第三章 多元线性回归

3.1 多元线性回归模型

一、多元线性回归模型的一般形式

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p+\varepsilon$$

$$\begin{cases} E(\varepsilon)=0 \\ \mathrm{var}(\varepsilon)=\sigma^2 \end{cases}$$

3.1 多元线性回归模型

一、多元线性回归模型的一般形式

对n组观测数据 $(x_{i1}, x_{i2},\ldots,x_{ip}; y_i)$, $i=1,2,\ldots,n$,
线性回归模型表示为:

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 \\ \quad \cdots\cdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \varepsilon_n \end{cases}$$

3.1 多元线性回归模型

一、多元线性回归模型的一般形式

写成矩阵形式为：$y=X\beta+\varepsilon$，　其中，

$$\mathbf{y}=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{X}=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{n\times(p+1)}$$

$$\boldsymbol{\beta}=\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad \boldsymbol{\varepsilon}=\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

3.1 多元线性回归模型

二、多元线性回归模型的基本假定

1. 解释变量$x_1,x_2,\ldots,x_p$是确定性变量,不是随机变量,且要求$\mathrm{rk}(\boldsymbol{X})=p+1<n$。

表明设计矩阵$\boldsymbol{X}$中的自变量列之间不相关,

$\mathbf{X}$是一满秩矩阵。

3.1 多元线性回归模型

二、多元线性回归模型的基本假定

2 .随机误差项具有0均值和等方差,即

$$\begin{cases} E(\varepsilon_i)=0, \quad i=1,2,\cdots,\ n \\ cov(\varepsilon_i,\varepsilon_j)=\begin{cases}\sigma^2\ , & i=j \\ 0\ , & i\neq j\end{cases} \end{cases} \qquad (i,j=1,2,\cdots,n)$$

这个假定称为Gauss-Markov条件

3.1 多元线性回归模型

二、多元线性回归模型的基本假定

3. 正态分布的假定条件为:

$$\begin{cases} \varepsilon_i \sim N(0,\sigma^2), i=1,2,\cdots,n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{ 相互独立} \end{cases}$$

用矩阵形式(3.5)式表示为:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

3.1 多元线性回归模型

二、多元线性回归模型的基本假定

在正态假定下:

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$$

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}$$

$$var(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$$

3.1 多元线性回归模型

三、多元线性回归方程的解释

y表示空调机的销售量,

x_1表示空调机的价格,

x_2表示消费者可用于支配的收入。

$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$

$E(y)=\beta_0+\beta_1 x_1+\beta_2 x_2$

在x_2保持不变时, 有 $\dfrac{\partial E(y)}{\partial x_1}=\beta_1$

在x_1保持不变时, 有 $\dfrac{\partial E(y)}{\partial x_2}=\beta_2$

3.1 多元线性回归模型

三、多元线性回归方程的解释

考虑国内生产总值GDP和三次产业增加值的关系，

$$\text{GDP}=x_1+x_2+x_3$$

现在做GDP对第二产业增加值x_2的一元线性回归，得回归方程

$$\hat{y}=5\,289.9+1.855\,4x_2$$

3.1 多元线性回归模型

年份	GDP	第一产业增加值x_1	第二产业增加值x_2	*第三产业增加值*x_3
1990	18 547.9	5 017.0	7 717.4	5 813.5
1991	21 617.8	5 288.6	9 102.2	7 227.0
1992	26 638.1	5 800.0	11 699.5	9 138.6
1993	34 634.4	6 882.1	16 428.5	11 323.8
1994	46 759.4	9 457.2	22 372.2	14 930.0
1995	58 478.1	11 993.0	28 537.9	17 947.2
1996	67 884.6	13 844.2	33 612.9	20 427.5
1997	74 462.6	14 211.2	37 222.7	23 028.7
1998	78 345.2	14 552.4	38 619.3	25 173.5
1999	82 067.5	14 472.0	40 557.8	27 037.7
2000	89 468.1	14 628.2	44 935.3	29 904.6
2001	97 314.8	15 411.8	48 750.0	33 153.0
2002	105 172.3	16 117.3	52 980.2	36 074.8
2003	117 390.2	16 928.1	61 274.1	39 188.0
2004	136 875.9	20 768.1	72 387.2	43 720.6

3.1 多元线性回归模型

三、多元线性回归方程的解释

建立GDP对x_1和x_2的回归，得二元回归方程

$$\hat{y} = 2\,914.6 + 0.607\,x_1 + 1.709\,x_2$$

你能够合理地解释两个回归系数吗？

3.2 回归参数的估计

一、回归参数的普通最小二乘估计

最小二乘估计要寻找 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$,…，$\hat{\beta}_p$,使得

$$Q(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\cdots,\hat{\beta}_p)=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})^2$$

$$=\min_{\beta_0,\beta_1,\beta_2,\cdots,\beta_p}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$$

3.2 回归参数的估计

一、回归参数的普通最小二乘估计

$$
\begin{cases}
\left.\dfrac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip}) = 0 \\
\left.\dfrac{\partial Q}{\partial \beta_1}\right|_{\beta_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{i1} = 0 \\
\left.\dfrac{\partial Q}{\partial \beta_2}\right|_{\beta_2=\hat{\beta}_2} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{i2} = 0 \\
\cdots\cdots \\
\left.\dfrac{\partial Q}{\partial \beta_p}\right|_{\beta_p=\hat{\beta}_p} = -2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})x_{ip} = 0
\end{cases}
$$

3.2 回归参数的估计

一、回归参数的普通最小二乘估计

经整理后得用矩阵形式表示的正规方程组

$$X'(y - X\hat{\beta}) = 0$$

移项得 $$X'X\hat{\beta} = X'y$$

当$(X'X)^{-1}$ 存在时，即得回归参数的最小二乘估计为：

$$\hat{\beta} = (X'X)^{-1}X'y$$

3.2 回归参数的估计

二、回归值与残差

称 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$ 为回归值

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

称为帽子矩阵，其主对角线元素记为h_{ii}，则

3.2 回归参数的估计

二、回归值与残差

$$tr(H)=\sum_{i=1}^{n}h_{ii}=p+1$$

此式的证明只需根据迹的性质tr(AB)=tr(BA),因而

$$\mathrm{tr}(\mathrm{H})=\mathrm{tr}(\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')=\mathrm{tr}(\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1})$$

$$=\mathrm{tr}(\mathrm{I}_{p+1})=p+1$$

3.2 回归参数的估计

二、回归值与残差

$$e = y - \hat{y} = y - Hy = (I - H)y$$

$$\begin{aligned}\mathrm{cov}(e, e) &= \mathrm{cov}((I-H)Y, (I-H)Y) \\ &= (I-H)\,\mathrm{cov}(Y, Y)\,(I-H)' \\ &= \sigma^2 (I-H) I_n (I-H)' = \sigma^2 (I-H)\end{aligned}$$

得 $D(e_i)=(1-h_{ii})\sigma^2$， $i=1,2,\ldots,n$

3.2 回归参数的估计

二、回归值与残差

$$\text{得 } E(\sum_{i=1}^{n} e_i^2) = \sum_{i=1}^{n} D(e_i) = (n-p-1)\sigma^2$$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} SSE = \frac{1}{n-p-1} (\mathrm{e}'\mathrm{e}) = \frac{1}{n-p-1} \sum_{i=1}^{n} e_i^2$$

是σ^2的无偏估计

3.2 回归参数的估计

三、回归参数的最大似然估计

$$y \sim N(X\beta, \sigma^2 I_n)$$

似然函数为

$$L = (2\pi)^{-n/2}\left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right)$$

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

等价于使$(y-X\beta)'(y-X\beta)$达到最小,这又完全与OLSE一样

3.2 回归参数的估计

例3.1 国际旅游外汇收入是国民经济发展的重要组成部分，影响一个国家或地区旅游收入的因素包括自然、文化、社会、经济、交通等多方面的因素，本例研究第三产业对旅游外汇收入的影响。《中国统计年鉴》把第三产业划分为12个组成部分，分别为x_1农林牧渔服务业，x_2地质勘查水利管理业，x_3交通运输仓储和邮电通信业，x_4批发零售贸易和餐饮业，x_5金融保险业，x_6房地产业，x_7社会服务业，x_8卫生体育和社会福利业，x_9教育文化艺术和广播，x_{10}科学研究和综合艺术，x_{11}党政机关，x_{12}其他行业。采用1998年我国31 个省、市、自治区的数据，以国际旅游外汇收入（百万美元）为因变量y，以如上12 个行业为自变量做多元线性回归，数据见表3.1，其中自变量单位为亿元人民币。

3.2 回归参数的估计

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-205.388	117.019		-1.755	.096
	x1	-1.438	22.913	-.012	-.063	.951
	x2	2.622	18.599	.023	.141	.889
	x3	3.297	2.468	.749	1.336	.198
	x4	-.946	1.298	-.312	-.729	.476
	x5	-5.521	4.514	-.963	-1.223	.237
	x6	4.068	3.960	.760	1.027	.318
	x7	4.162	5.079	.446	.819	.423
	x8	-15.404	10.835	-.520	-1.422	.172
	x9	17.338	8.374	1.038	2.071	.053
	x10	9.155	10.168	.221	.900	.380
	x11	-10.536	5.622	-.780	-1.874	.077
	x12	1.370	5.006	.042	.274	.787

a. Dependent Variable: y

3.3 参数估计量的性质

性质1 $\hat{\boldsymbol{\beta}}$ 是随机向量y的一个线性变换。

$$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}$$

性质2 $\hat{\boldsymbol{\beta}}$ **是β的无偏估计。**

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}) &= E((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}) \\
&= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}(\mathrm{y}) \\
&= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}(\mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\boldsymbol{\beta} = \boldsymbol{\beta}
\end{aligned}$$

3.3 参数估计量的性质

性质 3　$D(\hat{\boldsymbol{\beta}})=\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

$$D(\hat{\boldsymbol{\beta}})=\operatorname{cov}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}})$$

$$=\mathrm{E}\left((\hat{\boldsymbol{\beta}}-\mathrm{E}(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}}-\mathrm{E}(\hat{\boldsymbol{\beta}}))'\right)=\mathrm{E}\left((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\right)$$

$$=\mathrm{E}\left(\left((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}-\boldsymbol{\beta}\right)\left((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}-\boldsymbol{\beta}\right)'\right)$$

$$=E\left(\left((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta}\right)\left((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta}\right)'\right)$$

$$=E\left((\boldsymbol{\beta}+(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta})(\boldsymbol{\beta}+(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\varepsilon}-\boldsymbol{\beta})'\right)$$

$$=E\left((\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\boldsymbol{\varepsilon}\ \boldsymbol{\varepsilon}'\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\right)=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}(\boldsymbol{\varepsilon}\ \boldsymbol{\varepsilon}')\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}$$

$$=(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{E}(\sigma^2\mathrm{I}_n)\mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}=\sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

3.3 参数估计量的性质

当p=1时

$$\mathrm{X'X}=\begin{pmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{pmatrix}$$

$$(\mathrm{X'X})^{-1}\sigma^2=\frac{\sigma^2}{|\mathrm{X'X}|}\begin{pmatrix} \sum_{i=1}^{n}x_i^2 & -\sum_{i=1}^{n}x_i \\ -\sum_{i=1}^{n}x_i & \mathrm{n} \end{pmatrix}=\begin{pmatrix} \frac{\sigma^2}{\mathrm{nL_{xx}}}\sum_{i=1}^{n}x_i^2 & -\frac{\bar{x}}{L_{xx}}\sigma^2 \\ -\frac{\bar{x}}{L_{xx}}\sigma^2 & \frac{\sigma^2}{\mathrm{L_{xx}}} \end{pmatrix}$$

3.3 参数估计量的性质

性质4 Gauss-Markov定理

预测函数 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{10} + \hat{\beta}_2 x_{20} + \cdots + \hat{\beta}_p x_{p0}$

是 $\hat{\boldsymbol{\beta}}$ 的线性函数

Gauss-Markov定理

在假定$\mathbf{E(y)=X\beta}$, $\mathbf{D(y)=\sigma^2 I_n}$时,$\boldsymbol{\beta}$的任一线性函数 $C\boldsymbol{\beta}$的最小方差线性无偏估计(Best Lnear Unbiased Estimator简记为BLUE)为$\mathbf{c}'\hat{\boldsymbol{\beta}}$,其中$\mathbf{c}$是任一p+1维向量, $\hat{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的最小二乘估计。

3.3 参数估计量的性质

第一，取常数向量c的第j（j=0,1,…,n）个分量为1，其余分量为0，这时G-M定理表明最小二乘估计是β_j的最小方差线性无偏估计。

第二，可能存在$y_1, y_2, \ldots, y_n$的非线性函数，作为$C'\beta$的无偏估计，比最小二乘估计$C'\hat{\beta}$的方差更小。

第三，可能存在$C'\beta$的有偏估计量，在某种意义（例如均方误差最小）下比最小二乘估计$C'\hat{\beta}$更好。

第四，在正态假定下，$C'\hat{\beta}$是$C'\beta$的最小方差无偏估计。也就是说，既不可能存在$y_1, y_2, \ldots, y_n$的非线性函数，也不可能存在$y_1, y_2, \ldots, y_n$的其它线性函数，作为$C'\hat{\beta}$的无偏估计，比最小二乘估计$C'\beta$方差更小。

3.3 参数估计量的性质

性质5 cov（ $\hat{\boldsymbol{\beta}}$ ， e）=0

此性质说明 $\hat{\boldsymbol{\beta}}$ 与e不相关, 在正态假定下等价于与e独立, 从而与 $SSE = \mathbf{e}'\mathbf{e}$ 独立。

性质6 在正态假设 $\mathbf{y} \sim N(\mathbf{X\beta}, \mathbf{I_n}\sigma^2)$ 时

(1) $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X}^{-1})\sigma^2)$ 时

(2) $SSE/\sigma^2 \sim \chi^2(n-p-1)$

3.4 回归方程的显著性检验

一、F检验

$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$

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

$$SST = SSR + SSE$$

$$F = \frac{SSR / p}{SSE /(n - p - 1)}$$ 当H_0成立时服从 $F(p, n - p - 1)$

3.4 回归方程的显著性检验

一、F检验

方差来源	自由度	平方和	均方	F值	P值
回归 残差 总和	p n-p-1 n-1	SSR SSE SST	SSR/p SSE/(n-p-1)	$\frac{SSR/p}{SSE/(n-p-1)}$	P(F>F值) =P值

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11685742	12	973811.87	10.482	.000[a]
	Residual	1672296.2	18	92905.347		
	Total	13358039	30			

a. Predictors: (Constant), X12, X10, X1, X2, X4, X6, X11, X3, X8, X9, X7, X5

b. Dependent Variable: Y

3.4 回归方程的显著性检验

二、回归系数的显著性检验

$H_{0j}:\beta_j=0, \quad j=1,2,\ldots,p$

$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$

记 $(\mathbf{X}'\mathbf{X})^{-1}=(c_{ij})\ i,j=0,1,2,\ldots,p$

构造t统计量 $t_j = \dfrac{\hat{\beta}_j}{\sqrt{c_{jj}}\,\hat{\sigma}}$

其中 $\hat{\sigma} = \sqrt{\dfrac{1}{n-p-1}\sum_{i=1}^{n} e_i^2} = \sqrt{\dfrac{1}{n-p-1}\sum_{i=1}^{n} (y_i-\hat{y}_i)^2}$

3.4 回归方程的显著性检验

二、回归系数的显著性检验（剔除x_1）

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-204.406	112.889		-1.811	.086
X2	2.406	17.793	.021	.135	.894
X3	3.386	1.968	.769	1.720	.102
X4	-.955	1.255	-.316	-.761	.456
X5	-5.568	4.333	-.971	-1.285	.214
X6	4.096	3.829	.765	1.070	.298
X7	4.012	4.370	.430	.918	.370
X8	-15.120	9.584	-.510	-1.578	.131
X9	17.175	7.747	1.028	2.217	.039
X10	9.488	8.442	.229	1.124	.275
X11	-10.692	4.911	-.792	-2.177	.042
X12	1.352	4.865	.041	.278	.784

3.4 回归方程的显著性检验

二、回归系数的显著性检验

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-201.681	102.070		-1.976	.059
X3	3.618	.813	.822	4.449	.000
X8	-21.615	7.345	-.729	-2.943	.007
X9	27.854	4.232	1.667	6.582	.000
X11	-17.253	2.779	-1.278	-6.209	.000

3.4 回归方程的显著性检验

二、回归系数的显著性检验

从另外一个角度考虑自变量x_j的显著性。

y对自变量$x_1,x_2,\ldots,x_p$线性回归的残差平方和为SSE，回归平方和为SSR，在剔除掉x_j后，用y对其余的p-1个自变量做回归，记所得的残差平方和为$SSE_{(j)}$，回归平方和为$SSR_{(j)}$，则

自变量x_j对回归的贡献为$\Delta SSR_{(j)}=SSR-SSR_{(j)}$，称为$x_j$的偏回归平方和。由此构造偏F统计量

3.4 回归方程的显著性检验

二、回归系数的显著性检验

$$F_j = \frac{\Delta SSR_{(j)}/1}{SSE/(n-p-1)}$$

当原假设H_{0j}：$\beta_j=0$成立时，（3.42）式的偏F统计量F_j服从自由度为(1, n-p-1)的F分布，此F检验与（3.40）式的t检验是一致的，可以证明$F_j=t_j^2$

3.4 回归方程的显著性检验

三、回归系数的置信区间

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\,\hat{\sigma}} \sim t(n-p-1)$$

可得β_j的置信度为1-α的置信区间为：

$$(\hat{\beta}_j - t_{\alpha/2}\sqrt{c_{jj}}\,\hat{\sigma},\quad \hat{\beta}_j + t_{\alpha/2}\sqrt{c_{jj}}\,\hat{\sigma})$$

3.4 回归方程的显著性检验

四、拟合优度

决定系数为： $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

y关于$x_1, x_2, \ldots, x_p$的样本复相关系数

$$R = \sqrt{R^2} = \sqrt{\frac{SSR}{SST}}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.935[a]	.875	.791	304.80378

a. Predictors: (Constant), x12, x10, x1, x2, x4, x6 x8, x9, x7, x5

3.5 中心化和标准化

一、中心化

经验回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\cdots+\hat{\beta}_px_p$

经过样本中心 $(\bar{x}_1,\bar{x}_2,\cdots,\bar{x}_p;\bar{y})$

将坐标原点移至样本中心，即做坐标变换：

$$x'_{ij}=x_{ij}-\bar{x}_j,\qquad y'_i=y_i-\bar{y}$$

回归方程转变为：$\hat{y}'=\hat{\beta}_1x'_1+\hat{\beta}_2x'_2+\cdots+\hat{\beta}_px'_p$

回归常数项为 $\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2-\cdots-\hat{\beta}_p\bar{x}_p$

3.5 中心化和标准化

二、标准化回归系数

当自变量的单位不同时普通最小二乘估计的回归系数不具有可比性，例如有一回归方程为:

$$\hat{y} = 200 + 2000x_1 + 2x_2$$

其中x_1的单位是吨, x_2的单位是公斤

3.5 中心化和标准化

二、标准化回归系数

样本数据的标准化公式为：

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_j}{\sqrt{L_{jj}}},\qquad y_i^{*}=\frac{y_i-\bar{y}}{\sqrt{L_{yy}}}$$

得标准化的回归方程 $\hat{y}^{*}=\hat{\beta}_1^{*}x_1^{*}+\hat{\beta}_2^{*}x_2^{*}+\cdots+\hat{\beta}_p^{*}x_p^{*}$

$$\beta_j^{*}=\frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j,\ \mathrm{j}=1,\cdots,\mathrm{p}$$

3.5 中心化和标准化

二、标准化回归系数

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	-201.681	102.070		-1.976	.059
X3	3.618	.813	.822	4.449	.000
X8	-21.615	7.345	-.729	-2.943	.007
X9	27.854	4.232	1.667	6.582	.000
X11	-17.253	2.779	-1.278	-6.209	.000

标准化
回归系数

3.6 相关阵与偏相关系数

一、样本相关阵

自变量样本相关阵

$$\mathrm{r} = \left(\mathrm{X}^*\right)' \mathrm{X}^*$$

$$\mathrm{r} = \begin{pmatrix} 1 & \mathrm{r}_{12} & \cdots & \mathrm{r}_{1p} \\ r_{21} & 1 & \cdots & \mathrm{r}_{2p} \\ \cdots & & & \\ r_{p1} & \mathrm{r}_{p2} & \cdots & 1 \end{pmatrix}$$

增广的样本相关阵为：

$$\tilde{\mathrm{r}} = \begin{pmatrix} 1 & r_{y1} & \mathrm{r}_{y2} & \cdots & \mathrm{r}_{yp} \\ \mathrm{r}_{1y} & 1 & \mathrm{r}_{12} & \cdots & \mathrm{r}_{1p} \\ \mathrm{r}_{2y} & r_{21} & 1 & \cdots & \mathrm{r}_{2p} \\ \cdots & & & \cdots & \\ \mathrm{r}_{py} & r_{p1} & \mathrm{r}_{p2} & \cdots & 1 \end{pmatrix}$$

3.6 相关阵与偏相关系数

一、样本相关阵

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12
Y	1.000	0.260	0.342	0.580	0.479	0.518	0.530	0.741	0.379	0.575	0.673	0.257	0.038
X1	0.260	1.000	0.640	0.691	0.738	0.582	0.519	0.663	0.691	0.719	0.150	0.758	0.301
X2	0.342	0.640	1.000	0.773	0.658	0.502	0.464	0.602	0.660	0.686	0.118	0.760	0.337
X3	0.580	0.691	0.773	1.000	0.934	0.742	0.710	0.885	0.867	0.889	0.314	0.855	0.457
X4	0.479	0.738	0.658	0.934	1.000	0.780	0.743	0.887	0.926	0.892	0.348	0.849	0.437
X5	0.518	0.582	0.502	0.742	0.780	1.000	0.989	0.740	0.790	0.850	0.630	0.705	0.515
X6	0.530	0.519	0.464	0.710	0.743	0.989	1.000	0.703	0.753	0.821	0.646	0.666	0.493
X7	0.741	0.663	0.602	0.885	0.887	0.740	0.703	1.000	0.781	0.834	0.541	0.649	0.190
X8	0.379	0.691	0.660	0.867	0.926	0.790	0.753	0.781	1.000	0.931	0.404	0.906	0.548
X9	0.575	0.719	0.686	0.889	0.892	0.850	0.821	0.834	0.931	1.000	0.569	0.895	0.533
X10	0.673	0.150	0.118	0.314	0.348	0.630	0.646	0.541	0.404	0.569	1.000	0.241	0.155
X11	0.257	0.758	0.760	0.855	0.849	0.705	0.666	0.649	0.906	0.895	0.241	1.000	0.613
X12	0.038	0.301	0.337	0.457	0.437	0.515	0.493	0.190	0.548	0.533	0.155	0.613	1.000

3.6 相关阵与偏相关系数

二、偏判定系数

当其他变量被固定后,给定的任两个变量之间的相关系数,叫偏相关系数。

偏相关系数可以度量p+1个变量$y, x_1, x_2, \ x_p$之中任意两个变量的线性相关程度, 而这种相关程度是在固定其余p−1个变量的影响下的线性相关。

3.6 相关阵与偏相关系数

二、偏判定系数

偏判定系数测量在回归方程中已包含若干个自变量时，再引入某一个新的自变量后y的剩余变差的相对减少量，它衡量y的变差减少的边际贡献。

3.6 相关阵与偏相关系数

二、偏判定系数

以x_1表示某种商品的销售量，

x_2表示消费者人均可支配收入，

x_3表示商品价格。

从经验上看，销售量x_1与消费者人均可支配收入x_2之间应该有正相关，简单相关系数r_{12}应该是正的。但是如果你计算出的r_{12}是个负数也不要感到惊讶，这是因为还有其它没有被固定的变量在发挥影响，例如商品价格x_3在这期间大幅提高了。反映固定x_3后x_1与x_2相关程度的偏相关系数$r_{12;\,3}$会是个正数。

3.6 相关阵与偏相关系数

1．两个自变量的偏判定系数

二元线性回归模型为：$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i$

记SSE（x_2）是模型中只含有自变量x_2时y的残差平方和，SSE(x_1，x_2)是模型中同时含有自变量x_1和x_2时y的残差平方和。因此模型中已含有x_2时再加入x_1使y的剩余变差的相对减小量为：

$$r^2_{y1;2}=\frac{SSE(x_2)-SSE(x_1,x_2)}{SSE(x_2)}$$

此即模型中已含有x_2时，y与x_1的偏判定系数。

3.6 相关阵与偏相关系数

1．两个自变量的偏判定系数

同样地，模型中已含有x_1时，y与x_2的偏判定系数为：

$$r_{y2;1}^2 = \frac{SSE(x_1) - SSE(x_1, x_2)}{SSE(x_1)}$$

3.6 相关阵与偏相关系数

2.一般情况

在模型中已含有$x_2, ..., x_p$时，y与x_1的偏判定系数为：

$$r^2_{y1;2,\cdots,p} = \frac{SSE(x_2,\cdots,x_p) - SSE(x_1,x_2,\cdots,x_p)}{SSE(x_2,\cdots,x_p)}$$

3.6 相关阵与偏相关系数

三、偏相关系数

偏判定系数的平方根称为偏相关系数，

其符号与相应的回归系数的符号相同。

例3.2 研究北京市各经济开发区经济发展与招商投资的关系，因变量y为各开发区的销售收入（百万元），选取两个自变量，

x_1为截至1998年底各开发区累计招商数目，

x_2为招商企业注册资本（百万元）。

表中列出了至1998年底招商企业注册资本x_2在5亿至50亿元的15个开发区的数据。

3.6 相关阵与偏相关系数

三、偏相关系数

北京开发区数据

x1	x2	y	x1	x2	y
25	3547.79	553.96	7	671.13	122.24
20	896.34	208.55	532	2863.32	1400
6	750.32	3.1	75	1160	464
1001	2087.05	2815.4	40	862.75	7.5
525	1639.31	1052.12	187	672.99	224.18
825	3357.7	3427	122	901.76	538.94
120	808.47	442.82	74	3546.18	2442.79
28	520.27	70.12			

3.6 相关阵与偏相关系数

三、偏相关系数

偏相关系数表

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-327.04	218.001		-1.500	.159			
	x1	2.036	.438	.594	4.649	.001	.807	.802	.534
	x2	.468	.123	.485	3.799	.003	.746	.739	.436

a. Dependent Variable: y

3.6 相关阵与偏相关系数

三、偏相关系数

用y与x_1做一元线性回归时，x_1能消除y的变差SST的比例为

$$r_{y1}^2 = (0.807)^2 = 0.651 = 65.1\%$$

再引入x_2时，x_2能消除剩余变差SSE（X_1）的比例为

$$r_{y2;1}^2 = (0.739)^2 = 0.546 = 54.6\%$$

因而自变量x_1和x_2消除y变差的总比例为

$1-(1-r_{y1}^2)(1-r_{y2;1}^2)$ =1-(1-0.651)(1-0.546)=0.842=84.2%。

这个值84.2%恰好是y对x_1和x_2二元线性回归的判定系数R^2

3.6 相关阵与偏相关系数

三、偏相关系数

对任意p个变量$x_1,x_2,\ldots,x_p$定义它们之间的偏相关系数

$$r_{12;3,\cdots,p}=\frac{-\Delta_{12}}{\sqrt{\Delta_{11}\cdot\Delta_{22}}}$$

其中符号Δ_{ij}表示相关阵第i行第j列元素的代数余子式

验证 $$r_{12;3}=\frac{r_{12}-r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

3.7 本章小结与评注

例3.3 中国民航客运量的回归模型。

y—民航客运量(万人),

x_1—国民收入(亿元),

x_2—消费额(亿元),

x_3—铁路客运量(万人),

x_4—民航航线里程(万公里),

x_5—来华旅游入境人数(万人)。

根据《1994年统计摘要》获得1978-1993年统计数据

3.7 本章小结与评注

年份	y	x1	x2	x3	x4	x5
1978	231	3010	1888	81491	14.89	180.92
1979	298	3350	2195	86389	16.00	420.39
1980	343	3688	2531	92204	19.53	570.25
1981	401	3941	2799	95300	21.82	776.71
1982	445	4258	3054	99922	23.27	792.43
1983	391	4736	3358	106044	22.91	947.70
1984	554	5652	3905	110353	26.02	1285.22
1985	744	7020	4879	112110	27.72	1783.30
1986	997	7859	5552	108579	32.43	2281.95
1987	1310	9313	6386	112429	38.91	2690.23
1988	1442	11738	8038	122645	37.38	3169.48
1989	1283	13176	9005	113807	47.19	2450.14
1990	1660	14384	9663	95712	50.68	2746.20
1991	2178	16557	10969	95081	55.91	3335.65
1992	2886	20223	12985	99693	83.66	3311.50
1993	3383	24882	15949	105458	96.08	4152.70

3.7 本章小结与评注

Correlations

		y	x1	x2	x3	x4	x5
y	Pearson Correlatio	1	.989**	.985**	.227	.987**	.924**
	Sig. (2-tailed)		.000	.000	.398	.000	.000
	N	16	16	16	16	16	16
x1	Pearson Correlatio	.989**	1	.999**	.258	.984**	.930**
	Sig. (2-tailed)	.000		.000	.335	.000	.000
	N	16	16	16	16	16	16
x2	Pearson Correlatio	.985**	.999**	1	.289	.978**	.942**
	Sig. (2-tailed)	.000	.000		.278	.000	.000
	N	16	16	16	16	16	16
x3	Pearson Correlatio	.227	.258	.289	1	.213	.504*
	Sig. (2-tailed)	.398	.335	.278		.428	.046
	N	16	16	16	16	16	16
x4	Pearson Correlatio	.987**	.984**	.978**	.213	1	.882**
	Sig. (2-tailed)	.000	.000	.000	.428		.000
	N	16	16	16	16	16	16
x5	Pearson Correlatio	.924**	.930**	.942**	.504*	.882**	1
	Sig. (2-tailed)	.000	.000	.000	.046	.000	
	N	16	16	16	16	16	16

**.Correlation is significant at the 0.01 level (2-tailed).

*.Correlation is significant at the 0.05 level (2-tailed).

3.7 本章小结与评注

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.999[a]	.998	.997	49.49240

a. Predictors: (Constant), x5, x3, x4, x2, x1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13818877	5	2763775.354	1128.303	.000[a]
	Residual	24494.981	10	2449.498		
	Total	13843372	15			

a. Predictors: (Constant), x5, x3, x4, x2, x1

b. Dependent Variable: y

3.7 本章小结与评注

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	450.909	178.078		2.532	.030
	x1	.354	.085	2.447	4.152	.002
	x2	-.561	.125	-2.485	-4.478	.001
	x3	-.007	.002	-.083	-3.510	.006
	x4	21.578	4.030	.531	5.354	.000
	x5	.435	.052	.564	8.440	.000

a. Dependent Variable: y