第九章　一元线性回归

在社会经济中，各种经济变量之间联系紧密，一些因素影响另一些因素的现象十分普遍。这些经济因素间存在的这种因果关系可以被人们利用从而指导、控制或预测经济活动的发展。回归分析就是统计学中研究变量间关系的一种重要方法。

本章将重点学习统计关系与回归分析的基本概念，一元线性回归模型的建立与总离差平方和的分解和样本相关系数计算方法，一元线性回归显著性检验与模型适合性分析方法。以及 E（Y）的区间估计和因变量 Y 的预测方法。

9.1 回归分析的基本概念

9.1.1 因变量(Y)与自变量(X)之间的关系

根据因变量与自变量之间的关系不同，可以分为两种类型：

一种是函数关系，即对两个变量 X，Y 来说，当 X 值确定后，Y 值按照一定的规律唯一确定，即形成一种精确的关系。例如:微积分学中所研究的一般变量之间的函数关系就属于此种类型。

另一种是统计关系，即当 X 值确定后，Y 值不是唯一确定的，但大量统计资料表明，这些变量之间还是存在着某种客观的联系。例如：图 9.1 在直角坐标平面上，标出了 10 个观测点的坐标位置，他们表示以家庭为单位，某种商品年需求量与该商品价格之间的 10 对调查数据。

9.1.2 回归分析

回归分析(Regression Analysis)就是应用统计方法，对大量的观测数据进行整理、分析和研究，从而得出反映事物内部规律性的一些结论。

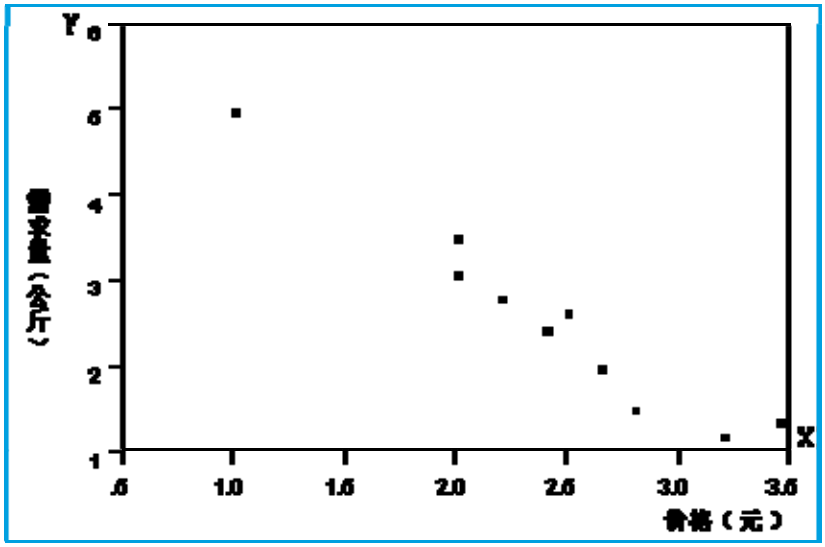


图 9-1 商品需求量与价格的关系

9.2 一元线性回归模型

9.2.1 统计关系的两个特征

统计关系具有以下两个特征：

(1) 因变量 Y 随自变量 X 有规律的变化，而统计关系直线描述了这一变化的趋势。

(2)观测点散布在统计关系直线的周围，此种情况说明 Y 的变化除了受自变量 X 影响以外，还受其他因素的影响。

因此试图建立这样一个回归模型，通过对此模型所作的一些假设，可以体现出上述统计关系所刻划的特征。

9.2.2 一元线性回归模型假设

根据统计关系特征，可以进行下述假设：

(1)对于自变量的每一水平 X，存在着 Y 的一个概率分布；

(2)这些 Y 的概率分布的均值，有规律的随 X 变化而变化

9.2.3 一元线性回归模型

若 Y 与 X 具有统计关系而且是线性的，则可以建立下述一元线性回归模型：

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i \qquad (i=1,2,\cdots,n) \qquad (9\text{-}1)$$

其中，(X_i, Y_i)表示(X,Y)的第i个观测值，β_0, β_1为参数，$\beta_0+\beta_1X_i$为反映统计关系直线的分量，ε_i为反映在统计关系直线周围散布的随机分量 $\varepsilon_i \sim N(0, \sigma^2)$。

对于任意X_i值有：

⑴Y_i服从正态分布

⑵$E(Y_i)=\beta_0+\beta_1X_i$；

⑶$\sigma^2(Y_i)=\sigma^2$；

⑷各Y_i间相互独立 $Y_i \sim N(\beta_0+\beta_1X_i, \sigma^2)$ 。

散点图中需求量与价格之间线性统计关系的回归模型，具体描述如图9-2。这里给出价格为X=2与X=3时，需求量Y的概率分布。根据以上回归模型的假设，当X=2时，此时观测到的需求量 Y=3，该值是对应于X这一水平的Y的一次随机抽取结果。

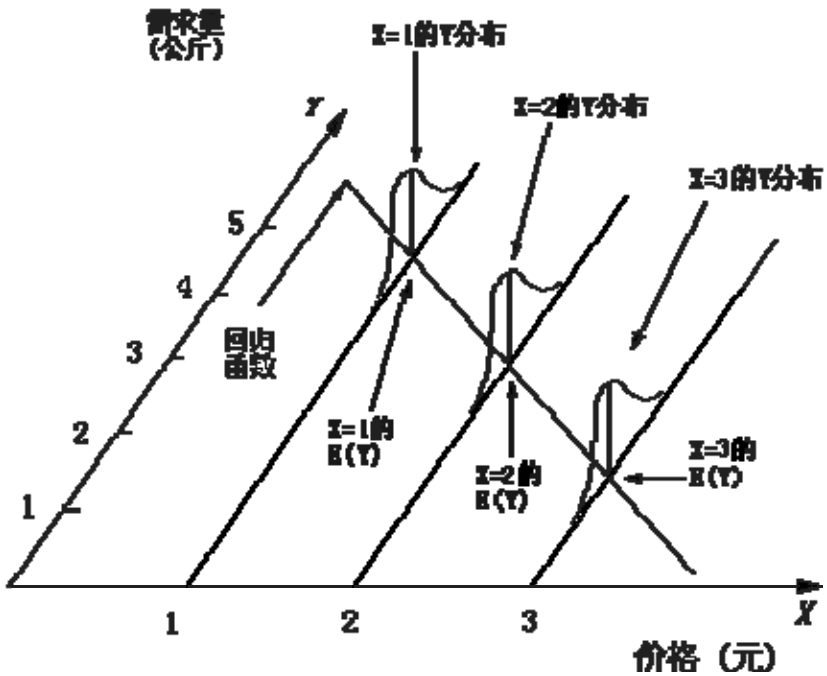


图 9-2 需求量与价格的线性统计关系

9.2.4 一元线性回归方程

对于图 9-1，在坐标直角平面上，标出了 10 个观测点的坐标位置，他们表示以家庭为单位，某种商品年需求量与该商品价格之间的 10 对调查数

据。若Y与X之间为线性关系，要想描述其关系，可以有无数条直线，需要在其中选出一条最能反映Y与X之间关系规律的直线，即对于式 9-1，要适当选取 β_0和 β_1。因此要根据样本数据采用最小二乘法对参数 β_0和 β_1进行估计。

设 β_0和 β_1的估计值为b_0和b_1，则可建立一元线性回归模型如下:

$$\hat{Y} = b_0 + b_1 X \tag{9-2}$$

假设找到一条回归直线如图 9-3 所示。一般而言，所求的b_0和b_1应能使每个样本观测点(X_i, Y_i)与回归直线之间的偏差尽可能小，即使观察值与拟合值的偏差平方和Q达到最小。

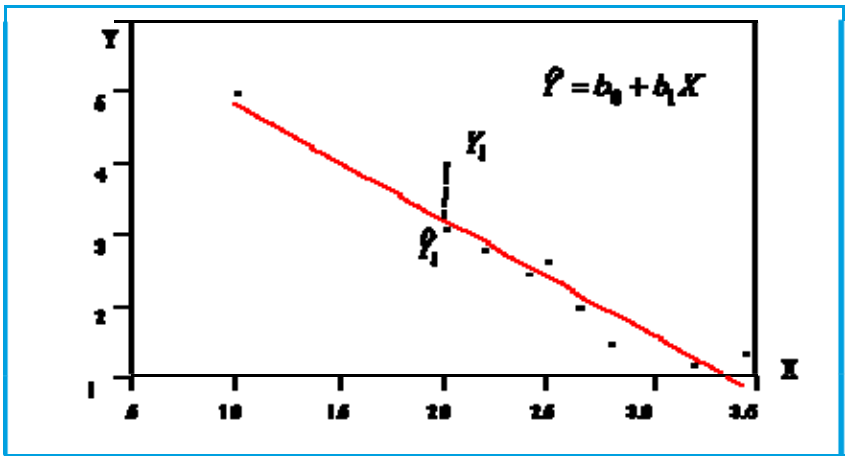


图 9-3 回归方程原理图

令 $Q = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$ （9-3）

使Q达到最小值的b_0和b_1称为最小二乘估计量。

显然，Q是b_0和b_1的二元函数，根据微积分中极值的必要条件，首先分别求Q关于b_0和b_1的偏导数：

$$\frac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]$$

$$\frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)] X_i$$

然后令这两个偏导数等于零，整理后得正规方程组

$$nb_0 + b_1\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

$$b_0\sum_{i=1}^{n} X_i + b_1\sum_{i=1}^{n} X_i^{\,2} = \sum_{i=1}^{n} X_i Y_i \tag{9-4}$$

解此方程组得到

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\ &= \frac{\sum_{i=1}^{n} X_i Y_i - \dfrac{(\sum X_i)(\sum Y_i)}{n}}{\sum_{i=1}^{n} X_i^{\,2} - \dfrac{(\sum X_i)^2}{n}} \end{aligned} \tag{9-5}$$

$$b_0 = \overline{Y} - b_1\overline{X} \tag{9-6}$$

9.2.5 最小二乘估计量b_0, b_1的特性

(1) 线性特性

线性特性是指参数估计量b_0, b_1,分别是样本观测值Y_i的线性组合，或者b_0和b_1分别是变量Y_i的线性参数。在统计学里，具有此种性质的估计量称为线性估计。

由(9-5)式可以得出

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

令 $C_i = \dfrac{X_i - \overline{X}}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

则 $b_1 = \sum_{i=1}^{n} C_i Y_i$ (9-7)

这表明b_1是Y_i的线性组合，即估计量b_1为线性估计。同理，由(9-6)和(9-7)式可以得到

$$b_0 = \sum_{i=1}^{n} k_i Y_i \tag{9-8}$$

其中

$$k_i = \frac{1}{n} - C_i \overline{X}$$

因此b_0也是线性估计。

(2) 无偏性

无偏性是指b_0和b_1分别是β_0和β_1的无偏估计，可以证明

$E(b_0) = \beta_0$ $E(b_1) = \beta_1$

9.3 总平方和分解

9.3.1 总平方和分解

由 $Y_i - \overline{Y} = Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y}$

可以得到

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

其中，$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}) = 0$

所以

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \tag{9-9}$$

定义总离差平方和 $SSTO=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$

它表示没有 X 的影响，单纯考察数据中 Y 的变动情况。

定义回归平方和 $SSR=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$

表示各 $\hat{Y}_i$ 的变动程度，该变动是由于回归直线中各X_i 的变动所引起的，并且通过X对Y的线性影响表现出来。

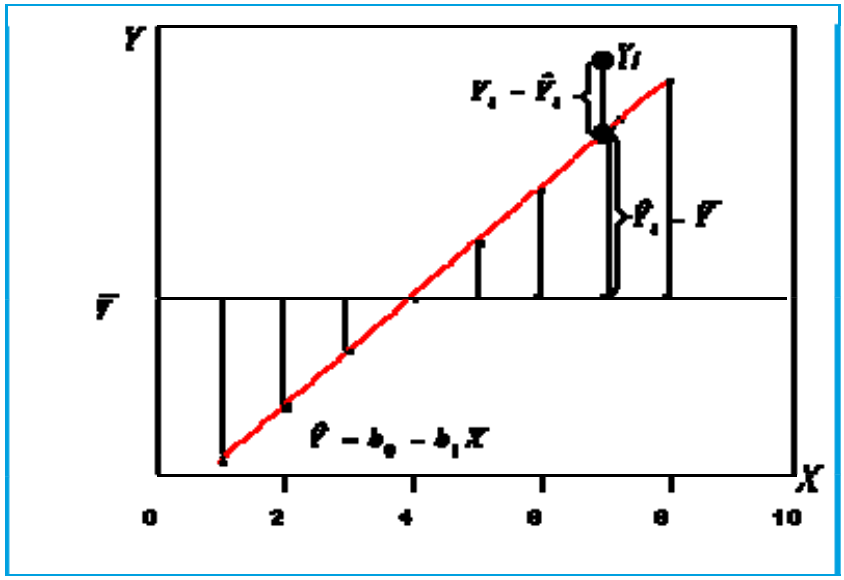


图 9-4 总平方和分解图

定义误差平方和 $SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$

表示各Y_i围绕所拟合的回归直线的变动程度，*SSTO*=*SSR*+*SSE*。

易得到：

$$SSTO=\sum_{i=1}^{n}Y_i^2-\frac{(\sum_{i=1}^{n}Y_i)^2}{n}$$

$$SSR = b_1^2[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}]$$

$SSE=SSTO-SSR$

9.3.2 自由度的分解

总平方和*SSTO*含有n个离差$Y_i - \overline{Y}_i$，这些离差项之间有一个约束条件，即$\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$，所以*SSTO*的自由度$f_t$为n-1。误差平方和*SSE*含有n个偏差$Y_i - \hat{Y}_i$，但由于估计参数$\beta_0$和$\beta_1$时用了两个正规方程（式 9-4），对于$Y_i - \hat{Y}_i$有两个约束条件，故*SSE*的自由度$f_E$为n-2。

回归平方和SSR的自由度f_R为 1，这是因为回归函数中有两个参数，而偏差$\hat{Y}_i - \overline{Y}$有一个约束条件$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}) = 0$，故自由度是 2-1。

因此，自由度的分解可以表示为

n-1=1+（n-2）

$f_T=f_R+f_E$

9.3.3 回归均方与误差均方

我们定义平方和除以它相应的自由度为均方。根据上面讨论的结果，在回归分析中，有两个均方，即回归均方与误差均方。回归均方记为 *MSR*，误差均方记为 *MSE*。

$$MSR = \frac{SSR}{1} \tag{9-10}$$

$$MSE = \frac{SSE}{n-2} \tag{9-11}$$

9.4 样本确定系数与样本相关系数

9.4.1 样本确定系数

当X与Y具有因果关系时，我们常把由于X的变动影响Y的变动的程度，说成是由X这一因素解释Y的变动时能解释多少；即，Y的总变差中能被X解释的那部分所占的比率，所占的比率愈大，说明X与Y相关的程度愈紧密。因此，我们定义下式

$$r^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{9-12}$$

为简单确定系数。r^2的取值范围为$0 \le r^2 \le 1$。

当样本的全部观察值都落在所拟和的回归直线上，这时SSE=0，$SSR=SSTO-SSE=SSTO$；所以$r^2=1$。

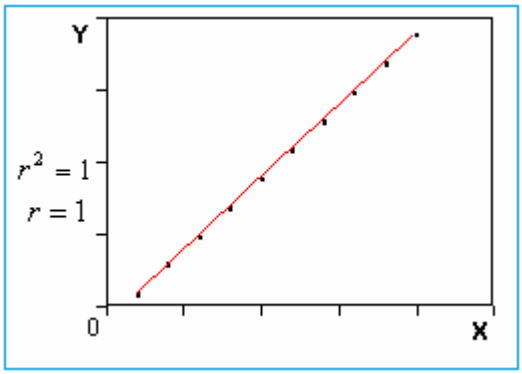
当X与Y无关，Y的变差完全由于不确定因素（或随机因素）引起，此时，$SSR=0$；所以，$r^2=0$。

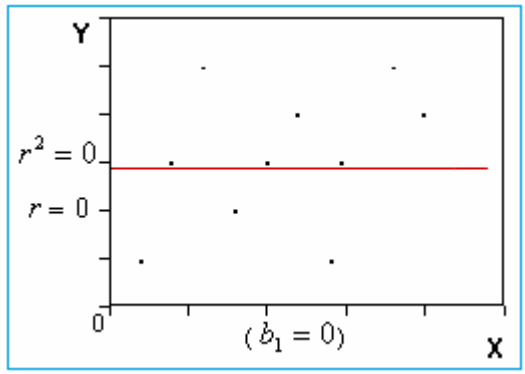
一般地，r^2常介于 0 与 1 之间，r^2愈接近 1，说明Y与X线性相关程度愈高。

9.4.2 样本相关系数

为了既能描述Y对于X的线性相关程度，又能描述随X变化Y变化的方向，常采用另一种尺度，即r^2的平方根，称为样本相关系数，定义如下：

$r = \pm\sqrt{r^2}$ 按定义可以导出 $r = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$ 它与b_1具有相同的分子，且r与b_1的分母均为正，故r与b_1有相同的符号。$-1 \le r \le 1$。各种情况如图 9-5 所示。





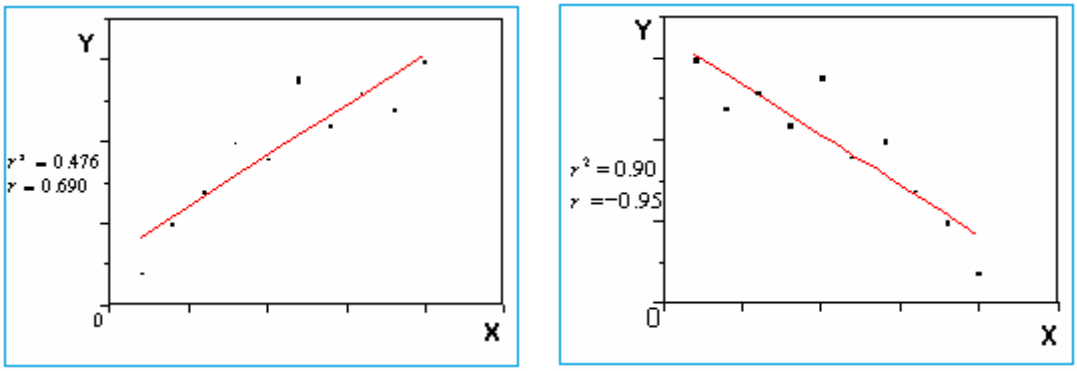


图 9-5 简单确定系数图示

9.5 一元线性回归显著性检验

根据所得到的样本数据，采用最小二乘法总是可以拟合一条直线来描述Y和X之间的关系。但是，样本资料具有随机性，因此，我们需要判断Y与X之间是否确实存在着线性关系，也就是需要判断Y对X的回归函数是否确实是一条直线。在回归函数$E(Y)=\beta_0+\beta_1X$中，如果$\beta_1=0$，则对于X的一切水平$E(Y)=\beta_0$. 说明Y的变化与X的变化无关，因而，我们不能通过X去预测Y。所以，对模型$Y_i=\beta_0+\beta_1X_i+\varepsilon_i$ 检验$\beta_1=0$ 是否成立，等价于检验Y与X之间是否存在线性关系。

9.5.1 b_1的抽样分布

为了检验$\beta_1=0$ 是否成立，需要构造一个合适的统计量，因此，首先讨论b_1的抽样分布。

因为b_1具有线性特性，即b_1是观测值Y_i的线性组合，而Y_i是正态分布的随机变量，且相互独立，故b_1也是服从正态分布的随机变量。

又因b_1具有无偏性，即其均值$E(b_1)=\beta_1$。

以下可以证明，b_1的方差 $\sigma^2(b_1)=\dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$ (9-13)

因为 $b_1=\sum_{i=1}^{n}C_iY_i$，且Y_i相互独立，其中 $C_i=\dfrac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$

$$\sigma^2(b_1)=\sum_{i=1}^{n}\sigma^2(C_iY_i)=\sum_{i=1}^{n}C_i^2\sigma^2(Y_i)=\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$$

所以，b_1是服从 $N(\beta_1,\dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2})$ 的随机变量。

9.5.2 F 检验

在一元线性回归中，为了检验Y对于X线性关系的统计显著性，对 β_1进行F检验

1^0 提出假设：H_0：$\beta_1=0$，H_1：$\beta_1\neq0$。若原假设成立表明Y与X无显著线性关系。

2^0 构造并计算统计量：$F=\dfrac{SSR/f_R}{SSE/f_E}$ (9-14)

3^0 查F分布临界值表，得临界值 $F_\alpha(1,n-2)$

4^0 比较：若F< $F_\alpha(1,n-2)$ 接受H_0，认为Y与X不存在一元线性关系。

表 9-1 方差分析表

变差来源	平方和	自由度	均方差	F比
回归	$SSR=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	1	$MSR=\frac{SSR}{1}$	$F=\frac{MSR}{MSE}$
误差	$SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	n-2	$MSE=\frac{SSE}{n-2}$	
总和	$SSTO=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	n-1	----------	----------

若F> $F_{\alpha}(1,n-2)$ 拒绝H_0，认为Y与X存在一元线性关系。

将各部分计算结果集中列于方差分析表内如表 9-1 所示

9.5.3 t 检验

1° 提出假设 H_0：$\beta_1=0$，H_1：$\beta_1\neq0$。若原假设成立表明Y与X无显著线性关系。

2° 构造并计算统计量：$t=\frac{b_1}{s(b_1)}$ (9-15)

其中 $s(b_1)=\sqrt{\frac{MSE}{\sum(X_i-\bar{X})^2}}$

3° 查 t 分布临界值表,得临界值 $t_{\alpha/2}(n-2)$

4° 比较:

若 $|t|<t_{\alpha/2}(n-2)$，接受H_0,认为Y与X不存在一元线性关系。

若 $|t|>t_{\alpha/2}(n-2)$，拒绝H_0

9.5.4 利用样本相关系数进行统计检验

1° 提出假设

H_0: $\rho = 0$

H_1: $\rho \neq 0$ （ρ:为总体Y与X的线性相关系数）

2° 计算简单相关系数γ

3° 查相关系数临界值表,得临界值 $\gamma_\alpha(n-2)$

4° 比较:

若$|\gamma| < \gamma_\alpha$，接受H_0，认为Y与X不存在一元线性关系。

若$|\gamma| > \gamma_\alpha$，拒绝H_0

9.6 模型适合性分析

在对一元线性回归模型的适合性进行分析时,由于误差项ε是不可观测或测量的, 需借助残差$e_i = Y_i - \hat{Y}_i$的图像,来考察模型是否存在以下情况：异方差性和自相关性。

9.6.1 误差项的异方差性检验

若ε_i不具有常数方差,称模型存在异方差性。此时,残差如图 9-6 所示，数据点呈现发散或收敛趋势。 在此种情况下,最小二乘法失效,因此需按照一定方法对数据进行变换,在计量经济学课程中,对此有详细讲述。

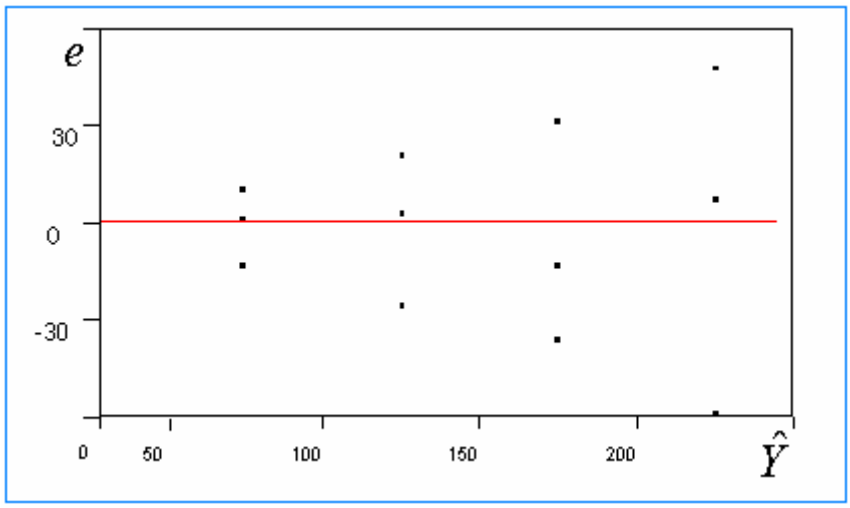


图 9-6 误差项具有异方差性的残差图

2、误差项的自相关性检验

如果观测值是来自一个时间序列的样本，则很可能出现误差项 ε_i 是不独立的，将残差 e_t 与时间 t 作残差图，将呈现出有规则的变化趋势。

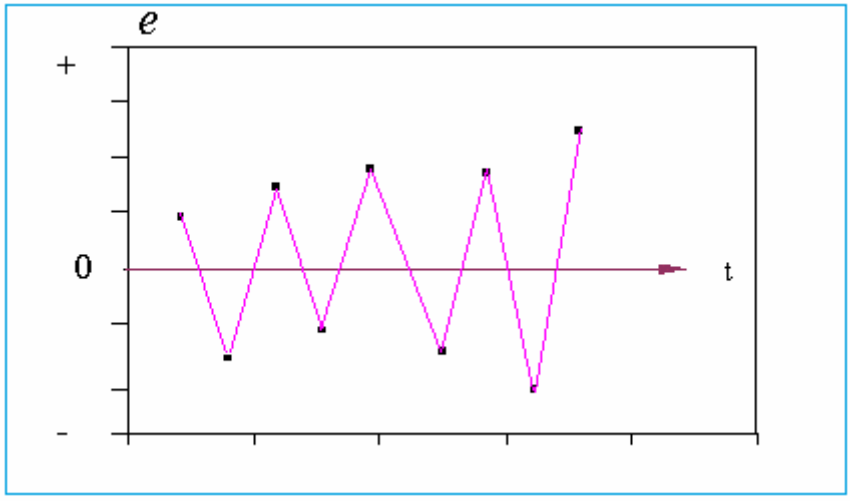


图 9-7 误差项具有负自相关性的残差图

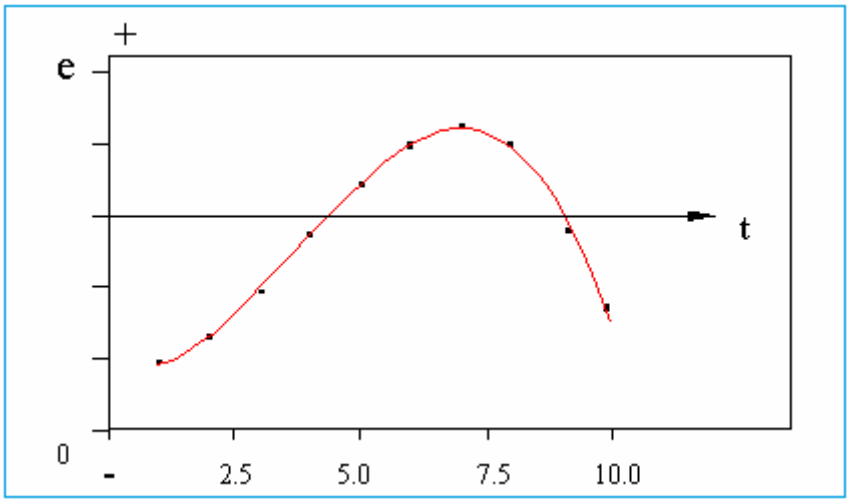


图 9-8 误差项具有正自相关性的残差图

我们称模型存在自相关(Autocorrelation)现象，也需按一定方法对数据进行修正，这在计量经济学课程中也有详细论述。

9.7 E(Y)的区间估计

因为误差项 ε_i 的方差（或 Y_i 的方差）σ^2 通常是未知的，因此，$\hat{Y}_0$ 的方差 $\sigma^2(\hat{Y}_0)$ 也是未知的。由于 *MSE* 是 σ^2 的无偏估计量，故可以用 *MSE* 代替 σ^2，从而得到 $\hat{Y}_0$ 的估计方差，记为 $S^2(\hat{Y}_0)$。

$$S^2(\hat{Y}_0) = MSE[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}] \tag{9-16}$$

可以证明，对于一元线性模型，以下随机变量

$$t = \frac{\hat{Y}_0 - E(Y_0)}{S(Y_0)} \tag{9-17}$$

服从自由度为n-2 的 t分布。因此可以得到，在置信度为 1- α 时，E(Y_0) 的置信区间为：

$$\hat{Y}_0 - t_{\alpha/2}(n-2)S(\hat{Y}_0) \le E(Y_0) \le \hat{Y}_0 + t_{\alpha/2}(n-2)S(\hat{Y}_0) \tag{9-18}$$

其中：$S(\hat{Y}_0) = \sqrt{MSE[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}]}$

9.8 因变量 Y 的预测

设：$d_0 = Y_0 - \hat{Y}_0$ (9-19)

因为 ε_i（i=1, 2, …, n）是相互独立的，故 Y_0 与 $\hat{Y}_0$ 也是相互独立的，所以有：

$$\sigma^2(d_0) = \sigma^2(Y_0 - \hat{Y}_0) = \sigma^2(Y_0) + \sigma^2(\hat{Y}_0) = \sigma^2 + \sigma^2(\hat{Y}_0) \tag{9-20}$$

$\sigma^2(\hat{Y}_0)$ 中也含有 σ^2，而 σ^2 未知，故用它的无偏估计量 MSE 代替，于是得到 $\sigma^2(d_0)$ 的无偏估计量，记为 $S^2(d_0)$。

$$\begin{aligned} S^2(d_0) &= MSE + S^2(\hat{Y}_0) \\ &= MSE[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}] \end{aligned} \tag{9-21}$$

可以证明，随机变量

$$t = \frac{Y_0 - \hat{Y}_0}{\sigma\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}} \tag{9-22}$$

服从自由度为n-2 的t分布。在给定置信度 1-α情况下，因变量单个值Y_0的置信区间为：

$$\hat{Y}_0 - t_{\alpha/2}(n-2)S(d_0) \le Y_0 \le \hat{Y}_0 + t_{\alpha/2}(n-2)S(d_0) \tag{9-23}$$

习 题

1.有 10 个同类企业的生产性固定资产价值和工业总产值资料如下：

企业编号	生产性固定资产价值（万元）	工业总产值（万元）
1	318	524
2	910	1019
3	200	638
4	409	815
5	415	913
6	502	928
7	314	605
8	1210	1516
9	1022	1219
10	1225	1624
合计	6525	9801

a.说明两变量之间的相关方向；

b.建立直线回归方程；

c.计算估计标准差；

d.估计生产性固定资产（自变量）为 1100 万元时总产值（因变量）的可能值。

2.下表中的数据是主修信息系统专业并获得企业管理学士学位的学生，毕业后的月薪（用 y 表示）和他在校学习时的总评分（用 x 表示）。由这些数据估计的回归方程是 $\hat{y} = 1290.5 + 581.1x$ 。

总评分	月薪/元	总评分	月薪/元
2.6	2800	3.2	3000
3.4	3100	3.5	3400
3.6	3500	2.9	3100

a.计算 SST，SSR 和 SSE 。

b.计算样本确定系数r^2。请对拟合优度做出评述。

c.样本相关系数的数值是多少？

3．根据上一练习

a.t 检验是否标明在总评分和月薪之间存在一个显著的关系？你的结论是什么？取 α=0.05。

b.利用 F 检验，检验变量间的显著关系。你的结论是什么？取 α =0.05。
c.做出 ANOVA 表。

4. 若 x 表示公寓住宅的年租金（千元），y 表示该公寓住宅的销售价格（千元），一家房地产公司的营业部门对 x 和 y 之间的关系进行回归分析。采集了近期出售的一些公寓住宅建筑的数据，经过计算机处理得到的输出如下。

回归方程是 $\hat{y} = 20.0 + 7.21x$			
预测量	系数	标准差	t 值
常数	20.000	3.2213	6.21
x	7.210	1.3626	5.29
方差分析			
来源	自由度	平方和	
回归	1	41587.3	
误差	7		
总计	8	51984.1	

a.样本中有多少公寓住宅建筑？
b.写出估计的回归方程。
c.s(b_1)的值是多少？

5.某公司采集了市场上办公用房的空闲率和租金率的数据。对于 18 个选取的销售地区，下面是这些地区的中心商业区的综合空闲率（%）和平均租金率（元/平米）的数据。

地区编号	综合空闲（%）	平均租金率（元/平方米）
1	21.9	18.54
2	6.0	33.70
3	22.8	19.67
4	18.1	21.01
5	12.7	35.09
6	14.5	19.41
7	20.0	25.28
8	19.2	17.02
9	16.0	24.04
10	6.6	31.42
11	15.9	18.74
12	9.2	26.76
13	19.7	27.72
14	20.0	18.20
15	8.3	25.00

16	17.1	29.78
17	10.8	37.03
18	11.1	28.64

a.用横轴表示空闲率，对这些数据画出散点图。
b.这两个变量之间能显示出什么关系吗？
c.求出在办公用房的综合空闲率已知时，能用来预测平均租金率的估计的回归方程。
d.在 0.05 显著性水平下检验关系的显著性。
e.估计的回归方程对数据的拟合好吗？请作出解释。
f.在一个综合空闲率是 25%的中心商业区，预测该市场的期望租金率。
g.若有某市的中心商业区，综合空闲率是 11.3%，预测该市中心商业区的期望租金率。