6.2 一元线性回归

本章我们只讨论回归函数为线性函数的情形(包括能转化为线性函数的情形)——称为线性回归.我们从只含一个自变量 X(因变量总是一个,记为 Y)的情况开始,称为一元线性回归.这个情况在数学上的处理足够简单,便于对回归分析的一些概念作进一步的说明.这样,假定回归模型为

$$Y = b_0 + b_1 X + e \tag{2.1}$$

其中 b_0, b_1 为未知参数.b_0 称为常数项或截距,b_1 则称为回归系数,或更确切地,称为 Y 对 X 的回归系数.e 为随机误差,如在 6.1 节中已解释过的,假定

$$E(e) = 0, 0 < \mathrm{Var}(e) = \sigma^2 < \infty \tag{2.2}$$

误差方差 σ^2 未知,在 6.1 节中我们曾解释过这个参数的意义及其重要性.

现设对模型(2.1)中的变量 X, Y 进行了 n 次独立观察,得样本

$$(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n) \tag{2.3}$$

据(2.1),这样本的构造可由方程

$$Y_i = b_0 + b_1 X_i + e_i, i = 1, \cdots, n \tag{2.4}$$

来描述.这里 e_i,是第 i 次观察时随机误差 e 所取之值,它是不能观察的.由于各次观察独立及(2.2)对随机变量 $e_1, e_2, \cdots, e_n$,有:

$e_1, \cdots, e_n$ 独立同分布,

$$E(e_i) = 0, \mathrm{Var}(e_i) = \sigma^2, i = 1, \cdots, n \tag{2.5}$$

以后我们还将进一步要求 e_i 遵从正态分布.

(2.4)与(2.5)结合,给出了样本(2.3)的概率性质.它是对理论模型(2.1)进行统计分析推断的依据.以此之故,在统计学著作中,往往更着重(2.4)+(2.5),把它称为一元线性回归模型,而理论模型(2.1)只起一个背景的作用.当然,理解(2.4)和(2.5)是以理解(2.1)为基础的.

以上的叙述是假定,回归函数已依据某种考虑选定了——在此选为线性形式.在实际工作中,这当然是一个要研究的问题.在某种稀少的场合下,回归函数的形式可根据某种理论上的结果给出.例如,从物理学知道,在一定温度(X)的范围内,一条金属杆之长(Y)大体上为 X 的线性函数.这时选择线性回归有充分根据.在多数应用问题中,不存在这样充分的理论根据,而在很大的程度上要依靠数据本身.例如,若数据(2.3)的散点图呈图 6.1 的形状,则选取线性回归函数似是妥当的.反之,若散点图呈现图 6.2(a)或 6.2(b)的形状,则回归函数似以取为二次多项式或指数函数为宜.在实际工作中,也常使用变量变换法.即在散点图与直线趋势差距较大时,设法对自变量以至因变量进行适当的变换,使变换后的散点图更接近于直线,这样就可以对变换后的新变量进行线性回归分析,再回到原变量.在一元的情况,由于散点图可资参考,在回归函数的选择上就有较大的操作余地.对多元(多个自变量)的

情况,问题就麻烦得多,选择余地也较小.

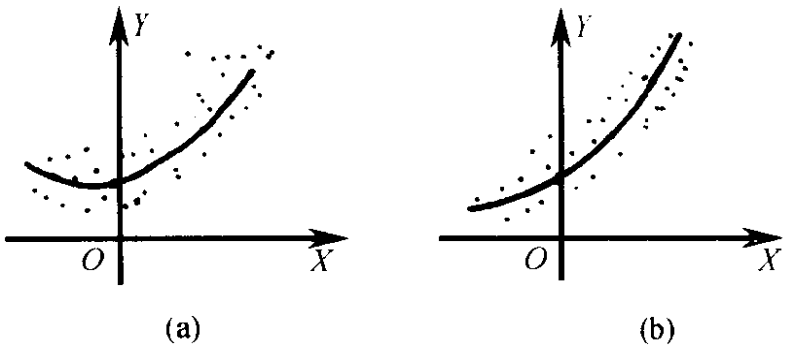

图 6.2

交代了这些之后,我们回到起先的出发点——(2.4)和(2.5).今后总用 $\overline{X}$ 和 $\overline{Y}$ 分别记 X_i 和 Y_i 的算术平均.以前我们曾指出:把自变量 X 视为非随机的,故 $X_1,\cdots,X_n$,以及 $\overline{X}$,就简单地是已知常数.因此,可以把模型(2.4)改写为

$$Y_i = \beta_0 + \beta_1(X_i - \overline{X}) + e_i, i = 1,\cdots,n \tag{2.6}$$

其关系是:

$$\beta_1 = b_1, \beta_0 = b_0 + b_1\overline{X} \tag{2.7}$$

故如估计出了 β_0 和 β_1,则由(2.7)就得到 b_0 和 b_1 的估计.改写为(2.6)的好处将在以后见到.这里注意到一点,即 β_1 后的因子 $X_i - \overline{X}$ 对 $i=1,\cdots,n$ 求和为 0.故把(2.4)改写为(2.6)有时称为模型的"中心化".

6.2.1 β_0 和 β_1 的点估计——最小二乘法

现在我们要在模型(2.6)和(2.5)之下,利用数据(2.3)去估计 β_0 和 β_1.假定我们用 α_0 和 α_1 去估计 β_0 和 β_1.我们要定出一个准则,以衡量由此所导致的偏差.我们从预测的眼光来看这个问题,如用 α_0 和 α_1,则回归函数 $\beta_0+\beta_1(x-\overline{X})$ 将用 $\alpha_0+\alpha_1(x-\overline{X})$ 去估计之.利用它在 X_i 点作预测,结果为

$$\hat{Y}_i = \alpha_0 + \alpha_1(X_i - \overline{X}), i = 1,\cdots,n \tag{2.8}$$

但我们已实际观察到：在 $X=X_i$ 处 Y 之取值为 Y_i，这样就有偏离 $Y_i-\hat{Y}_i, i=1,\cdots,n$. 我们当然希望这些偏离愈小愈好：衡量这些偏离大小的一个合理的单一指标为它们的平方和（通过平方去掉符号的影响，若简单求和，则正负偏离抵消了）：

$$Q(\alpha_0,\alpha_1)=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}[Y_i-\alpha_0-\alpha_1(X_i-\overline{X})]^2 \tag{2.9}$$

由此考虑得出以下的估计法则：找 α_0,α_1 之值，使(2.9)达到最小，以之作为 β_0,β_1 的估计. 利用多元函数求极值的方法，这只要解方程组

$$\frac{\partial Q}{\partial \alpha_0}=-2\sum_{i=1}^{n}(Y_i-\alpha_0-\alpha_1(X_i-\overline{X}))=0 \tag{2.10}$$

$$\frac{\partial Q}{\partial \alpha_1}=-2\sum_{i=1}^{n}(X_i-\overline{X})[Y_i-\alpha_0-\alpha_1(X_i-\overline{X})]=0 \tag{2.11}$$

由(2.10)解出 α_0，将解代入(2.11)，解出 α_1. 我们将这解分别记为 $\hat{\beta}_0$ 和 $\hat{\beta}_1$：

$$\hat{\beta}_0=\overline{Y} \tag{2.12}$$

$$\hat{\beta}_1=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\Big/\sum_{i=1}^{n}(X_i-\overline{X})^2$$

$$=\sum_{i=1}^{n}(X_i-\overline{X})Y_i\Big/\sum_{i=1}^{n}(X_i-\overline{X})^2 \tag{2.13}$$

“使(2.9)达到最小”这个估计方法，称为“最小二乘法”，这个重要的方法一般归功于德国大数学家高斯在 1799—1809 年间的工作*. 这个方法在数理统计学中有广泛的应用. 其好处之一在于计算简便，且如我们即将看到的，这方法导出的估计颇有些良好的性

---

* 法国数学家勒让德于 1805 年发表了这个方法. 高斯声称在 1799 年开始使用这个方法，但见诸文字是 1809 年.

质.其中之一是,如从公式(2.12)和(2.13)看到的,估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 都是 $Y_1,\cdots,Y_n$ 的线性函数,即形如 $c_{n1}Y_1+\cdots+c_{nn}Y_n$ 的函数,其中 $c_{n1},\cdots,c_{nn}$ 都是常数*.

利用模型的假定(2.6),(2.5),从公式(2.12)和(2.13)很容易推出最小二乘估计 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的一些性质:

1.$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别是 β_0 和 β_1 的无偏估计.

事实上,由(2.6)和(2.5),知 $E(Y_i)=\beta_0+\beta_1(X_i-\overline{X})$.故

$$E(\hat{\beta}_0)=\frac{1}{n}\sum_{i=1}^{n}E(Y_i)=\frac{1}{n}\sum_{i=1}^{n}[\beta_0+\beta_1(X_i-\overline{X}]=\beta_0$$

$$\begin{aligned}E(\hat{\beta}_1)&=\sum_{i-1}^{n}(X_i-\overline{X})E(Y_i)/\sum_{i=1}^{n}(X_i-\overline{X})^2\\&=\sum_{i-1}^{n}(X_i-\overline{X})[\beta_0+\beta_1(X_i-\overline{X})]/\sum_{i=1}^{n}(X_i-\overline{X})^2\\&=\beta_1\end{aligned}$$

2.$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差分别为:

$$\mathrm{Var}(\hat{\beta}_0)=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i)=n\sigma^2/n^2=\sigma^2/n \tag{2.14}$$

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1)&=\sum_{i=1}^{n}(X_i-\overline{X})^2\mathrm{Var}(Y_i)\Big/\Big[\sum_{i=1}^{n}(X_i-\overline{X})^2\Big]^2\\&=\sigma^2\Big/\sum_{i=1}^{n}(X_i-\overline{X})^2\end{aligned} \tag{2.15}$$

这里用到了 $Y_1,\cdots,Y_n$ 独立,$\mathrm{Var}(cY_i)=c^2\mathrm{Var}(Y_i)$,$\mathrm{Var}(c+e_i)=\mathrm{Var}(e_i)=\sigma^2$,$c$ 为常数.从(2.15)式我们得到一点启发.在第四

---

* 对 $\hat{\beta}_1$ 而言,系数 $c_{n1},\cdots,c_{nn}$ 与样本值 $X_1,\cdots,X_n$ 有关.但此处我们把 X 视为非随机的,因此它不影响 $c_{n1},\cdots,c_{nn}$ 为常数这个论断.若 X 也是随机变量,则情况就变得复杂.

章中我们已论述过，在无偏估计中，方差小者为优，如今 $\hat{\beta}_1$ 为无偏估计而其方差与

$$S_x^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{2.16}$$

成反比，故 S_x^2 愈大愈好. 而要 S_x^2 大，样本点 $X_1, \cdots, X_n$ 必须尽量散开一些. 这意味着当 X 之取值可以由我们选定时，我们不应把它们取在一小范围内，而最好让它们跨越较大之范围. 当然，这也要有个限度，不要把试验点取到没有实用意义的区域内去. 因为范围过大，线性回归与实际回归函数的差距会增加.

3. $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的协方差为 0：

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = 0 \tag{2.17}$$

事实上，$\hat{\beta}_0 - E(\hat{\beta}_0) = \sum_{i=1}^{n}(Y_i - EY_i)/n = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1(X_i - \overline{X}))/n = \sum_{i=1}^{n} e_i/n$，而

$$\begin{aligned}\hat{\beta}_1 - E\hat{\beta}_1 &= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - EY_i)\Big/\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \sum_{i=1}^{n}(X_i - \overline{X})e_i\Big/\sum_{i=1}^{n}(X_i - \overline{X})^2\end{aligned}$$

于是，利用 $E(e_ie_j) = E(e_i)E(e_j) = 0$ 当 $i \neq j$，而 $E(e_i^2) = \mathrm{Var}(e_i) = \sigma^2$，得

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) &= E\left[(\hat{\beta}_0 - E\hat{\beta}_0)(\hat{\beta}_1 - E\hat{\beta}_1)\right] \\ &= n^{-1}\Big[\sum_{i=1}^{n}(X_i - \overline{X})^2\Big]^{-1}\sigma^2\sum_{i=1}^{n}(X_i - \overline{X}) = 0\end{aligned}$$

这个性质指出：$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 不相关（见第三章，定理 3.2 下面的说明）. 它显示了中心化的好处：如果考虑原模型(2.1)中参数 b_0，b_1 的最小二乘估计 $\hat{b}_0, \hat{b}_1$（见下），则二者并非不相关.

由 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 不相关一般不能推出它们独立(第三章例 3.1). 但是,如果 $e_1,\cdots,e_n$ 服从正态分布,则 $Y_1,\cdots,Y_n$ 也服从正态分布. $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 作为 $Y_1,\cdots,Y_n$ 的线性函数,也服从正态分布*(第二章例 4.8). 因此在这种情况下,由 $\hat{\beta}_0,\hat{\beta}_1$ 不相关可推出它们独立(见第三章 2.3 节末尾).

由 β_0,β_1 的最小二乘估计 $\hat{\beta}_0,\hat{\beta}_1$,通过变换(2.7),即得模型(2.1)中的 b_0,b_1 的最小二乘估计分别为

$$\hat{b}_0 = \hat{\beta}_0 - \hat{b}_1\overline{X} = \overline{Y} - \hat{b}_1\overline{X}, \hat{b}_1 = \hat{\beta}_1 \qquad (2.18)$$

它们分别是 b_0 和 b_1 的无偏估计. 利用上述 $\hat{\beta}_0,\hat{\beta}_1$ 的方差协方差公式,不难算出 $\hat{b}_0,\hat{b}_1$ 的方差和 $\hat{b}_0$ 及 $\hat{b}_1$ 的协方差,细节留给读者.

$\hat{\beta}_0,\hat{\beta}_1$ 还有些更深刻的性质. 例如,若误差服从正态分布,则它们分别是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的最小方差无偏估计(见 4.3 节). 这个事实的证明超出本书范围之外.

6.2.2 残差与误差和方差 σ^2 的估计

仍以 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 记 β_0 和 β_1 的最小二乘估计. 则在 $X = X_i$ 处,因变量 Y 的预测值为 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1(X_i - \overline{X})$,而 Y 的实际观察值为 Y_i,二者之差

$$\delta_i = Y_i - \hat{Y}_i, i = 1,\cdots,n \qquad (2.19)$$

称为"残差".

残差的作用有二:一是当模型正确时,即(2.5)和(2.6)正确

---

* 更确切地,$(\hat{\beta}_0,\hat{\beta}_1)$的联合分布为二维正态分布.

时，它可以提供误差方差 σ^2 之一估计. 理由很清楚：用 $\hat{Y}_i$ 预测 Y_i，其精度取决于随机误差的大小，即误差方差的大小，误差方差愈大，预测愈不易准确，而残差（绝对值）就倾向于取大值. 反之则倾向于取小值. 往下我们证明

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\delta_i^2 \tag{2.20}$$

是 σ^2 的一个无偏估计.

为证明这个事实，注意

$$Y_i - \hat{Y}_i = \beta_0 + \beta_1(X_i - \overline{X}) + e_i - \hat{\beta}_0 - \hat{\beta}_1(X_i - \overline{X})$$

以及

$$\beta_0 - \hat{\beta}_0 = \beta_0 - \overline{Y} = \beta_0 - \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1(X_i - \overline{X}) + e_i) = -\ \overline{e}$$

其中 $\overline{e} = (e_1 + \cdots e_n)/n$，而

$$\begin{aligned}
&\beta_1 - \hat{\beta}_1 \\
&= \beta_1 - \sum_{j=1}^{n}(X_j - \overline{X})(\beta_0 + \beta_1(X_j - \overline{X}) + e_j)\Big/\sum_{j=1}^{n}(X_j - \overline{X})^2 \\
&= -\sum_{j=1}^{n}(X_j - \overline{X})e_j\Big/\sum_{j=1}^{n}(X_j - \overline{X})^2
\end{aligned}$$

故

$$\delta_i = e_i - \overline{e} - (X_i - \overline{X})\sum_{j=1}^{n}(X_j - \overline{X})e_j\Big/\sum_{j=1}^{n}(X_j - \overline{X})^2$$

平方，对 $i = 1, \cdots, n$ 求和，注意

$$\sum_{i=1}^{n}\left[(X_i - \overline{X})\sum_{j=1}^{n}(X_j - \overline{X})e_j\Big/\sum_{j=1}^{n}(X_j - \overline{X})^2\right]^2$$

$$= \left(\sum_{j=1}^{n}(X_j - \overline{X})e_j\right)^2\Big/\sum_{j=1}^{n}(X_j - \overline{X})^2$$

$$\sum_{i=1}^{n}(e_i - \overline{e})(X_i - \overline{X}) = \sum_{i=1}^{n}(X_i - \overline{X})e_i$$

即得

$$\sum_{i=1}^{n}\delta_i^2 = \sum_{i=1}^{n}(e_i - \bar{e})^2 - \Big(\sum_{i=1}^{n}(X_i - \bar{X})e_i\Big)^2 \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{2.21}$$

因为 $e_1,\cdots,e_n$ 独立同分布，有均值 0 方差 σ^2，故据第四章例 3.2. 及第三章(2.2)式，有

$$E\Big(\sum_{i=1}^{n}(e_i - \bar{e})^2\Big) = (n-1)\sigma^2$$

$$E\Big(\sum_{i=1}^{n}(X_i - \bar{X})e_i\Big)^2 = \mathrm{Var}\Big(\sum_{i=1}^{n}(X_i - \bar{X})e_i\Big)$$

$$= \sum_{i=1}^{n}(X_i - \bar{X})^2\mathrm{Var}(e_i)$$

$$= \sigma^2\sum_{i=1}^{n}(X_i - \bar{X})^2$$

以此代入(2.21)，即得

$$E\Big(\sum_{i=1}^{n}\delta_i^2\Big) = (n-2)\sigma^2$$

于是证明了 $\hat{\sigma}^2$ 为 σ^2 的无偏估计.

$\sum_{i=1}^{n}\delta_i^2$ 称为残差平方和. 其一重要性质是：当 e_i 服从正态分布 $N(0,\sigma^2)$时，有

$$\sum_{i=1}^{n}\delta_i^2/\sigma^2 \sim \chi^2_{n-2} \tag{2.22}$$

证明见本章附录 A. 注意自由度 $n-2$，它比样本大小 n 少 2. 这是因为有两个未知参数 β_0 和 β_1 需要估计，用掉了两个自由度(参看第四章例 3.2 末尾处的说明).

残差平方和有下述便于计算的表达式：

$$\sum_{i=1}^{n}\delta_i^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})Y_i$$

$$= \sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2 - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})Y_i \tag{2.23}$$

此式之方便在于:在计算残差平方和时,一般已先算出了回归系数 β_1 的估计 $\hat{\beta}_1$ 及 $\overline{Y}$. 而在算 $\hat{\beta}_1$ 时,需要算出 $\sum_{i=1}^{n}(X_i-\overline{X})Y_i$,故只须再计算平方和 $\sum_{i=1}^{n}Y_i^2$ 即可. (2.23)式证明如下:

$$\sum_{i=1}^{n}\delta_i^2=\sum_{i-1}^{n}(Y_i-\overline{Y}-\hat{\beta}_1(X_i-\overline{X}))^2$$

$$=\sum_{i=1}^{n}(Y_i-\overline{Y})^2-2A+B$$

其中 $B=\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\overline{X})^2=\hat{\beta}_1\left(\hat{\beta}_1\sum_{i=1}^{n}(X_i-\overline{X})^2\right)=\hat{\beta}_1\sum_{i=1}^{n}(X_i-\overline{X})Y_i$,而

$$A=\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})\hat{\beta}_1=\hat{\beta}_1\sum_{i=1}^{n}(X_i-\overline{X})Y_i$$

于是得到(2.23)第一式. 由此得出第二式.

残差的另一方面的作用是用以考察模型中的假定(即(2.5)和(2.6))是否正确. 道理如下:因为在模型正确时,残差是误差的一种反映,因误差 $e_1,\cdots,e_n$ 为独立同分布,具有"杂乱无章"的性质,即不应呈现任何规律性. 因此,残差 $\delta_1,\cdots,\delta_n$ 也应如此. 如果残差 $\delta_1,\cdots,\delta_n$ 呈现出某种规律性,则可能是模型中某方面假定与事实不符的征兆. 例如,若随着 X_i 增大 $|\delta_i|$ 有上升的趋势,这可能反映模型(2.1)中误差 e 的方差与 X 之值有关且随 X 之值上升而增加. 又如,设想回归函数为二次函数,则由图 6.3(l 为经验回归直线)可看出,当 X_i 很大或很小时,δ_i 取正号,而当 X_i 为中间值时,δ_i 取负号. 如出现这种情况,就可以怀疑线性假定有问题.

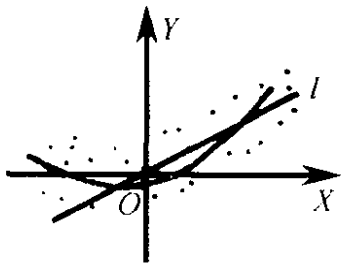

图 6.3

这种通过残差去考察回归模型是否正确的作法,叫做"回归诊断". 它已发展为回归分析的一个分支. 本书不能仔细讨论这方面

的问题，有兴趣的读者可参考陈希孺、王松桂著《近代回归分析》第二章，及张启锐著《实用回归分析》第四章.

6.2.3 区间估计和预测

本段我们在(2.5)和(2.6)的基础上加上假定：误差 e 服从正态分布，因此，现在(2.5)强化为

$$e_1, e_2, \cdots, e_n \text{ 独立同分布}. e_i \sim N(0, \sigma^2) \tag{2.24}$$

先考虑 $\hat{\beta}_1$. 前已指出，它是 $Y_1, \cdots, Y_n$ 的线性函数，有均值 β_1 方差 $\sigma^2 S_x^{-2}$，S_x^2 见(2.16)式，因此

$$(\hat{\beta}_1 - \beta_1)/(\hat{\sigma} S_x^{-1}) \sim N(0,1) \tag{2.25}$$

这个结果尚不能用于 β_1 的区间估计，因为 σ 未知，按 6.2.2 的结果，以 $\hat{\sigma}$(见 2.20)代替(2.25)中的 σ. 可以证明，经过这一代替，正态分布变为 t 分布(证明见附录 B)

$$(\hat{\beta}_1 - \beta_1)/(\hat{\sigma} S_x^{-1}) \sim t_{n-2} \tag{2.26}$$

这个结果就可以用来作 β_1 的区间估计或置信上、下界，因为 $(\hat{\beta}_1 - \beta_1)/(\hat{\sigma} S_x^{-1})$ 起了枢轴变量的作用，按 4.4 节中的方法，得到：

1. 置信系数为 $1-\alpha$ 的 β_1 的置信区间，为

$$[\hat{\beta}_1 - \hat{\sigma} S_x^{-1} t_{n-2}(\alpha/2), \hat{\beta}_1 + \hat{\sigma} S_x^{-1} t_{n-2}(\alpha/2)]$$

2. 置信系数为 $1-\alpha$ 的 β 的置信上、下界，分别为

$$\hat{\beta}_1 + \hat{\sigma} S_x^{-1} t_{n-2}(\alpha) \text{ 和 } \hat{\beta}_1 - \hat{\sigma} S_x^{-1} t_{n-2}(\alpha)$$

对截距 β_0 也一样做，也可以由下文对回归函数 $\beta_0 + \beta_1(x - \overline{X})$ 的区间估计中，令 $x = \overline{X}$ 得到.

对回归函数 $m(x) = \beta_0 + \beta_1(x - \overline{X})$，其点估计 $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1(x - \overline{X})$ 也是 $Y_1, \cdots, Y_n$ 的线性函数，因此在(2.24)的假定下，它也服从正态分布，其均值为 $m(x)$，而其方差 $\lambda(x)$，根据

(2.24),(2.25),及 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 独立,为

$$\begin{aligned}\lambda(x) &= \mathrm{Var}(\hat{\beta}_0) + (x-\overline{X})^2\mathrm{Var}(\hat{\beta}_1)\\ &= \sigma^2(1/n + (x-\overline{X})^2/S_x^2)\end{aligned}$$

于是得到$(\hat{m}(x)-m(x))/\sqrt{\lambda(x)} \sim N(0,1)$. $\hat{\sigma}$代σ,可以证明

$$(\hat{m}(x)-m(x))/(\hat{\sigma}(1/n + (x-\overline{X})^2/S_x^2)^{1/2}) \sim t_{n-2} \tag{2.27}$$

由此得出:

1. 置信系数为 $1-\alpha$ 的$m(x)$的置信区间为

$$[\hat{m}(x) - \hat{\sigma}(1/n + (x-\overline{X})^2/S_x^2)^{1/2}t_{n-2}(\alpha/2)$$

$$\hat{m}(x) + \hat{\sigma}(1/n + (x-\overline{X})^2/S_x^2)^{1/2}t_{n-2}(\alpha/2)]$$

2. 置信系数为 $1-\alpha$ 的$m(x)$的置信上下界,分别为 $\hat{m}(x)\pm\hat{\sigma}(1/n+(x-\overline{X})^2/S_x^2)^{1/2}t_{n-2}(\alpha)$(+号为上界).

这个区间之长 $2\hat{\sigma}(1/n+(x-\overline{X})^2/S_x^2)^{1/2}t_{n-2}(\alpha/2)$与 x 有关. x 愈接近X样本的中心 $\overline{X}$,则$(x-\overline{X})^2$ 愈小而区间长度就愈小.就是说,在估计回归函数 $m(x)$时,愈靠近样本 X 中心点处愈精确.这从理论上指明了我们在前面提到过的一点事实:当我们需要在自变量 X 的某个范围内使用回归方程时,应当把观察点 X_1,…,X_n 尽量取在这个范围内.如图 6.4,l 为由样本点配出的经验回归直线,l_1 和 l_2 分别是 $m(x)$的置信区间上、下端随 x 变化时划出的曲线.在 x 轴上的$\overline{X}$ 附近 l_1 和 l_2 相距较近,而当 x 离 $\overline{X}$ 愈远时,曲线愈分开.如图,在 x 轴的 x_0 处,A 点的纵坐标是回归函数$m(x_0)$的点估计 $\hat{m}(x_0)$,而 A_1,A_2 点的纵坐标,则分别是

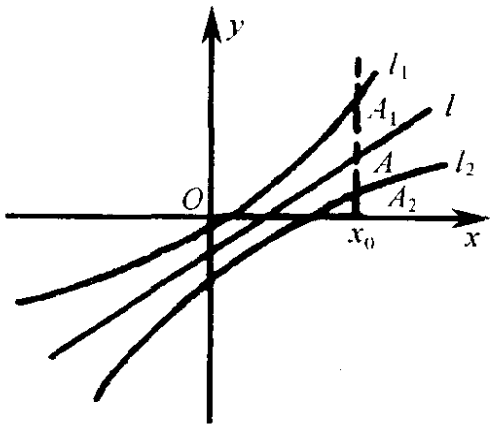

图 6.4

$m(x_0)$的置信区间的上、下两端点.曲线 l_1, l_2 只能在这个意义上去理解,而不能说,"理论回归直线落在 l_1, l_2 之间"的概率为 $1-\alpha$.因为,理论回归直线落在 l_1, l_2 之间,相当于说对一切 x_0 同时成立:$m(x_0)$落在通过 x_0 与纵轴平行的直线在 l_1, l_2 截出的两点的纵坐标之间*.

下面来考察 Y 的区间预报.假定要在自变量 X 的给定值 x_0 处预报 Y 之值 Y_0.前已说过(见 6.1 节),就用 $\hat{m}(x_0)$作为 Y_0 的预报值.考虑差 $\eta = Y_0 - \hat{m}(x_0)$.它是 $Y_1, \cdots, Y_n$ 和 Y_0 的线性函数,故仍为正态分布.因 $E(Y_0) = m(x_0), E[\hat{m}(x_0)] = m(x_0)$,有 $E(\eta) = 0$.为考虑其方差,注意 $Y_1, \cdots, Y_n$ 和 Y_0 独立,故 $\hat{m}(x_0)$与 Y_0 也独立,因此有

$$\begin{aligned}\mathrm{Var}(\eta) &= \mathrm{Var}(Y_0) + \mathrm{Var}(\hat{m}(x_0)) \\ &= \sigma^2(1 + 1/n + (x - \overline{X})^2/S_x^2)\end{aligned}$$

仿以前的做法,用 σ 的估计值 $\hat{\sigma}$ 代替 σ,得

$$\eta/(\hat{\sigma}(1 + 1/n + (x - \overline{X})^2/S_x^2)^{1/2}) \sim t_{n-2}$$

于是得到:不等式

$$\begin{aligned}&\hat{m}(x_0) - \hat{\sigma}(1 + 1/n + (x - \overline{X})^2/S_x^2)^{1/2} t_{n-2}\left(\frac{\alpha}{2}\right) \leqslant Y_0 \\ &\leqslant \hat{m}(x_0) + \hat{\sigma}(1 + 1/n + (x - \overline{X})^2/S_x^2)^{1/2} t_{n-2}\left(\frac{\alpha}{2}\right)\end{aligned} \tag{2.28}$$

其左右两端(所构造的区间)就是 Y_0 的置信系数为 $1-\alpha$ 的区间预测.应注意的是:与以前我们讲过的区间估计不同,此处的 Y_0 并不是一个未知的参数,其本身也有随机性.

---

* 理论上可以证明:把 l_1, l_2 之间夹出的区域放大一点,即把 l_1 往上推一点,l_2 往下推一点,就可以满足这要求,具体说,应以方程为 $y = \hat{m}(x) \pm \hat{\sigma}(1/n + (x - \overline{X})^2 S_x^2)(2F_{2,n-2}(\alpha))^{\frac{1}{2}}$的曲线代替 l_1, l_2(l_1 为 + 号).由第二章习题 29 可知,这个范围比(2.28)规定的范围宽一些.

比较(2.27)和(2.28),我们看出 $m(x_0)$的区间估计与 Y_0 的区间预测的另一点不同之处:$m(x_0)$的区间估计之长为 $2\hat{\sigma}(1/n+(x-\overline{X})^2/S_x^2)^{1/2}t_{n-2}(\alpha/2)$. 当 n 很大时,$\hat{\sigma}$ 接近于 σ,$t_{n-2}(\alpha/2)$接近 $u_{\alpha/2}$,这两部分保持有界*,另一个因子中,$1/n\to 0$.另一个因子,只要试验点 $X_1,\cdots,X_n$ 不过分集中于一处,以使 $\sum_{i=1}^{n}(X_i-\overline{X})^2\to\infty$,就可以证明$(x-\overline{X})^2/S_x^2\to 0$(习题5(b)).这样,上述区间之长将随 $n\to\infty$ 而趋于0. Y_0 的区间预测则不然,其长度表达式中含因子$(1+1/n+(x-\overline{X})^2/S_x^2)^{1/2}$. 随着 $n\to\infty$,其值总大于1,故不论你有多少样本,区间预测的精度仍有一个界限.这个道理我们在前面已解释过:预测问题中包含了一个无法克服的随机误差项.

6.2.4 假设检验

最有兴趣的假设检验问题是:检验原假设

$$H_0:\beta_1=c \tag{2.29}$$

其中 c 是一个给定的常数,对立假设为 $H_0:\beta_1\neq c$.尤其是 $c=0$ 的情况.因为,$\beta_1=0$ 表示回归函数 $m(x)$为一常数 β_0,与 x 无关.如果 $H_0:\beta_1=0$ 被接受了,则意味着我们接受如下的说法:所选定的自变量 X 其实对因变量 Y 无影响,故研究二者之间的关系也就没有意义了.

(2.29)的检验很容易利用(2.26)作出:

$$\varphi:\text{当}\ |\hat{\beta}_1-c|\leqslant\hat{\sigma}S_x t_{n-2}(\alpha/2)\ \text{时接受}\ H_0,\text{不然就否定}\ H_0 \tag{2.30}$$

这个检验 φ 有水平 α.单边假设 $\beta_1\leqslant c$,或 $\beta_1\geqslant c$ 的检验也类似地作出.

---

* 由于 $\hat{\sigma}$ 是随机的,它只是在"依概率收敛"的意义上接近 σ,故 $\hat{\sigma}$ 也有很小的可能性远远偏离 σ,甚至变得很大.只是当 n 很大时这种机会很小.

对截距 β_0 的检验也类似地作出. 例如, $\beta_0 = 0$ 的假设意味着回归直线通过原点,我们把细节留给读者.

例 1.1 从某大学男生中随机抽取 10 名,测得其身高(米)和体重(公斤)的数值为

(1.71,65),(1.63,63),(1.84,70),(1.90,75),(1.58,60)

(1.60,55),(1.75,64),(1.78,69),(1.80,65),(1.64,58)

以身高 X 为自变量,并把它看成非随机的,而以体重 Y 为因变量.假定回归为线性的.算出

$$\overline{X} = (1.71 + 1.63 + \cdots + 1.64)/10 = 1.723$$

$$\overline{Y} = (65 + 63 + \cdots + 58)/10 = 64.4$$

$$S_x^2 = (1.71 - 1.723)^2 + \cdots + (1.64 - 1.723)^2$$

$$= 0.1062$$

$$\sum_{i=1}^{10}(X_i - \overline{X})Y_i = (1.71 - 1.723) \times 65 + \cdots$$

$$+ (1.64 - 1.723) \times 58 = 5.268$$

由(2.12),(2.13),得出 β_0 和 β_1 的最小二乘估计值分别为

$$\hat{\beta}_0 = 64.4, \hat{\beta}_1 = 5.268/0.1062 = 49.6$$

经验回归方程为

$$y = 64.4 - 49.6(x - 1.723) = -21.06 + 49.6x$$

当 $x = 1.62$ 时 $Y = 59.29$. 这有两个解释,一是对身高为 1.62 米的学生,其平均体重的点估计为 59.29 公斤;二是如随机抽到一个学生量出其身高为 1.62 米,则以 59.29 公斤为其体重的预测值.

可按(2.23)式计算残差平方和.为此算出

$$\sum_{i=1}^{10}(Y_i - \overline{Y})^2 = (65 - 64.4)^2 + \cdots + (58 - 64.4)^2 = 316.4$$

因此按(2.23)式算出

$$\sum_{i=1}^{10}\delta_i^2 = 316.4 - 49.6 \times 5.268 = 54.39$$

由此得出误差方差 σ^2 的估计值

$$\hat{\sigma}^2 = 54.39/(10-2) = 6.799, \hat{\sigma} = 2.61$$

取 $\alpha=0.05$. 查 t 分布表,得 $t_{n-2}(\alpha/2)=t_8(0.025)=2.306$

于是用(2.27)和(2.28),得到回归函数 $m(x)=\beta_0+\beta_1(x-\overline{X})$的置信区间,以及在 x 点处 Y 的取值 y 的预测区间,分别为(置信系数都是 0.95)

$$-21.06+49.6x-2.61\left(0.1+\frac{(x-1.723)^2}{0.1062}\right)^{1/2}\times 2.306 \leqslant m(x)$$
$$\leqslant -21.06+49.6x+2.61\left(0.1+\frac{(x-1.723)^2}{0.1062}\right)^{1/2}\times 2.306$$

以及

$$-21.06+49.6x-2.61\left(1.1+\frac{(x-1.723)^2}{0.1062}\right)^{1/2}\times 2.306 \leqslant y$$
$$\leqslant -21.06+49.6x+2.61\left(1.1+\frac{(x-1.723)^2}{0.1062}\right)^{1/2}\times 2.306$$

对 $x=1.62$,上述两个区间分别是

$$-21.06+49.6x\times 1.62 \pm 2.691 = [56.6, 62.0]$$
$$-21.06+49.6x\times 1.62 \pm 6.343 = [53.0, 65.6]$$

可见,预测的精度比估计回归函数的精度差得多.

再考虑假设(2.29)的检验. 在此例中,取 $c=0$ 是没有意义的. 因为体重明摆着与身高有关,如检验假设 $\beta_1=0$,即使接受了,我们也只能归因于样本大小 n 太小,也不大会认为 $\beta_1=0$ 真可以被接受. 可以考虑的假设是 c 取一个合理的数字,例如 $c=50,40$ 之类. "$c=50$"这个假设可理解为:在另一城市一所大学曾作过较大规模的测量,在那里比较确切地估出 $\beta_1=50$. 现在换了一个城市,情况有无改变? 由于这样一种提法,且 50 这个数字先天地有一定的根据,在并无比较显著的证据的情况下,我们不愿轻易地认为 50 这个数字不适用于这间大学. 因此,取一个较小的水平,例如 $\alpha=0.05$,就要算比较恰当了. 具体检验可按(2.30). 算出

$$\hat{\sigma}S_x t_{n-2}(\alpha/2) = 2.61\times\sqrt{0.1062}\times 2.306 = 1.96$$

令 $|\hat{\beta}_1-c|=|49.6-50|=0.4<1.96$,故应接受原假设 $\beta_1=50$.

如原假设为 $\beta_1=52$,则被否定了.

现在有这样的问题:一方面用我们的数据估出 β_1 为 49.6,另一方面,按以往资料可以接纳 $\beta_1=50$,应取何者为好? 这就要分析情况,如果以往资料可以认为是与当前资料同质的,比方说,两校都是在全国范围招生,其学生的地域构成大体接近,则有充分理由认为,当前的 β_1 与以往的 β_1 应差不多.考虑到以往的 β_1 是依据大量数据算出,而当前的 β_1 只根据 10 个数据,我们觉得,取以往的 β_1 也许更合适(如果 $\beta_1=50$ 被否定,自又当别论).反之,如两校都是地方性的,其学生来源以本地居多,而两地身高体重在关系上又有差别,则我们就可能倾向于采用当前值了.

这个例子也许并不十分典型,但有关的考虑对其他应用问题也是适用的.统计学是一种帮助我们对数据进行分析的工具,其应用不能脱离对实际问题的背景的考虑.不加区别地机械地使用公式,难免导致与实际背离的结果.

6.2.5 几个有关问题

以上我们对一元线性回归(且随机误差服从正态分布的情况)的统计分析作了较仔细的论述.在这一段中,我们提出几点在使用这些方法时值得注意的事情.

1. 回归系数的解释问题

设想我们建立了回归方程

$$y = a + bx \tag{2.31}$$

一般地把回归系数 b 的意义解释为:当自变量 X 增加或减少 1 单位时,平均地说,Y 增加或减少 b 单位.这个解释对不对? 我们说,也对也不对,要看具体情况而定.

首先一个问题是 X 的变化区间.在实际应用中,真正的回归方程一般总是与线性方程有一定的偏离.在不很大的范围内,这种偏离也许不很大,不致对应用造成影响.一般总是在这个意义上,我们把回归方程认定为线性的.

日后在应用中,如果自变量值 x 超出了上述范围,则回归方

程(2.31)可能已不再成立.这时 X 增加 1 单位是否使 Y 平均增加 b 单位的论断,也就不能成立了.例如,若 X 为每亩施肥量而 Y 为每亩的产量.可以相信,在 X 的一个合理的范围内,Y 的平均值大致随 X 线性地增长.但一超出一定的范围,例如施肥量过大时,进一步增加施肥不仅不能导致增产,反而可能导致减产.

就是自变量之值处在合理的范围内时,回归系数意义的解释仍可能有问题.分两种情况来讨论.一种情况是 X 之值在试验中可由人指定(如上述施肥量).这时,只要在日后的应用中情况与你建立回归方程时大体相同——这主要指的是 X 以外的因素对 Y 的影响要相当,则上述解释,即 X 增减 1 单位时 Y 平均增减 b 单位,是正确的,否则就不见得正确.仍拿上面那个例子来说,设想在建立方程(2.31)而进行的试验中,所用的田地都是底肥很不充足的,而日后你把它用到底肥很充足的田地上;或者,在试验中用的是深耕(这对肥料吸收有利),而日后用到浅耕的田地上,则结果就不见得正确了.

如果自变量 X 是与 Y 一起观察所得,而不能事先由人控制,则情况更加复杂.在这种情况下,除了满足 X 必须处在合理范围内这个限制外,还必须注意,X 值必须是在"自然而然地"产生而不是人为地制造出来的情况下,上述解释才有效.举一个极端的例子.设把 X 作为体重而 Y 作为身高,则在 X 一定的范围内,仍可建立线性回归方程(2.31),比方说,$\hat{b}=0.02$.这意味着体重每增减 1 公斤,身高平均约增长 2 厘米.假如你观察一个正在长身体的青年人,在某时刻你量得他体重 X 为 52 公斤,身高 158 厘米.过若干时候他体重长到 54 公斤,你预测他身高 162 厘米左右,这个用法正确.因为你只是一个被动的观察者,并未设法去影响这个进程.反之,如果你用强力减肥法使一个胖子在两星期内体重下降 5 公斤,而预测他身高将下降 10 厘米左右,则恐怕不见得正确.因为 X 值的改变出于你人为的干预,违反了 X, Y 之间的关系的自然进程.再举一个例子:统计资料显示人的文化水平的提高导致出生率降低.但如某个国家孤立地进行提高人的文化水平的工作,就不

一定能导致出生率预期的降低.这是因为人口出生率是由一系列的经济社会和文化习惯等条件决定的.单抽出文化水平这个因子,其实是将它作为一个综合因子来看待.故如它的改变确实是显示了这种综合条件的改善,则应有利于出生率的降低.反之,如果其他条件(经济、社会等)并无改变甚至有了恶化,而只孤立地提高文化这个因子,则背离了建立回归方程的前提了.

2. 回归方程的外推

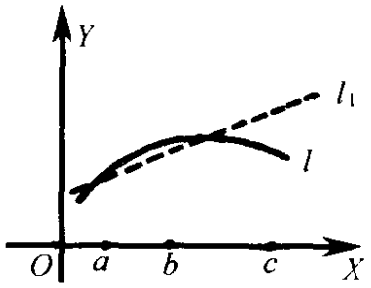

图 6.5

所谓外推,就是在建立回归方程时所用的自变量数据的范围之外去使用回归方程(如果在自变量数据的范围之内使用,就叫做内插).一般都是不主张对回归方程作外推使用的,原因我们在以前已提过了,即理论上回归方程一般并非严格的直线.例如,回归方程是曲线 l,如果你在 $a \leqslant x \leqslant b$ 这个范围内使用,则直线 l_1 可充分好地代表它,但如外推至 c 点,则与实际情况有较大的差距了(图 6.5).

当然,也不能说外推在任何情况下都不行.在某种很特殊的情况下,回归方程为线性这一点有充分的理论根据,这时外推应不致导致太大的偏差.其次,如外推距离不太远,问题一般也不会很大.在没有把握而情况允许时,可以做一些试验,以考察一下回归方程在拟应用的范围内符合的程度如何.

3. 回归方程不可逆转使用

在自变量 X 和因变量 Y 都是随机的场合,往往可以把其中任一个取为自变量.人的身高体重就是一个例子.这时就存在两个回归方程,如都为线性的,则分别有形状

$$y = a + bx, \quad x = c + dy \tag{2.32}$$

有趣的是,这两个方程并不一致.意思是,若你把(2.32)的第一个方程 $y = a + bx$ 对 x 解出得 $x = -a/b + y/b$,则这方程不一定就是(2.32)第二个方程,对实际数据配出的经验回归直线,也是这个

情况.设有了数据$(X_1, Y_1), \cdots, (X_n, Y_n)$,把 X 作为自变量配出回归方程(用最小二乘法,下同)$y = \hat{a} + \hat{b}x$,与把 Y 作为自变量配出的回归方程 $x = \hat{c} + \hat{d}y$ 不一定相同,且一般不相同.

因此,在人的身高(X)体重(Y)这个例子中,如你的目的是通过身高预测体重,则你应取 Y 为因变量,以建立回归方程 $y = a + bx$.如果什么时候你忽然需要通过体重预测身高,则你并不能利用上述方程去作,而必须从头做起,取 X 为因变量,用最小二乘法配出方程 $x = c + dy$.后一方程用于从 y 预测 x.

表面上看这一点颇使人感到难以理解,细想之下,道理其实不难.为方便计,设(X, Y)的联合分布为二维正态分布 $N(a, b, \sigma_1^2, \sigma_2^2, \rho)$,则如在第 2 章(见该章(3.10)式)中所证明的,Y 对 X 的回归方程为

$$(y - b) = \rho\sigma_2\sigma_1^{-1}(x - a) \tag{2.33}$$

而 X 对 Y 的回归方程则为

$$(x - a) = \rho\sigma_1\sigma_2^{-1}(y - b) \tag{2.34}$$

除非 $\rho^2 = 1$,即 X, Y 之间有严格的线性关系,(2.33)与(2.34)不一样,因为,由(2.33)得$(x - a) = \rho^{-1}\sigma_1\sigma_2^{-1}(y - b)$,除非 $\rho^2 = 1$,这与(2.34)不同.这样看来,理论上这二者本不一致.因此,由数据所配出两个经验回归方程,也不会一致了.

这个论点从理论上说清楚了问题.但在直观上,人们可能仍觉得有些难以理解.为说明这一点,考察这样一个情况:相关系数 $\rho > 0$但很小.这时,X, Y 有些关系,但关系很微弱:一者的变化只引起另一者很小的变化.因此,在两个回归关系 $y = a + bx$ 和 $x = c + dy$ 中,系数 b, d 都很接近 0.这样二者就必然不一致了.因由 $y = a + bx$ 得出 $x = a_1 + b_1 y$,其中 $b_1 = b^{-1}$.b_1 很大,因为 b 很小,故 b_1 不可能与 d 一致.

但应注意:我们强调回归方程不能逆转使用是指用于预测而言,如用于控制则另当别论.比如,建立了 Y 对 X 的回归方程 $y = a + bx$.为要把 Y 之值控制在 y_0 使其误差尽量小,自变量 X 应取

何值？那要从 $y_0 = a + bx$ 解出 $x = (y_0 - a)/b$. 当然，用于控制的情况应当是自变量 X 之值能由人选择时，这时不存在作 X 对 Y 之回归的问题.

4. 在本节的讨论中，我们都是在自变量 X 为非随机的假定下进行的. 而在应用中，又不时遇到 X 也是随机的情况，而我们也就当作 X 为非随机，仍使用本节导出的公式，这样做在理论上到底可以不可以？

这问题的仔细分析比较复杂，不能在这里详细给出了. 我们只指出两点：一是若 (X, Y) 的联合分布为二维正态 $N(a, b, \sigma_1^2, \sigma_2^2, \rho)$，则有关回归系数的点估计，区间估计，回归函数的区间估计与区间预测，回归系数的检验等公式，全都合用，但 $\hat{\beta}_1, \hat{\beta}_1$ 的方差公式已不适用（$\hat{\beta}_1$ 的方差表达式中含 X_i，因此处 X_i 也是随机变量，这是不可以的）. $\hat{\sigma}^2$ 仍是模型(2.1)中的误差 e 的方差的无偏估计，但这个方差应是给定 X 时 Y 的条件分布之方差，即 $\sigma_2^2(1 - \rho^2)$（见第二章(3.9)式）. 因此在这一场合，X 为随机变量并不影响方法的使用. 我们之所以能不顾 X 是否随机而使用本节导出的公式，主要就是基于这个理由. 二是若 (X, Y) 的分布不是正态时，虽说回归系数点估计的公式仍可用，但其他一切已不再成立了.