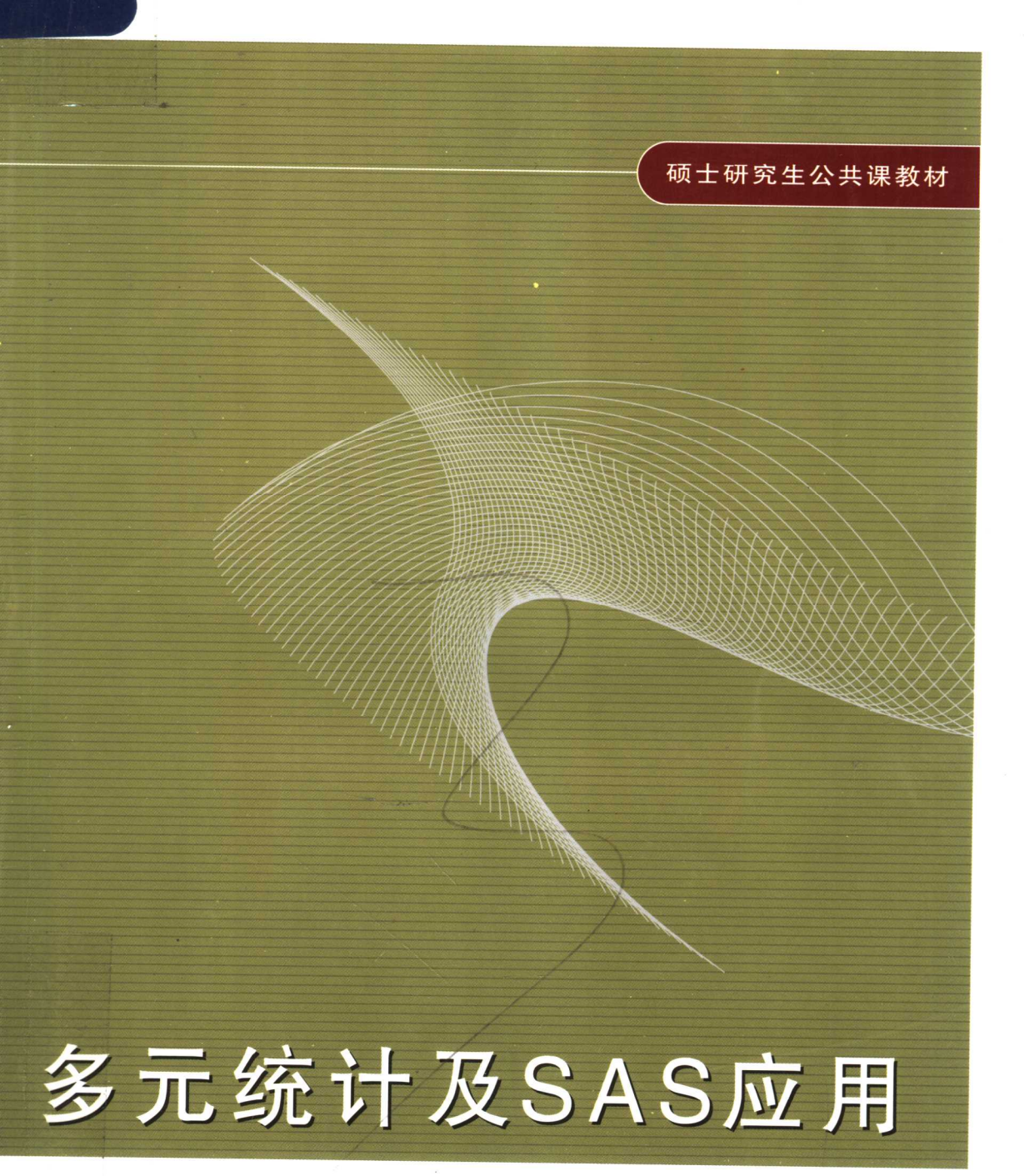
硕士研究生公共课教材



多元统计及SAS应用

余家林　肖枝洪　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

硕士研究生公共课教材

多元统计及SAS应用

余家林　肖枝洪　编著

图书在版编目(CIP)数据

多元统计及SAS应用/余家林,肖枝洪编著.—武汉:武汉大学出版社,2008.1

ISBN 978-7-307-06045-6

Ⅰ.多… Ⅱ.①余… ②肖… Ⅲ.多元分析:统计分析—应用软件,SAS—研究生—教材 Ⅳ.O212.4-39

中国版本图书馆CIP数据核字(2007)第195359号

责任编辑:杨 华 责任校对:黄添生 版式设计:詹锦玲

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)

(电子邮件:wdp4@whu.edu.cn 网址:www.wdp.com.cn)

印刷:湖北民政印刷厂

开本:720×1000 1/16 印张:14.375 字数:252千字

版次:2008年1月第1版 2008年1月第1次印刷

ISBN 978-7-307-06045-6/O·377 定价:26.00元

内 容 提 要

多元统计是数理统计学的一个分支。它根据多因素多指标试验与观测所得到的数据资料，对研究对象的特征及内在规律进行估计与推断，应用十分广泛。本教材包括多元线性回归、多元线性相关、多元非线性回归、回归的试验设计与分析、聚类分析、判别分析、主成分分析、因子分析及 SAS 的应用等内容。本书既可作为非数学专业硕士研究生多元统计课程的教材，也可作为科技工作者的参考文献。

前　言

多元统计是非数学专业硕士研究生教学计划中普遍开设的一门公共基础课，各学校各专业讲授的内容大体一致。随着硕士研究生入学水平与课题研究水平的提高，亟需一本相适应的教材，既能加强理论基础、帮助研究生熟悉多元统计原理，又能介绍近代流行的统计分析软件，使研究生在处理试验数据的过程中摆脱复杂计算的困扰。

由我们合编的《多元统计及 SAS 应用》是近几年来硕士研究生优质课程立项研究的一项成果。作为非数学专业硕士研究生的教材，编入了多元线性回归、多元线性相关、多元非线性回归、回归的试验设计与分析、聚类分析、判别分析、主成分分析、因子分析及 SAS 的应用等内容。讲课及上机实习可控制在 60 课时以内。

在编写中，我们特别注意说明统计方法的实际背景，详细讲述用统计方法解决实际问题的思路，对于应用 Statistical Analysis System（简称 SAS）所得到统计分析结果，则尽可能地与实际计算步骤一一对照，使初学者能够知其所以然。考虑到专业与课时设置的不同，本教材力求简明扼要，重点突出，通俗易懂，便于自学，例题与习题都在常识的范围之内。

教材中第一章、第二章、第三章由余家林编写，第四章、第五章、第六章由肖枝洪编写。本教材的出版得到华中农业大学研究生处及武汉大学出版社的大力支持，在此表示衷心的谢意。由于编者的水平所限，不妥之处难以避免，敬请读者和使用本教材的同行学友批评指正。

编　者

2007 年 10 月 9 日

目　录

第一章　多元线性回归

多元线性回归是一元线性回归的发展，可用来研究因变量取值与自变量取值的内在联系，建立多元线性回归方程. 在讲述多元线性回归的计算及其应用之前，为了承上启下，在 1.1 节中对一元线性回归的计算及其应用作了回顾. 熟悉这些内容的读者，不妨自 1.2 节开始读起.

1.1　一元线性回归

1.1.1　一元线性回归的概念

设自变量 x 的观测值 x_i 及因变量 y 对应的观测值 y_i 满足关系式

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\cdots,n,$$

式中，$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 是相互独立且都服从正态分布 $N(0,\sigma^2)$ 的随机变量.

根据最小二乘法，由 n 组观测值 (x_i,y_i) 确定参数 β_0 及 β_1 的估计值 b_0 及 b_1 后，所得到的估计式 $\hat{y} = b_0 + b_1 x$ 称为一元线性回归方程. 建立一元线性回归方程的过程以及对回归方程所作的显著性检验，称为一元线性回归分析或一元线性回归.

如果将 x_i 代入一元线性回归方程，记 $\hat{y}_i = b_0 + b_1 x_i$，则 $\hat{y}_i$ 与 y_i 之间的偏差平方和

$$Q = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - b_0 - b_1 x_i)^2.$$

由 $\dfrac{\partial Q}{\partial b_0} = 0$ 及 $\dfrac{\partial Q}{\partial b_1} = 0$ 可得到方程组

$$\begin{cases} nb_0 + b_1 \sum\limits_i x_i = \sum\limits_i y_i, \\ b_0 \sum\limits_i x_i + b_1 \sum\limits_i x_i^2 = \sum\limits_i x_i y_i. \end{cases}$$

解这个方程组，即可算出 b_0 及 b_1. 根据最小二乘法，b_0 及 b_1 的值使上述偏

差平方和 Q 取最小值. 称这个方程组为一元线性回归的正规方程组，b_0 为回归常数或截距，b_1 为回归系数.

注：前面曾假设 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立且都服从正态分布 $N(0,\sigma^2)$. 在建立回归方程的过程中，这两个假设都没有用到. 在对回归方程作显著性检验或进行区间预测时，将根据这两个假设导出检验统计量的分布.

1.1.2 一元线性回归参数的确定

由 n 组观测值 (x_i,y_i) 确定参数 β_0 及 β_1 的估计值 b_0 及 b_1 是一元线性回归的关键. 根据一元线性回归的正规方程组可以导出

$$b_1=\frac{l_{xy}}{l_{xx}},\quad b_0=\bar{y}-b_1\bar{x},$$

式中，$\bar{x}=\frac{1}{n}\sum_i x_i$，$\bar{y}=\frac{1}{n}\sum_i y_i$，

$$l_{xy}=\sum_i(x_i-\bar{x})(y_i-\bar{y})=\sum_i x_iy_i-\frac{1}{n}\sum_i x_i\sum_i y_i,$$

$$l_{xx}=\sum_i(x_i-\bar{x})^2=\sum_i x_i^2-\frac{1}{n}\Big(\sum_i x_i\Big)^2.$$

称 l_{xy} 为 x 与 y 的离均差乘积和，l_{xx} 为 x 的离均差平方和.

记 $l_{yy}=\sum_i(y_i-\bar{y})^2=\sum_i y_i^2-\frac{1}{n}\Big(\sum_i y_i\Big)^2$，则 l_{yy} 为 y 的离均差平方和.

进一步，由正规方程组的第一个方程可以导出

$$\sum_i(b_0+b_1x_i)=\sum_i y_i \quad 及 \quad b_0+b_1\bar{x}=\bar{y}.$$

因此有结论：① $\sum_i\hat{y}_i=\sum_i y_i$，$\frac{1}{n}\sum_i\hat{y}_i=\bar{y}$，② 当 $x=\bar{x}$ 时，$\hat{y}=\bar{y}$.

这说明，将 x 的 n 个观测值 x_i 代入回归方程所得到的 n 个估计值 $\hat{y}_i$ 的平均值等于 $\bar{y}$，将 $\bar{x}$ 代入回归方程所得到的估计值 $\hat{y}$ 也等于 $\bar{y}$.

1.1.3 一元线性回归的矩阵表示

作一元线性回归时，自变量 x 及因变量 y 的观测值 x_i 及 y_i 所满足的关系式

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i,\quad i=1,2,\cdots,n$$

又称为一元线性回归模型.

若记

$$\boldsymbol{y}=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix}1&x_1\\1&x_2\\\vdots&\vdots\\1&x_n\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\\beta_1\end{pmatrix},\quad \boldsymbol{\varepsilon}=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix},$$

则上述模型的矩阵表示为 $\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，且

$$E(\boldsymbol{\varepsilon})=\begin{pmatrix}E(\varepsilon_1)\\E(\varepsilon_2)\\\vdots\\E(\varepsilon_n)\end{pmatrix}=\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}=\boldsymbol{0},\quad \mathrm{Cov}(\boldsymbol{\varepsilon})=\begin{pmatrix}\sigma^2&0&\cdots&0\\0&\sigma^2&\cdots&0\\\vdots&\vdots&&\vdots\\0&0&\cdots&\sigma^2\end{pmatrix}=\sigma^2\boldsymbol{I}.$$

正规方程组的矩阵表示为

$$\begin{pmatrix}n&\sum\limits_i x_i\\\sum\limits_i x_i&\sum\limits_i x_i^2\end{pmatrix}\begin{pmatrix}b_0\\b_1\end{pmatrix}=\begin{pmatrix}\sum\limits_i y_i\\\sum\limits_i x_i y_i\end{pmatrix},$$

其中

$$\begin{pmatrix}n&\sum\limits_i x_i\\\sum\limits_i x_i&\sum\limits_i x_i^2\end{pmatrix}=\boldsymbol{X'X},\quad \begin{pmatrix}\sum\limits_i y_i\\\sum\limits_i x_i y_i\end{pmatrix}=\boldsymbol{X'y}.$$

若记 $\boldsymbol{b}=\begin{pmatrix}b_0\\b_1\end{pmatrix}$，则正规方程组可进一步用矩阵表示为 $\boldsymbol{X'Xb}=\boldsymbol{X'y}$，当 $\boldsymbol{X'X}$ 的逆矩阵存在时，正规方程组的解 $\boldsymbol{b}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$，式中，

$$(\boldsymbol{X'X})^{-1}=\begin{pmatrix}\dfrac{\sum\limits_i x_i^2}{nl_{xx}}&\dfrac{-\sum\limits_i x_i}{nl_{xx}}\\\dfrac{-\sum\limits_i x_i}{nl_{xx}}&\dfrac{1}{l_{xx}}\end{pmatrix}.$$

在统计分析软件 SAS 的输出中，将正规方程组的增广矩阵

$$(\boldsymbol{X'X},\boldsymbol{X'y})\quad 或\quad \begin{pmatrix}n&\sum\limits_i x_i&\sum\limits_i y_i\\\sum\limits_i x_i&\sum\limits_i x_i^2&\sum\limits_i x_i y_i\end{pmatrix}$$

表示为下列形式的加边增广矩阵

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{pmatrix} \quad 或 \quad \begin{pmatrix} n & \sum_i x_i & \sum_i y_i \\ \sum_i x_i & \sum_i x_i^2 & \sum_i x_i y_i \\ \sum_i y_i & \sum_i x_i y_i & \sum_i y_i^2 \end{pmatrix},$$

将$(\boldsymbol{X}'\boldsymbol{X})^{-1}$，$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 及 SSE 表示为矩阵

$$\begin{pmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{b} \\ \boldsymbol{b}' & \text{SSE} \end{pmatrix}.$$

1.1.4 回归方程的显著性检验

离均差平方和 $l_{yy}=\sum_i (y_i-\bar{y})^2$ 表示 n 个观测值 y_i 之间的差异. 当各个 y_i 已知时，l_{yy} 是一个定值，作回归方程的显著性检验时，称它为总平方和，也记作 SST 或 SS_{tot}.

以下证明：$\text{SST}=\sum_i (y_i-\hat{y}_i)^2+\sum_i (\hat{y}_i-\bar{y})^2$. 因为

$$\begin{aligned} \sum_i (y_i-\bar{y})^2 &= \sum_i (y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2 \\ &= \sum_i (y_i-\hat{y}_i)^2+\sum_i (\hat{y}_i-\bar{y})^2 \\ &\quad +2\sum_i (y_i-\hat{y}_i)(\hat{y}_i-\bar{y})^2, \end{aligned}$$

最后面的一项可写为

$$\begin{aligned} &2\sum_i (y_i-b_0-b_1x_i)(b_0+b_1x_i-\bar{y}) \\ &=2\sum_i (y_i-\bar{y}+b_1\bar{x}-b_1x_i)(\bar{y}-b_1\bar{x}+b_1x_i-\bar{y}) \\ &=2b_1\sum_i (y_i-\bar{y})(x_i-\bar{x})-2b_1^2\sum_i (x_i-\bar{x})^2 \\ &=2b_1(l_{xy}-b_1l_{xx})=0, \end{aligned}$$

因此，

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

式中，$\sum_i (y_i-\hat{y}_i)^2$ 是 y_i 与 $\hat{y}_i$ 之间的偏差平方和，通过回归已经达到了最小值，称 $\sum_i (y_i-\hat{y}_i)^2$ 为剩余平方和，记作 SSE 或 SS_{res}.

而 $\sum_i(\hat{y}_i-\bar{y})^2$ 表示 n 个 $\hat{y}_i$ 之间的差异，是将 x_i 代入回归方程得到 $\hat{y}_i$ 造成的，称 $\sum_i(\hat{y}_i-\bar{y})^2$ 为回归平方和，记作 SSR 或 SS_{reg}.

由等式 SST = SSE + SSR 可以对 SSR 的意义作下列分析：

如果 SSR 的数值较大，SSE 的数值便比较小，说明回归的效果好. 如果 SSR 的数值较小，SSE 的数值便比较大，说明回归的效果差.

根据对 $y_i=\beta_0+\beta_1x_i+\varepsilon_i$ 中的 ε_i 所作的两个假设可以证明：

当原假设 H_0 为 $\beta_1=0$ 并且 H_0 成立时，

$$\frac{SST}{\sigma^2}\sim\chi^2(n-1),\quad \frac{SSR}{\sigma^2}\sim\chi^2(1),\quad \frac{SSE}{\sigma^2}\sim\chi^2(n-2),$$

且 SSR 与 SSE 相互独立，$F=\dfrac{SSR/1}{SSE/(n-2)}\sim F(1,n-2)$，$\hat{\sigma}^2=MSE=\dfrac{SSE}{n-2}$ 为 σ^2 的无偏估计量.

因此，给出显著性水平 α，将 F 与 $F_\alpha(1,n-2)$ 进行比较，当 $F>F_\alpha$ 时放弃 H_0，称回归方程显著；否则接受 H_0，称回归方程不显著.

注：对回归方程作显著性检验的基本思想与方法类似于方差分析，在 SAS 输出的结果中检验的过程与结果将用方差分析表来显示.

计算 SSR 及 SSE 的公式为

$$SSR=b_1l_{xy},\quad SSE=l_{yy}-SSR.$$

这里，

$$\begin{aligned}SSR&=\sum_i(\hat{y}_i-\bar{y})^2=\sum_i(b_0+b_1x_i-b_0-b_1\bar{x})^2\\&=b_1^2\sum_i(x_i-\bar{x})^2=b_1^2l_{xx}=b_1l_{xy}.\end{aligned}$$

1.1.5　相关系数与决定系数

由 SSR,SSE 及 b_1 的计算公式可推出

$$SSE=l_{yy}\left(1-b_1\frac{l_{xy}}{l_{yy}}\right)=l_{yy}\left(1-\frac{l_{xy}^2}{l_{xx}l_{yy}}\right).$$

若记 $r=\dfrac{l_{xy}}{\sqrt{l_{xx}l_{yy}}}$，则

$$SSE=l_{yy}(1-r^2),\quad SSR=r^2l_{yy},\quad l_{xy}=r\sqrt{l_{xx}l_{yy}}.$$

因此，当 $|r|$ 大时，SSE 小，SSR 大，变量 x 与 y 的线性关系密切；当 $|r|$ 小时，SSE 大，SSR 小，变量 x 与 y 的线性关系不密切.

当 $r>0$ 时，$b_1>0$，$\hat{y}$ 随 x 的增加而增加，x 与 y 的线性相关关系为正相关；当 $r<0$ 时，$b_1<0$，$\hat{y}$ 随 x 的增加而减少，x 与 y 的线性相关关系为负相关.

称 r 为变量 x 与 y 的相关系数.

至于 r^2 也有很重要的实际意义. 根据

$$r^2=\frac{l_{xy}^2}{l_{xx}l_{yy}}=\frac{b_1 l_{xy}}{l_{yy}}=\frac{\mathrm{SSR}}{\mathrm{SST}},$$

可以将 r^2 解释为 SSR 在 SST 中所占的比率，也就是 SST 中可以用线性关系来说明的部分在 SST 中所占的比率.

称 r^2 为变量 x 与 y 的决定系数.

对相关系数作显著性检验时，可以由

$$F=\frac{\mathrm{SSR}}{\mathrm{SSE}/(n-2)}=\frac{r^2}{(1-r^2)/(n-2)}$$

作 F 检验.

也可以先查相关系数检验专用的临界值，再将 $|r|$ 与临界值进行比较，然后作出 r 是否显著的结论.

$|r|$ 的临界值是将上述统计量变形为 $|r|=\sqrt{\dfrac{F}{F+(n-2)}}$ 后，将 F 检验的临界值 $F_\alpha(1,n-2)$ 代入计算的结果. 2.1 节的表 2-1 中列出了 $|r|$ 的部分临界值.

1.1.6 一元线性回归方程的应用

1. 点预测

由 n 组观测值建立一元线性回归方程 $\hat{y}=b_0+b_1x$ 后，给定 $x=x_0$，即可由回归方程求出 $y_0=\beta_0+\beta_1x_0+\varepsilon_0$ 的点估计值 $\hat{y}_0=b_0+b_1x_0$. 在应用学科中，称 $\hat{y}_0$ 为 y_0 的点预测值.

理论上已经证明：

当 $x=x_0$，$y=y_0$，$\hat{y}_0=b_0+b_1x_0$ 时，

$$E(b_0)=\beta_0,\quad E(b_1)=\beta_1,\quad E(\hat{y}_0)=E(y_0),$$

且统计量

$$y_0-\hat{y}_0\sim N\left(0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]\right).$$

因此，当 n 比较大，x_0 与 $\bar{x}$ 比较接近时，$y_0-\hat{y}_0$ 的方差比较小，点预测的效果比较好.

2. 区间预测

统计量

$$t=\frac{y_0-\hat{y}_0}{\sqrt{\mathrm{MSE}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]}}\sim t(n-2),$$

由置信水平 $1-\alpha$ 确定 $P\{|t|<t_\alpha(n-2)\}=1-\alpha$ 中的临界值 $t_\alpha(n-2)$ 后，若记

$$\delta=t_\alpha(n-2)\sqrt{\mathrm{MSE}\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{l_{xx}}\right]},$$

则 $P\{|y_0-\hat{y}_0|<\delta\}=1-\alpha$，$(\hat{y}_0-\delta,\hat{y}_0+\delta)$ 便是 $x=x_0$ 时 y_0 的预测区间，而 δ 为预测区间的半径.

当 n 及 $t_\alpha(n-2)$ 一定，$\bar{x}$ 及 SSE 也一定时，预测区间的大小由 $|x_0-\bar{x}|$ 决定，要得到比较精确的预测必须 x_0 与 $\bar{x}$ 比较接近，最好不要超出建立回归方程时 x 的取值范围.

当 $n\to+\infty$ 时，$l_{xx}\to+\infty$，$\delta\approx t_\alpha(n-2)\sqrt{\mathrm{MSE}}$.

因此，若用回归方程进行预测，则当 n 比较小时，只能内插，不能外推；当 n 比较大时，既能内插，又能外推.

注：如果对两个随机变量 X 与 Y 作一元线性回归分析，只需满足下列条件，上述检验与预测方法仍然适用：① 在给定 X_i 时，Y_i 的条件分布是正态分布，并且相互独立，其条件均值为 $\beta_0+\beta_1X_i$，条件方差为 σ^2；② X_i 是独立随机变量，其概率分布不涉及参数 β_0,β_1 与 σ^2.

1.1.7　一元线性回归的实例

【例 1.1】　棉花红铃虫第一代产卵高峰日百株卵量 x（粒）与百株累计卵量 y（粒）的 8 组观测数据如表 1-1（承蒙邝幸泉提供），试建立一元线性回归方程并作回归方程的显著性检验. 如果令 $x_0=20$，试求点预测值及置信水平 $\alpha=95\%$ 的置信区间.

表 1-1　　棉花红铃虫第一代卵量的观测数据

i	1	2	3	4	5	6	7	8
x_i	14.3	14.0	69.3	22.7	7.3	8.0	1.3	7.9
y_i	46.3	30.7	144.6	69.2	16.0	12.3	2.7	26.3

解 (1) 在表 1-2 中计算或用计算器的双变数统计运算程序算出 x 与 y 的平方和及乘积和

$$\sum_i x_i^2 = 5\,899.66, \quad \sum_i x_i = 144.8, \quad \bar{x} = 18.1,$$

$$\sum_i y_i^2 = 29\,890.25, \quad \sum_i y_i = 348.1, \quad \bar{y} = 43.512\,5,$$

$$\sum_i x_i y_i = 13\,109.99.$$

表 1-2 **计算 x 与 y 的平方和及乘积和**

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	14.3	46.3	204.49	2 143.69	662.09
2	14.0	30.7	196.00	942.49	429.80
3	69.3	144.6	4 802.49	20 909.16	10 020.78
4	22.7	69.2	515.29	4 788.64	1 570.84
5	7.3	16.0	53.29	256.00	116.80
6	8.0	12.3	64.00	151.29	98.40
7	1.3	2.7	1.69	7.29	3.51
8	7.9	26.3	62.41	691.69	207.77
$\sum_i$	144.8	348.1	5 899.66	29 890.25	13 109.99

(2) 计算 x 与 y 的离均差平方和及离均差乘积和

$$l_{xx} = \sum_i x_i^2 - \frac{1}{n}\Big(\sum_i x_i\Big)^2 = 3\,278.78,$$

$$l_{xy} = \sum_i x_i y_i - \frac{1}{n}\sum_i x_i \sum_i y_i = 6\,809.38,$$

$$l_{yy} = \sum_i y_i^2 - \frac{1}{n}\Big(\sum_i y_i\Big)^2 = 14\,743.548\,75.$$

(3) 计算回归系数与回归常数

$$b_1 = \frac{l_{xy}}{l_{xx}} = 2.076\,802\,957,$$

$$b_0 = \bar{y} - b_1\bar{x} = 5.922\,366\,479,$$

所求的回归方程为

$$\hat{y} = 5.922\,4 + 2.076\,8\,x.$$

(4) 作回归方程的显著性检验

$$\mathrm{SST} = l_{yy} = 14\,743.548\,75,$$

$$\mathrm{SSR} = b_1 l_{xy} = 14\,141.740\,52,$$

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSR} = 601.808\,23,$$

$$F = 140.99, \quad F_{0.01}(1,6) = 13.7, \quad F > F_{0.01}(1,6),$$

故当显著性水平 $\alpha = 0.01$ 时回归方程是显著的.

此回归方程及其显著性表明：棉花红铃虫第一代产卵高峰日百株卵量 x 与百株累计卵量 y 之间的线性关系极其显著. 回归系数 $b_1 > 0$ 表明：当 x 增加时，$\hat{y}$ 也跟随着增加，且 $\hat{y}$ 增加的速度约为 x 增加速度的 2.08 倍.

如果令 $x_0 = 20$，则点预测值

$$\hat{y}_0 = b_0 + b_1 x_0 = 5.922\,4 + 2.076\,8 \times 20 = 47.458\,4,$$

$$\delta = t_\alpha(n-2)\sqrt{\mathrm{MSE}\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{l_{xx}}\right]}$$

$$= 2.446\,91 \times \sqrt{\frac{601.808\,23}{8-2}\left[1 + \frac{1}{8} + \frac{(20-18.1)^2}{3\,278.78}\right]}$$

$$= 26.005\,2,$$

置信水平 $\alpha = 95\%$ 的置信区间为(21.453 2,73.463 6).

1.1.8　应用 SAS 作一元线性回归

Statistical Analysis System 简称 SAS，可用来分析数据和编写报告. 它是美国 SAS 研究所的产品，在国际上被誉为标准软件. 它的显示管理系统有三个主要的窗口：

(1) 编辑窗口(PROGRAM EDITOR) 是编辑程序和数据文件的地方.

(2) 日志窗口(LOG) 是记录程序的运行情况并显示 ERROR 信息的地方.

(3) 输出窗口(OUTPUT) 是程序运行结果暂时存放的地方.

进入 SAS 的显示管理系统后，便出现供挑选的菜单，用鼠标点击其中的 Window，再点击所要进入的窗口名，即可进入选定的窗口.

SAS 程序通常可划分为数据步与过程步：

(1) 输入待分析的数据，建立 SAS 数据文件，称为数据步.

(2) 调用 SAS 内部的批处理程序分析 SAS 数据文件中的数据，称为过程步.

可以根据需要编写多个数据步或过程步，但是每一个数据步都要以 DATA 语句开始，每一个过程步都要以 PROC 语句开始，程序的最后要以 RUN 语句结束.

提交程序可点击以“运行”为标志的按钮.

应用SAS作例1.1中一元线性回归的程序为

```
data ex;input x y @@;
cards;
14.3 46.3 14 30.7 69.3 144.6 22.7 69.2
7.3 16 8 12.3 1.3 2.7 7.9 26.3 20 .
(最后的一组数据20.表示令x0=20,准备要计算点预测值)
;
proc gplot;
plot y*x;(以x为横坐标、y为纵坐标画散点图及回归方程所对应的回归直线)
symbol i=rl v=dot;
proc reg;model y=x;run;
```

在SAS的OUTPUT窗口,将输出以下几项主要的结果:

① 以x为横坐标、y为纵坐标的散点图及回归方程所对应的回归直线,如图1-1所示.

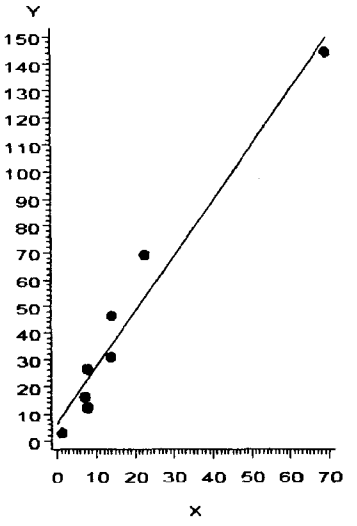


图1-1 (x_i, y_i)的散点图及回归直线

② 回归方程作显著性检验的方差分析表.

```
Model: MODEL1
Dependent Variable: Y
Analysis of Variance
```