

第 4 章 一元线性回归

4.1 一元线性回归模型

为什么在青少年时期要选择上学？

除了满足好奇心、求知欲及个人成长外，一个重要原因是教育能提高未来的收入水平。

如何从理论上解释教育投资的回报率(returns to schooling)?

Mincer (1958)提出基于效用最大化的理性选择模型：

个体选择多上一年学，则需推迟一年挣钱(另需交学费)；为弥补其损失，市场均衡条件要求给予受教育多者更高的未来收入。

由此可得工资对数与教育年限的线性关系：

$$\ln w = \alpha + \beta s \quad (4.1)$$

$\ln w$ 为工资对数， s 为教育年限(schooling)，而 α 与 β 为参数。

α 为截距项，表示当教育年限为 0 时的工资对数水平，因为 $\ln w = \alpha + \beta \cdot 0 = \alpha$ 。

β 为斜率，表示教育年限对工资对数的边际效应，即每增加一年教育，将使工资增加百分之几，因为对方程(4.1)两边求导可得

$$\beta = \frac{d \ln w}{ds} = \frac{\frac{dw}{w}}{\frac{ds}{s}} \approx \frac{\frac{\Delta w}{w}}{\frac{\Delta s}{s}} \quad (4.2)$$

教育年限只是影响工资的因素之一。严格来说，方程(4.1)应为

$$\ln w = \alpha + \beta s + \text{其他因素} \quad (4.3)$$

将其他因素记为 ε ，则有

$$\ln w = \alpha + \beta s + \varepsilon \quad (4.4)$$

方程(4.4)即劳动经济学(labor economics)中著名的明瑟方程(the Mincer equation)的基本形式(Mincer, 1974)。

但多上一年学，究竟能使未来收入提高百分之几？

这取决于参数 β 的取值。明瑟模型并未提供关于 α 与 β 具体取值的信息。

对于这种定量问题(quantitative question)，只有通过数据才能给出定量回答(quantitative answer)。

需要用计量经济学方法，通过样本数据来估计未知参数 α 与 β 。

明瑟模型推断工资对数与教育年限为线性关系，此预言是否与现实数据相符？

使用数据集 `grilic.dta` 来考察，此数据集包括 758 位美国年轻男子的教育投资回报率数据。

先看此数据集的变量 `s` 与 `lnw` 的前 10 个观测值

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

```
. list s lnw in 1/10
```

	s	lnw
1.	12	5.9
2.	16	5.438
3.	14	5.71
4.	12	5.481
5.	9	5.927
6.	9	4.804
7.	18	6.512
8.	15	5.808
9.	12	5.737
10.	18	6.382

为了考察工资对数与教育年限的关系，画二者的散点图，并在图上画出离这些样本点最近的“回归直线”，参见图 4.1。

```
. twoway scatter lnw s || lfit lnw s
```

其中，“lfit”表示“linear fit”，即线性拟合。

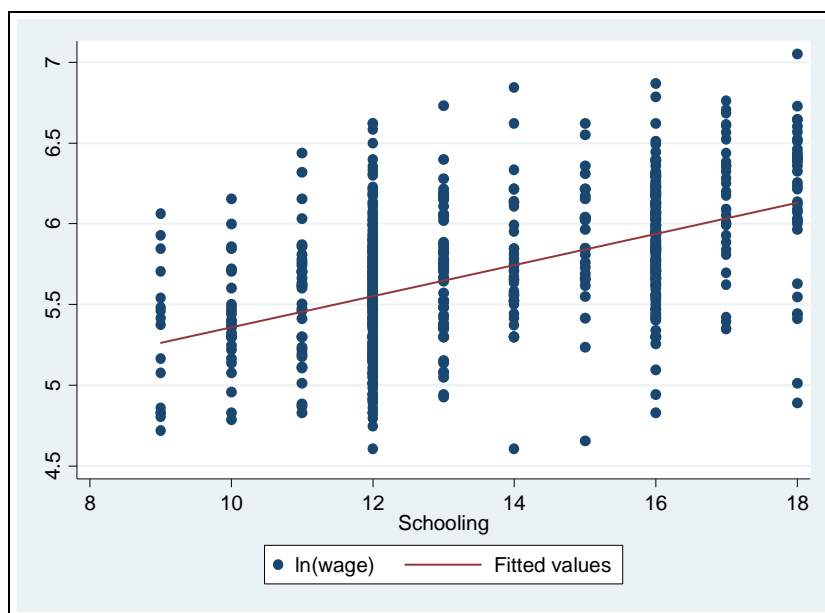


图 4.1 工资对数与教育年限的散点图与线性拟合

工资对数与教育年限正相关，似乎存在线性关系，在形式上与明瑟方程相一致。

更一般地，假设从总体随机抽取 n 位个体，则一元线性回归模型可写为

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (i = 1, \dots, n) \quad (4.5)$$

y_i : 被解释变量(dependent variable, regressand)

x_i : 解释变量(explanatory variable, independent variable, regressor)

α : 截距项(intercept)或常数项(constant); β : 斜率(slope)

α 与 β : 统称“回归系数”(regression coefficients)或“参数”(parameters)。

ε_i : “误差项” (error term)或“扰动项” (disturbance), 包括遗漏的其他因素、变量的测量误差、回归函数的设定误差(比如, 忽略了非线性项)以及人类行为的内在随机性等。

除 x_i 以外, 影响 y_i 的所有其他因素都在 ε_i 中。

下标 i 表示个体 i , 比如第 i 个人, 第 i 个企业, 第 i 个国家等。

i 的取值为 $1, \dots, n$, 其中 n 为“样本容量” (sample size)。

方程 (4.5) 右边的确定性部分为 $\alpha + \beta x_i$, 称为总体回归线 (population regression line) 或总体回归函数 (population regression function, 简记 PRF)。

方程(4.5)假设总体回归函数为线性，可视为一阶近似(忽略二次项及高阶项)。

模型 $y_i = \alpha + \beta x_i + \varepsilon_i$ 也称为数据生成过程 (Data Generation Process, 简记 DGP), 参见图 4.2。

从数据生成的角度来看，随机变量 x_i 与 ε_i 首先从相应的概率分布中抽取观测值(observation)。

确定 x_i 与 ε_i 的取值后，根据方程 $y_i = \alpha + \beta x_i + \varepsilon_i$ 生成 y_i 的取值。

由于 ε_i 通常无法观测(unobservable)，故研究者只知道 (x_i, y_i) 。

计量经济学的主要任务之一就是通过数据 $\{x_i, y_i\}_{i=1}^n$ 来获取关于总体参数 (α, β) 的信息。

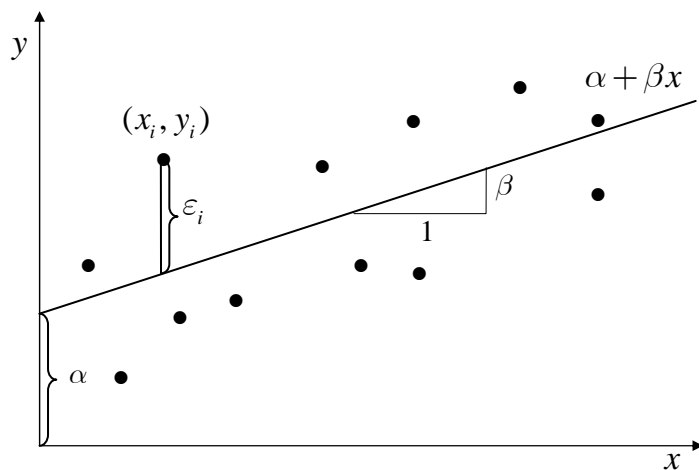


图 4.2 数据生成过程

4.2 OLS 估计量的推导

如何根据观测值 $\{x_i, y_i\}_{i=1}^n$ 来估计总体回归直线 $\alpha + \beta x_i$?

希望在 (x, y) 平面上找到一条直线, 使得此直线离所有这些点(观测值)最近, 参见图 4.3。

在此平面上, 任意给定一条直线, $y_i = \hat{\alpha} + \hat{\beta}x_i$ (其中, $\hat{\alpha}$ 读为 alpha hat, $\hat{\beta}$ 读为 beta hat), 计算每个点(观测值)到这条线的距离, $e_i \equiv y_i - \hat{\alpha} - \hat{\beta}x_i$, 称为“残差”(residual)。

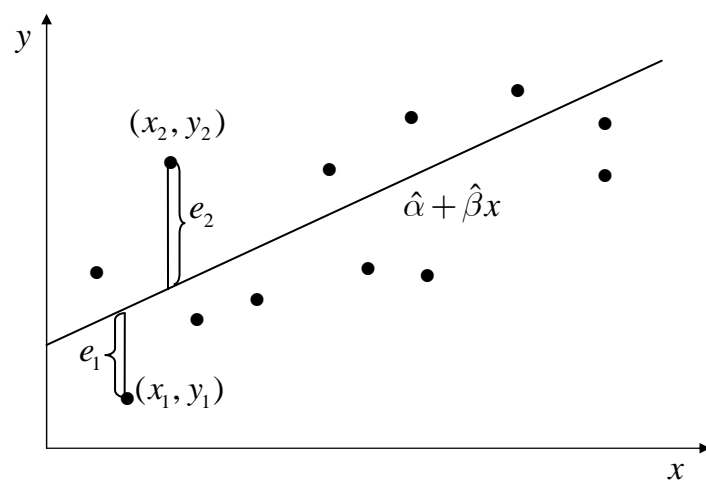


图 4.3 残差平方和最小化

如直接把残差加起来， $\sum_{i=1}^n e_i$ ，会出现正负相抵的现象。解决方法之一为使用绝对值，即 $\sum_{i=1}^n |e_i| = \sum_{i=1}^n |y_i - \hat{\alpha} - \hat{\beta}x_i|$ 。

但绝对值不易运算(无法微分), 故考虑其平方,

$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$, 称为“残差平方和”(Sum of Squared Residuals, 简记 SSR; 或 Residual Sum of Squares, 简记 RSS)。

“普通最小二乘法”(Ordinary Least Squares, 简记 OLS)就是选择 $\hat{\alpha}, \hat{\beta}$, 使得残差平方和最小化。可将 OLS 的目标函数写为

$$\min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \quad (4.6)$$

OLS 是线性回归模型的基本估计方法。

此最小化问题的一阶条件为

$$\begin{cases} \frac{\partial}{\partial \hat{\alpha}} \sum_{i=1}^n e_i^2 = -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \\ \frac{\partial}{\partial \hat{\beta}} \sum_{i=1}^n e_i^2 = -2 \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta} x_i) x_i = 0 \end{cases} \quad (4.7)$$

消去方程左边的“-2”可得

$$\left\{ \begin{array}{l} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \\ \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)x_i = 0 \end{array} \right. \quad (4.8)$$

对上式各项分别求和，移项可得

$$\left\{ \begin{array}{l} n\hat{\alpha} + \hat{\beta}\sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ \hat{\alpha}\sum_{i=1}^n x_i + \hat{\beta}\sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{array} \right. \quad (4.9)$$

这是有关估计量 $\hat{\alpha}, \hat{\beta}$ 的二元一次线性方程组，称为“正规方程组”(normal equations)。从方程组(4.9)的第 1 个方程可得

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad (4.10)$$

其中， $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ ， $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ 。

将表达式(4.10)代入方程组(4.9)的第 2 个方程可得

$$(\bar{y} - \hat{\beta}\bar{x}) \sum_{i=1}^n x_i + \hat{\beta} \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \quad (4.11)$$

合并同类项，移项可得：

$$\hat{\beta} \left(\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i \right) = \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i \quad (4.12)$$

使用关系式 $\sum_{i=1}^n x_i = n\bar{x}$ ，求解 $\hat{\beta}$ ：

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \quad (4.13)$$

上式可写为更直观的离差形式：

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (4.14)$$

OLS 估计量要有定义，必须分母 $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$ 。

解释变量 x_i 应有所变动，不能是常数，是对数据的最基本要求。

如果 x_i 没有任何变化，则相同的 x_i 取值将对应于不同的 y_i 取值，无法估计 x 对 y 的作用，参见图 4.4。

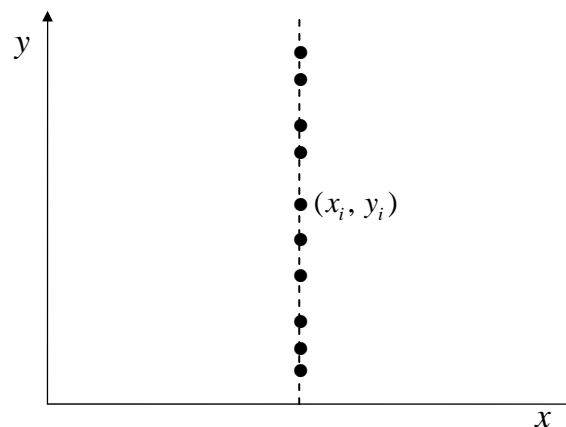


图 4.4 解释变量 x 没有变化的情形

根据方程(4.10)与(4.14)，可求解 OLS 估计量 $\hat{\alpha}, \hat{\beta}$ ，得到 $\hat{y} \equiv \hat{\alpha} + \hat{\beta}x_i$ ，称为样本回归线(sample regression line)或样本回归函数(sample regression function，简记 SRF)，参见图 4.5。

从方程(4.10)可知， $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$ ，即样本回归线一定经过 (\bar{x}, \bar{y}) 。

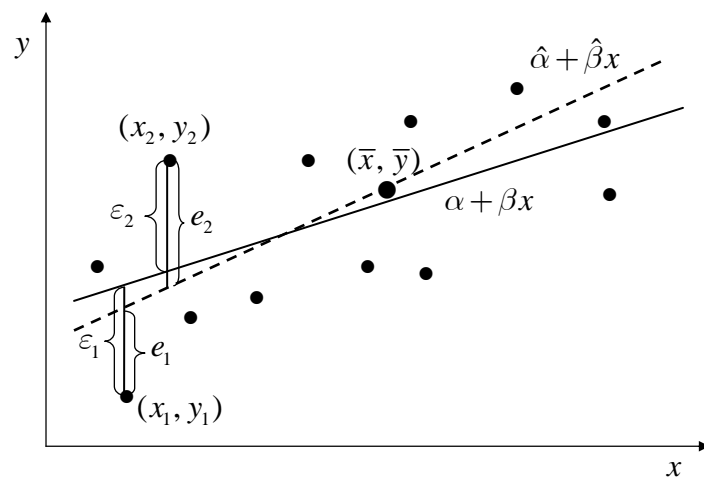


图 4.5 总体回归线与样本回归线

4.3 OLS 的正交性

定义被解释变量 y_i 的“拟合值”(fitted value)或“预测值”(predicted value)为

$$\hat{y}_i \equiv \hat{\alpha} + \hat{\beta} x_i \quad (4.15)$$

可将残差写为

$$e_i = y - (\hat{\alpha} + \hat{\beta}x_i) = y_i - \hat{y}_i \quad (4.16)$$

根据正规方程组(4.8):

$$\begin{cases} \sum_{i=1}^n e_i = 0 \\ \sum_{i=1}^n x_i e_i = 0 \end{cases} \quad (4.17)$$

写为向量内积的形式:

$$(1 \quad \cdots \quad 1) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = 0, \quad (x_1 \quad \cdots \quad x_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = 0 \quad (4.18)$$

定义常数向量、残差向量、解释向量以及拟合值向量为

$$\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}_{n \times 1}, \quad \mathbf{e} \equiv \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} \quad (4.19)$$

则方程(4.18)可写为

$$\mathbf{1}'\mathbf{e} = 0, \quad \mathbf{x}'\mathbf{e} = 0 \quad (4.20)$$

故残差向量 \mathbf{e} 与常数向量 $\mathbf{1}$ 正交，而且 \mathbf{e} 也与解释向量 \mathbf{x} 正交。

将常数项视为取值都为 1 的解释变量，而 α 为此变量的系数。
残差向量与所有解释变量(包括 $\mathbf{1}$ 与 \mathbf{x})正交。

残差向量 \mathbf{e} 也与拟合值向量 $\hat{\mathbf{y}}$ 正交，因为

$$\hat{\mathbf{y}}'\mathbf{e} \equiv (\hat{y}_1 \quad \cdots \quad \hat{y}_n) \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \sum_{i=1}^n \hat{y}_i e_i = \sum_{i=1}^n (\hat{\alpha} + \hat{\beta} x_i) e_i = \hat{\alpha} \underbrace{\sum_{i=1}^n e_i}_{=0} + \hat{\beta} \underbrace{\sum_{i=1}^n x_i e_i}_{=0} = 0$$

(4.21)

OLS 残差与解释变量及拟合值的正交性是 OLS 的重要特征，为推导证明提供了方便。

比如，考虑方程(4.16)， $e_i = y_i - \hat{y}_i$ ，将两边对 i 加总，并除以 n 可得：

$$0 = \frac{1}{n} \sum_{i=1}^n e_i = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{y} - \bar{\hat{y}} \quad (4.22)$$

其中， $\bar{\hat{y}} \equiv \frac{1}{n} \sum_{i=1}^n \hat{y}_i$ 。故被解释变量的均值恰好等于拟合值的均

值，即

$$\bar{y} = \bar{\hat{y}} \quad (4.23)$$

4.4 平方和分解公式

被解释变量可分解为相互正交的两个部分，即

$$y_i = \hat{y}_i + e_i \quad (4.24)$$

如回归方程有常数项(通常都有)，则被解释变量的离差平方和 $\sum_{i=1}^n (y_i - \bar{y})^2$ (Total Sum of Squares, TSS)可分解为

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^n e_i^2}_{\text{RSS}} \quad (4.25)$$

方程(4.25)称为“平方和分解公式”，将 $\sum_{i=1}^n (y_i - \bar{y})^2$ 分解为两部分。

右边第一项为 $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ ，由于 $\bar{y} = \bar{\hat{y}}$ (被解释变量的均值等于拟合值的均值)，故可写为 $\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$ ，即可由模型解释的部分，称为 Explained Sum of Squares (ESS)。

右边第二项为残差平方和 $\sum_{i=1}^n e_i^2$ (Residual Sum of Squares, RSS)，是模型所无法解释的部分。

平方和分解公式能够成立，正是由于 OLS 的正交性。

证明：将离差 $(y_i - \bar{y})$ 写为 $(y_i - \hat{y}_i + \hat{y}_i - \bar{y})$ ，则可将 TSS 写为

$$\begin{aligned}\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (e_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^n e_i^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + 2 \sum_{i=1}^n e_i (\hat{y}_i - \bar{y})\end{aligned}\quad (4.26)$$

只需证明交叉项 $\sum_{i=1}^n e_i (\hat{y}_i - \bar{y}) = 0$ 即可，而这由 OLS 的正交性

所保证：

$$\sum_{i=1}^n e_i (\hat{y}_i - \bar{y}) = \sum_{i=1}^n e_i \hat{y}_i - \sum_{i=1}^n e_i = 0 - 0 = 0 \quad (4.27)$$

如没有常数项，则无法保证 $\sum_{i=1}^n e_i = 0$ ，故平方和分解公式不成立。

4.5 拟合优度

OLS 的样本回归线为离所有样本点最近的直线。

但究竟离这些样本点有多近？

希望有绝对的度量，以衡量样本回归线对数据的拟合优良程度。

在有常数项的情况下，根据平方和分解公式，可将被解释变量的离差平方和分解为模型可以解释与不可解释的部分。

如果模型可以解释的部分所占比重越大，则样本回归线的拟合程度越好。

定义 拟合优度(goodness of fit) R^2 为

$$R^2 \equiv \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4.28)$$

拟合优度 R^2 也称可决系数(coefficient of determination)。

在有常数项的情况下, 拟合优度等于被解释变量 y_i 与拟合值 \hat{y}_i 之间相关系数的平方, 即 $R^2 = [\text{Corr}(y_i, \hat{y}_i)]^2$, 故记为 R^2 。

显然, $0 \leq R^2 \leq 1$ 。

R^2 越高, 则样本回归线对数据的拟合程度越好。

如果 $R^2=1$ ，则解释变量 x 可以完全解释 y 的变动。此时，残差平方和 $\sum_{i=1}^n e_i^2 = 0$ ，所有残差均为 0，故所有样本点都在样本回归线上。

如果 $R^2=0$ ，则解释变量 x 对于解释 y 没有任何帮助。此时， $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 0$ ，故对于任何个体 i ，都有 $\hat{y}_i \equiv \bar{y}$ ；故样本回归线为水平线，与 x 轴平行。这意味着 $\hat{\beta}=0$ ，无论 x 如何变动，对 y 都没有影响。

如果 $0 < R^2 < 1$ ，则为介于以上两种极端的中间情形，即 x 可以解释 y 的一部分，但无法解释其余部分。

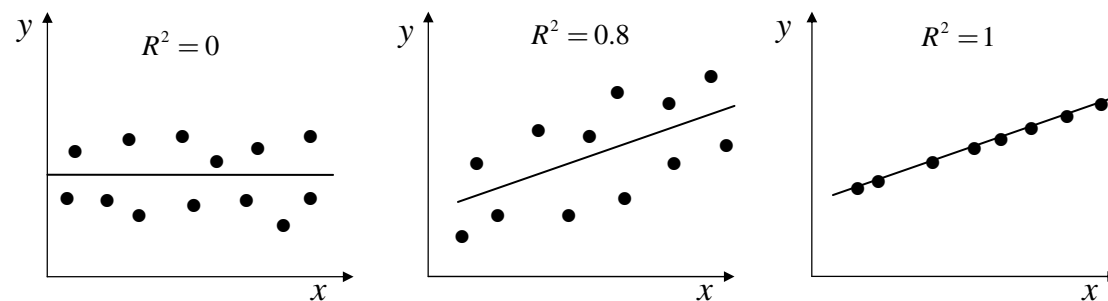


图 4.6 拟合优度的三种情形示意图

R^2 只是反映拟合程度的好坏，除此外并无太多意义。

评估回归方程是否显著，应使用 F 检验(R^2 与 F 统计量也有联系)。

4.6 无常数项的回归

偶尔也进行无常数项的回归，或许是经济理论的要求，也可能在模型变换时消去常数项。

无常数项的回归必然经过原点，也称为“经过原点的回归”(regression through the origin)。

此时，一元线性回归模型可写为

$$y_i = \beta x_i + \varepsilon_i \quad (i = 1, \dots, n) \quad (4.29)$$

依然进行 OLS 估计，最小化残差平方和：

$$\min_{\hat{\beta}} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}x_i)^2 \quad (4.30)$$

一阶条件为

$$\frac{d}{d\hat{\beta}} \sum_{i=1}^n e_i^2 = -2 \sum_{i=1}^n (y_i - \hat{\beta}x_i)x_i = 0 \quad (4.31)$$

消去方程左边的“-2”可得

$$\sum_{i=1}^n (y_i - \hat{\beta}x_i)x_i = 0 \quad (4.32)$$

求解 $\hat{\beta}$ 可得

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \quad (4.33)$$

方程(4.33)与有常数项回归的表达式类似。

如果回归模型无常数项,则平方和分解公式不成立,不宜使用 R^2 来度量拟合优度。

但即使没有常数项, OLS 也仍满足正交性, 因为正规方程(组)(4.32)的表达式基本不变。

记 $e_i \equiv y_i - \hat{\beta}x_i$, 则正规方程(4.32)可写为

$$\sum_{i=1}^n x_i e_i = 0 \quad (4.34)$$

记拟合值 $\hat{y}_i \equiv \hat{\beta}x_i$, 容易证明残差仍与拟合值正交:

$$\sum_{i=1}^n \hat{y}_i e_i = \sum_{i=1}^n \hat{\beta}x_i e_i = \hat{\beta} \sum_{i=1}^n x_i e_i = \hat{\beta} \cdot 0 = 0 \quad (4.35)$$

仍可利用 OLS 的正交性将 $\sum_{i=1}^n y_i^2$ 分解为:

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n (\hat{y}_i + e_i)^2 = \sum_{i=1}^n \hat{y}_i^2 + 2 \underbrace{\sum_{i=1}^n \hat{y}_i e_i}_{=0} + \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \hat{y}_i^2 + \sum_{i=1}^n e_i^2 \quad (4.36)$$

$\sum_{i=1}^n \hat{y}_i^2$ 为可由模型解释的部分，而 $\sum_{i=1}^n e_i^2$ 为模型不可解释的部分。

定义非中心 R^2 (uncentered R^2):

$$R_{uc}^2 \equiv \frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} \quad (4.37)$$

如果无常数项，Stata 汇报的 R^2 正是 R_{uc}^2 。

R_{uc}^2 与 R^2 的定义不同，二者不具有可比性，但在 Stata 中都称为“R-squared” (在无常数项回归时汇报 R_{uc}^2)。

4.7 一元回归的 Stata 实例

在 Stata 中，进行一元回归的命令为

```
. regress y x, noconstant
```

其中，“y”为被解释变量，“x”为解释变量，选择项“noconstant”表示无常数项(默认有常数项)。

使用数据集 `grilic.dta`，将工资对数(`lnw`)对教育年限(`s`)进行一元回归。

```
. use grilic.dta,clear
. reg lnw s
```

Source	SS	df	MS	Number of obs = 758		
Model	35.2039946	1	35.2039946	F(1, 756) = 255.70		
Residual	104.082155	756	.137674809	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.2527		
				Adj R-squared = 0.2518		
				Root MSE = .37105		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.0966245	.0060425	15.99	0.000	.0847624	.1084866
_cons	4.391486	.0821136	53.48	0.000	4.230288	4.552684

“Coef.”表示回归系数(Coefficient)，而“_cons”表示常数项(constant)。可将样本回归线写为

$$\widehat{\ln w} = 4.391 + 0.097 s \quad (4.38)$$

$\widehat{\ln w}$ 表示被解释变量 $\ln w$ 的拟合值或预测值，而 $\hat{\alpha}=4.391$ ， $\hat{\beta}=0.097$ 。

教育投资的回报率为 9.7%，即每增加一年教育，平均可提高收入 9.7%。

上表左上部显示，TSS (Total)为 139.28615，其中可解释部分 ESS (Model)为 35.2039946，而不可解释部分 RSS (Residual)为 104.082155。

上表右上部显示， R^2 (R-squared)为 0.2527，即教育年限约可解释工资对数 25%的变动。其他统计指标将在第 5 章介绍。

如想进行无常数项的回归，可输入命令：

```
. reg lnw s,noc
```

Source	SS	df	MS	Number of obs = 758		
				F(1, 757) =36727.24		
Model	24154.3906	1	24154.3906	Prob > F = 0.0000		
Residual	497.855977	757	.657669719	R-squared = 0.9798		
				Adj R-squared = 0.9798		
Total	24652.2466	758	32.5227528	Root MSE = .81097		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.4154001	.0021676	191.64	0.000	.4111449	.4196553

无常数项回归的 R^2 高达 0.9798。

但无常数项的 R^2 与有常数项的 R^2 并不可比, 后者更可信(更具经济意义)。

教育投资回报率上升为 41.54%, 显然不合理。

在有常数项的回归中, 常数项在 1% 水平上显著不为 0, 故此例应包括常数项。

4.8 Stata 命令运行结果的存储与调用

所有 Stata 命令可分为两种，即 e-类命令(e-class commands)与 r-类命令(r-class commands)。

e-类命令为“估计命令”(estimation commands)，比如“regress”；所有其他命令为 r-类命令，比如，“summarize”。

r-类命令的运行结果都存储在“r()”，可以通过输入命令“return list”来显示，比如

```
. use grilic.dta,clear  
. sum s
```

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

. return list

```

scalars:
           r(N) = 758
      r(sum_w) = 758
      r(mean) = 13.40501319261214
      r(Var) = 4.981058057949899
      r(sd) = 2.231828411403955
      r(min) = 9
      r(max) = 18
      r(sum) = 10161

```

上表列出了在运行命令“sum s”之后，Stata 所存储的结果，其中包括未显示的“r(Var)” (方差)、“r(sum)” (求和)等。

可调用这些结果来作进一步计算。比如，为了计算“变异系数” (coefficient of variation, 即标准差除以平均值)，可使用命令：

```
. display r(sd)/r(mean)
```

```
.16649207
```

另一方面，e-类命令的运行结果都存储在“e()”，可以通过输入命令“ereturn list”来显示，比如，

```
. reg lnw s
```

Source	SS	df	MS	Number of obs = 758		
Model	35.2039946	1	35.2039946	F(1, 756) = 255.70		
Residual	104.082155	756	.137674809	Prob > F = 0.0000		
Total	139.28615	757	.183997556	R-squared = 0.2527		
				Adj R-squared = 0.2518		
				Root MSE = .37105		
lnw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
s	.0966245	.0060425	15.99	0.000	.0847624	.1084866
_cons	4.391486	.0821136	53.48	0.000	4.230288	4.552684

. ereturn list

```
scalars:
      e(N) = 758
      e(df_m) = 1
      e(df_r) = 756
      e(F) = 255.7039662336329
      e(r2) = .2527458374860054
      e(rmse) = .3710455612613365
      e(mss) = 35.20399459202199
      e(rss) = 104.0821552499956
      e(r2_a) = .2517574060541087
      e(ll) = -323.0498302841153
      e(ll_0) = -433.4714451849197
      e(rank) = 2

macros:
      e(cmdline) : "regress lnw s"
      e(title) : "Linear regression"
      e(marginsok) : "XB default"
      e(vce) : "ols"
      e(depvar) : "lnw"
      e(cmd) : "regress"
      e(properties) : "b V"
      e(predict) : "regres_p"
      e(model) : "ols"
      e(estat_cmd) : "regress_estat"

matrices:
      e(b) : 1 x 2
      e(V) : 2 x 2

functions:
      e(sample)
```

上表列出了运行命令 `reg` 后 Stata 存储的结果，包括标量 (scalars)、宏(macros)、矩阵(matrices，即系数矩阵 $e(b)$ 与协方差矩阵 $e(V)$)，以及函数(functions)。

4.9 总体回归函数与样本回归函数：蒙特卡罗模拟

为直观理解总体回归函数(PRF)与样本回归函数的关系(SRF)，使用蒙特卡罗法进行模拟。

所谓“蒙特卡罗法”(Monte Carlo Methods, MC)，是通过计算机模拟，从总体抽取大量随机样本的计算方法。

考虑如下数据生成过程(DGP)或总体回归模型:

$$y_i = 1 + 2x_i + \varepsilon_i \quad (i = 1, \dots, 30) \quad (4.39)$$

解释变量 $x_i \sim N(3, 2^2)$, 扰动项 $\varepsilon_i \sim N(0, 3^2)$, 样本容量为 $n = 30$ 。

从 $N(3, 2^2)$ 随机抽取 30 个解释变量 x_i 的观测值, 并从 $N(0, 3^2)$ 随机抽取 30 个扰动项 ε_i 的观测值。

根据总体回归模型(4.39)计算相应的被解释变量 y_i 。

把 y_i 对 x_i 进行回归, 得到样本回归函数(SRF), 并与总体回归函数(PRF)进行比较。

在 Stata 命令窗口依次输入如下命令：

```
. clear                (删除内存中已有数据)
. set obs 30            (确定随机抽样的样本容量为 30)
. set seed 10101        (指定随机抽样的“种子”为 10101)
. gen x = rnormal(3,4)   (得到服从  $N(3, 2^2)$  分布的随机样本，记为  $x$ )
. gen e = rnormal(0,9)   (得到服从  $N(0, 3^2)$  分布的随机样本，记为  $e$ )
. gen y = 1 + 2*x + e    (计算被解释变量  $y$ )
. reg y x               (把  $y$  对  $x$  进行 OLS 回归)
```

其中，命令“set seed 10101”用来确定随机数的初始值(称为“种子”，可任意设置，此处设为 10101)，以便再次模拟时得到完全一样的结果。

Source	SS	df	MS	Number of obs = 30		
Model	2362.04948	1	2362.04948	F(1, 28) = 28.26		
Residual	2340.54765	28	83.5909875	Prob > F = 0.0000		
Total	4702.59714	29	162.158522	R-squared = 0.5023		
				Adj R-squared = 0.4845		
				Root MSE = 9.1428		
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x	2.355635	.4431423	5.32	0.000	1.447899	3.26337
_cons	-1.641879	2.52587	-0.65	0.521	-6.815888	3.532131

由于样本容量仅为 30，故存在一定的抽样误差。

斜率的真实值为 2，而样本估计值为 2.36；截距项的真实值为 1，而样本估计值为-1.64，符号相反(但不显著)。

把总体回归函数、散点图与样本回归函数画在一起, 参见图 4.7。

```
. twoway function PRF=1+2*x,range(-5 15) ||  
scatter y x || lfit y x,lpattern(dash)
```

选择项 “range(-5 15)” 用于指定画图的横轴范围介于-5 与 15 之间; 默认为 0 与 1 之间, 即 “range(0 1)”。

选择项 “lpattern(dash)” 表示画虚线, 默认画实线。

实线为总体回归函数(PRF); 而虚线为样本回归线(SRF), 即被解释变量的拟合值。SRF 似乎比较接近于 PRF。

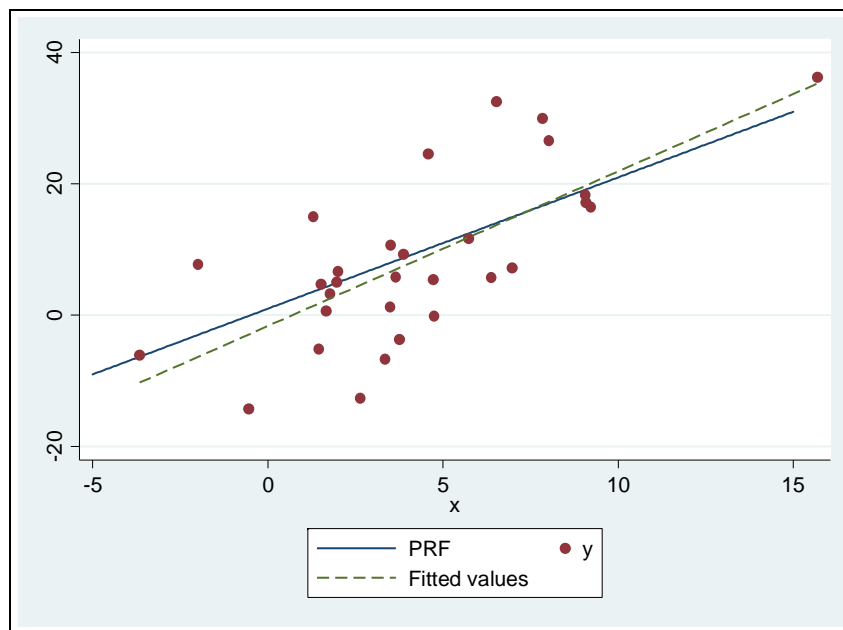


图 4.7 总体与样本回归线的蒙特卡罗模拟

如使用不同的随机数种子再次抽样, 将得到不同的 **SRF**; 而 **PRF** 始终不变。