第 17 章 多元线性回归分析

本章重点大纲：

17.1 多元线性回归分析参数的点估计

17.2 回归系数的区间估计与检定

17.3 多元回归分析模式的适合性

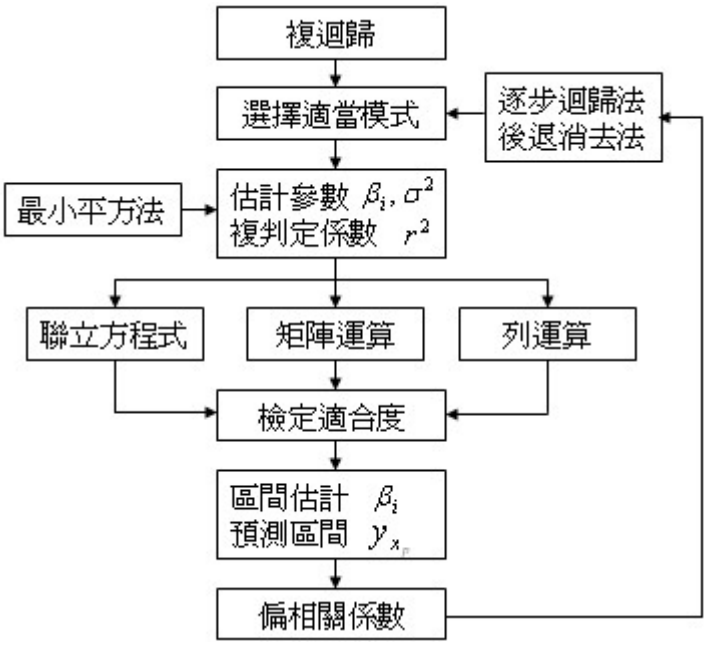
17.4 偏相关系数

17.5 因变量与期望值的预测区间与区间估计

17.6 利用列运算作估计与检定

17.7 选择适当多元回归模式

17.8 中文统计应用



17.1 多元线性回归分析参数的点估计

多元线性回归分析(multiple linear regression analysis)是分析两个以上自变量对因变量的直线相关关系。多元线性回归分析的数学模式如下：

$$Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}+\varepsilon_i \qquad i=1,\ldots,n$$

假设条件是：

1. $\beta_0,\beta_1,\cdots,\beta_k$ 为未知参数。(k 个控制变量)
2. x_{ij} 是第 j 个控制变量（人为选择变量，自变量）的第 i 个样本数据值，没有误差。
   $(y,,x_{i1},x_{i2},\cdots,x_{ik})$ 为一组样本数据。
3. ε_i 是误差项，随机变量，独立，期望值为 0，方差未知但相同。
   $E(\varepsilon_i)=0,\ V(\varepsilon_i)=\sigma^2,\ Cov(\varepsilon_i,\varepsilon_j)=0 \qquad i\neq j$
4. Y_i 是随机变量，独立，期望值为 $\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik}$，方差未知但相同。
   $E(Y_i)=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik},\ V(Y_i)=\sigma^2,\ Cov(Y_i,Y_j)=0 \qquad i\neq j$

以上模式可以用矩阵来表示：

$$Y=X\beta+\varepsilon$$

$$Y=\begin{pmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{pmatrix} \quad X=\begin{pmatrix}1 & x_{11} & x_{12} & \cdots & x_{1k}\\1 & x_{21} & x_{22} & \cdots & x_{2k}\\\vdots & \vdots & \vdots & \ddots & \vdots\\1 & x_{n1} & x_{n2} & \cdots & x_{nk}\end{pmatrix}$$

$$\beta=\begin{pmatrix}\beta_0\\\beta_1\\\vdots\\\beta_k\end{pmatrix} \quad \varepsilon=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix} \quad X'=\begin{pmatrix}1 & 1 & \cdots & 1\\x_{11} & x_{21} & \cdots & x_{n1}\\\vdots & \vdots & \ddots & \vdots\\x_{1k} & x_{2k} & \cdots & x_{nk}\end{pmatrix}$$

X' 是 X 的转置矩阵(transpose)。Y 是 n×1 的矩阵，X 是 n×(k+1) 的矩阵，β 是(k+1)×1 的矩阵，ε 是 n×1 的矩阵，X' 是 (k+1)×n 的矩阵。

β 的估计量 b 是：$b=\begin{pmatrix}b_0\\b_1\\\vdots\\b_k\end{pmatrix}$

β 的点估计 b，是利用最小平方法(least squares method)，这是使误差项的平方和最小的估计值。求得下列联立方程式：$X'Xb=X'Y$

所以 β 的估计量 b 是：$b=(X'X)^{-1}X'Y$

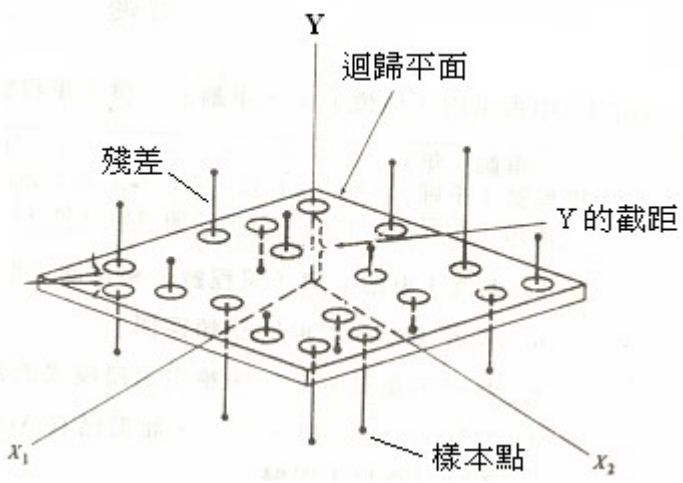


图 17.1　最小平方和多元回归

计算方差分析的各项平方和 ：

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=SS_E+SS_R=Y'Y-n\bar{Y}^2$$

$$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=b'X'Y-n\bar{Y}^2$$

$$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=Y'Y-b'X'Y$$

回归模式方差(未解释方差) σ^2 的估计量：

$$MS_E=\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\frac{SS_E}{n-k-1}$$

多元判定系数 r^2 (coefficient of multiple determination)是判定所有$\mathbf{X_i}$变数对 Y 是否有(直线的)关系。修正判定系数r_a^2 (adjusted r^2)定义如下：

$$r^2=\frac{SS_R}{SS_T}=1-\frac{SS_E}{SS_T}\qquad r_a^2=1-\left(\frac{n-1}{n-k-1}\right)\frac{SS_E}{SS_T}$$

估计量 b 的共变量矩阵(covariance matrix)是：

$$Cov(b)=\begin{pmatrix}V(b_0) & Cov(b_0,b_1) & \cdots & Cov(b_0,b_k)\\ Cov(b_1,b_0) & V(b_1) & \cdots & Cov(b_1,b_k)\\ \vdots & \vdots & \ddots & \vdots\\ Cov(b_k,b_0) & Cov(b_k,b_1) & \cdots & V(b_k)\end{pmatrix}=\sigma^2(X'X)^{-1}$$

$$Cov(b_i,b_i)=V(b_i)$$

共变数矩阵 Cov(b)的估计量：

$$S_b^2=\begin{pmatrix}S_{b_0}^2 & S_{b_0b_1} & \cdots & S_{b_0b_k}\\ S_{b_1b_0} & S_{b_1}^2 & \cdots & S_{b_1b_k}\\ \vdots & \vdots & \ddots & \vdots\\ S_{b_kb_0} & S_{b_kb_1} & \cdots & S_{b_k}^2\end{pmatrix}=MS_E(X'X)^{-1}$$

例题 17.1：食品公司随机抽样 16 个超级市场，比较广告与销售量的关系。自变量 X_1 是信箱媒体广告，自变量 X_2 是销售点试吃广告，因变量 Y 是广告期间销售量。进行多元回归分析。

超级市场 Sample	媒体广告 X_1 ($0,000)	销售点广告 x_2 ($0,000)	销售量 y$00,000
1	2	2	8.74
2	2	3	10.53
3	2	4	10.99
4	2	5	11.97
5	3	2	12.74
6	3	3	12.83
7	3	4	14.69
8	3	5	15.30
9	4	2	16.11
10	4	3	16.31
11	4	4	16.46
12	4	5	17.69
13	5	2	19.65
14	5	3	18.86
15	5	4	19.93
16	5	5	20.51

解答：

$$Y=\begin{pmatrix}8.74\\10.53\\10.99\\11.97\\12.74\\12.83\\14.69\\15.30\\16.11\\16.31\\16.46\\17.69\\19.65\\18.86\\19.93\\20.51\end{pmatrix}\qquad X=\begin{pmatrix}1&2&2\\1&2&3\\1&2&4\\1&2&5\\1&3&2\\1&3&3\\1&3&4\\1&3&5\\1&4&2\\1&4&3\\1&4&4\\1&4&5\\1&5&2\\1&5&3\\1&5&4\\1&5&5\end{pmatrix}$$

$$X'X=\begin{pmatrix}16&56&56\\56&216&196\\56&196&216\end{pmatrix}\qquad (X'X)^{-1}=\begin{pmatrix}1.2875&-0.175&-0.175\\-0.175&0.05&0\\-0.175&0&0.05\end{pmatrix}$$

$$b=(X'X)^{-1}X'Y=\begin{pmatrix}1.2875&-0.175&-0.175\\-0.175&0.05&0\\-0.175&0&0.05\end{pmatrix}\begin{pmatrix}243.31\\912.17\\865.70\end{pmatrix}=\begin{pmatrix}2.13437\\3.02925\\0.70575\end{pmatrix}$$

回归方程式为： $y = 2.13437 + 3.02925x_1 + 0.70575x_2$

$$SS_T=\sum_{i=1}^{n}(y_i-\bar{y})^2=SS_E+SS_R=Y'Y-n\bar{Y}^2=197.25$$

$$SS_R=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=b'X'Y-n\bar{Y}^2=193.49$$

$$SS_E=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=Y'Y-b'X'Y=3.76$$

回归模式方差(未解释方差)，即 σ^2 的估计量：

$$MS_E=\frac{SS_E}{n-k-1}=0.289$$

多元判定系数 r^2：　$r^2=\frac{SS_R}{SS_T}=1-\frac{SS_E}{SS_T}=0.981$

$$r_a^2=1-\left(\frac{n-1}{n-k-1}\right)\frac{SS_E}{SS_T}=1-\left(\frac{15}{13}\right)\frac{3.76}{197.25}=0.978$$

所以这个回归模式的解释能力相当高。

共变量矩阵 Cov(b)的估计量：

$$S^2{}_b=0.289\begin{pmatrix}1.2875 & -0.175 & -0.175\\ -0.175 & 0.05 & 0\\ -0.175 & 0 & 0.05\end{pmatrix}$$

17.2　回归系数的区间估计与检定

从本节以后检定所需要的假设条件，除了多元回归分析的原来的四个假设以外，还要加上常态分配的假设，即：　$\varepsilon_i\sim N(0,\sigma^2)$。所以：

1. 回归参数 β_i 的估计量 b_i，是一个随机变量，根据以上假设，b_i 是一个常态分配，

   $b_i\sim N\left(\beta_i,\sigma^2(X'X)^{-1}_{(i+1)(i+1)}\right)$

   $(X'X)^{-1}_{(i+1)(i+1)}$ 是矩阵 $(X'X)^{-1}$ 的第 (i+1) 列第 (i+1) 行。

2. $S_{b_i}=\sqrt{MS_E(X'X)^{-1}_{(i+1)(i+1)}}$

3. β_i 的区间估计，β_i 的 1－α 信赖区间：

   $$b_i-t_{\frac{\alpha}{2},n-k-1}S_{b_i}\le\beta_i\le b_i+t_{\frac{\alpha}{2},n-k-1}S_{b_i}$$

4. β_i 的检定：

   双尾检定 $\begin{cases}H_0^I:\beta_i=c\\ H_A^I:\beta_i\neq c\end{cases}$　左尾检定 $\begin{cases}H_0^{II}:\beta_i\ge c\\ H_A^{II}:\beta_i<c\end{cases}$　右尾检定 $\begin{cases}H_0^{III}:\beta_i\le c\\ H_A^{III}:\beta_i>c\end{cases}$

计算检定值 $t=\dfrac{b_i-c}{s_{b_i}}$

若 $|t|\geq t_{\frac{\alpha}{2},n-k-1}$，则否定双尾检定 H_0^I。

若 $t<-t_{\alpha,n-k-1}$，则否定左尾检定 H_0^{II}。

若 $t>t_{\alpha,n-k-1}$，则否定右尾检定 H_0^{III}。

5. σ 的推论，MS_E 与 b 是独立的，而且：$\dfrac{(n-k-1)MS_E}{\sigma^2}\sim\chi^2_{n-k-1}$

σ^2 的信赖区间：$\left(\dfrac{(n-k-1)MS_E}{\chi^2_{\frac{\alpha}{2},n-k-1}},\dfrac{(n-k-1)MS_E}{\chi^2_{1-\frac{\alpha}{2},n-k-1}}\right)$

例题 17.2：食品公司广告与销售量的多元回归分析。

β_0 的 95% 信赖区间：

$$b_0\pm t_{\frac{\alpha}{2},n-k-1}S_{b_0}=2.13437\pm t_{0.025,13}\sqrt{0.289(1.2875)}=2.13437\pm 1.3176$$

β_1 的 95% 信赖区间：

$$b_1\pm t_{\frac{\alpha}{2},n-k-1}S_{b_1}=3.02925\pm t_{0.025,13}\sqrt{0.289(0.05)}=3.02925\pm 0.2596$$

β_2 的 95% 信赖区间：

$$b_2\pm t_{\frac{\alpha}{2},n-k-1}S_{b_2}=0.70575\pm t_{0.025,13}\sqrt{0.289(0.05)}=0.70575\pm 0.2596$$

17.3 多元回归分析模式的适合性

利用方差分析 F 分配检定 $H_0:\beta_1=\beta_2=\cdots=\beta_k=0$ 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio
回归模式 Regression	k	SS_R	$MS_R=\dfrac{SS_R}{k}$	$F=\dfrac{MS_R}{MS_E}$
误差 Error	n-k-1	SS_E	$MS_E=\dfrac{SS_E}{n-k-1}$	
总和 Total	n-1	SS_T		

$F=\dfrac{MS_R}{MS_E}=\dfrac{r^2/k}{(1-r^2)/(n-k-1)}$ 若 $F\geq F_{\alpha,k,n-k-1}$，则否定 H_0。

例题 17.3：食品公司广告与销售量的多元回归分析。

利用方差分析 F 分配检定 $H_0: \beta_1=\beta_2= 0$ 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio	$F_{0.05,2,13}$
回归模式 Regression	2	SS_R ＝193.49	MS_R＝96.745	F＝334.75	3.81
误差 Error	13	SS_E ＝3.76	MS_E＝0.289		
总和 Total	15	SS_T ＝197.25			

因为 F＝334.75$\geq F_{0.05,2,13}=3.81$，所以否定 H_0。

17.4 偏相关系数

偏判定系数 (partial determination coefficient) $\rho^2_{yk\cdot\{\text{all except } y,k\}}$ 是当其它自变量已经对 y 作回归方程式，变量 x_k 再加入回归式，可以增加对 y 变异的解释程度。

偏相关系数 (partial correlation coefficient) $\rho_{yk\cdot\{\text{all except } y,k\}}$是当其它变量保持常数，变量 y 与变量 x_k 的相关系数。换言之，$\rho_{y1\cdot23}$是当变量 x_2 与 x_3 保持常数，其它变量不考虑时，变量 y 与变量 x_1 的相关系数。 另外一种说法是，(1)x_1 对 x_2 与 x_3 回归的残差(residual)，与(2)y 对 x_2 与 x_3 回归的残差(residual)，两者之间的相关系数。样本偏判定系数 $r^2_{yk\cdot\{\text{all except } y,k\}}$是偏判定系数 $\rho^2_{yk\cdot\{\text{all except } y,k\}}$的估计量。样本偏相关系数 $r_{yk\cdot\{\text{all except } y,k\}}$是偏相关系数 $\rho_{yk\cdot\{\text{all except } y,k\}}$的估计量。

令 $SS_R(X_1)$为 Y 对 X_1 回归的「已解释变异」平方和。$SS_R(X_1,X_2,X_3)$为 Y 对 X_1，X_2 与 X_3 回归的「已解释变异」平方和。

令 $SS_E(X_1)$为 Y 对 X_1 回归的误差(残差,「未解释变异」)平方和(residual sum of square)。 $SS_E(X_1,X_2,X_3)$为 Y 对 X_1，X_2 与 X_3 回归的误差(残差，「未解释变异」)平方和。

$SSR(X_1)+SSE(X_1)=SS_R(X_1,X_2)+SS_E(X_1,X_2)=\ldots=SS_T=\sum(Y-\overline{Y})^2$

令 $SS_R(X_2 \mid X_1)$为已有 X_1 对 Y 作回归，加入 X_2 当作回归自变量，可以增加回归的「已解释变异」平方和；$SS_R(X_3 \mid X_1,X_2)$为已有 X_1,X_2 对 Y 作回归，如果再加入 X_3 当作回归自变量，可以增加回归的「已解释变异」平方和。

$SS_R(X_2 \mid X_1)=SS_R(X_1,X_2)-SS_R(X_1)=SS_E(X_1)-SS_E(X_1,X_2)$

$SS_R(X_3 \mid X_1,X_2)=SS_R(X_1,X_2,X_3)-SS_R(X_1,X_2)=SS_E(X_1,X_2)-SS_E(X_1,X_2,X_3)$

因变数 Y 和其它自变数 X_i 的偏判定系数 $r^2_{y4\cdot123}$或记作 $r^2_{YX_4\cdot X_1X_2X_3}$ 的计算公式如下：

$$r^2_{y3\cdot124}=\frac{SS_R(X_3 \mid X_1,X_2,X_4)}{SS_E(X_1,X_2,X_4)}=\frac{SS_E(X_1,X_2,X_4)-SS_E(X_1,X_2,X_3,X_4)}{SS_E(X_1,X_2,X_4)}=\frac{SS_R(X_1,X_2,X_3,X_4)-SS_R(X_1,X_2,X_4)}{SS_T-SS_R(X_1,X_2,X_4)}$$

$$0 \le r^2_{y3\cdot124} \le 1$$

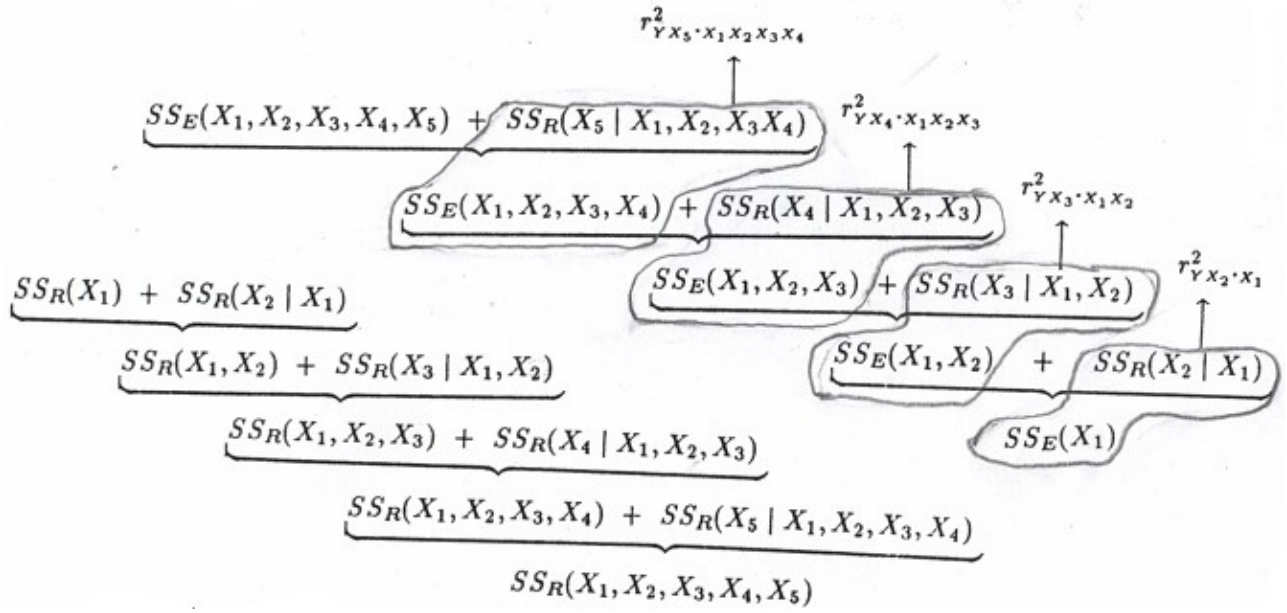


图 17.3 偏相关系数

变异来源 Source	平方和 SS	自由度 df
回归	$SS_R(X_1,X_2,X_3)$	3
	$SS_R(X_3 \mid X_1,X_2)$	1
	$SS_R(X_1,X_2)$	2
	$SS_R(X_2 \mid X_1)$	1
	$SS_R(X_1)$	1
残差	$SS_E(X_1,X_2,X_3)$	n－4
总和	SS_T	n－1

$$g=\begin{pmatrix} g_0 \\ g_1 \\ g_2 \\ \vdots \\ g_k \end{pmatrix}=\left(\sum_{i=1}^{n} y_i \quad \sum_{i=1}^{n} x_{i1}y_i \quad \sum_{i=1}^{n} x_{i2}y_i \quad \cdots \quad \sum_{i=1}^{n} x_{ik}y_i\right)'$$

$$SS_R(X_1)=\hat{\beta}_0 g_0+\hat{\beta}_1 g_1-n\bar{y}^2 \qquad SS_R(X_1,\cdots,X_k)=\sum_{j=0}^{k}\hat{\beta}_j g_j-n\bar{y}^2$$

$$SS_R(X_2 \mid X_1)=\hat{\beta}_2 g_2 \qquad SS_R(X_r \mid X_1,\cdots,X_{r-1})=\hat{\beta}_r g_r$$

$$SS_R(X_{r+1},\cdots,X_k \mid X_1,\cdots,X_r)=\sum_{j=r+1}^{k}\hat{\beta}_j g_j$$

$\hat{\beta}_i=b_i$ 是所有自变量放入回归模式(全模式 full model)的估计参数。

当回归方程式已有 $X_1,X_2,...,X_{k-1}$ 自变量，如果加入新的自变量 X_k，但是使 $SS_R(X_1,X_2,...,X_{k-1})$增加到 $SS_R(X_1,X_2,...,X_{k-1},X_k)$，增加很少，即 $SS_R(X_k \mid X_1,X_2,...,X_{k-1})$很小，则自变数 X_k 不值得再加入回归方程式。

檢定 $\begin{cases} H_0: 已有X_1對Y迴歸，自變數X_2不值得再加入迴歸: \rho_{YX_2 \cdot X_1}=0 \\ H_A: 已有X_1對Y迴歸，自變數X_2值得再加入迴歸: \rho_{YX_2 \cdot X_1} \neq 0 \end{cases}$

检定的统计量(n 是样本数据的数目)：

$$t^*=\frac{r_{YX_2 \cdot X_1}\sqrt{n-3}}{\sqrt{1-r^2_{YX_2 \cdot X_1}}}$$

若 $|t^*|>t_{\frac{\alpha}{2},n-3}$，则否定 H_0

檢定 $\begin{cases} H_0: X_1,X_2,\cdots,X_{k-1}對Y迴歸，X_k不值得再加入: \rho_{YX_k \cdot X_1,\cdots,X_{k-1}}=0 \\ H_A: X_1,X_2,\cdots,X_{k-1}對Y迴歸，X_k值得再加入: \rho_{YX_k \cdot X_1,\cdots,X_{k-1}} \neq 0 \end{cases}$

检定的统计量(n 是样本数据的数目)：

$$t^* = \frac{r_{YX_k \bullet X_1 X_2 \cdots X_{k-1}} \sqrt{n-k-1}}{\sqrt{1-r^2_{YX_k \bullet X_1 X_2 \cdots X_{k-1}}}}$$

若 $|t^*| > t_{\frac{\alpha}{2},n-k-1}$，则否定 H_0

检定 H_0：$\beta_{r+1}=\beta_{r+2}=\ldots=\beta_k=0$

计算 $F = \dfrac{SS_R(X_{r+1},\cdots,X_k \mid X_1,\cdots X_r)/(k-r)}{SS_E(X_1,\cdots,X_k)/(n-k-1)}$

若 $F \geq F_{\alpha,k-r,n-k-1}$，则否定 H_0

例题 17.4：抽查货运公司十天的行车记录，得行驶哩数(X_1)，货运数量 (X_2)，车种(X_3)，及行驶时间(Y)的数据，经输入计算机而得各差异平方和(SS)及自由度(df)分别为：

变异来源 Source	平方和 SS	自由度 df
回归	SS_R(X1,X2,X3) ＝270	3
	SS_R(X1,X2) ＝250	2
	SS_R(X3 ｜ X1,X2)＝20	1
残差	SS_E(X1,X2,X3) ＝30	6

1. 求偏相关系数 $r_{YX_3 \cdot X_1 X_2}$
2. 若回归模型已有 Y 与 X_1,X_2，试以显着水平 α＝0.05，检定 X_3 是否值得引入回归模型？

解答：从自由度可以看出样本数 n＝10

$$r^2_{YX_3 \cdot X_1 X_2} = \frac{SS_R(X_3 \mid X_1 X_2)}{SS_T - SS_R(X_1,X_2)} = \frac{20}{(270+30)-250} = 0.4$$

$$r_{YX_3 \cdot X_1 X_2} = \sqrt{0.4} = 0.632 \qquad t^* = \frac{r_{YX_3 \cdot X_1 X_2}\sqrt{n-4}}{\sqrt{1-r^2_{YX_3 \cdot X_1 X_2}}} = 2.00$$

因为 $|t^*| = 2.00 < t_{0.025,6} = 2.447$，接受 H_0，X_3 不值得引入回归模型。

17.5 因变量与其期望值的预测区间与区间估计

1. 对于某个特定值 X_p ＝[1 x_{1p} … x_{kp}]，未知数 y_p 的期望值 $\mu_{y_p} = X_p'\beta$，其估计值为 $\hat{y}_p = X_p'b$
2. $S_{\hat{y}_p} = \sqrt{MS_E\left(X_p'(X'X)^{-1}X_p\right)}$
3. $S_{y_p - \hat{y}_p} = \sqrt{MS_E\left(1 + X_p'(X'X)^{-1}X_p\right)}$

4. $\mu_{y_p}=E(\hat{y}_p)$的区间估计，μ_{y_p}的 1－α 信赖区间：

$$\hat{y}_p - t_{\frac{\alpha}{2},n-k-1}S_{\hat{y}_p} \le \mu_{y_p} \le \hat{y}_p + t_{\frac{\alpha}{2},n-k-1}S_{\hat{y}_p}$$

5. 对于某个特定值 X_p，未知数 y_p 的 1－α 预测区间：

$$\hat{y}_p - t_{\frac{\alpha}{2},n-k-1}S_{y_p-\hat{y}_p} \le y_p \le \hat{y}_p + t_{\frac{\alpha}{2},n-k-1}S_{y_p-\hat{y}_p}$$

例题 17.5：食品公司广告与销售量的多元回归分析。

对于特定值 $X_p=[1\ 6\ 2]$，未知数 y_p 的期望值，其$\mu_{y_p}=X_p'\beta$，其估计值为

$$\hat{y}_p=X_p'b=2.13437+3.02925(6)+0.70575(2)=21.72$$

$$X_p'(X'X)^{-1}X_p=\begin{pmatrix}1\\6\\2\end{pmatrix}\begin{pmatrix}1.2875 & -0.175 & -0.175\\ -0.175 & 0.05 & 0\\ -0.175 & 0 & 0.05\end{pmatrix}(1\quad 6\quad 2)=0.4875$$

$$S_{\hat{y}_p}=\sqrt{MS_E\left(X_p'(X'X)^{-1}X_p\right)}=\sqrt{0.289(0.4875)}=0.375$$

$$S_{y_p-\hat{y}_p}=\sqrt{MS_E\left(1+X_p'(X'X)^{-1}X_p\right)}=\sqrt{0.289(1+0.4875)}=0.655$$

μ_{y_p}的 95%信赖区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}S_{\hat{y}_p}=21.72\pm t_{0.025,13(0.375)}=21.72\pm 0.81$$

对于某个特定值 X_p，未知数 y_p 的 1－α 预测区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}S_{y_p-\hat{y}_p}=21.72\pm t_{0.025,13}(0.655)=21.72\pm 1.415$$

17.6　利用列运算作估计与检定

以上多元线性回归分析的估计，要利用矩阵的乘法及反矩阵。现在我们介绍一种以矩阵基本列运算为主的计算。假设有因变量 y 与自变量 x_i, i＝1,…,k。

$Y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_kx_{ik}+\varepsilon_i$　　i＝1,…,n

为了方便，以下表示 x_i 的注标都简化为 i，例如：S_{ii} 原本应该记作$S_{x_ix_i}$。首先计算下列平方和：

$$S_{ii}=\sum_{h=1}^{n}(x_{hi}-\bar{x}_i)^2=\sum_{h=1}^{n}x_{hi}^2-n\bar{x}_i^2=(n-1)S_i^2$$

S_i^2＝第 i 个自变数方差

$$S_{ij}=S_{ji}=\sum_{h=1}^{n}(x_{hi}-\bar{x}_i)(x_{hj}-\bar{x}_j)=\sum_{h=1}^{n}x_{hi}x_{hj}-n\bar{x}_i\bar{x}_j$$

$$S_{iy}=\sum_{h=1}^{n}(x_{hi}-\bar{x}_i)(y_h-\bar{y})=\sum_{h=1}^{n}x_{hi}y_h-n\bar{x}_i\bar{y}$$

$$S_{yy}=\sum_{h=1}^{n}(y_h-\bar{y})^2=\sum_{h=1}^{n}y_h^2-n\bar{y}^2$$

(一). 将上述平方和写成下列矩阵：

$$S=\left(\begin{array}{cccc|c|cccc} S_{11} & S_{12} & \cdots & S_{1k} & |S_{1y}| & 1 & 0 & \cdots & 0 \\ S_{12} & S_{22} & \cdots & S_{2k} & |S_{2y}| & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{1k} & S_{2k} & \cdots & S_{kk} & |S_{ky}| & 0 & 0 & \cdots & 1 \end{array}\right)$$

利用列运算(row operations)，将上述矩阵转换成下列型式：

$$T=\left(\begin{array}{cccc|c|cccc} 1 & 0 & \cdots & 0 & |b_1| & p_{11} & p_{12} & \cdots & p_{1k} \\ 0 & 1 & \cdots & 0 & |b_2| & p_{21} & p_{22} & \cdots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & |b_k| & p_{k1} & p_{k2} & \cdots & p_{kk} \end{array}\right)$$

上述矩阵的中间列 $b_1,b_2,\ldots, b_k$ 是回归参数 $\beta_1,\beta_2,\ldots,\beta_k$ 的估计值。

β_0 的估计值 b_0 为：$b_0=\bar{y}-\sum b_i\bar{x}_i$

$SS_T=S_{yy}$ $\quad SS_R=\sum_{i=1}^{k}b_iS_{iy}$ $\quad SS_E=S_{yy}-\sum_{i=1}^{k}b_iS_{iy}$

$$MS_E=\frac{S_{yy}-\sum_i b_iS_{iy}}{n-k-1}$$

多元判定系数 r^2(coefficient of multiple determination)是：

$$r^2=\frac{SS_R}{SS_T}=\frac{\sum b_iS_{iy}}{S_{yy}}$$

(二). 利用方差分析 F 分配检定 $H_0:\beta_1=\beta_2=\ldots=\beta_k=0$ 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio
回归模式 Regression	k	$SS_R=\sum_{i=1}^{k}bS_{iy}$	$MS_R=\frac{SS_R}{k}$	$F=\frac{MS_R}{MS_E}$
误差 Error	n−k−1	$SS_E=S_{yy}-\sum_{i=1}^{k}bS_{iy}$	$MS_E=\frac{SS_E}{n-k-1}$	
总和 Total	n−1	$SS_T=S_{yy}$		

$$F=\frac{MS_R}{MS_E}=\frac{\sum_i b_iS_{iy}/k}{(S_{yy}-\sum_i b_iS_{iy})/(n-k-1)}=\frac{r^2/k}{(1-r^2)/n-k-1}$$

若 $F\geq F_{\alpha,k,n-k-1}$，则否定 H_0。

(三). β_i 的区间估计，β_i 的 1−α 信赖区间：

$$b_i-t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_Ep_{ii}}\leq\beta_i\leq b_i+t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_Ep_{ii}}$$

p_{ii} 是矩阵 T 中的 p_{ii}(矩阵对角线的值)

(四). $\beta_i-\beta_j$ 的区间估计，$\beta_i-\beta_j$ 的 1－α 信赖区间：

$$b_i-b_j \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E(p_{ii}+p_{jj}-2p_{ij})}$$

(五). 对于某个特定值 $X_p=[1,x_{1p},\ldots,x_{kp}]$，未知数 y_p 的期望值 $\mu_{y_p}=X_p'\beta$，其估计值为 $\hat{y}_p=X_p'b$。

$\mu_{y_p}=E(\hat{y}_p)$ 的区间估计，μ_{y_p} 的 1－α 信赖区间：

$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E[1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)]}$，当 k＝2，则：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-3}\sqrt{MS_E\left[1/n+p_{11}(x_{1p}-\bar{x}_1)^2+2p_{12}(x_{1p}-\bar{x}_1)(x_{2p}-\bar{x}_2)+p_{22}(x_{2p}-\bar{x}_2)^2\right]}$$

(六). 对于某个特定值 X_p，未知数 y_p 的 1－α 预测区间：

$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E[1+1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)]}$

当 k＝2，则：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-3}\sqrt{MS_E\left[1+1/n+p_{11}(x_{1p}-\bar{x}_1)^2+2p_{12}(x_{1p}-\bar{x}_1)(x_{2p}-\bar{x}_2)+p_{22}(x_{2p}-\bar{x}_2)^2\right]}$$

(七). 每两个自变数 x_i 与 x_j，其样本相关系数为 r_{ij}；一个自变量 x_i 与因变量 y，其样本相关系数 r_{yi}：

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \qquad r_{yi}=\frac{s_{iy}}{\sqrt{s_{ii}s_{yy}}}$$

(八). 偏相关系数(partial correlation coefficients)$\rho_{yi\cdot j}$：

$$\rho_{yi\cdot j}=\frac{\rho_{yi}-\rho_{yj}\rho_{ij}}{\sqrt{(1-\rho_{yj}^2)(1-\rho_{ij}^2)}}$$

(九). 样本偏相关系数(sample partial correlation coefficients) $r_{yi\cdot j}$ 是偏相关系数 $\rho_{yi\cdot j}$ 的估计值。

$$r_{yi\cdot j}=\frac{r_{yi}-r_{yj}r_{ij}}{\sqrt{(1-r_{yj}^2)(1-r_{ij}^2)}}$$

令矩阵 R 为 k 个变量(包括自变量与因变量)的样本偏相关系数的矩阵：

$$R=\begin{pmatrix}1 & r_{12} & \cdots & r_{1k}\\ r_{21} & 1 & \cdots & r_{2k}\\ \vdots & \vdots & \vdots & \vdots\\ r_{k1} & r_{k2} & \cdots & 1\end{pmatrix} \qquad R^{-1}=\begin{pmatrix}r^{11} & r^{12} & \cdots & r^{1k}\\ r^{21} & r^{22} & \cdots & r^{2k}\\ \vdots & \vdots & \vdots & \vdots\\ r^{k1} & r^{k2} & \cdots & r^{kk}\end{pmatrix}$$

矩阵 R^{-1} 为 R 的反矩阵：

样本偏相关系数 $r_{jk\cdot(\text{all except } j,k)}$ 是当其它变量保持常数，变量 x_j 与变量 x_k 的相关系数。

$$r_{jk\cdot(\text{all except } j,k)}=\frac{-r^{jk}}{\sqrt{r^{jj}r^{kk}}}$$

例题 17.6： 下列是随机抽样 5 个家庭的父亲，母亲，和长子的教育年数。若以长子的教育年数为因变数，父亲和母亲的教育年数为自变量，进行多元回归分析。

家庭	父亲	母亲	长子

Family	x_1	x_2	y	x_1^2	x_2^2	x_1x_2	x_1y	x_2y	y^2
1	10	10	11	100	100	100	110	110	121
2	11	9	12	121	81	99	132	108	144
3	8	11	12	64	121	88	96	132	144
4	9	12	12	81	144	108	108	144	144
5	12	8	8	144	64	96	96	64	64
总和	50	50	55	510	510	491	542	558	617
平均	10	10	11						

解答：$S_{11}=\sum(x_1-\bar{x}_1)^2=\sum x_1^2-n(\bar{x}_1)^2=510-500=10$

$S_{12}=S_{21}=\sum(x_1-\bar{x}_1)(x_2-\bar{x}_2)=\sum x_1x_2-n\bar{x}_1\bar{x}_2=491-500=-9$

$S_{22}=\sum(x_2-\bar{x}_2)^2=\sum x_2^2-n(\bar{x}_2)^2=510-500=10$

$S_{1y}=\sum(x_1-\bar{x}_1)(y-\bar{y})=\sum x_1y-n\bar{x}_1\bar{y}=542-550=-8$

$S_{2y}=\sum(x_2-\bar{x}_2)(y-\bar{y})=\sum x_2y-n\bar{x}_2\bar{y}=558-550=8$

$S_{yy}=\sum(y-\bar{y})^2=\sum y^2-n(\bar{y})^2=617-605=12$

$$\left(\begin{array}{cc|c|cc} S_{11} & S_{12} & S_{1y} & 1 & 0 \\ S_{12} & S_{22} & S_{2y} & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{cc|c|cc} 10 & -9 & -8 & 1 & 0 \\ -9 & 10 & 8 & 0 & 1 \end{array}\right) \quad \left(\begin{array}{cc|c|cc} 1 & -0.9 & -0.8 & 0.1 & 0 \\ -9 & 10 & 8 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{cc|c|cc} 1 & -0.9 & -0.8 & 0.1 & 0 \\ 0 & 1.9 & 0.8 & 0.9 & 1 \end{array}\right) \quad \left(\begin{array}{cc|c|cc} 1 & -0.9 & -0.8 & 0.1 & 0 \\ 0 & 1 & 0.421 & 0.474 & 0.526 \end{array}\right)$$

$$\left(\begin{array}{cc|c|cc} 1 & 0 & -0.421 & 0.527 & 0.474 \\ 0 & 1 & 0.421 & 0.474 & 0.526 \end{array}\right)$$

所以 $b_1=-0.421$，$b_2=0.421$

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=11-(-0.421)(10)-(0.421)(10)=11$$

最小平方和的回归方程式：

$$\hat{y}=11-0.421x_1+0.421x_2$$

利用方差分析 F 分配检定 $H_0:\beta_1=\beta_2=0$ 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio
回归模式 Regression	k=2	$SS_R=\sum_i b_iS_{iy}=6.736$	$MS_R=3.368$	$F=1.28$
误差 Error	n−k−1=2	$SS_E=5.264$	$MS_E=2.632$	
总和 Total	n−1=4	$SS_T=S_{yy}=12$		

多元判定系数 $r^2=\frac{\sum b_iS_{iy}}{S_{yy}}=\frac{6.736}{12}=0.5613$

利用方差分析 F 分配检定 H_0:$\beta_1=\beta_2=0$ 的假设。

检定的 $F=1.28<F_{0.05,2,2}=19.0$，接受 H_0。

所以父母亲的教育年数和长子的教育年数无关。

例题 17.7：下列是野生动物鸟园想知道鸟的实际数目，随机抽样 6 个小区域，先用观查看到鸟的数目，在用听的听到鸟的声音的数目，最后将鸟赶出计算实际的鸟的数目。如果实际鸟的数目是因变量；看到鸟的数目与听到声音的数目 为自变量，进行多元回归分析。

区域	看到	听到	实际						
	x_1	x_2	y	x_1^2	x_2^2	x_1x_2	x_1y	x_2y	y^2
1	0	1	2	0	1	0	0	2	4
2	4	5	8	16	25	20	32	40	64
3	2	3	3	4	9	6	6	9	9
4	3	6	6	9	36	18	18	36	36
5	1	0	1	1	0	0	1	0	1
6	2	3	4	4	9	6	8	12	16
总和	12	18	24	34	80	50	65	99	130
平均	2	3	4						

解答： $S_{11}=10$ $S_{12}=S_{21}=14$ $S_{22}=26$ $S_{1y}=17$ $S_{2y}=27$ $S_{yy}=34$

$$\begin{pmatrix} S_{11} & S_{12} & |S_{1y}| & 1 & 0 \\ S_{12} & S_{22} & |S_{2y}| & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 10 & 14 & |17| & 1 & 0 \\ 14 & 26 & |27| & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & |1.0| & 0.40625 & -0.21875 \\ 0 & 1 & |0.5| & -0.21875 & 0.15625 \end{pmatrix}$$

所以 $b_1=1.0, b_2=0.5$

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=4.0-1.0(2)-(0.5)(3)=0.5$$

最小平方和的回归方程式 ：

$$\hat{y}=0.5+1.0x_1+0.5x_2$$

利用方差分析 F 分配检定 H_0:$\beta_1=\beta_2=0$ 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio
回归模式 Regression	2	$SS_R=30.5$	$MS_R=15.25$	$F=13.07$
误差 Error	3	$SS_E=3.5$	$MS_E=1.1667$	
总和 Total	n-1=5	$SS_T=S_{yy}=34$		

多元判定系数 $r^2=\frac{\sum b_i S_{iy}}{S_{yy}}=\frac{30.5}{34}=0.897$

利用方差分析 F 分配检定 $H_0:\beta_1=\beta_2=0$ 的假设，

检定的 $F=13.07>F_{0.05,2,3}=9.552$，所以否定 H_0。

检定 $H_0:\beta_1=0$ 的假设：

检定的 $t=\frac{b_1}{\sqrt{MS_E p_{11}}}=\frac{1}{\sqrt{(1.1667)(0.40625)}}=1.4525$

$t = 1.4525<t_{0.025,3}=3.182$，所以接受 H_0。

检定 $H_0:\beta_2=0$ 的假设：

检定的 $t=\frac{b_2}{\sqrt{MS_E p_{22}}}=\frac{0.5}{\sqrt{(1.1667)(0.15625)}}=1.1711$

$t =1.1711 < t_{0.025,3}=3.182$，所以接受 H_0。

β_1 的 1－α 信赖区间：

$$b_1 - t_{0.025,3}\sqrt{(1.1667)(0.40625)} \le \beta_1 \le b_1 + t_{0.025,3}\sqrt{(1.1667)(0.40625)}$$

对于某个特定值 $x_1=3, x_2=5, X_p=[1,3,5]$，未知数 y_p 的 期望值 $\mu_{y_p}=X_p'\beta$，

$$\hat{y}_p=(1,3,5)(0.5,1.0,0.5)'=6$$

$\mu_{y_p}=E(\hat{y}_p)$的区间估计，μ_{y_p}的 1－α 信赖区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E\left[1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)\right]}$$

$$=\hat{y}_p \pm t_{\frac{\alpha}{2},n-3}\sqrt{MS_E\left[1/n+p_{11}(x_{1p}-\bar{x}_1)^2+2p_{12}(x_{1p}-\bar{x}_1)(x_{2p}-\bar{x}_2)+p_{22}(x_{2p}-\bar{x}_2)^2\right]}$$

$$=6\pm 3.182\sqrt{1.1667[1/6+0.40625(3-2)^2+2(-0.21875)(3-2)(5-3)+0.15625(5-3)^2]}$$

$=6\pm5.727$

特定值 $x_1=3,x_2=5,X_p=[1,3,5]$，未知数 y_p 的 1－α 预测区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E\left[1+1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)\right]}=6\pm 6.025$$

例题 17.8：下列是随机抽样 5 个超市某产品的定价(X_1)，广告(X_2)，和销售量 (Y)，进行多元回归分析。

超市	定价	广告	销售量
	x_1	x_2	y
1	9.8	7.0	6.6
2	11.4	8.0	8.2
3	9.0	9.0	9.0
4	8.2	10.0	10.4
5	10.6	11.0	11.8
总和	49.0	45.0	46.0
平均	9.8	9.0	9.2

解答：n＝5, $\bar{x}_1$＝9.8, $\bar{x}_2$＝9.0, $\bar{y}$＝9.2

S_{11}＝6.4 S_{12}＝S_{21}＝－1.6 S_{22}＝10 S_{1y}＝－1.9 S_{2y}＝12.6 S_{yy}＝16.0

$$\begin{pmatrix} S_{11} & S_{12} & S_{1y} & 1 & 0 \\ S_{12} & S_{22} & S_{2y} & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 6.4 & -1.6 & -1.9 & 1 & 0 \\ -1.6 & 10.0 & 12.6 & 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0.01888 & 0.16276 & 0.02604 \\ 0 & 1 & 1.26302 & 0.02604 & 0.10417 \end{pmatrix}$$

所以 b_1＝0.01888,b_2＝1.26302

$$b_0=\bar{y}-b_1\bar{x}_1-b_2\bar{x}_2=9.2-0.01888(9.8)-1.26302(9.0)=-2.3522$$

最小平方和的回归方程式 ：

$$\hat{y}=-2.3522+0.01888x_1+1.26302x_2$$

利用方差分析 F 分配检定 H_0 : β_1＝β_2＝0 的假设：

多元回归分析的方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio
回归模式 Regression	2	SS_R＝15.878	MS_R＝7.939	F＝130.148
误差 Error	2	SS_E＝0.122	MS_E＝0.061	
总和 Total	n-1＝4	$SS_T=S_{yy}$＝16		

利用方差分析 F 分配检定 H_0:β_1＝β_2 ＝0 的假设 ：

检定的 F＝130.148 > $F_{0.05,2,2}$＝19.0，所以否定 H_0

检定 H_0 : β_1＝0 的假设：

检定的 $t=\frac{b_1}{\sqrt{MS_E p_{11}}}=\frac{0.01888}{\sqrt{(0.16276)(0.061)}}=0.189$

t＝0.189<$t_{0.025,2}$＝4.303，所以接受 H_0

检定 H_0 : β_2＝0 的假设：

检定的 $t=\frac{b_2}{\sqrt{MS_E p_{22}}}=\frac{1.26302}{\sqrt{(0.10417)(0.061)}}=15.844$

t＝15.844>$t_{0.025,2}$＝4.303，所以否定 H_0

对于某个特定值 x_1＝9, x_2＝10, X_p＝[1,9,10]，未知数 y_p 的期望值 $\mu_{y_p}=X_p'\beta$，

$$\hat{y}_p=(1,\ 9,\ 10)(-2.3522,\ 0.01888,\ 1.26302)'=10.4$$

$\mu_{y_p}=E(\hat{y}_p)$ 的区间估计，μ_{y_p}的 1－α 信赖区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E\left[1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)\right]}$$

$$=\hat{y}_p \pm t_{\frac{\alpha}{2},n-3}\sqrt{MS_E\left[1/n+p_{11}(x_{1p}-\bar{x}_1)^2+2p_{12}(x_{1p}-\bar{x}_1)(x_{2p}-\bar{x}_2)+p_{22}(x_{2p}-\bar{x}_2)^2\right]}$$

$$=10.4 \pm 4.305\sqrt{0.061[1/5+0.16276(9-9.8)^2+2(-0.02604)(9-9.8)(10-9.0)+0.10417(10-9.0)^2]}$$

$$=10.4\pm 0.713$$

特定值 $x_1=9, x_2=10, X_p=[1,9,10]$，未知数 y_p 的 1－α 预测区间：

$$\hat{y}_p \pm t_{\frac{\alpha}{2},n-k-1}\sqrt{MS_E\left[1+1/n+\sum_i\sum_j p_{ij}(x_{ip}-\bar{x}_i)(x_{jp}-\bar{x}_j)\right]} =10.4\pm 1.280$$

样本相关系数 r_{yi}：$r_{12}=\frac{s_{12}}{\sqrt{s_{11}s_{22}}}=\frac{-16}{\sqrt{6.4(10)}}=-0.2$

$$r_{y1}=\frac{s_{1y}}{\sqrt{s_{11}s_{yy}}}=\frac{-1.9}{\sqrt{6.4(16.0)}}=-0.188$$

$$r_{y2}=\frac{s_{2y}}{\sqrt{s_{22}s_{yy}}}=\frac{12.6}{\sqrt{10(16.0)}}=0.996$$

样本偏相关系数(sample partial correlation coefficients) $r_{yi.\ j}$

$$r_{y1\cdot 2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}}=\frac{-0.188-(0.996)(-0.200)}{\sqrt{(1-0.992)(1-0.04)}}=0.128$$

$$r_{y2\cdot 1}=\frac{r_{y2}-r_{y1}r_{12}}{\sqrt{(1-r_{y1}^2)(1-r_{12}^2)}}=\frac{0.996-(-0.188)(-0.200)}{\sqrt{(1-0.035)(1-0.04)}}=0.996$$

17.7 选择适当多元回归模式

选择适当多元回归模式，有两种方法：一是后退消除法(Backward elimination procedure)，另一是逐步回归法(Stepwise Regression Procedure)。

(一). 后退消除法(Backward elimination procedure)是将所有自变量放入多元回归模式，然后删除不适当的无关自变量，一直到找出最适当的多元回归模式。

1. 将所有自变量 x_i 放入多元回归模式，检定是否所有 β_i 全部等于 0。

   若是，则停止，所有自变量 x_i 都与因变量 y 无关。

   若否，则到下一步。

2. 计算 $F_i=\frac{b_i^2}{MS_E p_{ii}}$

3. 选取最小的 F_i，令 $F_j=\min_i\{F_i\}$

   若 $F_j \geqq F_{\alpha,1,n-k-1}$，保留 x_j，结束。

   若 $F_j < F_{\alpha,1,n-k-1}$，删除 x_j，计算新的

$$b_i^*=b_i-\frac{p_{ij}b_j}{p_{jj}}$$

$$p_{ii}^{*} = p_{ii} - \frac{p_{ij}^{2}}{p_{jj}}$$

$$p_{il}^{*} = p_{il} - \frac{p_{ij}p_{lj}}{p_{jj}}$$

4. 令 $b_i = b_i^{*}$，$p_{ii} = p_{ii}^{*}$，$p_{ij} = p_{ij}^{*}$。回到第 2 步。

(二). 逐步回归法(Stepwise Regression Procedure)是一种前进选择法(Forward selection procedure)是先将一个自变量放入多元回归模式，然后再加入适当的有关自变量，一直到找出最适当的多元回归模式。

1. 计算所有自变量与因变量的样本相关系数 $r_{y1}, r_{y2}, \ldots, r_{yk}$。

选取最大的 r_{yi}，令 $r_{yj} = \max_i \{r_{yi}\}$

计算回归方程式： $\hat{y} = b_0 + b_j x_j$

检定 $H_0: \beta_j = 0$

若接受 H_0，则停止，所有自变量 x_i 都与因变量 y 无关。

若否，则到下一步。

2. 计算 $r_{yi.\ j}$, $i \neq j$

选取最大的 $r_{yi.\ j}$，令 $r_{yl\bullet j} = \max_i \{r_{yi\bullet j}\}$

计算回归方程式：$\hat{y} = b_0 + b_j x_j + b_l x_l$

检定 $H_0: \beta_j = \beta_l = 0$

若接受 H_0，则停止，利用回归方程式：$\hat{y} = b_0 + b_j x_j$。

若否，则到下一步。

3. 分别检定$H_0^1: \beta_l = 0$，及$H_0^2: \beta_j = 0$

若接受H_0^1，则停止，利用回归方程式：$\hat{y} = b_0 + b_j x_j$。

若否定H_0^1，接受H_0^2，则删除 x_j，到第 4 步。

若否定H_0^1及H_0^2，到第 4 步 。

4. 令 R 代表目前多元回归模式中的自变量。

对所有 $i \notin R$，计算 $r_{yi.\ R}$。

选取最大的 $r_{yi.\ R}$，令 $r_{yp\bullet R} = \max_i \{r_{yi\bullet R}\}$

R＝R+{p}

计算回归方程式：$\hat{y} = b_0 + \sum_{i \in R} b_i x_i$

检定 H_0：所有的 $i \in R$ $\beta_i = 0$。

若接受 H_0，则停止。令 R＝R－{p}，利用回归方程式：$\hat{y} = b_0 + \sum_{i \in R} b_i x_i$

若否，则到下一步。

5. 对所有的 $i \in R$，分别检定$H_0^i: \beta_i = 0$

若有接受 H_0^j，则删除 x_j，R＝R－{j}，计算回归方程式：

$\hat{y} = b_0 + \sum_{i \in R} b_i x_i$，回到第 4 步。

若否定所有的 H_0^i，回到第 4 步。

例题 17.9： 下列 三个自变量与因变量，进行后退消除法：

样本	x_1	x_2	x_3	y
1	8.4	8.0	1.0	35
2	2.0	6.5	8.5	10
3	3.5	6.2	6.5	9
4	10.4	5.0	1.5	30
5	6.5	6.5	7.5	20
6	6.2	7.3	4.5	23
7	12.4	6.4	4.0	28
8	7.0	6.0	10.0	8
9	5.8	6.1	3.0	29
10	3.0	5.4	11.0	4
11	6.0	7.3	4.5	18
12	5.5	6.6	5.5	14
13	9.0	6.5	2.5	32
14	1.1	5.8	7.0	6

S_{11}＝131.36　S_{12}＝3.81　S_{22}＝7.86　S_{33}＝122.50

S_{13}＝－79.55　S_{23}＝－10.85　S_{1y}＝343.90　S_{2y}＝38.60

S_{3y}＝－383.00　S_{yy}＝1466.00

$$\left(\begin{array}{ccc|c|ccc} S_{11} & S_{12} & S_{13} & S_{1y} & 1 & 0 & 0 \\ S_{12} & S_{22} & S_{23} & S_{2y} & 0 & 1 & 0 \\ S_{13} & S_{23} & S_{33} & S_{3y} & 0 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} 131.36 & 3.81 & -79.55 & 343.90 & 1 & 0 & 0 \\ 3.81 & 7.86 & -10.85 & 38.60 & 0 & 1 & 0 \\ -79.55 & -10.85 & 122.50 & -383.00 & 0 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} 1 & 0.029004 & -0.605588 & 2.617996 & 0.007613 & 0 & 0 \\ 0 & 7.749494 & -8.542711 & 28.625484 & -0.029004 & 1 & 0 \\ 0 & -8.542711 & 74.325500 & -174.738391 & 0.605588 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} 1 & 0 & -0.573615 & 2.510859 & 0.007721 & -0.003743 & 0 \\ 0 & 1 & -1.102357 & 3.693846 & -0.003743 & 0.129041 & 0 \\ 0 & 0 & 64.908379 & -143.182936 & 0.573614 & 1.102357 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|c|ccc} 1 & 0 & 0 & 1.245509 & 0.012790 & 0.005999 & 0.008837 \\ 0 & 1 & 0 & 1.262129 & 0.005999 & 0.147762 & 0.016983 \\ 0 & 0 & 1 & -2.205914 & 0.008837 & 0.016983 & 0.015406 \end{array}\right)$$

所以：$b_1=1.246, b_2=1.262, b_3=-2.206$

检定 $H_0:\beta_1=\beta_2=\beta_3=0$

$SS_R=b_1S_{1y}+b_2S_{2y}+b_3S_{3y}=1.246(343.90)+1.262(38.60)-2.206(-383.00)=1322.1106$

方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio	$F_{0.05,3,10}$	R^2
回归模式 Regression	3	$SS_R=1322.11$	$MS_R=440.7$	$F=30.63$	3.708	0.902
误差 Error	10	$SS_E=143.89$	$MS_E=14.39$			
总和 Total	n-1=13	$SS_T=S_{yy}=1466$				

所以否定 $H_0:\beta_1=\beta_2=\beta_3=0$

假设	$t=\frac{b_i}{\sqrt{p_{ii}s^2}}$	$F=\frac{b_i^2}{(p_{ii}s^2)}$
$H_0^1:\beta_1=0$	$\frac{1.246}{\sqrt{(0.012790)14.39}}=2.904$	8.435
$H_0^2:\beta_2=0$	$\frac{1.262}{\sqrt{(0.147762)14.39}}=0.866$	0.749
$H_0^3:\beta_3=0$	$\frac{-2.206}{\sqrt{(0.015406)14.39}}=-4.684$	21.951

因为 $t_{0.05,10}=2.228$, $F_{0.05,1,10}=4.965$，所以 X_2 要退出回归方程式。

新的回归方程式：$y_i=\beta_0^*+\beta_1^*x_1+\beta_3^*x_3$。

$$b_1^*=b_1-\frac{p_{12}b_2}{p_{22}}=1.246-\frac{(0.005999)(1.262)}{0.147762}=1.195$$

$$b_3^*=b_3-\frac{p_{32}b_2}{p_{22}}=-2.206-\frac{(0.016983)(1.262)}{0.147762}=-2.351$$

$$p_{11}^*=p_{11}-\frac{p_{12}^2}{p_{22}}=0.012790-\frac{(0.005999)^2}{0.147762}=0.012546$$

$$p_{33}^*=p_{33}-\frac{p_{32}^2}{p_{22}}=0.015406-\frac{(0.016983)^2}{0.147762}=0.013454$$

$$p_{13}^*=p_{13}-\frac{p_{12}p_{32}}{p_{22}}=0.008837-\frac{(0.005999)(0.016983)}{0.147762}=0.008148$$

$$SS_R^*=b_1^*S_{1y}+b_3^*S_{3y}=1.195(343.9)-2.351(-383.0)=1311.39$$

方差分析表如下：

变数来源 Source	自由度 df	平方和 SS	平均平方和 MS	F 比值 F-ratio	$F_{0.05,2,11}$	R^2

回归模式 Regression	2	$SS_R=1311.39$	$MS_R=655.70$	$F=46.64$	3.982	0.894
误差 Error	11	$SS_E=154.61$	$MS_E=14.06$			
总和 Total	13	$SS_T=S_{yy}=1466$				

所以否定$H_0:\beta_1^*=\beta_3^*=0$

假设	$t=\frac{b_i}{\sqrt{p_{ii}s^2}}$	$F=\frac{b_i^2}{(p_{ii}s^2)}$
$H_0^1:\beta_1^*=0$	$\frac{1.195}{\sqrt{(0.012546)14.06}}=2.845$	8.10
$H_0^3:\beta_3^*=0$	$\frac{-2.351}{\sqrt{(0.013454)14.06}}=-5.405$	29.217

所以 x_1 与 x_3 都可以在回归模式，但是 X_3 的 F 值最大，所以 X_3 是影响因变量 Y 最大的自变量。如果回归模式只有 X_3，则 R^2 是：

$$R^2=S_{3y}^2/S_{33}S_{yy}=0.817$$

由上述 R^2，如果回归模式只有 X_3，则 X_3 解释回归的变异程度有 81.7%。如果回归模式已经有 X_3 再加入 X_1，则 X_1 解释回归的变异程度有(89.4－81.7)%＝7.7%。如果回归模式已经有 X_1 与 X_3 再加入 X_2，则 X_2 解释回归的变异程度只有(90.2－89.4) %＝0.8%。

17.8 中文统计应用

一 多元回归

中文统计应用在多元回归，我们以例题 15.1 为范例，说明如下：

1. 开启中文统计功能列表，再开启已存入资料的旧文件如 chap15-1.xls，或输入新数据。
2. 选择「回归与相关分析」之下的「回归(简单回归与多元回归」，出现选择窗体，在(1) 输入变量 Y 范围：输入 A1:A15，或用鼠标选取，(2) 输入变量 X 范围：输入 B1:D15，(3) 标计：点选，(4) 选择输出范围。
3. 得到「多元回归」的结果。

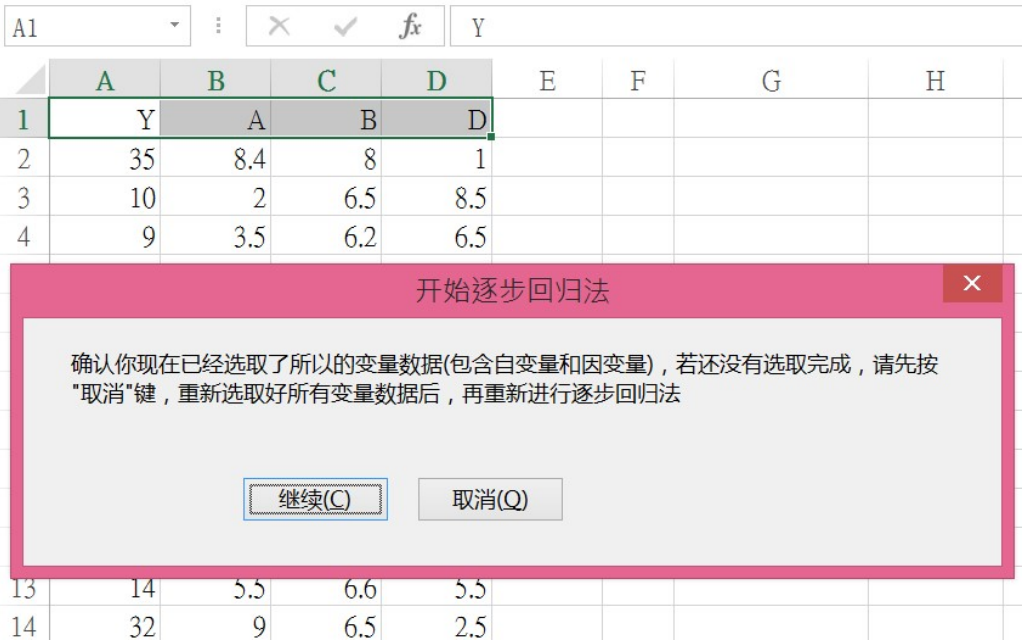
	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	回归统计						
4	Multiple R	0.949588					
5	R Square	0.901717					
6	Adjusted R	0.872232					
7	标准误差	3.79582					
8	观测值	14					
9							
10	方差分析						
11		df	SS	MS	F	gnificance F	
12	回归分析	3	1321.918	440.6392	30.58242	2.38E-05	
13	残差	10	144.0825	14.40825			
14	总计	13	1466				
15							
16		Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	15.3328	11.61043	1.320605	0.216057	-10.5369	41.20245
18	A	1.245509	0.429287	2.901341	0.015797	0.288997	2.202021
19	B	1.26213	1.459108	0.865001	0.40731	-1.98897	4.513225
20	D	-2.20592	0.471146	-4.68204	0.000865	-3.2557	-1.15615

二　逐步回归

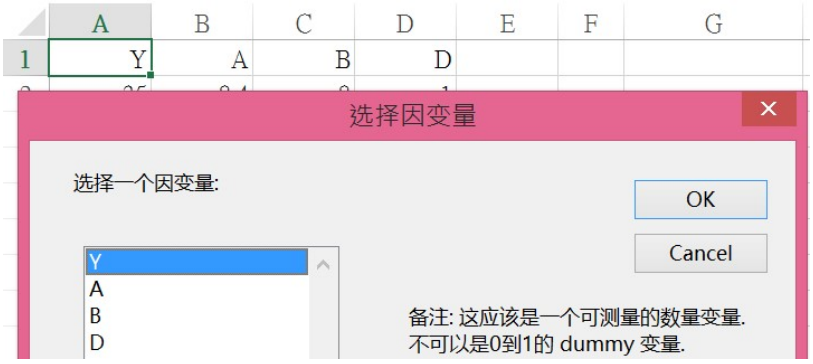
中文统计应用在多元回归，我们以例题 15.1 为范例，说明如下：

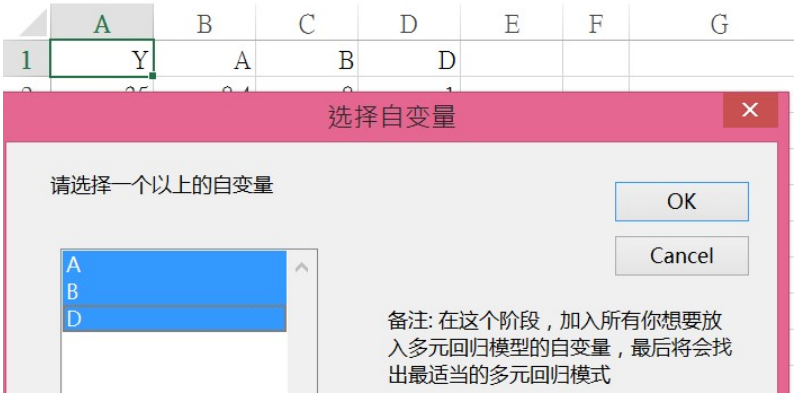
1. 开启中文统计功能列表，再开启已存入资料的旧文件如 chap15-1.xls，或输入新数据。

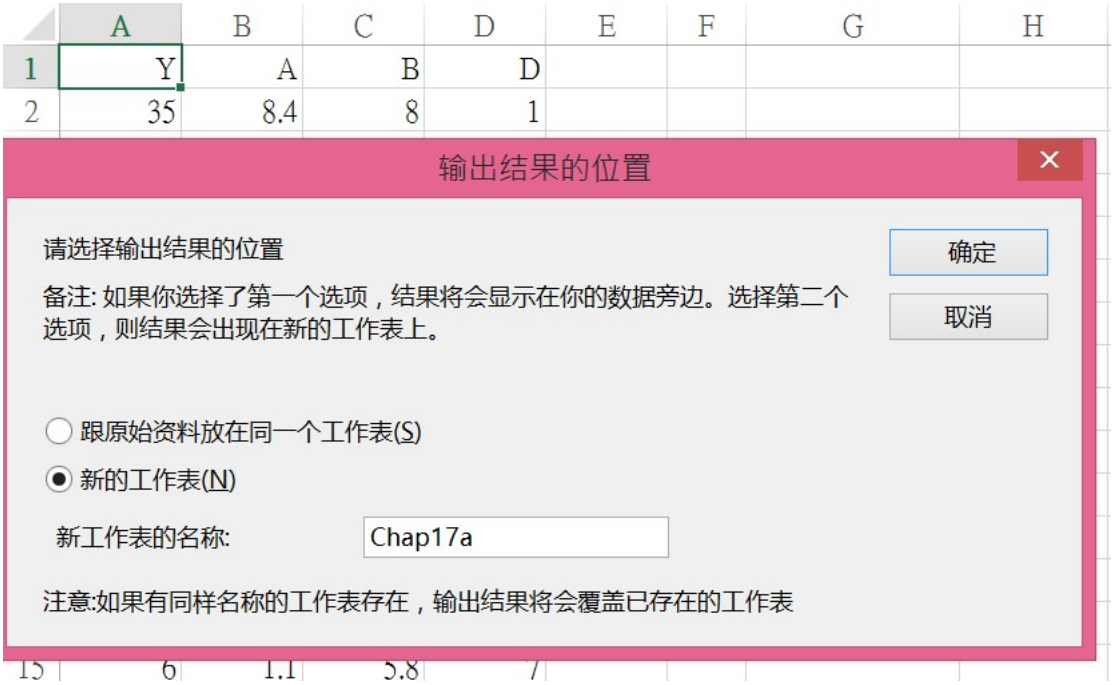
2. 标记变量名称 1. 不能用底线(_)开始，2. 不能用数字 (例如 20)，3. 不能用 Excel 表格名称(例如 X1 或 A6)，4. 可以用一个英文字母(例如 A 或 Y)，但不能用 R 或 C，5. 不要用符号，6. 可以用空格，但会转成底线，例如 X 1 会转成 X_12.
3. 选取因变量和自变量，将这些变量「标示」起来。



4. 选择「回归与相关分析」之下的「逐步回归法」，出现选择窗体，「开始逐步回归法」，按「继续」。
5. 在「选择因变量」，标示「Y」，按「OK」。
6. 在「选择自变量」，标示「A,B,D」，按「OK」。





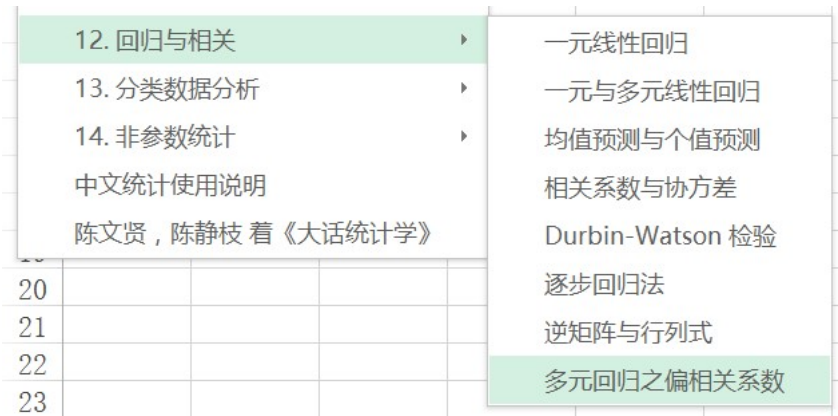


7. 得到「逐步回归」的结果。

1	逐步回归法的结果					
2						
3	步骤 1 - Entering 变量: D					
4						
5	结论					
6		R的倍数	0.9038			
7		决定系数R平方	0.8168			
8		调整后的R平方	0.8016			
9		标准误差	4.7306			
10						
11	ANOVA Table					
12		来源	自由度	平方和	平均平方和	F值	p值
13		可解释的	1	1197.4612	1197.4612	53.5101	0.0000
14		不可解释的	12	268.5388	22.3782		
15						
16	回归模型的系数					
17			系数	标准差	t值	p值	
18		常数	36.1959	2.6692	13.5607	0.0000	
19		D	-3.1265	0.4274	-7.3151	0.0000	

21	步骤 2 - Entering 变量: A					
22						
23	结论			Change	% Change		
24		R的倍数	0.9457	0.0419	%4.6		
25		决定系数R平方	0.8944	0.0775	%9.5		
26		调整后的R平方	0.8752	0.0736	%9.2		
27		标准误差	3.7521	-0.9784	-%20.7		
28							
29	ANOVA Table						
30		来源	自由度	平方和	平均平方和	F值	p值
31		可解释的	2	1311.1369	655.5684	46.5653	0.0000
32		不可解释的	11	154.8631	14.0785		
33							
34	回归模型的系数						
35			系数	标准差	t值	p值	
36		常数	24.5260	4.6205	5.3081	0.0002	
37		D	-2.3510	0.4352	-5.4018	0.0002	
38		A	1.1943	0.4203	2.8416	0.0160	

二 多元回归之



偏相关系数

	A	B	C	D	E	F	G	H
1	多元回归之偏相关系数							
2	Y	已加入的X		欲加入的X				
3	Y	X1	X2	X3				
4	2.6	5.1	7.8	114				
5	2.6	4.9	8	116			数据的数目	67
6	2.7	5.1	8.1	117			独立变量的数目 (≦6)	3
7	3	5.1	8.1	122			已经在回归模式的变量数目	2
8	2.9	5.1	8.1	124			欲加入回归模式的变量数目	1
9	3.1	5.2	8.1	128			检定的 α =	0.05
10	3.2	5.1	8.3	132				
11	3.7	5.2	8.8	133			SSE(已有变量)	8.7476
12	3.6	5.3	8.9	133			SSE(全部变量)	6.9956
13	3.4	5.4	9.1	134				
14	3.7	5.7	9.2	135			偏判定系数	0.2003
15	3.6	5.7	9.5	136			偏相关系数	0.4475
16	4.1	5.9	10.3	140				
17	3.5	5.8	10.6	147			检定是否加入回归模式 F =	15.7772
18	4.2	5.7	11.3	150			F 临界值	3.9934
19	4.3	5.8	12.1	151			检定 p 值 =	0.0002
20	4.2	6	12	151				

习题

1. 中古车的「估价」、「车龄」、与「里程数」，资料如下：

车龄(年)	X_1	1	2	2	3	3	4	6	7	8	10
里程数(千哩)	X_2	8.1	17.0	12.6	18.4	19.5	29.2	40.4	51.6	62.6	80.1
估价(千元)	Y	5.45	4.80	5.00	4.00	3.70	3.20	3.15	2.69	1.90	1.47

(1). 求以「车龄」与「里程数」来预测「价格」的方程式

(2). 求 β_1 与 β_2 之 95%信赖区间

(3). 求 r^2，并以 $\alpha=0.5$，检定回归模式的适合性

(4). 计算 $r_{YX_2 \cdot X_1}$，以 $\alpha=0.5$，如果已有 X_1 对 Y 回归，自变量 X_2 是否值得再加入回归

(5). 以 $\alpha=0.5$，检定 $H_0: \beta_1 = \beta_2 = 0$

2. 班级的 20 位学生。调查每位学生两次期中考(x_1, x_2)，GPA(x_3)，以及期末考(y)的成绩。得知下列的资料：

x_1	x_2	x_3	y	x_1	x_2	x_3	y
87	25	2.9	60	93	60	3.2	44
100	84	3.3	80	92	69	3.1	53
91	52	3.5	73	100	86	3.6	86
85	60	3.7	83	80	67	3.5	59
56	76	2.8	33	100	96	3.8	81
81	28	3.1	65	69	51	2.8	20
85	67	3.1	53	80	75	3.6	64
96	83	3.0	68	74	70	3.1	38
79	60	3.7	88	79	66	2.9	77
96	69	3.7	89	95	83	3.3	47

(1). 以 x_1, x_2, x_3 对 y 作直线回归模式，并计算 r^2

(2). 以 $\alpha = 0.5$，检定回归模式的适合性

(3). 不考虑 GPA，以两次期中考成绩，对 y 作回归模式，并计算 r^2

(4). 以 GPA 对 y 作回归模式，并计算 r^2

(5). 计算 $r_{YX_3 \cdot X_1 X_2}$，以 $\alpha=0.5$，如果已有 X_1,X_2 对 Y 回归，自变量 X_3 是否值得再加入回归

(6). 以 $\alpha = 0.5$，检定 $H_0: \beta_1 = \beta_2 = 0$

3.下列四个变量数据：

Y	X1	X2	X3
10.8	12.9	10.3	9.5
12.1	14.3	11.4	12.2
12.9	12.6	8.3	8.4
9.5	12.9	12.3	10.8
12.4	14.8	10.3	9.4
9.5	13.7	9.4	8.7
11.2	14.2	11.6	11.3
7.2	12.3	8.7	9.0
11.7	13.0	10.4	9.1
11.0	13.3	9.0	8.0
6.8	11.1	10.6	9.4
9.1	12.8	10.1	9.2
11.2	12.8	9.6	10.4
9.8	13.8	11.1	9.9
9.0	13.2	10.0	10.2

(1). 以 X_1, X_2, X_3 对 Y 作全模式回归，并计算 r^2

(2). 如果 $X_1=13$, $X_2=10$, $X_3=10$，则计算 Y 预测值的 95%的信赖区间，及实际值的 95%的预测区间

(3). 不考虑 X_2 与 X_3，以 X_1 的数据，对 Y 作回归模式，并计算 r^2

(4). 不考虑 X_1，以 X_2 与 X_3 的数据，对 Y 作回归模式，并计算 r^2

(5). 以 $\alpha=0.5$，检定 $H_0:\beta_1=\beta_2=\beta_3=0$

(6). 以 $\alpha=0.5$，检定 $H_0:\beta_1=\beta_2=0$

(7). 计算 $r_{YX_3\cdot X_1X_2}$，以 $\alpha=0.5$，如果已有 X_1 与 X_2 对 Y 回归，自变量 X_3 是否值得再加入回归

4. 有关美国国会大选执政党席位，与该年总统声望，及实质所得变化，可能有相关：

Y＝以前 8 次选举平均席位，执政党国会席位的「损失率」

X_1＝当年执政党总统的「声望」

X_2＝与前一年比较，「实质所得」的变化

年	Y	X_1	X_2
1946	7.3%	32%	－$40
1950	2.0	43	100
1954	2.3	65	－10
1958	5.9	56	－10
1962	－0.8	67	60
1966	1.7	48	100

(1). 以 X_1, X_2 对 Y 作回归模式，并计算 r^2

(2). 不考虑 X_1，以 X_2 的数据，对 Y 作回归模式，并计算 r^2

(3). 不考虑 X_2，以 X_1 的数据，对 Y 作回归模式，并计算 r^2

(4). 如果 $X_1=60\%, X_2=\$50$，则计算 Y 预测值的 95%的信赖区间，及实际值的 95%的预测区间

(5). 在 1970 年 $X_1=56\%, X_2=\$70, Y=1.0\%$，其实际值是否在 95%的预测区间

(6). 计算 $r_{YX_2\cdot X_1}$，以 $\alpha=0.5$，如果已有 X_1 对 Y 回归，自变量 X_2 是否值得再加入回归

(7). 以 $\alpha=0.5$，检定 $H_0:\beta_1=\beta_2=0$

5. 下列三个变量数据：

Y	X_1	X_2
74.2	3.0	8.5
68.1	2.7	7.1
50.2	2.6	5.9
68.3	3.2	8.6
71.6	2.4	13.3
68.9	3.2	7.4
59.6	3.1	6.8
39.8	2.4	3.8
86.6	3.5	9.1
71.2	2.8	12.6

(1). 以 X_1, X_2 对 Y 作全模式回归，并计算 r^2

(2). 如果 $X_1=3.0, X_2=9.0$，则计算 Y 预测值的 95%的信赖区间，及实际值的 95%的预测区间

(3). 不考虑 X_2，以 X_1 的数据，对 Y 作回归模式，并计算 r^2

(4). 以 $\alpha=0.5$，检定 $H_0:\beta_1=\beta_2=0$

(5). 计算 $r_{YX_2\cdot X_1}$，以 $\alpha=0.5$，如果已有 X_1 对 Y 回归，自变量 X_2 是否值得再加入回归

6. 某位学生利用最小平方法，进行实质国民生产毛额 Y 对时间的直线回归，得到模式：Y＝ 264.3＋18.77t。最近 26 年，t＝1, 2 ,...,26 之 y 值列于下面。检查他的计算是否有错误。就数据如下，关于线性多元回归模式中的假设，那一项不合？(注意：回归方法通常不适用于此类型的数据，应采时间数列分析)。

年(i)	1	2	3	4	5	6	7	8	9
y_i	309.9	323.7	324.1	355.3	383.4	395.1	412.8	407	438
$\hat{y}$	283.1	301.9	320.6	339.4	358.2	376.9	395.7	414.5	433.2
残差	26.8	21.8	3.5	15.9	25.2	18.2	17.1	-7.5	4.8

年(i)	10	11	12	13	14	15	16	17	18
y_i	446.1	452.5	447.3	475.9	487.7	497.2	529.8	551	581.1
$\hat{y}$	452.0	470.8	489.5	508.3	527.1	545.8	564.6	583.4	602.1
残差	-5.9	-18.3	-42.2	-32.4	-39.4	-48.6	-34.8	-32.4	-21

7. 下为残差与预测值的数据(括号内的数字表时间) ：

预测值	2.2(9)	3.1(6)	2.5(1)	2.3(7)	3.6(7)	3.6(14)	2.6(8)
残差	-1	-2	3	-3	-1	5	0
预测值	2.5(3)	3.0(12)	3.2(4)	2.9(11)	3.3(2)	2.7(10)	3.2(5)
残差	0	3	-2	2	-5	0	1

(1). 绘出残差与预测值以及残差与时序的图形。

(2). 判断是否违背基本假设？

8. 某保险精算师想发展一模式估计保险人的寿命，与一些从业医师谈过后，他认为保险人的寿命(y)应该与每周运动时数(x_1)、胆固醇(x_2)、超过标准血压点数(x_3)，从过去死亡的保险人中随机抽样 40 位，得到以下保险人寿命的回归模式:

	系数	标准误	t 统计
截距	55.8	11.8	4.729
x_1	1.79	0.44	4.068
x_2	-0.021	0.011	-1.909
x_3	-0.016	0.014	-1.143

标准误 ＝9.47　　R 平方 ＝22.5%

ANOVA

Source of Variation	自由度	*SS*	*MS*	*F*
回归	3	936	312	3.477
残差	36	3230	89.722	
总和	39	4166		

请根据这个保险人寿命回归模式回答以下的问题。

(1) 请用 5%的显着水平检定上述保险人寿命回归模式的有效性。

(2) 请问以 1%显着水平是否有足够的证据显示每周运动时数(x_1)与保险人的寿命(y)长短有线性关系?

(3) 请问以 5%显着水平是否有足够的证据显示胆固醇(x_2)与保险人的寿命(y)长短有线性关系?

(4) 请问以 5%显着水平是否有足够的证据显示超过标准血压点数(x_3)与保险人的寿命(y)长短有线性关系?

(5) 请计算与解释上述简保险人寿命回归模式的决策系数与调整后决策系数。

(6) 请解释上述简保险人寿命回归模式中每周运动时数(x_1)、胆固醇(x_2)、与超过标准血压点数(x_3)的 3 个系数。

(7) 请估计一般每周运动时数为 10 小时、胆固醇为 200、超过标准血压点数为 40 点之保险人寿命的点估计与 95%信赖区间。

(8) 请估计某位每周运动时数为 10 小时、胆固醇为 200、超过标准血压点数为 40 点之保险人寿命的点估计与 95%信赖区间。

9. 某小区大学的教师想知道会影响学习成绩的因素有哪些?他认为最后的学习成绩(以 100 分计)(y)应该与逃学时数(x_1)、迟交作业次数(x_2)、期中考分数(以 100 分计)(x_3)，从过去修课学生中随机抽样 50 位，得到以下最后学习成绩的回归模式:

	系数	标准误	t 统计
截距	41.6	17.8	2.337
x_1	-3.18	1.66	-1.916
x_2	-1.17	1.13	-1.035
x_3	0.63	0.13	4.846

标准误 = 13.74　　R 平方= 30.0%

ANOVA

Source of Variation	自由度	*SS*	*MS*	*F*
回归	3	3716	1238.667	6.558
残差	46	8688	188.870	
总和	49	12404		

请根据这个最后学习成绩的回归模式回答以下的问题。

(1) 请用 5%的显着水平检定上述最后学习成绩回归模式的有效性。

(2) 请问以 5%显着水平是否有足够的证据显示逃学时数(x_1)与最后学习成绩(y)有线性关系?

(3) 请问以 5%显着水平是否有足够的证据显示迟交作业次数(x_2)与最后学习成绩(y)有线性关系?

(4) 请问以 5%显着水平是否有足够的证据显示超过期中考分数(x_3)与最后学习成绩(y)有线性关系?

(5) 请计算与解释上述最后学习成绩回归模式的决策系数与调整后决策系数。

(6) 解释上述最后学习成绩回归模式中逃学时数(x_1)、迟交作业次数(x_2)、期中考分数(以 100 分计)(x_3)的 3 个系数。

(7) 请估计一般逃学时数为 3 小时、迟交作业次数为 2、期中考分数为 75 分之最后学习成绩的点估计与 95%信赖区间。

(8) 请估计某位逃学时数为 3 小时、迟交作业次数为 2、期中考分数为 75 分之最后学习成绩的点估计与 95%信赖区间。

10. 某建筑公司刚买到一块地想用来盖住宅，而影响购屋人购买房屋大小的因素应该与每月家庭收入、家庭人数、家庭主要收入者的教育程度，购买房屋大小以 10 平方公尺计、家庭收入以千元计与家庭主要收入者的教育程度以受教育的年数计，从过去建筑公司所成交的买者中随机抽样 50 位，得到以下购买房屋大小的回归模式:

回归摘要输出

R 的倍数	0.865
R 平方	0.748
调整的 R 平方	0.726
标准误	5.195
观察值个数	50

ANOVA

	自由度	*SS*	*MS*	*F*	显着值
回归	4	3605.7736	901.4434	33.4081	0.0001
残差	45	1214.2264	26.9828		
总和	49	4820.0000			

	系数	标准误	t 统计	P-值
截距	-1.6335	5.8078	-0.281	0.7798
家庭收入	0.4485	0.1137	3.9545	0.0003
家庭人数	4.2615	0.8062	5.286	0.0001
教育程度	0.6517	0.4319	-1.509	0.1383

请根据这个购买房屋大小的回归模式回答以下的问题。

(1) 请用 5%的显着水平检定上述购买房屋大小回归模式的有效性。

(2) 请问以 5%显着水平是否有足够的证据显示每月家庭收入与购买房屋大小有线性关系?

(3) 请问以 5%显着水平是否有足够的证据显示迟交家庭人数与购买房屋大小有线性关系?

(4) 请问以 5%显着水平是否有足够的证据显示超过家庭主要收入者教育程度与购买房屋大小有线性关系?

(5) 请解释上述购买房屋大小回归模式的决策系数与调整后决策系数。

(6) 请解释上述购买房屋大小回归模式中每月家庭收入、家庭人数、家庭主要收入者教育程度的 3 个系数。

(7) 请估计一般家庭收入为 NT$60000、家庭人数为 4 人、家庭主要收入者教育程度为 16 年之购买房屋大小的点估计与 95%信赖区间。

(8) 请估计某位一般家庭收入为 NT$60000、家庭人数为 4 人、家庭主要收入者教育程度为 16 年之购买房屋大小的点估计与 95%信赖区间。

11. 某汽车经销商想知道训练一位销售员需要多少时间，x 代表销售员从开始到可以完全自主决策的训练时间(以天计)、y 代表销售员每周的销售金额(以千计)，从过去经销商所训练销售员中随机抽样 10 位，得到以下训练时间与销售金额的信息:

X	10	14	16	20	25	30	35	40	43	50
Y	12	20	23	27	36	45	40	28	26	25

请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释训练时间与销售金额是否有线性关系?

(2) 请找出上述问题的简单线性模式。

(3) 请计算与解释请上述简单线性模式的决策系数与相关系数。

(4) 依据散布图、决策系数与相关系数的结果，请问是否一致?

(5) 请找出上述问题的二次抛物线模式($y = \beta_0 + \beta_1 x + \beta_2 x^2$)。

(6) 请用 5%的显着水平检定上述二次抛物线模式的有效性。

(7) 请问以 5%显着水平是否有足够的证据显示训练时间与销售金额有线性关系?

(8) 请问以 5%显着水平是否有足够的证据显示训练时间的二次与销售金额大小有线性关系?

(9) 请解释上述二次抛物线模式的决策系数与调整后决策系数。

(10) 请解释上述二次抛物线模式中训练时间、训练时间二次的 2 个系数。

(11) 请用二次抛物线模式估计一般训练时间为 15 天之销售员销售金额的点估计与 95%信赖区间。

(12) 请用二次抛物线模式估计某位训练时间为 15 天之销售员销售金额的点估计与 95%信赖区间。

12. 某财务公司的分析师想发展一模式来估计国际金价的涨跌，她认为国际金价(y)应该与每桶原油价格(x_1)与银行利率(x_2)有关，而且她认为每桶原油价格与银行利率的互动(x_1x_2)也会影响国际金价，从过去半年的数据中随机抽样 20 天，得到以下国际金价的回归模式:

	系数	标准误	t 统计
截距	115.6	78.1	1.480
x_1	22.3	7.1	3.141
x_2	14.7	6.3	2.333
x_1x_2	-1.36	0.52	-2.615

标准误 = 20.9　　R 平方 = 55.4%

ANOVA

	自由度	SS	MS	F
回归	3	8661	2887.0	6.626
残差	16	6971	435.7	
总和	19	15632		

请根据这个国际金价回归模式回答以下的问题。

(1) 请用 5%的显着水平检定上述国际金价回归模式的有效性。

(2) 请问以 5%显着水平是否有足够的证据显示每桶原油价格(x_1)与国际金价(y)有线性关系?

(3) 请问以 5%显着水平是否有足够的证据显示银行利率(x_2)与国际金价(y)有线性关系?

(4) 请问以 5%显着水平是否有足够的证据显示每桶原油价格与银行利率的互动(x_1x_2)与国际金价(y)有线性关系?

(5) 请计算与解释上述国际金价回归模式的决策系数与调整后决策系数。

(6) 请解释上述国际金价回归模式中每桶原油价格(x_1)、银行利率(x_2)、与每桶原油价格与银行利率的互动(x_1x_2)的 3 个系数。

(7) 请估计一般每桶原油价格为 25 美元、银行利率为 6%之国际金价的点估计与 95%信赖区间。

(8) 请估计某天每桶原油价格为 25 美元、银行利率为 6%之国际金价的点估计与 95%信赖区间。

13. 某人力网站的分析师想发展一模式来估计不同专业人士(医师、牙医、律师)的收入水平，她认为专业人士的收入水平(y)应该与经验年资以年计(x_1)有关，而且她设定二元变量(x_2)为 1 代表是医师 0 代表非医师、与二元变量(x_3) 为 1 代表是牙医 0 代表非牙医，从过去半年的资料中随机抽样 125 名，得到以下专业人士收入水平的回归模式:

	系数	标准误	t 统计
截距	71.65	18.56	3.860
x_1	2.07	0.81	2.556
x_2	10.16	3.16	3.215
x_3	-7.44	2.85	-2.611

标准误 = 42.6　　R 平方 = 30.9%

ANOVA

	自由度	SS	MS	F
回归	3	98008	32669.333	18.008
残差	121	219508	1814.116	
总和	124	317516		

请根据这个专业人士收入水平回归模式回答以下的问题。

(1) 请用 5%的显着水平检定上述专业人士收入水平回归模式的有效性。

(2) 请问以 5%显着水平是否有足够的证据显示经验年资(x_1)与专业人士收入水平(y)有线性关系?

(3) 请问以 5%显着水平是否有足够的证据显示三种不同专业人士(医师、牙医、律师)的收入水平并不相同?

(4) 请计算与解释上述专业人士收入水平回归模式的决策系数与调整后决策系数。

(5) 请解释上述专业人士收入水平回归模式中经验年资(x_1)、三种不同专业人士(医师、牙医、律师)的 3 个系数。

(6) 请估计一般牙医且经验年资为 10 年专业人士收入水平的点估计与 95%信赖区间。

(7) 请估计某位牙医且经验年资为 10 年专业人士收入水平的点估计与 95%信赖区间。