硕士研究生公共课教材

多元统计与SAS应用

（第二版）

肖枝洪　余家林　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

硕士研究生公共课教材

多元统计与SAS应用

（第二版）

肖枝洪　余家林　编著

WUHAN UNIVERSITY PRESS
武汉大学出版社

图书在版编目(CIP)数据

多元统计与SAS应用/肖枝洪,余家林编著. —2版. —武汉：武汉大学出版社,2013.9
硕士研究生公共课教材
ISBN 978-7-307-11781-5

Ⅰ.多… Ⅱ.①肖… ②余… Ⅲ.①多元分析:统计分析—研究生—教材 ②统计分析—统计程序—研究生—教材 Ⅳ.①O212.4 ②C819

中国版本图书馆CIP数据核字(2013)第222278号

责任编辑:顾素萍 责任校对:黄添生 版式设计:马 佳

出版发行:**武汉大学出版社** (430072 武昌 珞珈山)
(电子邮件:cbs22@whu.edu.cn 网址:www.wdp.com.cn)
印刷:武汉中科兴业印务有限公司
开本:720×1000 1/16 印张:14.75 字数:260千字 插页:1
版次:2008年1月第1版 2013年9月第2版
2013年9月第2版第1次印刷
ISBN 978-7-307-11781-5 定价:28.00元

第二版前言

《多元统计与SAS应用》第一版自2008年发行以来，得到了许多读者的使用和赞扬，而且现已脱销。应广大读者的要求，我们在第一版的基础上进行修订后再出第二版。

第二版与第一版相比，基本框架不变，只是在部分章节上做了微调。本次修订了我们在教学中发现以及读者反映出来的问题。

感谢为本次修订提出宝贵意见的教师和读者。感谢武汉大学出版社的大力支持！

由于作者水平有限，书中的错误和疏漏之处可能还是存在，敬请读者提出宝贵意见，以便进一步修订和改进！

肖枝洪

于重庆理工大学花溪校区

2013年7月

第一版前言

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，急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）所得到的统计分析结果，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

第五章

第六章

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

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

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

$$\begin{cases} nb_0 + b_1 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i, \\ b_0 \sum\limits_{i=1}^{n} x_i + b_1 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

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

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

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

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 $l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{1}{n}\Big(\sum_{i=1}^{n}y_i\Big)^2$，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

$$\sum_{i=1}^{n}(b_0+b_1x_i)=\sum_{i=1}^{n}y_i\quad 及\quad b_0+b_1\bar{x}=\bar{y}.$$

因此有结论：

① $\sum_{i=1}^{n}\hat{y}_i=\sum_{i=1}^{n}y_i$，$\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\bar{y}$；

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

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

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

$$\begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix},$$

其中

$$\begin{pmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{pmatrix} = \boldsymbol{X'X}, \quad \begin{pmatrix} \sum\limits_{i=1}^{n} y_i \\ \sum\limits_{i=1}^{n} x_i y_i \end{pmatrix} = \boldsymbol{X'y}.$$

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，

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

在统计分析软件 SAS 的输出中，将正规方程组的增广矩阵

$$(\boldsymbol{X}'\boldsymbol{X},\boldsymbol{X}'\boldsymbol{y}) \quad 或 \quad \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

表示为下列形式的加边增广矩阵：

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{pmatrix} \quad 或 \quad \begin{pmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^2 \end{pmatrix},$$

将 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$，$\boldsymbol{b}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ 及 SSE（即剩余平方和，其定义见 1.1.4 节）表示为矩阵

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

1.1.4 回归方程的显著性检验

离均差平方和 $l_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2$ 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**，也记为 SST 或 $\mathrm{SS}_{\mathrm{tot}}$.

以下证明：$\mathrm{SST}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$. 因为

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

最后面的一项可写为

$$
\begin{aligned}
&2\sum_{i=1}^{n}(y_i-b_0-b_1x_i)(b_0+b_1x_i-\overline{y})\\
&\quad=2\sum_{i=1}^{n}(y_i-\overline{y}+b_1\overline{x}-b_1x_i)(\overline{y}-b_1\overline{x}+b_1x_i-\overline{y})\\
&\quad=2b_1\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})-2b_1^2\sum_{i=1}^{n}(x_i-\overline{x})^2\\
&\quad=2b_1(l_{xy}-b_1l_{xx})\\
&\quad=0,
\end{aligned}
$$

因此，

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

式中，$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 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 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 为**剩余平方和**，记为 SSE 或 SS_{res}.

而 $\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ 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 $\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$ 为**回归平方和**，记为 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

$$\mathrm{SSR}=b_1 l_{xy},\quad \mathrm{SSE}=l_{yy}-\mathrm{SSR}.$$

这里，

$$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(b_0+b_1x_i-b_0-b_1\bar{x})^2$$

$$=b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2=b_1^2 l_{xx}=b_1 l_{xy}.$$

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

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

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

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

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