

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

在这种情况下，扰动项$u=\gamma x_2+\varepsilon$与解释变量x_1不相关，因为

$$\mathrm{Cov}(x_1, u)=\mathrm{Cov}(x_1, \gamma x_2+\varepsilon)=\gamma\,\mathrm{Cov}(x_1, x_2)+\mathrm{Cov}(x_1, \varepsilon)=0+0=0 \tag{9.3}$$

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u中，可能增大扰动项的方差，从而影响 OLS 估计的精确度。

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

在这种情况下，根据大样本理论，OLS 估计不一致，称其偏差为**遗漏变量偏差**(omitted variable bias)。

这种与解释变量相关的遗漏变量，也称为**混杂变量**(confounder)。

如果不存在遗漏变量，或虽有遗漏变量但与解释变量不相关，则称为**非混杂性**(unconfounded 或 no confounding)。

存在遗漏变量本身并不要紧，但遗漏变量不能与解释变量相关。

遗漏变量偏差在实践中较常见，成为某些实证研究的致命伤。

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

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)种方法着眼于直接解决遗漏变量问题，即把遗漏的变量补上去。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，然后尽可能地去收集数据。

如果有些相关变量确实无法获得，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、孟岭生、施新政、吴斌珍(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)的方法来研究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。

但这种控制实验的方法在其他学科未必可行；比如，医学上对新药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(即使用老鼠做实验，老鼠之间仍然有差异)，故无法进行严格的控制实验。

当代统计学之父费雪(Ronald Fisher)提出**随机控制实验**(randomized controlled experiment)，简称“随机实验”。

考虑以下回归模型：

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

其中，x是完全随机地决定的(比如，通过抛硬币或电脑随机数)，故x与世界上的任何其他变量都相互独立。

由于x独立于ε，故$\mathrm{Cov}(x,\varepsilon)=0$，因此无论遗漏了多少解释变量，OLS 都是一致的。

如果x是取值为 0 或 1 的虚拟变量，则可将样本数据分为两组。

通常称“$x=1$”的那一组为**实验组**(experimental group)或“**处理组**”(treatment group)，比如医学实验中吃药的那组。

称“$x=0$”的那一组为**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性差别。

例 在农学中将地块随机分成三组(因为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？班级规模过大(师生比过低)可能影响任课老师对每位学生的关注，从而影响学习效果。

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(称为 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。

实验结果发现，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)，详见 Stock and Watson (2012) 第 13 章。

随机实验虽然说服力强，但通常成本较高。

另外一种实验方法为**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了实验组或控制组。

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？

为避免内生性(即解释变量与扰动项相关)，Card and Krueger (1994)考虑了一个自然实验。

在 1992 年，美国新泽西州通过法律将最低工资从每小时$4.25提高到$5.05，但在相邻的宾夕法尼亚州最低工资却保持不变。

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)。

该文收集了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？问题在于，当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？

解决遗漏变量偏差的第(iii)种方法为工具变量法，而第(iv)种方法为使用面板数据，将分别在第 10 章与第 12 章介绍。

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是需要说明，它是如何在存在遗漏变量的情况下避免遗漏变量偏差的。

如果无法令人信服地说明这一点，则其结果就是可疑的。

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

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

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

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

其中，加入了与被解释变量y无关的解释变量x_2。

由于真实参数$\gamma = 0$(x_2对y无影响)，故可将模型写为

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

其中，扰动项仍是原来的ε。

在方程(9.7)中，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$\mathrm{Cov}(x_2, \varepsilon) = 0$。

扰动项ε与所有解释变量均无关，故 OLS 仍然一致，即$\underset{n\to\infty}{\mathrm{plim}}\,\hat{\beta} = \beta$，$\underset{n\to\infty}{\mathrm{plim}}\,\hat{\gamma} = \gamma = 0$。

但引入无关变量后，由于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，然后逐渐增加解释变量。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$\mathbf{z}$。

这种方法的缺点是，小模型很可能存在遗漏变量偏差，这样系数估计量就不一致，t 检验、F 检验都将失效，因此很难确定该如何取舍变量。

与此相反，**由大到小**(general to specific)的建模方式从一个尽可能大的模型开始，收集所有可能的解释变量，然后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)。

这样做虽然冒着包含无关变量的危险，但其危害性毕竟没有遗漏变量严重。

实践中，一般很难找到所有与被解释变量相关的解释变量。

在实证研究中，常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

好的经济理论应用简洁的模型来很好地描述复杂的经济现实。

但这两个目标常常是矛盾的。

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

需要在模型的解释力与简洁性之间找到最佳的平衡。

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，需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$\overline{R}^2$：选择解释变量的个数K以最大化$\overline{R}^2$。

(2) **赤池信息准则**(Akaike Information Criterion，简记 AIC)：选择解释变量的个数 K，使得以下目标函数最小化：

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

其中，SSR 为残差平方和$\sum_{i=1}^{n} e_i^2$。

上式右边的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，简记 BIC)或**施瓦茨信息准则**(Schwarz Information Criterion，简记 SIC 或 SBIC)：选择解释变量的个数 K，使得以下目标函数最小化：

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

BIC 准则与 AIC 准则仅第二项有差别。

一般来说，$\ln n > 2$(除非样本容量很小)，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，常用信息准则来确定滞后阶数。

考虑以下 p 阶自回归模型(AR(p)，详见第 13 章)，

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

其中，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，设一个最大滞后期 $p_{\max}$，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(参见第 8 章)，考虑应该引入气温(*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
```

Akaike's information criterion and Bayesian information criterion

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

Note: BIC uses N = number of observations. See **[R] IC note**.

在以上模型中加入气温的一阶滞后项(L.temp)，重新进行估计。

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

Akaike's information criterion and Bayesian information criterion

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

Note: BIC uses N = number of observations. See **[R] IC note**.

增加解释变量 L.temp 后，AIC 与 BIC 都下降了。

进一步，引入气温的二阶滞后项(L2.temp)：

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

Akaike's information criterion and Bayesian information criterion

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

Note: BIC uses N = number of observations. See **[R] IC note**.

加入气温的二阶滞后项后，AIC 与 BIC 反而比仅包括气温的滞后项上升了。

仅包含气温的滞后项可以达到 AIC 与 BIC 的最小值。

从信息准则的角度，应包含气温的一阶滞后项，但不该引入更高阶的气温滞后项。

其次，使用序贯t规则，并假设$p_{\max}=2$，估计以下模型：

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

Source	SS	df	MS	Number of obs	=	28
				F(5, 22)	=	21.92
Model	.103722201	5	.02074444	Prob > F	=	0.0000
Residual	.020822754	22	.000946489	R-squared	=	0.8328
				Adj R-squared	=	0.7948
Total	.124544954	27	.004612776	Root MSE	=	.03077

consumption	Coefficient	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)。

因此，令 $\hat{p} = p_{\max} - 1 = 1$(即去掉 L2.temp)，重新进行估计。

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

Source	SS	df	MS	Number of obs	=	29
				F(4, 24)	=	28.98
Model	.103387183	4	.025846796	Prob > F	=	0.0000
Residual	.021406049	24	.000891919	R-squared	=	0.8285
				Adj R-squared	=	0.7999
Total	.124793232	28	.004456901	Root MSE	=	.02987

consumption	Coefficient	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。多元线性回归只能看作是非线性关系的一阶线性近似。

但二阶乃至高阶的非线性部分真的不重要吗？

如果存在非线性项，但被遗漏了，则会导致遗漏变量偏差，这是**模型设定误差**(specification error)的一种形式。

假设真实模型为

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

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

上式的解释变量与扰动项相关：

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

Ramsey's RESET 检验(Regression Equation Specification Error Test)(Ramsey, 1969)的基本思想是，如果怀疑非线性项被遗漏，那么就把非线性项引入方程，并检验其系数是否显著。

假设线性回归模型为

$$y = \alpha + \beta x_1 + \gamma x_2 + \varepsilon \tag{9.13}$$

记此回归的拟合值为

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_1 + \hat{\gamma} x_2 \tag{9.14}$$

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。

考虑以下辅助回归：

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

然后对$H_0 : \delta_2 = \delta_3 = \delta_4 = 0$作$F$检验，即检验拟合值$\hat{y}$的高次项系数是否联合为0。

如果拒绝H_0，说明模型中应有高次项；反之，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。

也可以直接将解释变量x_1与x_2的高次项放入辅助回归中，比如

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

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，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
```

“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 为例进行演示。

首先，打开数据集，并进行线性 OLS 回归。

```
. use grilic.dta,clear
. qui reg lnw s expr tenure smsa rns
```

然后，使用拟合值的高次项进行 RESET 检验。

```
. estat ovtest
```

```
Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of lnw

H0: 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。

下面，直接使用解释变量的高次项进行 RESET 检验。

```
. estat ovtest,rhs
```

```
Ramsey RESET test for omitted variables
Omitted: Powers of independent variables

H0: Model has no omitted variables

F(9, 743) =   2.03
 Prob > F = 0.0336
```

可在 5%的水平上拒绝原假设，即认为遗漏了高阶非线性项。

根据劳动经济学的知识，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
				F(6, 751)	=	69.65
Model	49.7985818	6	8.29976364	Prob > F	=	0.0000
Residual	89.487568	751	.11915788	R-squared	=	0.3575
				Adj R-squared	=	0.3524
Total	139.28615	757	.183997556	Root MSE	=	.34519

lnw	Coefficient	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*)却变得很不显著；因为二者存在多重共线性(参见下一节)。

再次使用解释变量的高次项进行 RESET 检验：

```
. estat ovtest,rhs
```

```
note: expr2 omitted because of collinearity.
note: expr2^2 omitted because of collinearity.

Ramsey RESET test for omitted variables
Omitted: Powers of independent variables

H0: Model has no omitted variables

F(11, 741) =   1.73
  Prob > F = 0.0626
```

可在 5%的水平上，接受“无遗漏变量”的原假设。

在本例中，最重要的模型设定误差乃是遗漏了对个人能力的度量，将在第 10 章进一步讨论。

9.6 多重共线性

如果在解释变量中，有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，故无法定义 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，在现实中较少出现。

即使出现，Stata 也会自动去掉多余的变量。

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对其余的解释变量$\{x_1, \cdots, x_{k-1}, x_{k+1}, \cdots, x_K\}$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，在所有线性无偏估计中方差最小，因为高斯-马尔可夫定理并未排除多重共线性的情形。

但 BLUE 只是保证 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$(\mathbf{X}'\mathbf{X})$不可逆。

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

$\mathbf{X}$中元素轻微变化就会引起$(\mathbf{X}'\mathbf{X})^{-1}$很大变化，导致 OLS 估计值$\hat{\boldsymbol{\beta}}$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，最后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，则不容易区分它们各自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

$$\mathrm{Var}(\hat{\beta}_k \mid \mathbf{X}) = \frac{\sigma^2}{(1-R_k^2)S_k} \qquad (9.17)$$

其中，$\sigma^2 \equiv \mathrm{Var}(\varepsilon)$为扰动项的方差；而$S_k \equiv \sum_{i=1}^{n}(x_{ik} - \overline{x}_k)^2$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，则很难准确地估计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$\hat{\beta}_k$(此时x_k与常数项构成严格多重共线性)。

从表达式(9.17)可知，方差$\mathrm{Var}(\hat{\beta}_k \mid \mathbf{X})$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，简记VIF)为

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

根据表达式(9.17)与(9.18)，可将方差写为

$$\mathrm{Var}(\hat{\beta}_k \mid \mathbf{X}) = \mathrm{VIF}_k \cdot \frac{\sigma^2}{S_k} \qquad (9.19)$$

其中，σ^2是扰动项本身的方差，S_k衡量变量x_k自己的波动，而VIF_k才真正度量由于变量x_k与其他解释变量的多重共线性所导致其方差$\mathrm{Var}(\hat{\beta}_k \mid \mathbf{X})$的膨胀程度。

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

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

判断是否存在多重共线性的一个经验规则是，$\{\text{VIF}_1, \cdots, \text{VIF}_K\}$的最大值不应超过 10。

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

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

可在 Stata 中通过画出函数(9.18)来考察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)
```

选择项“xtitle(R2)”指示横轴的标题为R2。

选择项“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。

选择项“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。

选择项“ylabel(10 100 200 300)”则表示在纵轴上10、100、200与300的位置给出标签，结果参见图9.1。

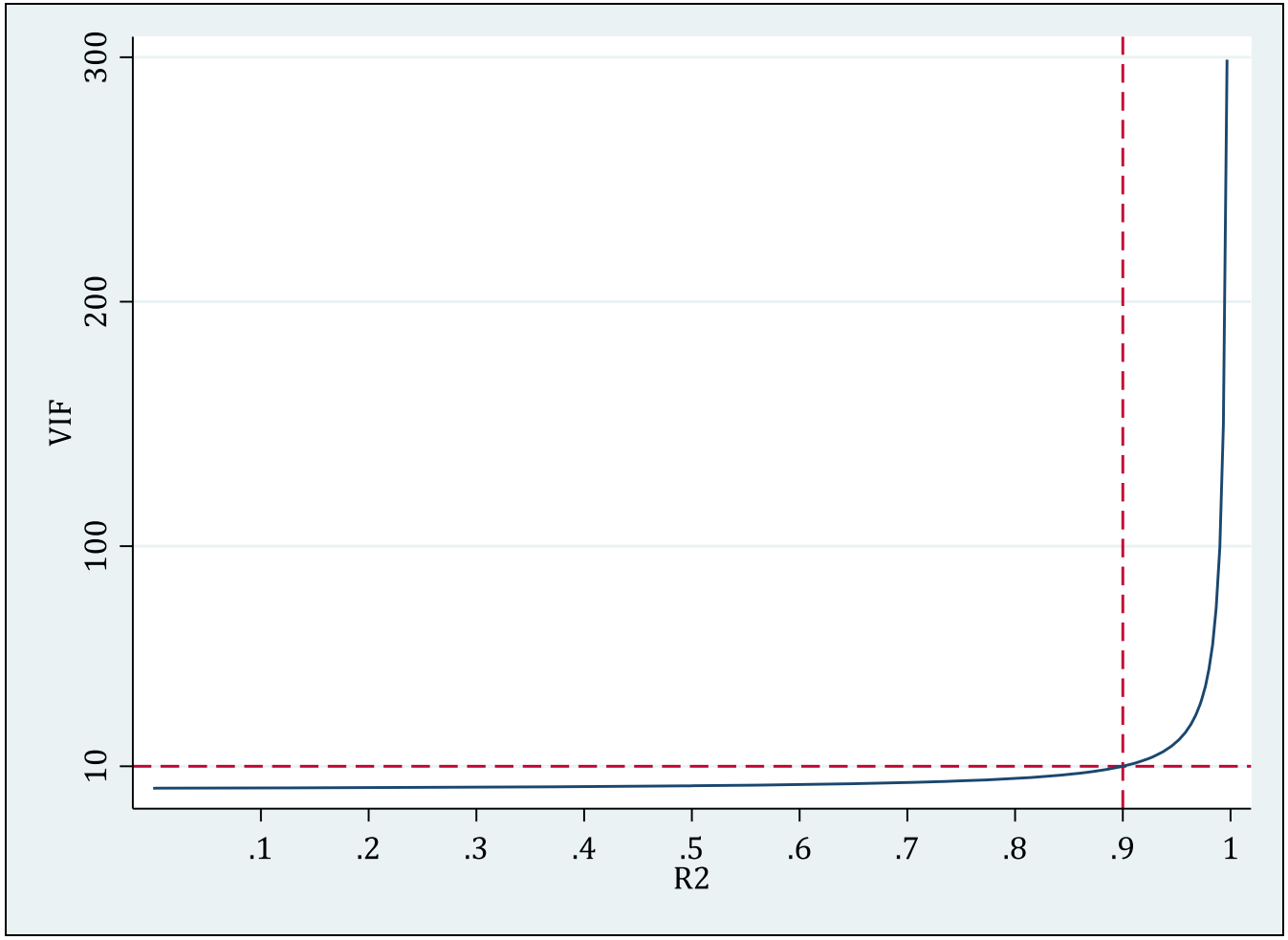

图 9.1 VIF_k 与 R_k^2 的关系

如果发现存在多重共线性，可采取以下处理方法。

(1) 如果不关心具体的回归系数，而只关心整个方程预测被解释变量的能力，则通常可不必理会多重共线性(假设整个方程是显著的)。

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) 如果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，则也可不必理会。

即使在有方差膨胀的情况下，这些系数依然显著；如果没有多重共线性，则只会更加显著。

(3) 如果多重共线性影响到所关心变量的显著性，则应设法进行处理。

比如，需要增大样本容量，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就是“无为而治”(do nothing)。

在 Stata 中作完回归后，可使用如下命令计算各变量的 VIF。

```
. estat vif
```

仍以数据集 `grilic.dta` 为例。

首先，考察线性回归的 VIF。

```
. use grilic.dta,clear
. qui reg lnw s expr tenure 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，故不必担心存在多重共线性。

如果在模型中引入解释变量的平方项，则容易引起多重共线性，因为x与x^2通常较相关。

考虑在上述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
				F(6, 751)	=	68.26
Model	49.1549871	6	8.19249785	Prob > F	=	0.0000
Residual	90.1311627	751	.120014864	R-squared	=	0.3529
				Adj R-squared	=	0.3477
Total	139.28615	757	.183997556	Root MSE	=	.34643

lnw	Coefficient	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，故存在多重共线性。

进一步，将 *s2* 对 *s* 进行回归。

```
. reg s2 s
```

Source	SS	df	MS	Number of obs	=	758
				F(1, 756)	>	99999.00
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	Coefficient	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

此回归的R^2高达 0.9939, 即变量 *s* 可以解释其平方项(*s2*) 99%的变动。

s 与 *s2* 所包含的信息基本相同，故会导致严重的多重共线性。

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

一个解决方法是将变量标准化，即减去均值，除以标准差：

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

其中，$\bar{x}$为变量x的样本均值，s_x为样本标准差，而$\tilde{x}$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。

继续上面的例子。

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
				F(6, 751)	=	68.26
Model	49.154987	6	8.19249783	Prob > F	=	0.0000
Residual	90.1311629	751	.120014864	R-squared	=	0.3529
				Adj R-squared	=	0.3477
Total	139.28615	757	.183997556	Root MSE	=	.34643

lnw	Coefficient	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。

计算各变量的方差膨胀因子。

```
. 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，故认为基本不存在多重共线性。

为了验证这一点，将 *sd2* 对 *sd* 进行回归：

```
. reg sd2 sd
```

Source	SS	df	MS	Number of obs	=	758
				F(1, 756)	=	159.77
Model	152.821515	1	152.821515	Prob > F	=	0.0000
Residual	723.111439	756	.956496612	R-squared	=	0.1745
				Adj R-squared	=	0.1734
Total	875.932954	757	1.1571109	Root MSE	=	.97801

sd2	Coefficient	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

此回归的R^2仅为0.1745，相对于 *s2* 对 s 回归的R^2(0.9939)而言，大大下降。

由于 *sd2* 在上面的回归中不显著，下面去掉 *sd2*，但保留 *sd*，再次进行回归：

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

Source	SS	df	MS	Number of obs	=	758
				F(5, 752)	=	81.75
Model	49.0478812	5	9.80957624	Prob > F	=	0.0000
Residual	90.2382686	752	.119997698	R-squared	=	0.3521
				Adj R-squared	=	0.3478
Total	139.28615	757	.183997556	Root MSE	=	.34641

lnw	Coefficient	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

注意到 *sd* 的回归系数为 0.2291，似乎偏高。

由于 *sd* 为标准化的变量，故 *sd* 变化一个单位，等价于 *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
				F(5, 752)	=	81.75
Model	49.0478814	5	9.80957628	Prob > F	=	0.0000
Residual	90.2382684	752	.119997697	R-squared	=	0.3521
				Adj R-squared	=	0.3478
Total	139.28615	757	.183997556	Root MSE	=	.34641

lnw	Coefficient	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

对比以上两表可知，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，它们可能对 OLS 的回归系数产生很大影响。

这些数据称为**极端值**、**离群值**(outliers) 或**高影响力数据**(influential data)，参见图 9.2。

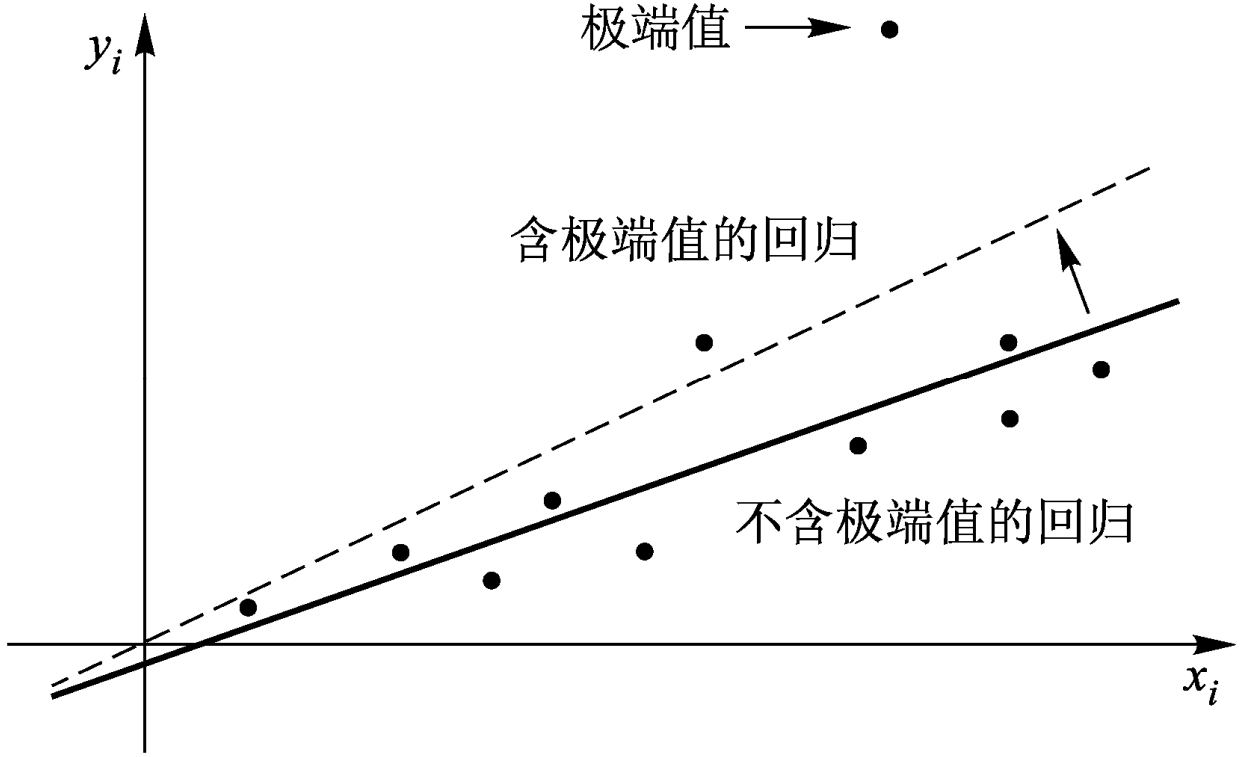

图 9.2　极端观测值对回归系数的影响

以数据集 nerlove.dta 为例。

首先，打开数据集，进行回归；然后人为地构造一个极端值，再进行回归，并比较回归结果。

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

Source	SS	df	MS	Number of obs	=	145
				F(4, 140)	=	437.90
Model	269.524728	4	67.3811819	Prob > F	=	0.0000
Residual	21.5420958	140	.153872113	R-squared	=	0.9260
				Adj R-squared	=	0.9239
Total	291.066823	144	2.02129738	Root MSE	=	.39227

lntc	Coefficient	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，然后再次进行回归。

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

“_n”表示第n个观测值，故“_n==1”表示第 1 个观测值。

```
. reg lntc lnq lnpl lnpk lnpf
```

Source	SS	df	MS	Number of obs	=	145
				F(4, 140)	=	0.92
Model	7.4424142	4	1.86060355	Prob > F	=	0.4551
Residual	283.624409	140	2.02588864	R-squared	=	0.0256
				Adj R-squared	=	-0.0023
Total	291.066823	144	2.02129738	Root MSE	=	1.4233

lntc	Coefficient	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，且所有系数都变得不显著，R^2也从 0.926 降为 0.0256($\overline{R}^2$则变为负数)。

所有这些变化都仅仅是因为在 145 个观测值中有一个观测值发生了变化，这正是“高影响力数据”(influential data)的含义。

将此人造极端值去掉，再对比回归结果。

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

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

lntc	Coefficient	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”了，无论回归系数与显著性水平都类似于没有极端值的原始全样本。

如何发现极端数据？

对于一元回归，可通过画(x, y)的散点图来直观地考察是否存在极端观测值。

但画图的方法对于多元回归则行不通。

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$\hat{\boldsymbol{\beta}}$为全样本的OLS估计值，而$\hat{\boldsymbol{\beta}}^{(i)}$为去掉第$i$个观测值后的OLS估计值。

我们关心$(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^{(i)})$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**杠杆值**(leverage)为

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

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：

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

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

所有观测数据的影响力lev_i满足：

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

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

影响力lev_i的平均值为$\left(K/n\right)$。

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

在 Stata 中作完回归后，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 的例子。

```
. use nerlove.dta,clear
. qui reg lntc lnq lnpl lnpk lnpf
. 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，似乎并不大。

下面来看 *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。如果使用命令“`sort lev`”，则只能按升序排列。

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

为了演示目的，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 lnpk lnpf
. 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，故存在高影响力的极端观测值。

如果发现存在极端数据，如何处理呢？

首先，应仔细检查是否因数据输入有误而导致极端观测值。

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

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。

比如，Mankiw, Romer and Weil(1992)以跨国数据研究经济发展的决定因素，同时汇报了非石油输出国家(全样本)与 OECD 国家(子样本)的回归结果(参见习题)。

9.8 虚 拟 变 量

如果使用**定性数据**(qualitative data)或**分类数据**(categorical data)，通常需要引入**虚拟变量**，即取值为 0 或 1 的变量。

比如，性别分男女，可定义“男性虚拟变量”：

$$D=\begin{cases}1, & \text{性别为男}\\ 0, & \text{性别为女}\end{cases} \tag{9.23}$$

若已定义男性虚拟变量，则无须再使用女性虚拟变量，否则将导致信息重复(因为性别非男即女)。

对于全球的五大洲，则需要四个虚拟变量，即

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

(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，则最多只能在回归方程中放入$(M-1)$个虚拟变量。

如果在回归方程中包含了M个虚拟变量，则会产生严格多重共线性。

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$(1\cdots1)'$(因为M类中必居其一)。

这种情况称为**虚拟变量陷阱**(dummy variable trap)。

由于 Stata 会自动识别严格多重共线性，这种担心已不重要。

如果模型中没有常数项，则可以放入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：

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

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

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

因此，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，比如进行以下回归：

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

解决方法之二是去掉常数项，即进行如下回归：

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

在模型中引入虚拟变量，会带来什么影响呢？

考虑一个有关中国经济的时间序列模型：

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

经济结构可能在 1978 年改革开放后有变化，故引入虚拟变量：

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

根据引入虚拟变量的方式，考虑以下两种情况。

(1) 仅仅引入虚拟变量D_t本身，则回归方程为

(2)

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

根据虚拟变量D_t在不同时期的不同取值，该模型等价于

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

仅仅引入虚拟变量相当于在不同时期给予不同的截距项，参见图 9.3。

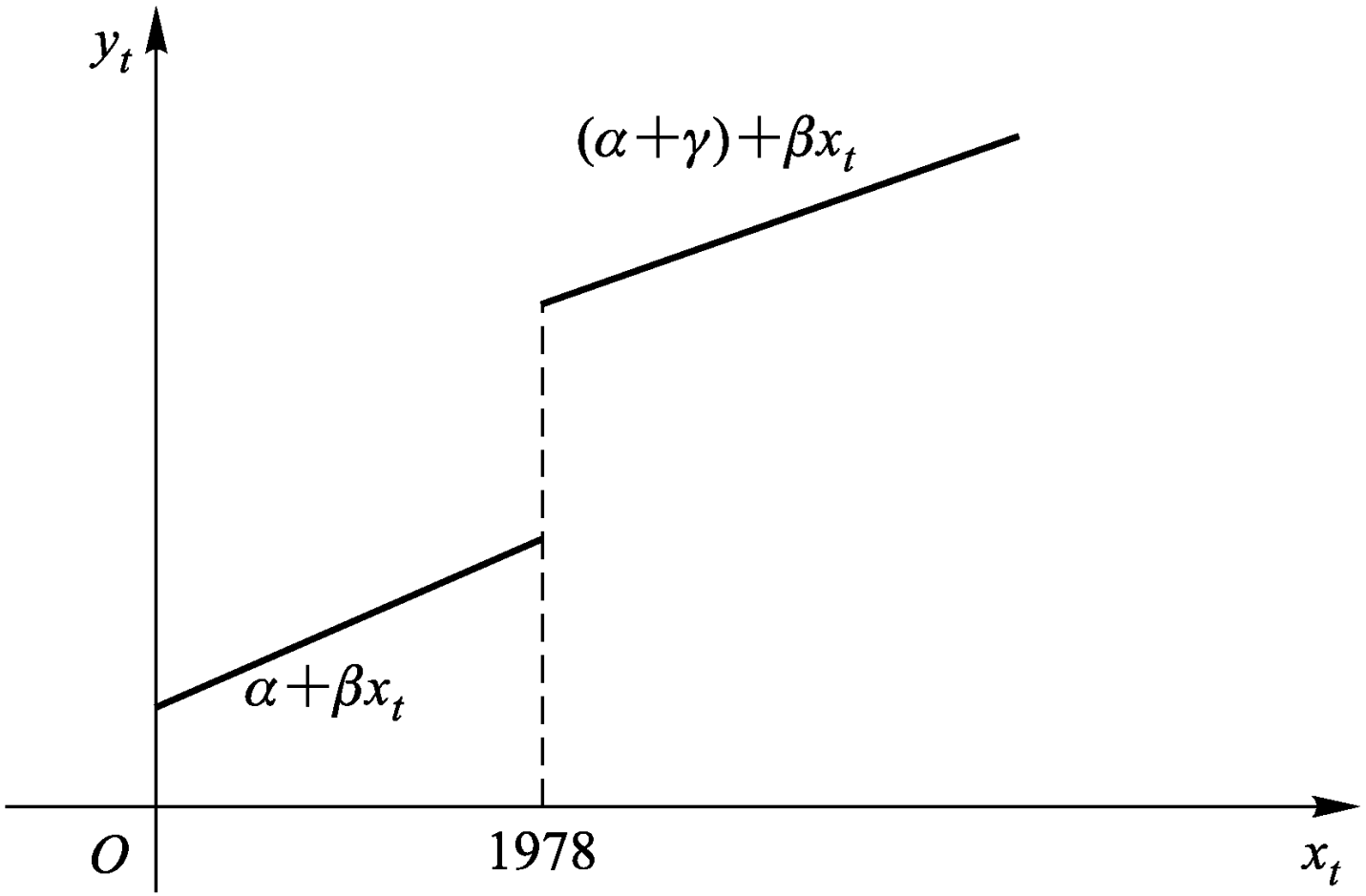

图 9.3　仅引入虚拟变量的效果

(2) 引入虚拟变量D_t，以及虚拟变量与解释变量的**交互项**(interaction term)$D_t x_t$，则回归方程为

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

该模型等价于

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

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(这种情形比较少见)。

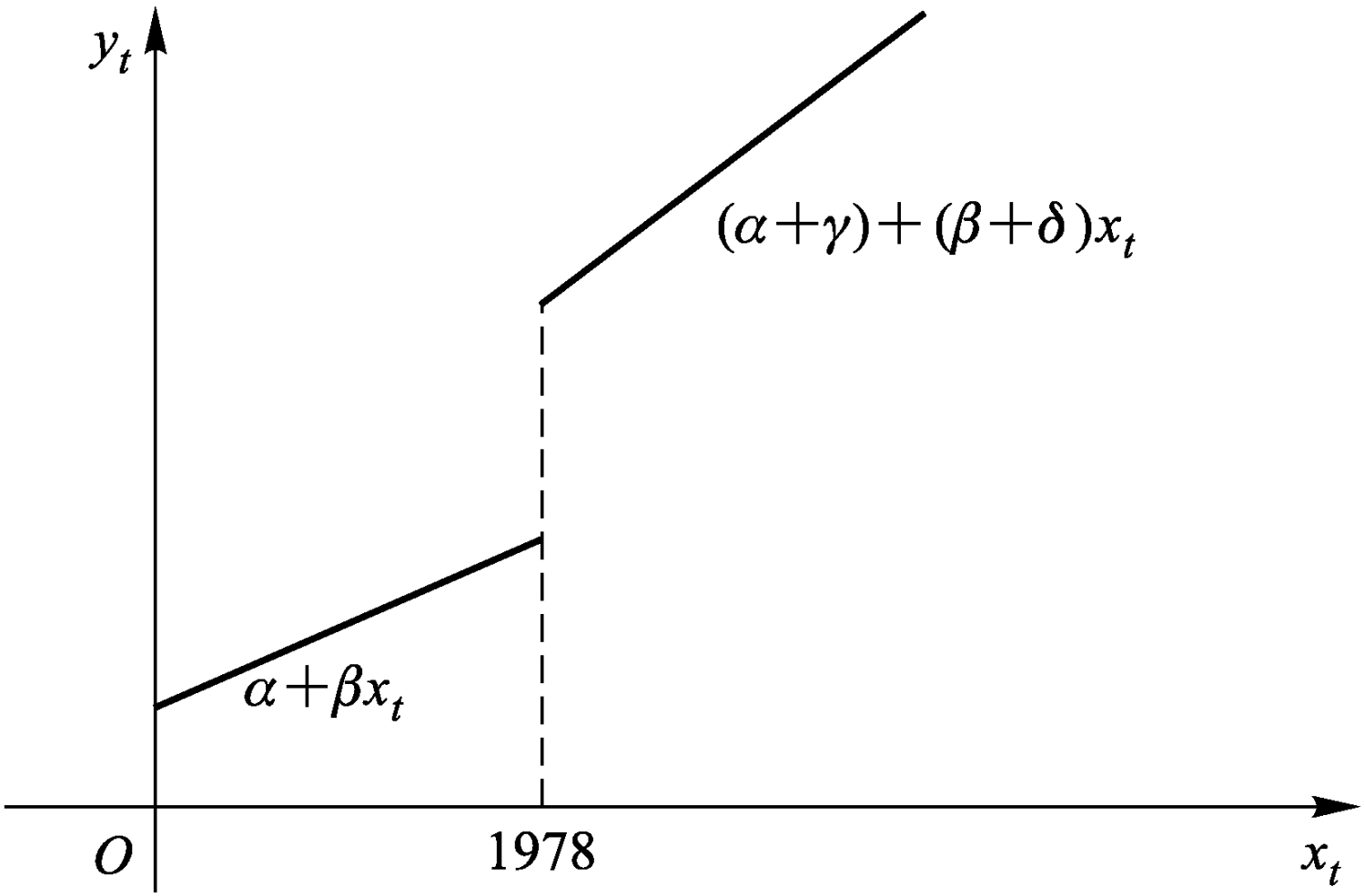

图 9.4　引入虚拟变量及其互动项的效果

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)
```

“( )”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*”，则可使用如下 Stata 命令：

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

“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。

由此生成的这些虚拟变量，将按照变量 *province* 的字母顺序而排序。

在进行回归时可使用变量的简略写法，比如：

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

其中，为了避免虚拟变量陷阱而略去了第一个省的虚拟变量 *prov1*。

与此相应的回归模型可写为

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \sum_{j=2}^{30} \delta_j \, prov_{ij} + \varepsilon_i \qquad (9.36)$$

其中， $prov_{ij}$为个体i的变量$prov_j(j = 2, \cdots, 30)$的取值。

根据上文的分析，(被略去虚拟变量而作为参照标准的)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)是重要的问题。

如果存在**结构变动**(structural break)，但未加考虑，也是一种模型设定误差。

在此，仅考虑结构变动的日期为已知的情形。

假设要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 $1950 \le t < 1978$，第 2 个时期为 $1978 \le t \le 2010$，则两个时期对应的回归方程可分别记为

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

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

需要检验的原假设为，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。

更一般地，如果有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，可对两个时期，即方程(9.37)与(9.38)，分别进行回归。

在有约束(即H_0成立)的情况下，可将模型合并为

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

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

此时，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)通过作以下三个回归来检验“无结构变动”的原假设。

首先，回归整个样本，$1950 \le t \le 2010$，得到残差平方和，记为SSR^*。

其次，回归第 1 部分子样本，$1950 \le t < 1978$，得到残差平方和SSR_1。

最后，回归第 2 部分子样本，$1978 \le t \le 2010$，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，分别进行回归则为“无约束 OLS”，其残差平方和为

$$\mathrm{SSR}=\mathrm{SSR}_1+\mathrm{SSR}_2 \qquad (9.40)$$

$\mathrm{SSR}^{*}\geq\mathrm{SSR}=\mathrm{SSR}_1+\mathrm{SSR}_2$，因为有约束 OLS 的拟合优度比无约束 OLS 更差。

如果H_0成立(无结构变动)，则$(\mathrm{SSR}^{*}-\mathrm{SSR}_1-\mathrm{SSR}_2)$应该比较小；即施加约束后，不应使得残差平方和上升很多。

反之，如果$(\mathrm{SSR}^{*}-\mathrm{SSR}_1-\mathrm{SSR}_2)$很大，则倾向于认为$H_0$不成立，即存在结构变动。

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

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

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。

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

套用上面的公式，可得检验结构变动的F统计量：

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

其中，n为样本容量，K为有约束回归的参数个数(含常数项)。对于此一元回归的例子，$K=2$。

检验结构变动的另一简便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。

比如，对于$K=2$的情形，可进行如下回归：

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

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

此检验所得 F 统计量与传统的邹检验完全相同。

虚拟变量法与邹检验是等价的。

与传统的邹检验相比，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)邹检验是在“球形扰动项”(同方差、无自相关)的假设下得到的，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)如果发现存在结构变动，邹检验并不提供究竟是截距项还是斜率变动的信息(至少需要再作一个邹检验)，而虚拟变量法则可同时提供这些信息。

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。数据来自国家统计局网站。

先看一下中国 1978—2013 年“居民人均消费”(*c*)与“人均国内生产总值”(*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)
```

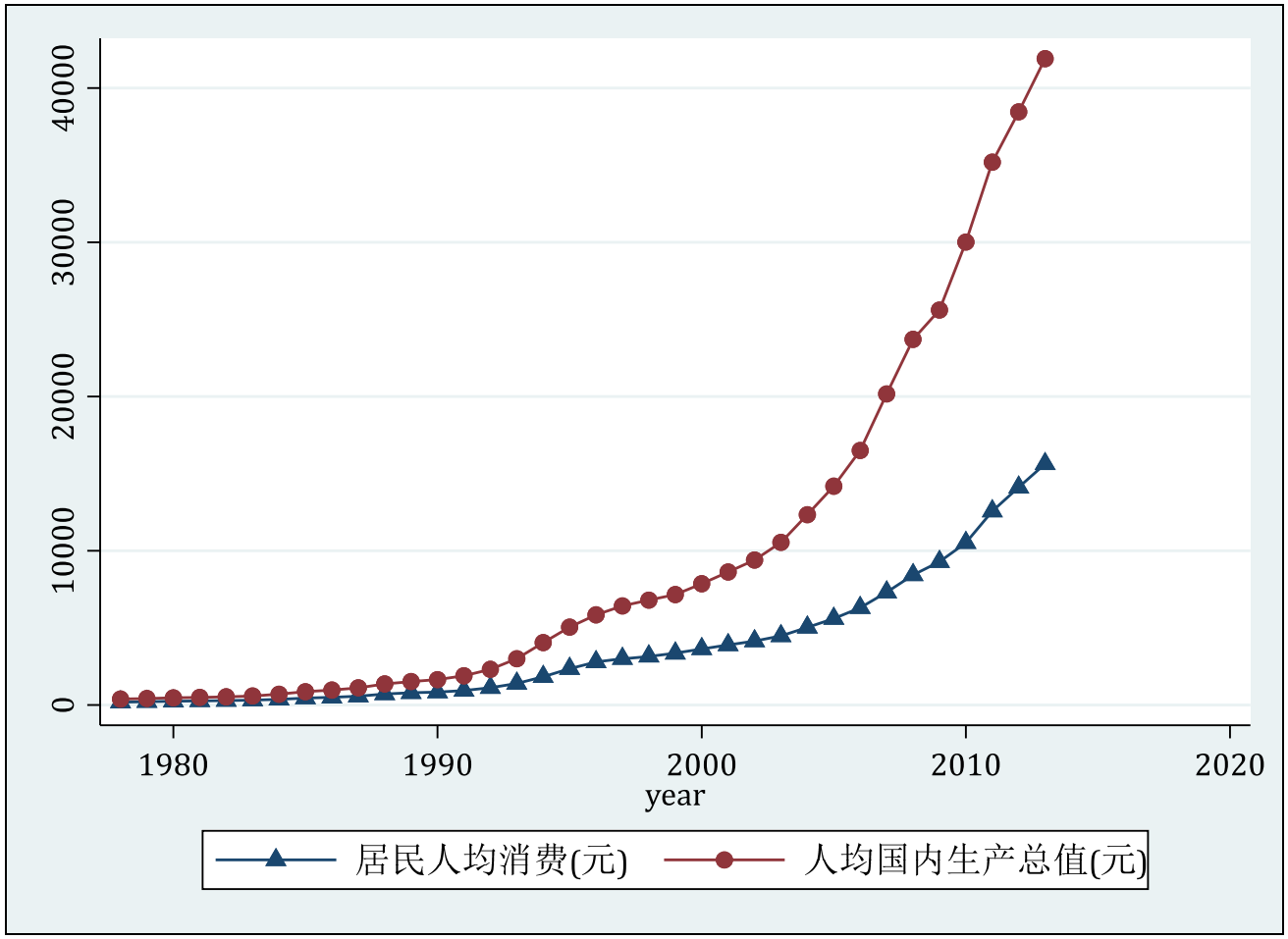

图 9.5　居民人均消费与人均国内生产总值的时间趋势

从图 9.5 可知，二者的走势具有较强相关性。但图 9.5 的右边有些空白区域，不够美观。

将上述命令略加改进，并在 1992 年处画一条垂直线(结果见图 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。

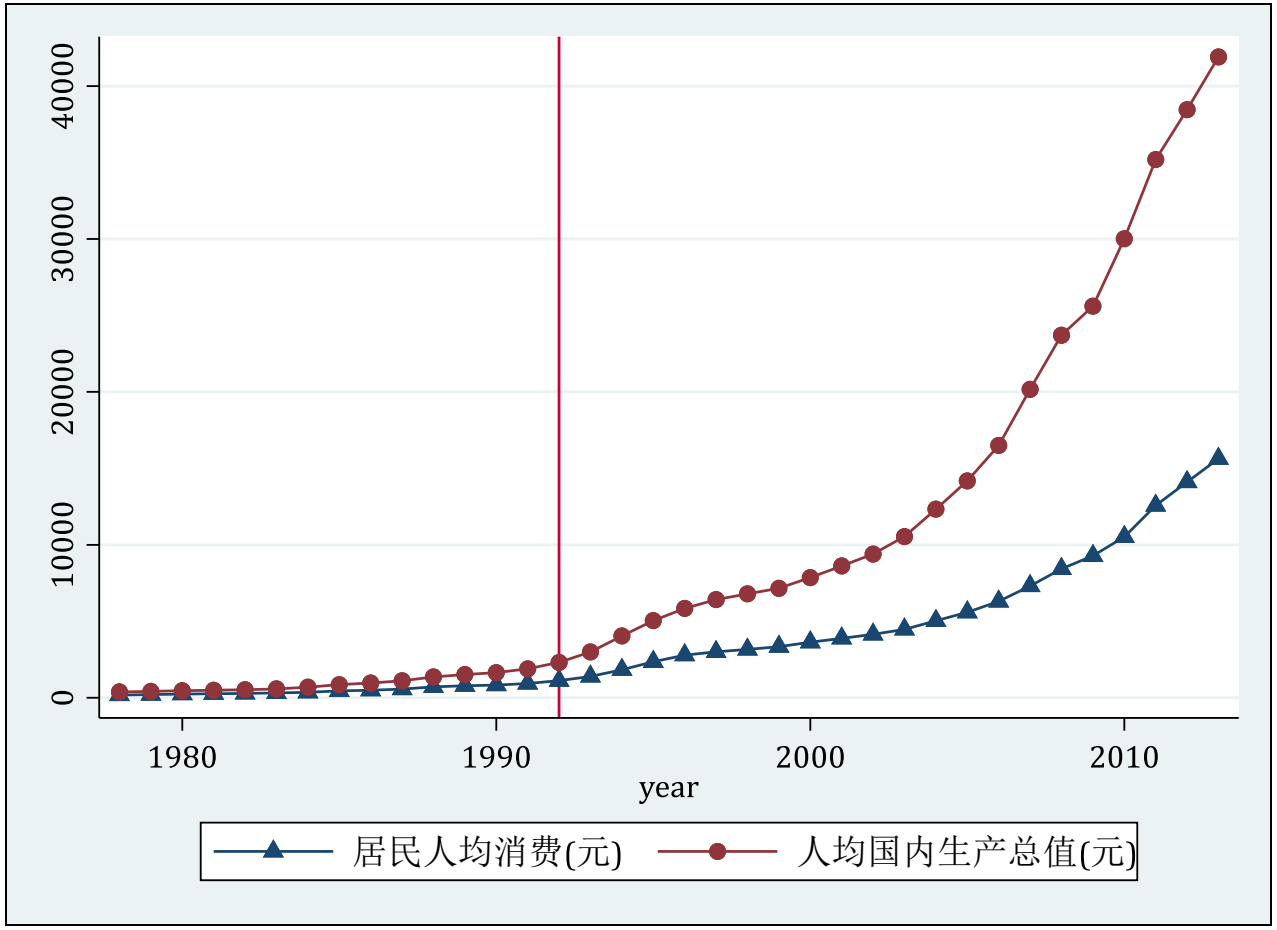
40000
30000
20000
10000
0
1980
1990
2000
2010
year
居民人均消费(元)
人均国内生产总值(元)

图 9.6　改进的趋势图

考察一个简单(粗糙)的消费函数：

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

首先，使用传统的邹检验(F 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，以获得其残差平方和：

```
. reg c y
```

Source	SS	df	MS	Number of obs	=	36
				F(1, 34)	=	7139.56
Model	617812224	1	617812224	Prob > F	=	0.0000
Residual	2942143.12	34	86533.6213	R-squared	=	0.9953
				Adj R-squared	=	0.9951
Total	620754367	35	17735839.1	Root MSE	=	294.17

c	Coefficient	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)
```

“scalar”表示标量，即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
				F(1, 12)	=	4344.64
Model	829125.648	1	829125.648	Prob > F	=	0.0000
Residual	2290.06599	12	190.838833	R-squared	=	0.9972
				Adj R-squared	=	0.9970
Total	831415.714	13	63955.0549	Root MSE	=	13.814

c	Coefficient	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
				F(1, 20)	=	4889.80
Model	366038781	1	366038781	Prob > F	=	0.0000
Residual	1497151	20	74857.5501	R-squared	=	0.9959
				Adj R-squared	=	0.9957
Total	367535932	21	17501711	Root MSE	=	273.6

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

```
. dis ((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
				F(3, 32)	=	4405.23
Model	619254926	3	206418309	Prob > F	=	0.0000
Residual	1499441.07	32	46857.5333	R-squared	=	0.9976
				Adj R-squared	=	0.9974
Total	620754367	35	17735839.1	Root MSE	=	216.47

c	Coefficient	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
```

使用虚拟变量法所得 F 统计量也为 15.39，与传统邹检验完全相同。

该检验的 p 值为 0.0000，故可在 1%水平上强烈拒绝“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(参见第 7 章与第 8 章)。

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

```
White's test
H0: Homoskedasticity
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

怀特检验的结果可在 5%的水平上拒绝“同方差”的原假设。

为了进行自相关的 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。

此模型的扰动项存在异方差与自相关，故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
```

故应将截断参数设为 3。下面进行 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	Coefficient	Newey–West std. err.	t	P>|t|	[95% conf.	interval]
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

然后，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
```

该检验的 p 值为 0.0000，故可在 1%水平上强烈拒绝“无结构变动”的原假设，即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

在现实数据中，有时会出现某些时期**数据缺失**(missing data)的情形，尤其是历史比较久远的数据。

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)的方法来补上缺失数据。

考虑最简单的情形。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

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

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)：

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

经整理可得

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

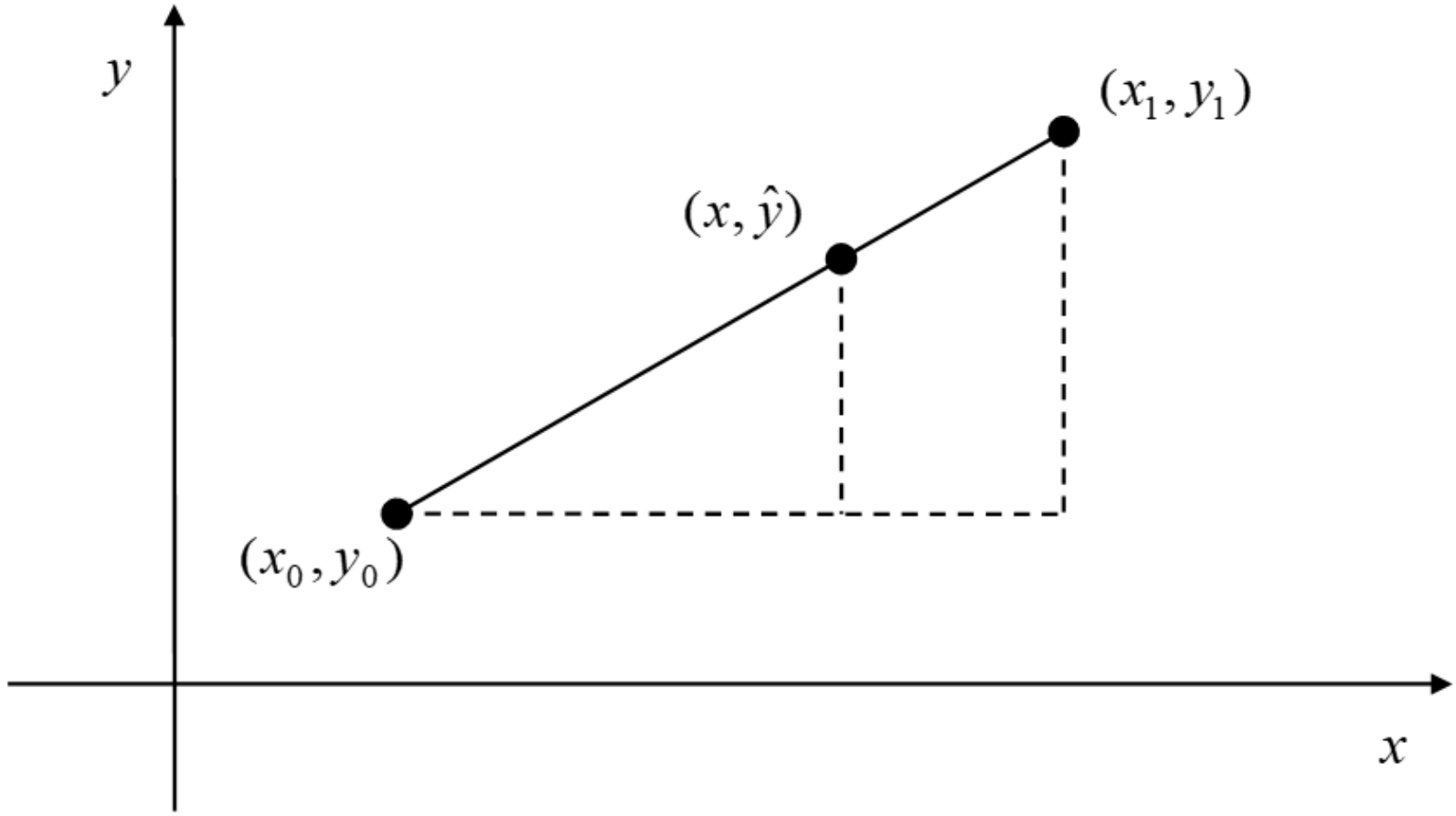

图 9.7 线性插值示意图

线性插值的基本假设是变量以线性的速度均匀地变化。

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

如果需要以原变量 y 进行回归，可将线性插值的对数值 $\widehat{\ln y}$ 再取反对数(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*。

继续以数据集 consumption.dta 为例。

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

为了演示目的，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，并记为 *y1*。

```
. gen y1=y

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

下面，直接用 *y1* 对 *year* 进行线性插值，并将结果记为 *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，并将结果记为 *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*，而且整体估计效果不如先取对数再插值的结果 *y3* (1980 年的结果是个例外)。

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 $\mathbf{X}$ 中某列的数值是另一列的很多倍，这可能使计算机在对 $(\mathbf{X}'\mathbf{X})^{-1}$ 进行数值计算时出现较大误差。

由于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。