第六章　回归分析

6.1　一元线性回归分析

一、一元线性回归模型

二、未知参数的估计

三、参数估计量的分布

四、参数β的显著性检验

五、预测和控制

0、回归分析的基本思想

变量之间的关系
- 确定性关系
- 相　关关系

$S = \pi r^2$　　确定性关系

身高和体重　　相关关系

相关关系的特征是:变量之间的关系很难用一种精确的方法表示出来.

确定性关系和相关关系的联系

由于存在测量误差等原因,确定性关系在实际问题中往往通过相关关系表示出来;另一方面,当对事物内部规律了解得更加深刻时,相关关系也有可能转化为确定性关系.

回归分析——处理变量之间的相关关系的一种数学方法,它是最常用的数理统计方法.

回归分析 → 线性回归分析 → 一元线性回归分析 / 多元线性回归分析

回归分析 → 非线性回归分析

一、一元线性回归的数学模型

回归分析的任务——根据试验数据估计回归函数;讨论回归函数中参数的点估计、区间估计;对回归函数中的参数或者回归函数本身进行假设检验;利用回归函数进行预测与控制等等.

问题的一般提法

对x的一组不完全相同的值 $x_1, x_2, \cdots, x_n$，设$Y_1, Y_2, \cdots, Y_n$分别是在 $x_1, x_2, \cdots, x_n$处对 Y的独立观察结果 .

称$(x_1, Y_1), (x_2, Y_2), \cdots, (x_n, Y_n)$是一个样本.

对应的样本值记为

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n).$$

利用样本来估计 Y关于x的回归函数 $\mu(x)$.

求解步骤

1.推测回归函数的形式

方法一　根据专业知识或者经验公式确定;

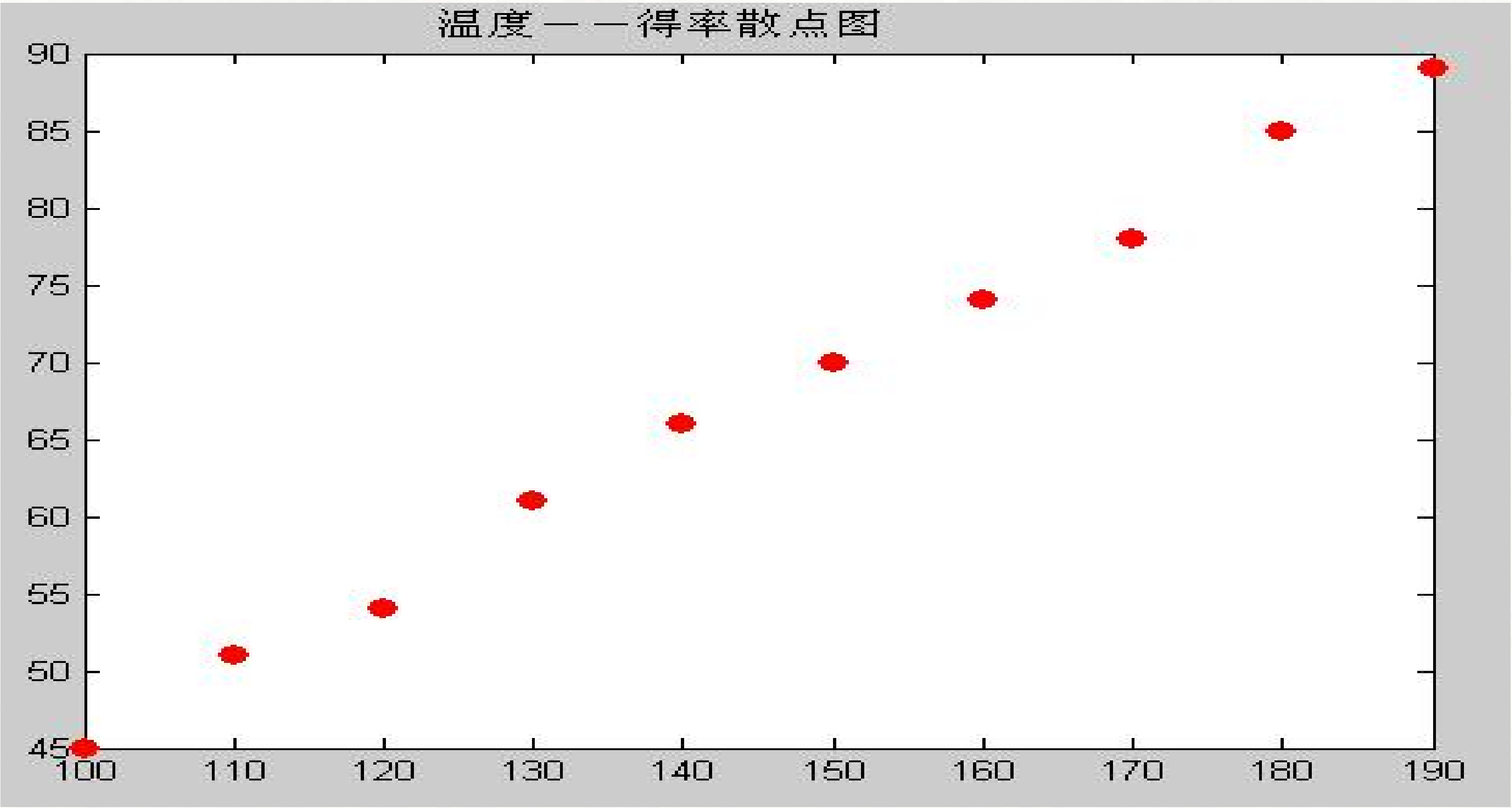
方法二　作散点图观察.

例1　为研究某一化学反应过程中,温度$x(^oC)$对产品得率$Y(\%)$的影响,测得数据如下.

温度$x(^oC)$	100	110	120	130	140	150	160	170	180	190
得率$Y(\%)$	45	51	54	61	66	70	74	78	85	89

用*MATLAB*画出散点图

```
x=100:10:190;y=[45,51,54,61,66,70,74,78,85,89];
plot(x,y,'.r')
```



观察散点图,$\mu(x)$具有线性函数$\alpha+\beta x$的形式.

2.建立回归模型

$\mu(x)=\alpha+\beta x$ 　一元线性回归问题

假设对于x的每一个值有$Y_i \sim N(\alpha+\beta x_i,\sigma^2)$,

α,β,σ^2都是不依赖于x的未知参数.

记$\varepsilon_i=Y_i-(\alpha+\beta x_i)$,那么

$$\left\{\begin{array}{l} Y_i=\underbrace{\alpha+\beta x_i}_{x\text{的线性函数}}+\underbrace{\varepsilon_i}_{\text{随机误差}}, \quad \varepsilon_i \sim N(0,\sigma^2). \\ \alpha,\beta,\sigma^2\text{是不依赖于}x\text{的未知参数}. \end{array}\right.$$

一元线性回归模型

二、未知参数的估计

$$Y=\alpha+\beta x+\varepsilon, \quad \varepsilon \sim N(0,\sigma^2).$$

对于样本$(Y_1,x_1),(Y_2,x_2),\cdots,(Y_n,x_n)$

$Y_i=\alpha+\beta x_i+\varepsilon_i, \varepsilon_i \sim N(0,\sigma^2)$，各$\varepsilon_i$相互独立.

于是$Y_i \sim N(\alpha+\beta x_i,\sigma^2), i=1,2,\cdots,n.$

需要对参数α，β及σ^2进行估计。

1. (α,β)的最小二乘估计(Least square estimation)

使得下式成立的(α, β)称为其最小二乘估计.

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

设 $Q(\alpha,\beta)=\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)^2$,求偏导可得

$$\frac{\partial Q(\alpha,\beta)}{\partial\alpha}=-2\sum_{i=1}^{n}(Y_i-\alpha-\beta x_i)=0$$

$$\frac{\partial Q(\alpha,\beta)}{\partial\beta}=-2\sum_{i=1}^{n}x_i(Y_i-\alpha-\beta x_i)=0$$

$$\left.\begin{aligned}\frac{\partial Q}{\partial \alpha} &= -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0\\ \frac{\partial Q}{\partial \beta} &= -2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i=0\end{aligned}\right\}$$

$$\left.\begin{aligned}n\alpha+\Big(\sum_{i=1}^{n}x_i\Big)\beta &= \sum_{i=1}^{n}y_i\\ \Big(\sum_{i=1}^{n}x_i\Big)\alpha+\Big(\sum_{i=1}^{n}x_i^2\Big)\beta &= \sum_{i=1}^{n}x_iy_i\end{aligned}\right\}\text{正规方程组}$$

由于系数矩阵满足 $\begin{vmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{vmatrix} \neq 0,$

则

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

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

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

若令

$$l_{xx}=\sum(x_i-\overline{x})^2=\sum x_i^2-n\overline{x}^2$$

$$l_{xy}=\sum(x_i-\overline{x})(y_i-\overline{y}),\quad l_{yy}=\sum(y_i-\overline{y})^2$$

$$\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x},$$

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{l_{xy}}{l_{xx}}.$$

最小二乘估计法

- 18世纪，欧拉，拉普拉斯等被对于怎样充分利用全部的观测结果 ,以期得到一个效率高的估计这个问题曾困扰。
- LSE是勒让德 (A. M . Legendre)于 1805年在其著作《计算慧星轨道的新方法》中提出，解决了此问题。
- 高斯在正态误差下利用LSE ,大大提高了 LSE在实用上的方便和广泛性。1809年，著作《关于绕日行星运动的理论》。
- 由于正态误差理论对这个方法的重要意义，归功于高斯。

从一种“事后诸葛亮”的眼光，我们现在看起来会觉得这个方法似乎平淡无奇，甚至是理所当然的．这正说明了创造性思维之可贵和不易．从一些数学大家未能在这个问题上有所突破，可以看出当时这个问题之困难．欧拉、拉普拉斯在许多很困难的数学问题上有伟大的建树，但在这个问题上未能成功．除了在思想上囿于“解方程”这一思维定势之外，也许还因为，这是一个实用性质的问题而非纯数学问题．解决这种问题，需要一种植根于实用而非纯数学精确性的思维．例如，按数学理论，容器以做成球形最省，但基于实际以至美学上的原因，在现实中有各种形状的容器存在．总之，从 LSE发现的历史中，使我们对纯数学和应用数学思维之间的差别，多少有一些启示．

---《最小二乘法的历史回顾与现状》

--- 陈希孺 院士

2. (α,β)的最大似然估计

根据 $Y_1, Y_2, \cdots, Y_n$ 的独立性可得到联合密　度函数为

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right].$$

L 取最大值等价于

$$Q(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

取最小值. 这就回到了最小二乘估计的情形。

注 在正态假设下参数的最小二乘估计等价于最大似然估计。

将$\hat{\alpha}$，$\hat{\beta}$代入$Y=\alpha+\beta x$，得

$\hat{Y}=\hat{\alpha}+\hat{\beta}x$　称其为Y关于x的线性回归方程

又由于$\hat{\alpha}=\bar{y}-\hat{\beta}\bar{x}$,

$$\hat{y}=\bar{y}+\hat{\beta}(x-\bar{x}) \longrightarrow \hat{y}-\bar{y}=\hat{\beta}(x-\bar{x})$$

显然，回归直线通过散点图的几何中心$(\bar{x},\bar{y})$.

注　$\hat{y}-\bar{y}=\hat{\beta}(x-\bar{x}) \longrightarrow \hat{y}_i-\bar{y}=\hat{\beta}(x_i-\bar{x}), i=1,2,\cdots,n.$

3. 未知参数σ^2的估计

$$Y=\alpha+\beta x+\varepsilon,\quad \varepsilon\sim N(0,\sigma^2).$$

$$E\{[Y-(\alpha+\beta x)]^2\}=E(\varepsilon^2)=D(\varepsilon)+[E(\varepsilon)]^2=\sigma^2$$

则$\hat{\sigma}^2$的估计为 $$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2$$

显然σ^2越小,用回归函数$\mu(x)=\alpha+\beta x$作为Y的近似导致的均方误差就越小.

$$\hat{y}_i=\hat{y}\Big|_{x=x_i}=\hat{\alpha}+\hat{\beta}x_i$$ $y_i-\hat{y}_i$ x_i处的残差

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2$$ **残差平方和**

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}[Y_i-(\hat{\alpha}+\hat{\beta}x_i)]^2=\frac{1}{n}Q_e$$

对于Q_e， $Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[(y_i-\overline{y})-(\hat{y}_i-\overline{y})]^2$

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

由于 $\hat{y}-\overline{y}=\hat{\beta}(x-\overline{x})$ $\Rightarrow$ $\hat{y}_i-\overline{y}=\hat{\beta}(x_i-\overline{x})$,

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

$$=\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})^2,$$

$$2\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{y})=2\sum_{i=1}^{n}(y_i-\overline{y})\cdot\hat{\beta}\cdot(x_i-\overline{x})$$

$$=2\hat{\beta}\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})=2\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2,$$

$$\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2,$$

则 $$Q_e=\sum_{i=1}^{n}(y_i-\overline{y})^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2-2\sum_{i=1}^{n}(y_i-\overline{y})(\hat{y}_i-\overline{y})$$

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

$$Q_e=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\overline{y})^2-\hat{\beta}^2\sum_{i=1}^{n}(x_i-\overline{x})^2$$

$$=\sum_{i=1}^{n}(y_i-\overline{y})^2-\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$$

因而

$$\hat{\sigma}^2=\frac{1}{n}Q_e=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2-\hat{\beta}^2\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2$$

此估计不一定是无偏估计

注：
$$\underbrace{\sum_{i=1}^{n}(y_i-\overline{y})^2}_{S_T}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{Q_e}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}_{S_R}$$

S_T	Q_e	S_R
总离差平方和	残差平方和	回归平方和

这里也有平方和分解

例2（p193例6.2） 设父亲和他们长子的身高分别为

$(x_i, y_i)(i=1,2,\cdots,12)$,其观测数据为

父亲身高x	65	63	67	64	68	62	70	66	68	67	69	71
长子身高y	68	66	68	65	69	66	68	65	71	67	68	70

求Y关于x的线性回归方程

解 回归方程为$\hat{Y}=\hat{\alpha}+\hat{\beta}x$ 将观测值代入正规方程

$$\begin{cases} n\hat{\alpha}+(\sum_{i=1}^{n} x_i)\hat{\beta}=\sum_{i=1}^{n} y_i \\ (\sum_{i=1}^{n} x_i)\hat{\alpha}+(\sum_{i=1}^{n} x_i^2)\hat{\beta}=\sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\begin{cases} 12\hat{\alpha} + 800\hat{\beta} = 811 \\ 800\hat{\alpha} + 53418\hat{\beta} = 54107 \end{cases}$$

求解得

$$\hat{\alpha} = 35.82 \qquad \hat{\beta} = 0.476$$

则Y关于x的线性回归方程为

$$\hat{Y} = 35.82 + 0.476x$$

这个例子表明：高个子的先代会有高个子的后代，但后代的增高并不与先代的增高等量。例如父亲身高超过祖父身高6in,则儿子的身高超过父亲的身高大约为3in。

历史上有趣的发现

十九世纪，英国著名的统计学家**F.Galton**及其弟子**K.Pearson** ,研究了**1078**对夫妇及其一个成年儿子的身高关系。他们以儿子身高作为纵坐标、夫妇平均身高为横坐标作散点图，结果发现二者的关系近似于一条直线。经计算得到了如下方程：

$$\hat{y} = 33.73 + 0.516x$$

由此方程可以看到：夫妇平均身高增加或减少一个单位，儿子的身高只增加或减少 0.516个单位。也就是说，子代的身高就不像父辈身高那样分化，而是逐渐向平均身高回归。Galton引进“回归”（regression)一词来表达这种变化关系。不过后来人们研究其它变量间的关系时，并没有发现如上所述的回归现象，但仍沿用“回归”的概念以纪念统计学家F.Galton。

三、参数估计的分布

为了对参数估计量进行检验，需要讨论它们的分布

1. $\hat{\beta}$的分布

$$\hat{\beta}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})Y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$=\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}Y_i=\sum_{i=1}^{n}a_iY_i$$

$\rightarrow a_i$

由于$Y_1, Y_2, \cdots, Y_n$相互独立，而且$Y_i \sim N(\alpha+\beta x_i, \sigma^2)$

因而$\hat{\beta}$服从正态分布。

$\hat{\beta}$期望和方差分别为

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

$$E\hat{\beta} = \sum_{i=1}^{n} a_i EY_i = \sum_{i=1}^{n} a_i(\alpha + \beta x_i) = \beta \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta$$

$$D\hat{\beta} = \sum_{i=1}^{n} a_i^2 DY_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2}{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

则 $\hat{\beta} \sim N(\beta, \sigma^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2)$

2. $\hat{\alpha}$的分布

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} Y_i \right) \bar{x}$$

$$= \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right] Y_i$$

因而$\hat{\alpha}$服从正态分布，其期望值为

$$E\hat{\alpha} = E\bar{Y} - E\hat{\beta}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \beta\bar{x} = \alpha$$

$$D\hat{\alpha}=\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^2 DY_i$$

$$=\left[\frac{1}{n}+\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2\bar{x}^2}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}\right]\sigma^2 \quad =\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2$$

则 $\hat{\alpha}\sim N\left(\alpha,\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2\right)$

3. 对$x=x_0$,回归方程$\hat{Y}_0=\hat{\alpha}+\hat{\beta}x_0$的分布

$$\hat{Y}_0=\hat{\alpha}+\hat{\beta}x_0=\sum_{i=1}^{n}\left[\frac{1}{n}-\frac{(x_i-\bar{x})\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]Y_i+\frac{\sum_{i=1}^{n}(x_i-\bar{x})x_0}{\sum_{i=1}^{n}(x_i-\bar{x})^2}Y_i$$

$$=\sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]Y_i$$

因而$\hat{Y}_0$服从正态分布，其期望值为

$$E\bar{Y}_0=E(\hat{\alpha}+\hat{\beta}x_0)=\alpha+\beta x_0$$

$$D(\hat{Y}_0)=\sum_{i=1}^{n}\left[\frac{1}{n}+\frac{(x_i-\bar{x})(x_0-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^2 DY_i$$

$$=\left[\frac{1}{n}+\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2(x_0-\bar{x})^2}{(\sum_{i=1}^{n}(x_i-\bar{x})^2)^2}\right]\sigma^2=\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2$$

则 $\hat{Y}_0 \sim N\left(\alpha+\beta x_0,\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2\right)$

4. $\hat{\sigma}^2$的分布 （复杂）

先计算$\hat{\sigma}^2$的期望，看$\hat{\sigma}^2$是否是σ^2的无偏估计。

$$\hat{\sigma}^2=\frac{1}{n}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$=\frac{1}{n}\left[\sum_{i=1}^{n}Y_i^2-n\bar{Y}^2-\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$E\hat{\sigma}^2=\frac{1}{n}\left[\sum_{i=1}^{n}EY_i^2-nE(\bar{Y})^2-E(\hat{\beta}^2)\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$=\frac{1}{n}\left[\sum_{i=1}^{n}EY_i^2-nE(\bar{Y})^2-(D\hat{\beta}+E^2(\hat{\beta}))\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$
=\frac{1}{n}\left[\sum_{i=1}^{n}EY_i^2-nE(\bar{Y})^2-(D\hat{\beta}+E^2(\hat{\beta}))\sum_{i=1}^{n}(x_i-\bar{x})^2\right]
$$

$$
=\frac{1}{n}\left[\sum_{i=1}^{n}[DY_i+(EY_i)^2]-n[D\bar{Y}+(E\bar{Y})^2]\right.
$$

$$
\left.-[D\hat{\beta}+E^2(\hat{\beta})]\sum_{i=1}^{n}(x_i-\bar{x})^2\right]
$$

$$
=\frac{1}{n}\left[\sum_{i=1}^{n}[\sigma^2+(\alpha+\beta x_i)^2]-n\left(\frac{\sigma^2}{n}+(\alpha+\beta\bar{x})^2\right)\right.
$$

$$
\left.-\left(\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\beta^2\right)\sum_{i=1}^{n}(x_i-\bar{x})^2\right]
$$

$$=\frac{1}{n}\left[(n-1)\sigma^2+\beta^2\sum_{i=1}^{n}(x_i-\bar{x})^2-\sigma^2-\beta^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

$$=\frac{1}{n}(n-2)\sigma^2=\frac{(n-2)}{n}\sigma^2 \neq \sigma^2$$

结论： $$\hat{\sigma}^2=\frac{1}{n}\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\hat{\beta}^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$$

不是σ^2的无偏估计！

设 $\hat{\sigma}^{*2}=\frac{n}{n-2}\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2=\frac{1}{n-2}Q_e$

$$=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\frac{\hat{\beta}^2}{n-2}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

则 $E\hat{\sigma}^{*2}=\sigma^2$ 无偏估计

对于 $\hat{\sigma}^{*2}=\dfrac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)^2=\dfrac{1}{n-2}Q_e$

$$=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\hat{\beta}^2\frac{1}{n-2}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

有如下结论说明其分布。

定理6. 1 假设(Y_i, x_i)满足线性回归模型的条件，则

$$\frac{(n-2)}{\sigma^2}\hat{\sigma}^{*2}=\frac{1}{\sigma^2}Q_e\sim\chi^2(n-2)$$

而且$\hat{\sigma}^{*2}$分别与$\hat{\alpha},\hat{\beta}$独立，其中$\hat{\sigma}^{*2}$是σ^2的无偏估计.

证明参见下一节多元回归理论

四、参数β的显著性检验

根据前三小节的理论，给定一组观测值，就可以得其相应的回归方程。但是二者是否具有此种相关关系，需要进行必要的检验.

通常检验一元线性回归模型是否成立，需要检验：

(1) 给定x时，Y服从正态分布且方差相等；

(2) 对于给定的范围，EY是x的线性函数；

(3) $Y_1, Y_2, \cdots, Y_n$相互独立.

本节主要讨论第二类问题，也就是等价于β是否为0.

三种等价的检验方法

(一)、t 检验

设 $Y_i=\alpha+\beta x_i+\varepsilon_i,\quad \varepsilon_i \sim N(0,\sigma^2).$

检验假设: $H_0:\beta=0,\quad H_1:\beta\neq 0.$

$$\hat{\beta}\sim N\left(\beta,\sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2\right),\ \frac{(n-2)\hat{\sigma}^{*2}}{\sigma^2}=\frac{Q_e}{\sigma^2}\sim\chi^2(n-2).$$

并且$\hat{\beta},\hat{\sigma}^{*2}$相互独立,因此

$$T=\frac{\hat{\beta}-\beta}{\hat{\sigma}^*}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2).$$

当H_0为真时$\beta=0$,此时$T=\dfrac{\hat{\beta}}{\hat{\sigma}^*}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2).$

则，H_0的拒绝域为

$$|t| = \frac{|\hat{\beta}|}{\hat{\sigma}^*}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \geq t_{\alpha/2}(n-2)$$

拒绝$H_0: \beta \neq 0$,认为回归效果显著.

接受$H_0: \beta = 0$,认为回归效果不显著.

对于给定的显著性水平α,可构造检验步骤如下：

(1)$H_0 : \beta = 0$;

(2)构造检验统计量$T = \hat{\beta}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 / \hat{\sigma}^*}$;

(3)对于给定的α,查分位数$t_{\alpha/2}(n-2)$;

(4)对给定的一组回归观测值，带入检验统计量计算得t,

如果$|t| \geq t_{\alpha/2}(n-2)$，则拒绝$H_0$，否则接受$H_0$.

回归效果不显著的原因分析

(1)影响Y取值的,除x及随机误差外还有其他不可忽略的因素;

(2)$E(Y)$与x的关系不是线性的;

(3)Y与x不存在关系.

例3(p197例6.3) 检验例2中的回归效果是否显著,取显著性水平为0.05.

解 对$\alpha=0.05, n-2=10$,查表得

查表得$t_{0.05/2}(n-2)=t_{0.025}(10)=2.2281$

$$|t|=3.128$$

$$|t|>t_{0.025}(10).$$

拒绝$H_0:\beta=0$,认为回归效果显著.

(二)、F检验　采用方差分析的思想

我们从数据出发研究各y_i不同的原因。

数据总的波动用总偏差平方和$S_T=\sum(y_i-\bar{y})^2=l_{yy}$表示。

引起各y_i不同的原因主要有两个因素：其一是H_0可能不真，$E(y)$随x的变化而变化，从而在每一个x的观测值处的回归值不同，其波动用回归平方和$S_R=\sum\left(\hat{y}_i-\bar{y}\right)^2$ 表示；

其二是其它一切因素，包括随机误差、x对$E(y)$的非线性影响等，这可用残差平方和$Q_e=\sum\left(y_i-\hat{y}_i\right)^2$表示。

满足平方和分解式：$S_T = Q_e + S_R$ 即

$$\sum(y_i-\bar{y})^2=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\bar{y})^2$$

$$=\sum(y_i-\hat{y}_i)^2+\hat{\beta}^2 l_{xx}$$

且 $E\sum(y_i-\bar{y})^2=(n-1)\sigma^2$，$\mathrm{E}\sum(y_i-\hat{y}_i)^2=(n-2)\sigma^2$.

再利用 $\hat{\beta}\sim N(\beta,\sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2)$

有 $E(S_R)=E\hat{\beta}^2 l_{xx}=l_{xx}(\sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2+\beta^2)=\sigma^2+\beta^2 l_{xx}$

上述结论归纳为如下定理。

定理 设$y_i=\alpha+\beta x_i+\varepsilon_i$，其中$\varepsilon_i$，…，$\varepsilon_n$相互独立，且$\varepsilon_i \sim N(0, \sigma^2)$，$i=1,\ldots,n$，沿上面的记号，有

$$E(S_R)=\sigma^2+\beta^2 l_{xx}$$

$$E\left(Q_e/(n-2)\right)=\sigma^2$$

$\hat{\sigma}^{*2}=Q_e/(n-2)$ 是σ^2的无偏估计。

进一步，有关S_R和Q_e的分布，有如下定理：

定理 设$y_1, y_2, \ldots, y_n$相互独立，且$y_i \sim N(\alpha+\beta x_i, \sigma^2)$，$i=1, \ldots, n$，则在上述记号下，有

（1）$Q_e/\sigma^2 \sim \chi^2(n-2)$，

（2）若H_0成立，则有$S_R/\sigma^2 \sim \chi^2(1)$

（3）S_R与$Q_e, \bar{Y}$独立（或$\hat{\beta}$与$Q_e, \bar{Y}$独立）。

如同方差分析那样，我们可以考虑采用F比

$$F=\frac{S_R}{Q_e/(n-2)}=\frac{S_R/1}{Q_e/(n-2)}$$

作为检验统计量：

在$\beta=0$时，$F\sim F(1,n-2)$，其中$f_R=1, f_e=n-2.$

对于给定的显著性水平a，拒绝域为

$F\geq F_a(1,n-2).$ 整个检验也可列成一张方差分析表。

注：也可利用T分布和F分布的关系得到F检验统计量。

$$F=\frac{S_R}{Q_e/(n-2)}=\frac{S_R/1}{Q_e/(n-2)}=T^2=\frac{\hat{\beta}\sum_{i=1}^{n}(x_i-\overline{x})^2}{\hat{\sigma}^{*2}}$$

例 3　合金的强度y (×107Pa) 与合金中碳的含量x (%) 有关。为研究两个变量间的关系。首先是收集数据，我们把收集到的数据记为(xi,yi),i=1,2,…,n。本例中，我们收集到12组数据，列于下表中

表：　合金钢强度y与碳含量x的数据

序号	x(%)	y ($\times 10^7$Pa)	序号	x(%)	y ($\times 10^7$Pa)
1	0.10	42.0	7	0.16	49.0
2	0.11	43.0	8	0.17	53.0
3	0.12	45.0	9	0.18	50.0
4	0.13	45.0	10	0.20	55.0
5	0.14	45.0	11	0.21	55.0
6	0.15	47.5	12	0.23	60.0

使用例中合金钢强度和碳含量数据，我们可求得回归方程，见下表.

计算表

$\sum x_i=1.90$	$n=12$	$\sum y_i=590.5$
$\overline{x}=0.1583$		$\overline{y}=49.2083$
$\sum x_i^2=0.3194$	$\sum x_i y_i=95.9250$	$\sum y_i^2=29392.75$
$n\overline{x}^2=0.3008$	$n\cdot\overline{x}\cdot\overline{y}=93.4958$	$n\overline{y}^2=29057.5208$
$l_{xx}=0.0186$	$l_{xy}=2.4292$	$l_{yy}=335.2292$
$\hat{\beta}=l_{xy}/l_{xx}=130.6022$ $\hat{\alpha}=\overline{y}-\overline{x}\hat{\beta}=28.5340$		

由此给出回归方程为: $\hat{y}=28.5340+130.6022$

在合金钢强度的例中，我们已求出了回归方程，这里我们考虑关于回归方程的显著性检验。经计算有

$$S_T = l_{yy} = 335.2292 \qquad f_T = 11$$

$$S_R = \hat{\beta}^2 l_{xx} = 130.6022^2 \times 0.0186 = 317.2589, \qquad f_R = 1$$

$$Q_e = S_T - S_R = 335.2292 - 317.2589 = 17.9703 \qquad f_e = 10$$

来源	平方和	自由度	均方和	F比
回归	S_R=317.2589	f_A=1	MS_R=317.2589	176.55
残差	Q_e=17.9703	f_e=10	MQ_e= 1.79703	
总和	S_T=335.2292	f_T=11		

若取α=0.01，则$F_{0.01}(1,10)$ =10<F，因此在显著性水平0.01下回归方程是显著的。

(三)、相关系数检验

一元线性回归方程是反映两个随机变量x与y间的线性相关关系，它的显著性检验还可通过对二维总体相关系数r的检验进行。

由于 $S_T = Q_e + S_R$ 即

$$\sum (y_i - \overline{y})^2 = \sum \left(y_i - \hat{y}_i \right)^2 + \sum \left(\hat{y}_i - \overline{y} \right)^2$$

$$令 r^2 = \frac{S_R}{S_T} = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \overline{y} \right)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2},$$

显然$0 \le r^2 \le 1$, r^2越接近1,

说明误差平方和越小，线性方程越显著。

又因为 $S_R=\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2=\hat{\beta}^2 l_{xx}$

而 $\hat{\beta}=\dfrac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\dfrac{l_{xy}}{l_{xx}}$,

所以 $S_R=\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2=\hat{\beta}^2 l_{xx}=\dfrac{l^2{}_{xy}}{l^2{}_{xx}}l_{xx}=\dfrac{l^2{}_{xy}}{l_{xx}}$

所以$r^2=\dfrac{S_R}{S_T}=\dfrac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\dfrac{l^2{}_{xy}}{l_{xx}}\dfrac{1}{l_{yy}}=\dfrac{l^2{}_{xy}}{l_{xx}l_{yy}}$,

则 $r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$

称上式为相关系数。$-1\le r\le 1$.

假设 $H_0: r=0, \quad H_1: r \neq 0$

所用的检验统计量为样本相关系数

$$r=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x_i-\overline{x})^2\sum(y_i-\overline{y})^2}}=\frac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$$

拒绝域为$W=\{|r|\geq c\}$，其中临界值c是

$\mathrm{H}_0: r=0$成立下r的分布的a分位数，

故应有$c=r_a(n-2)$.

r与F统计量$F=\dfrac{S_R}{Q_e/(n-2)}$之间的关系

$$r^2=\frac{S_R}{S_T}=\frac{S_R}{S_R+Q_e}=\frac{1}{1+\dfrac{Q_e}{S_R}}=\frac{1}{1+\dfrac{1}{\dfrac{S_R}{Q_e}}}$$

$$=\frac{1}{1+\dfrac{n-2}{\dfrac{S_R}{Q_e/(n-2)}}}=\frac{1}{1+\dfrac{n-2}{F}}=\frac{F}{F+n-2}\quad\uparrow\text{增函数}$$

$$r^2 = \frac{F}{F+(n-2)}$$

故可以从F分布的a分位数$F_a(1,n-2)$得到

r的α分位数为

$$c = r_\alpha(n-2) = \sqrt{\frac{F_\alpha(1,n-2)}{F_\alpha(1,n-2)+(n-2)}}$$

譬如，对$a=0.01$, $n=12$, $F_{0.01}(1,10)=10.04$

$$r_{0.01}(10)=\sqrt{\frac{10.04}{10.04+10}}=0.708$$

于是为了使用方便，人们已对$r_a(n-2)$编制了专门的表，见附表9.

以例3中数据为例，可以计算得到

$$r=\frac{2.4292}{\sqrt{0.0186\times 335.2292}}=0.9728$$

若取$a=0.01$，查附表9知$r_{0.01}(10)=0.708$，由于$0.9728>0.708$，因此，在显著性水平0.01下回归方程是显著的。

注： 在一元线性回归场合，三种检验方法是等价的：在相同的显著性水平下，要么都拒绝原假设，要么都接受原假设，不会产生矛盾。

F 检验可以很容易推广到多元回归分析场合，而其他二个则否，所以，F检验是最常用的关于回归方程显著性检验的检验方法。

五、预测

1. 系数β的置信区间

当回归效果显著时,对系数β作区间估计.

由 $T=\dfrac{\hat{\beta}-\beta}{\hat{\sigma}^*}\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sim t(n-2)$,

得系数β的置信水平为$1-\alpha$的置信区间为

$$\left(\hat{\beta}\ \pm\ t_{\alpha/2}(n-2)\times\frac{\hat{\sigma}^*}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right).$$

2. 回归函数函数值的点估计和置信区间

回归函数 $\mu(x_0)=\alpha+\beta x_0$的点估计:

由于$\hat{y}=\hat{\mu}(x)=\hat{\alpha}+\hat{\beta}x$

所以当$x=x_0$, $\hat{y}_0=\hat{\mu}(x_0)=\hat{\alpha}+\hat{\beta}x_0$,

估计量:$\hat{Y}_0=\hat{\alpha}+\hat{\beta}x_0$.

$\big\downarrow$

$\mu(x_0)=\alpha+\beta x_0$的点估计

回归函数 $\mu(x_0)=\alpha+\beta x_0$的区间估计：

由于 $$\hat{Y}_0 \sim N\left(\alpha+\beta x_0, \left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2\right)$$

则 $$\frac{\hat{Y}_0-(\alpha+\beta x_0)}{\sigma\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim N(0,1),$$

又因为 $\dfrac{(n-2)\hat{\sigma}^{*2}}{\sigma^2} \sim \chi^2(n-2)$,

$\hat{\sigma}^{*2}, \hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0$相互独立, 则

$$T = \frac{\hat{Y}_0 - (\alpha + \beta x_0)}{\hat{\sigma}^* \sqrt{\dfrac{1}{n} + \dfrac{(x_0 - \overline{x})^2}{\sum\limits_{i=1}^{n} (x_i - \overline{x})^2}}} \sim t(n-2),$$

于是$\mu(x_0)=\alpha+\beta x_0$的置信水平为$1-\alpha$的置信区间为

$$\left(\hat{\alpha}+\hat{\beta}x_0 \pm t_{\alpha/2}(n-2)\hat{\sigma}^* \sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right).$$

3. Y的观察值的点预测和预测区间

设Y_0是在$x = x_0$处对Y的观察结果.则Y_0与$Y_1,\ldots,Y_n$相互独立。

$$\hat{Y}_0 = \hat{\mu}(x_0) = \hat{\alpha} + \hat{\beta} x_0$$ 此式为Y_0的点预测

又因为 $$Y_0 - \hat{Y}_0 = Y_0 - (\hat{\alpha} + \hat{\beta} x_0)$$

由于Y_0与$\hat{Y}_0$相互独立，而且都服从正态分布，因而$Y_0 - \hat{Y}_0$亦服从正态分布，其期望值为

$$E(Y_0 - \hat{Y}_0) = EY_0 - E(\hat{\alpha} + \hat{\beta} x_0) = \alpha + \beta x_0 - \alpha - \beta x_0 = 0$$

$$D(Y_0-\hat{Y}_0)=DY_0+D(\hat{Y}_0)=\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2$$

$$Y_0-\hat{Y}_0\sim N\left(0,\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\sigma^2\right)$$

$\dfrac{(n-2)\hat{\sigma}^{*2}}{\sigma^2}\sim\chi^2(n-2)$, $\hat{\sigma}^{*2}$, $Y_0-\hat{Y}_0$相互独立, 则

$$T=\frac{Y_0-(\hat{\alpha}+\hat{\beta}x_0)}{\hat{\sigma}^*\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}}\sim t(n-2),$$

给定置信水平为$1-\alpha$, Y_0的预测区间为

$$\left(\hat{\alpha}+\hat{\beta}x_0\pm t_{\alpha/2}(n-2)\hat{\sigma}^*\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right)$$

设 $\delta(x_0)=t_{\alpha/2}(n-2)\hat{\sigma}^* \sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}$

于是 给定置信水平为$1-\alpha$, Y_0的预测上下限为

$$y_1(x_0)=\hat{Y}_0-\delta(x_0)$$

$$y_2(x_0)=\hat{Y}_0+\delta(x_0)$$

则这两条曲线形成带状域包含回归曲线.

例4(p198例6.4) (续例2)

(1)设$x_0 = 66.5, 1-\alpha = 0.95$

(2)设$x_0 = 70.3, 1-\alpha = 0.95$

试求出两种情形下，Y_0的置信上下限.

解 已知$n-2=10$, $t_{0.025}(10) = 2.2281$

(1) $\hat{y}_0 = 35.8 + 0.476 x_0 = 35.8 + 0.476 \cdot 65.5 = 66.998$

$$\delta(x_0) = \delta(66.5) = 2.2281 \times 1.40 \times \sqrt{1 + \frac{1}{12} + \frac{(66.5 - 800/12)^2}{(2.66)^2}}$$

$$\approx 2.2281 \times 1.40 \times 1.129 = 3.522$$

于是 给定置信水平为0.95, Y_0的预测区间为

$$(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0)) = (63.467, 70.52)$$

(2) 同理可得

$x_0 = 70.5$, 给定置信水平为0.95, Y_0的预测区间为

$$(\hat{y}_0 - \delta(x_0), \hat{y}_0 + \delta(x_0)) = (63.832, 74.924)$$

四、小结

1.回归分析的任务

研究变量之间的相关关系

2.一元线性回归的步骤

(1)推测回归函数; (2)建立回归模型;
(3)估计未知参数; (4)进行假设检验;
(5)预测与控制.

Thank You!