第 11 章 回归与相关分析

第 231 页

$$\sum y_i = nb_0 + b_1\sum x_{1i} + b_2\sum x_{2i} + \cdots + b_k\sum x_{ki}$$

$$\sum x_{1i}y_i = b_0\sum x_{1i} + b_1\sum x_{1i}^2 + b_2\sum x_{1i}x_{2i} \cdots + b_k\sum x_{1i}x_{ki}$$

$$\sum x_{2i}y_i = b_0\sum x_{2i} + b_1\sum x_{1i}x_{2i} + b_2\sum x_{2i}^2 + \cdots + b_k\sum x_{2i}x_{ki}$$

第 238 页

11.6.4 回归系数 β_1 的~~区间估计与检验~~

11.6.4 回归系数 β_1 的方差的讨论

第 241 页

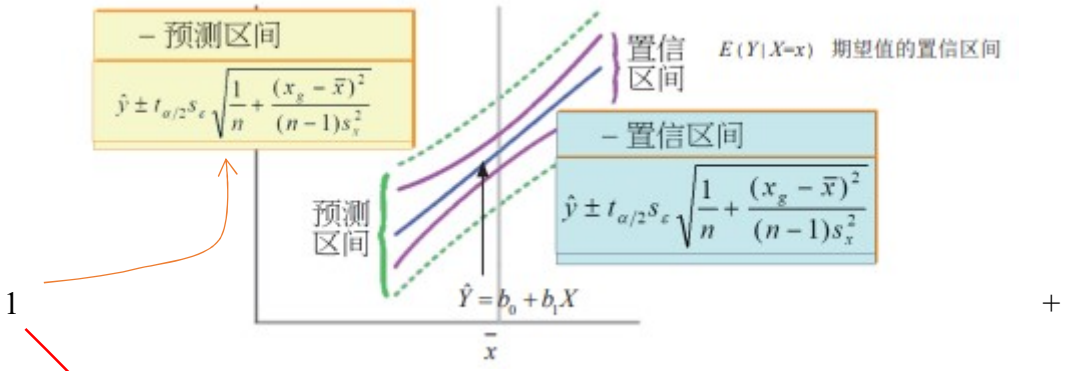

图 11.11 预测区间与期望值的置信区

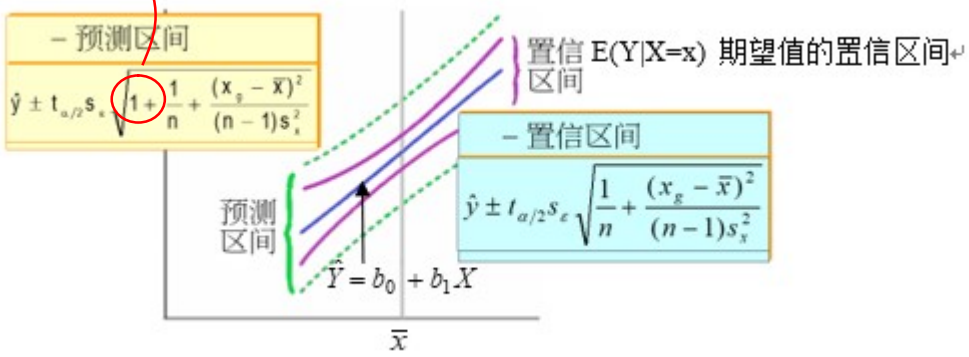

图 11.11 预测区间与期望值的置信区

11.3 简单线性回归分析参数的点估计

例题 11.1：美国汽车协会(AAA) 希望比较汽车引擎汽缸大小与每加仑汽油的哩程数的关系。自变量 X 是汽车引擎汽缸大小(立方英吋 cubic inches ,CI,1 CI = 16.387 cc)。因变量 Y 是每加仑汽油的哩程数(miles-per-gallon mpg，1 mpg = 0.425 每公升汽油的公里数)。现在抽样 8 辆不同厂牌的中小型车(compact car)，进行简单回归分析。

汽车 Model	引擎汽缸大小 x (CI)	每加仑汽油的哩程数 y (mpg)
Chevrolet Cavalier	121	30
Dutsun Nissan Stanza	120	31
Dodge Omni	97	34
Fort Escort	98	27
Mazda 626	122	29
Plymouth Horizon	97	34
Renault Alliance/Encore	85	38
Toyota Corolla	122	32

解答：先计算 S_{xy}, S_{xx}, S_{yy}，列表如下：

x	y	x^2	y^2	xy
121	30	14641	900	3630
120	31	14400	961	3720
97	34	9409	1156	3298
98	27	9604	729	2646
122	29	14884	841	3538
97	34	9409	1156	3298
85	38	72251	1444	3230
122	22	14884	1024	3904
862	255	94456	8211	27264

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 27264 - 8(\frac{862}{8})(\frac{225}{8}) = -212.25$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 94456 - 8(\frac{862}{8})^2 = 1575.5$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 8211 - 8(\frac{225}{8})^2 = 82.875$$

$$b_1 = \frac{S_{xy}}{S_{xx}} = -0.1347 \qquad b_0 = \bar{y} - b_1\bar{x} = 46.3889$$

$$\hat{y}_i = 463889 - 0.1347x_i$$

$$SS_E = S_{yy} - b_1 S_{xy} = 54.2849 \qquad MS_E = \frac{SS_E}{n-2} = 9.0475$$

MS_E 是 Y_i 标准差 σ 的点估计值，称作标准误。

请注意：点估计值 b_0，σ 的单位和 Y_i 的单位相同；b_1 的单位是 Y_i 的单位除以 X_i 的单位；相关系数 r 是没有单位。例如：Y_i 的单位从公斤改为公克(资料乘以 1000)，X_i 的单位从公尺改为公分(资料乘以 100)，则回归直线截距 b_0 和 σ 改为乘以 1000，回归直线斜率 b_1 改为乘以 10(=1000/100)；相关系数 r 没有改变，检验直线相关的 p 值也是不变。

11.4 相关分析

例题 11.2：例题 11.1，汽车引擎汽缸大小与每加仑汽油的哩程数。如果自变量 X 汽车引擎汽缸大小视作随机变量，则与因变量 Y 的样本相关系数：

$$r=\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}=\frac{-212.25}{\sqrt{(1575.5)(82.875}}=-0.5874 \qquad \text{判定系数} \quad r^2=0.3450$$

这个判定系数 $r^2=0.3450$ 并非很高。因为直线回归模型只能解释 35%的平方和 SS_T，而有 65%的平方和不能被直线回归模型解释。也许非直线回归模型如： $y=\beta_0+\beta_1 x+\beta_2 x^2+\varepsilon$ 解释 SS_T 的比率更高，换言之，非直线回归模型的判定系数可能更高。或者再加入其他变量，可以提高回归模型的解释能力。非直线回归模型与多变量的回归，请参考 CD 第 15 章复回归模型。

11.5 检验自变量与因变量是否相关

$$(Y_i-\bar{Y}) \quad = \quad (\hat{Y}_i-\bar{Y}) \quad + \quad (Y_i-\hat{Y}_i)$$

总差异 = 已解释差异 + 未解释差异(残差)

「总差异」是样本点和总平均的差异，「已解释差异」是回归直线和总平均的差异，「未解释差异」(残差)是样本点和回归直线的差异。

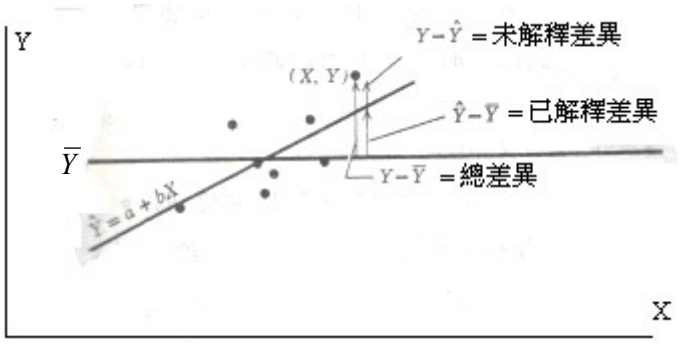

图 11.a 总差异、已解释差异、未解释差异

$$\sum(Y_i-\bar{Y})^2 \quad = \quad \sum(\hat{Y}_i-\bar{Y})^2 \quad + \quad \sum(Y_i-\hat{Y}_i)^2$$
$$SS_T \quad = \quad SS_R \quad + \quad SS_E$$

总变异 = 已解释变异 + 未解释变异(残差变异)

证明： $\sum(Y_i-\bar{Y})^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2$

$$\sum(Y_i-\bar{Y})^2=\sum[(Y_i-\hat{Y}_i)+(\hat{Y}_i-\bar{Y})]^2$$
$$=\sum(Y_i-\hat{Y}_i)^2+2\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})+\sum(\hat{Y}_i-\bar{Y})^2$$

$$\sum(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})=\sum(y_i-b_0-b_1x_i)(b_0+b_1x_i-\bar{y})$$
$$=b_0\sum(y_i-b_0-b_1x_i)+b_1\sum x_i(y_i-b_0-b_1x_i)-\bar{y}\sum(y_i-b_0-b_1x_i)=0$$

根据最小平方的正规方程式： $\sum(y_i - b_0 - b_1 x_i) = 0,\ \sum x_i(y_i - b_0 - b_1 x_i) = 0$

因此，简单回归的逻辑关系如下：

最小平方法 ➔ SS_E 最小 ➔ 常态方程式 ➔ 回归直线方程式 ➔ $SS_T = SS_R + SS_E$

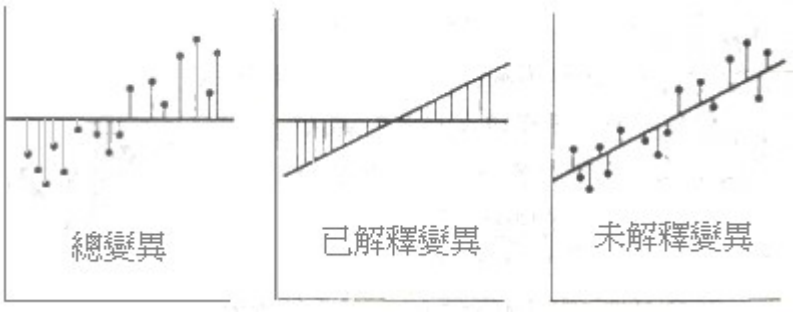

图 11.b 从方差分析检验回归变量的相关性

所以 $SS_T = SS_R + SS_E$ 并非都成立，只有在最小平方法的回归模式才成立。

当然，方差分析(ANOVA)也成立，因为方差分析是回归模式的一个特例。

11.6 方差的区间估计

例题 11.4：例题 11.1，汽车引擎汽缸大小与每加仑汽油的哩程数。

方差 σ^2 的 95% 置信区间：

$$\frac{SS_E}{\chi^2_{0.025,6}} = \frac{54.2849}{14.449} = 3.757 \le \sigma^2 \le \frac{SS_E}{\chi^2_{0.975,6}} = \frac{54.2849}{1.237} = 43.884$$

方差 σ^2 的点估计是 $MS_E = 9.0475$

方差 σ^2 的 95% 置信区间是 (3.757,43.884)

11.7 回归与相关分析的区间估计与检验

简单线性回归分析的数学模型： $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1,\ldots,n \quad \varepsilon \sim N(0,\sigma^2)$

从本节以后检验所需要的假设条件，除了回归分析的原来的四个假设以外，还要加上正态分布的假设，即：$\varepsilon_i \sim N(0,\sigma^2)$。

例题 11.5：例题 11.1，汽车引擎汽缸大小与每加仑汽油的哩程数。

β_0 的 95% 置信区间： $S_{b_0} = \sqrt{MS_E\left(\frac{\sum_i x_i^2}{nS_{xx}}\right)} = \sqrt{9.0475(\frac{94456}{8\times 1575.5})} = 8.234$

$$b_0 \pm t_{0.025,6} S_{b_0} = 46.3889 \pm 2.447(8.234) = 46.3889 \pm 20.149$$

β_1 的 95% 置信区间： $S_{b_1} = \sqrt{\frac{MS_E}{S_{xx}}} = \sqrt{\frac{9.0475}{1575.5}} = 0.076$

$$b_1 \pm t_{\frac{\alpha}{2},n-2} S_{b_1} = -0.1347 \pm 2.447(0.076) = -0.1347 \pm 0.1854$$

因为 β_1 的 95% 置信区间 $0 \in (-0.3201, 0.0507)$，所以我们接受 $H_0 : \beta_0 = 0$

11.7.3 相关系数的区间估计与检验

两个随机变量的样本相关系数 r，是总体相关系数 ρ 的估计量(或估计值，符号相同)。

r 是 ρ 的最大概似估计量，但是它不是 ρ 的无偏估计量。

1. 当 ρ 接近 +1，r 的分布是左偏分布。ρ 的置信区间，并非以 r 为中心左右对称，而是 r 的右边(大于的部份)较小，r 的左边(小于的部份)较大。请见图 11.10。

2. 当 ρ 接近 -1，r 的分布是右偏分布。ρ 的置信区间，并非以 r 为中心左右对称，而是 r 的右边(大于的部份)较大，r 的左边(小于的部份)较小。请见图 11.10。

3. 当 ρ 等于 0，r 的分布是 t 分布，自由度 n-2 。ρ 的置信区间，是以 r 为中心左右对称。

一. ρ 的 $1-\alpha$ 置信区间计算，要经过 Fisher 转换(Fisher's transformation)，其计算步骤如下：

(1) 计算样本相关系数 r

(2) 计算 Fisher 转换 $Z_r = \frac{1}{2}\log_e(\frac{1+r}{1-r})$ (查表 A.7)

(3) $l = Z_r - z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$ $\quad u = Z_r + z_{\frac{\alpha}{2}}\frac{1}{\sqrt{n-3}}$

(4) $\frac{e^{2l}-1}{e^{2l}+1} \le \rho \le \frac{e^{2u}-1}{e^{2u}+1}$

计算 e^{2l}，e^{2u} 可查表 A.15 $e^{-\lambda}$。注意：$Z_{-r} = -Z_r$ ， $e^{\lambda} = \frac{1}{e^{-\lambda}}$

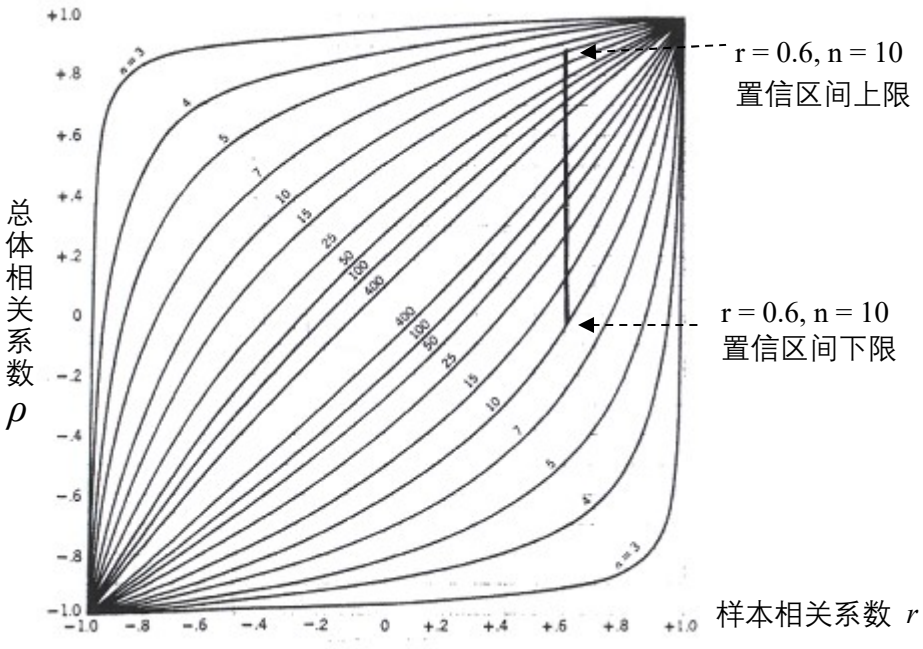

图 11.c 从样本相关系数 r 与样本量，找出总体相关系数 ρ 的 95%置信区间

二. ρ 的检验：

(1)检验： 双尾：$\begin{cases} H_0^I : \rho = 0 \\ H_1^I : \rho \neq 0 \end{cases}$ 左尾：$\begin{cases} H_0^{II} : \rho \geq 0 \\ H_1^{II} : \rho < 0 \end{cases}$ 右尾：$\begin{cases} H_0^{III} : \rho \leq 0 \\ H_1^{III} : \rho > 0 \end{cases}$

计算检验值 $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

若 $|t| \geq t_{\frac{\alpha}{2},n-2}$，则拒绝 H_0^I；若 $t < -t_{\alpha,n-2}$，则拒绝 H_0^{II}；若 $t > t_{\alpha,n-2}$， 则拒绝 H_0^{III}

(2) 检验：双尾：$\begin{cases} H_0^I : \rho = c \neq 0 \\ H_1^I : \rho \neq c \end{cases}$ 左尾：$\begin{cases} H_0^{II} : \rho \geq c \neq 0 \\ H_1^{II} : \rho < c \end{cases}$ 右尾：$\begin{cases} H_0^{III} : \rho \leq c \neq 0 \\ H_1^{III} : \rho > c \end{cases}$

计算检验值 $z^* = (Z_r - Z_c)\sqrt{n-3} = \dfrac{\sqrt{n-3}}{2}\log_e\left[\dfrac{(1+r)(1-c)}{(1-r)(1+c)}\right]$

$Z_r = \dfrac{1}{2}\log_e(\dfrac{1+r}{1-r})$ $Z_c = \dfrac{1}{2}\log_e(\dfrac{1+c}{1-c})$

若 $|z^*| \geq t_{\frac{\alpha}{2}}$，则拒绝 H_0^I；若 $z^* < -z_\alpha$，则拒绝 H_0^{II}；若 $z^* > z_\alpha$，则拒绝 H_0^{III}

例题 11.6： 例题 11.1 汽车引擎汽缸大小与每加仑汽油的哩程数。计算 ρ 的 95% 置信区间。

解答：利用 Fisher 转换： $r = -0.5874$，查表 A.7 $Z_r = \dfrac{1}{2}\log_e(\dfrac{1+r}{1-r}) = -0.678$

$l = -0.678 - (1.96)\dfrac{1}{\sqrt{8-3}} = -0.1555$ $u = -0.678 + (1.96)\dfrac{1}{\sqrt{8-3}} = 0.199$

$\dfrac{e^{-3.1}-1}{e^{-3.1}+1} \leq \rho \leq \dfrac{e^{0.40}-1}{e^{0.40}+1}$

查表 A.15 得到： $\dfrac{0.045-1}{0.045+1} \leq \rho \leq \dfrac{\frac{1}{0.67}-1}{\frac{1}{0.67}+1}$ $\dfrac{-0.955}{1.045} \leq \rho \leq \dfrac{0.493}{2.493}$

ρ 的 95% 置信区间： $-0.914 \leq \rho \leq 0.198$

11.7.4 两组变量相关系数的区间估计与检验

1. 两组变量的相关系数 ρ_1 与 ρ_2 的区间估计，要利用 Fisher 转换(Fisher's transformation) 。$\rho_1 - \rho_2$ 的 $1-\alpha$ 置信区间，计算步骤如下：

(1) 计算样本相关系数 r_1 与 r_2。

(2) 计算 $Z_{r1} = \dfrac{1}{2}\log_e(\dfrac{1+r_1}{1-r_1})$ 与 $Z_{r2} = \dfrac{1}{2}\log_e(\dfrac{1+r_2}{1-r_2})$

(3) $l = Z_{r1} - Z_{r2} - z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}$ $u = Z_{r1} - Z_{r2} + z_{\frac{\alpha}{2}}\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}$

(4) $\frac{e^{2l}-1}{e^{2l}+1} \le \rho_1 - \rho_2 \le \frac{e^{2u}-1}{e^{2u}+1}$

2. ρ 的检验：双尾：$\begin{cases} H_0^I : \rho_1 = \rho_2 \\ H_1^I : \rho_1 \neq \rho_2 \end{cases}$ 左尾：$\begin{cases} H_0^{II} : \rho_1 \ge \rho_2 \\ H_1^{II} : \rho_1 < \rho_2 \end{cases}$ 右尾：$\begin{cases} H_0^{III} : \rho_1 \le \rho_2 \\ H_1^{III} : \rho_1 > \rho_2 \end{cases}$

3. 计算检验值 $z^* = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$

若 $|z^*| \ge t_{\frac{\alpha}{2}}$，则拒绝 H_0^I；若 $z^* < -z_\alpha$，则拒绝 H_0^{II}；若 $z^* > z_\alpha$，则拒绝 H_0^{III}

例题 11.7：营养学者要研究肥胖症，有 28 对兄妹其中有一个肥胖者，他们体重的相关系数为 r = 0.64；有 23 对兄妹其中没有肥胖者，他们体重的相关系数为 r = 0.38。检验「有肥胖者家庭」与「一般家庭」，兄妹体重的相关系数是否相同？ (α = 0.05)

解答：查表 A.7，得到：$Z_{r1} = Z_{0.64} = 0.758$， $Z_{r2} = Z_{0.38} = 0.400$

计算检验值 $z^* = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} = \frac{0.758 - 0.400}{\sqrt{\frac{1}{25}+\frac{1}{20}}} = 0.193$

$z^* < z_{0.025} = 1.96$ ， 所以接受 H_0

11.8 残差分析

简单线性回归分析的数学模型： $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $i = 1, \ldots, n$ $\varepsilon_i \sim N(0, \sigma^2)$

残差(residual) ε_i 是随机变量，根据假设条件，残差是：1. 正态分布，2. 互相独立，3.平均数为 0，4. 有相同的方差。

所以，线性回归分析后，可能对样本残差 $e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$ 进行检查或检验，以确定回归模型是否满足假设条件。

标准化残差(Standardized Residual) = $\frac{e_i}{MS_E\sqrt{1-h_i}}$， $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$

(一)、检查样本残差是否正态分布：

1. 画出样本残差的直方图、计算偏态、峰态系数(第 2 章)。
2. 画出常态分数图，判定是否一直线(第 5 章)。
3. 利用卡方适合度检验(第 12 章)、或 Kolmogorov-Smirnov 检验、Lilliefors 检验(附录)。

(二)、检查样本残差是否有相同方差：画出样本残差 e_i 和 $\hat{y}_i$ 的二度空间散布图，或样本残差 e_i 和 x_i

的二度空间散布图。如果散布图是无规则的散布，则残差方差相同，如图 11.13(a)；图 11.13(b)的残差方差不等。

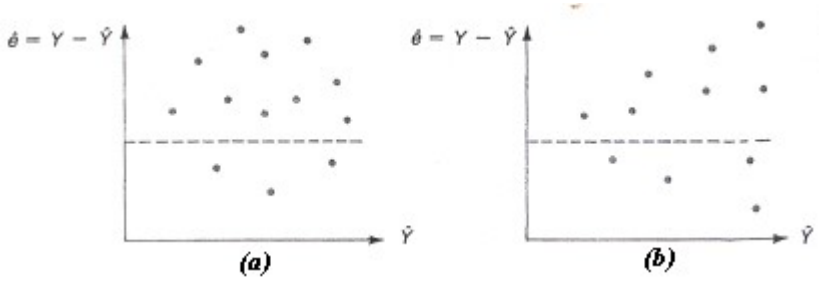

图 11.13 残差散布图

(三)、检查样本残差是否独立

Durbin-Watson 统计量是检查一时间序列资料，是否自相关(autocorrelation)。如果 X,Y 是以时间序列收集的数据，$e_i = y_i - \hat{y}_i$

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

$0 \le DW \le 4$，如果 $DW = 2$，则残差是无相关；若 $0 \le DW < 2$，则残差是正的自相关；若 $2 < DW \le 4$，则残差是负的自相关。当然，DW 越接近 0 或 4，自相关性越强。检验自相关是否显着：

检验：$\begin{cases} H_0 : \text{殘差無自相關} \\ H_1 : \text{殘差有自相關} \end{cases}$

计算 DW，利用中文统计 2.0，Durbin-Watson 检验法：

$d = DW$，$dL = d_L$，$dU = d_U$，$dL = d_L$，$dU = d_U$，$4 - dL = 4 - d_L$，$4 - dU = 4 - d_U$。

若 $DW < d_L$，则拒绝 H_0，有显着的正自相关。

若 $d_L \le DW \le d_U$，则无法判定。

若 $d_U < DW < 4 - d_U$，则接受 H_0，残差无自相关。

若 $4 - d_U \le DW \le 4 - d_L$，则无法判定。

若 $DW > 4 - d_L$，则拒绝 H_0，有显着的负自相关。

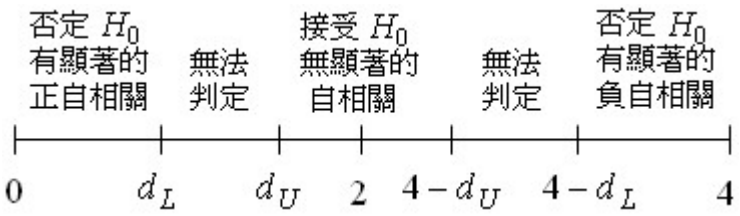

图 11.14 Durbin-Watson 检验法则

无母数统计的连检验(run test)，也可以检验残差的独立性(随机性)。

例题 11.8：例题 11.1 汽车引擎汽缸大小与每加仑汽油的哩程数的残差分析。

解答：利用中文统计 2.0：先执行「简单直线回归」，勾选残差；再选择 Durbin-Watson 检验法，输出残差范围，按「确定」。

	A	B	C
1	Durbin-Watson檢定法		
2			
3	顯著水準		0.05
4	主要自變數		1
5	觀察值		8
6	d		1.5595
7	dL		0.763
8	dU		1.332
9	4-dU		2.668
10	4-dL		3.237

$$d = DW = 1.5595,$$
$$dL = d_L = 0.763,$$
$$dU = d_U = 1.332,$$
$$4 - dL = 4d_L = 2.668,$$
$$4 - dU = 4d_U = 3.237$$

检验结果：

接受 H_0，残差无自相关。

习题

1. 已知下列 9 组(x,y) 值 ：

x	1	1	1	2	3	3	4	5	5
y	9	7	8	10	15	12	19	24	21

(1) 绘出散布图。

(2) 计算 $\bar{x}$，$\bar{y}$，S_{xx}，S_{yy}，S_{xy}

(3) 求最小平方的直线回归的方程式，并在散布图中绘出此直线。

(4) 求对应于 x= 3 时，预测值 y。

(5) 计算样本相关系数。

(6) 检验回归模型的直线关系。

(7) 求残差平方和，绘出残差散布图，进行残差分析。

(8) 估计误差方差，计算 MS_E

(9) 求回归直线之斜率的 90% 置信区间。

(10) 求对应于 x = 4 时，因变量 y 的期望值的 90% 置信区间。

(11) 求对应于 x = 6 时，因变量 y 的 90% 预测区间。

2. 某项实验欲研究塑料纤维强度(y)与聚合体大小(x)之间的关系。进行 15 次实验，求得 (x,y)值，并计算出下列的统计数量：

$$\bar{x} = 8.3,\ \bar{y} = 54.8 \qquad S_{xx} = 5.6,\ S_{xy} = -12.4,\ S_{yy} = 38.7$$

(1) 求最小平方的回归直线的方程式。

(2) 以 $\alpha = 0.05$，检验 $\beta \geq -2$。

(3) 估计 x=10 时，y 的期望值，并求其 95% 置信区间。

(4) 将 y 的总变异分解为二部份：可解释变异与未解释变异。

(5) 以回归直线解释 y 的变异的百分比为何？

(6) x 与 y 之间的样本相关系数为何？

3. 已知下列 (x,y) 的资料：

x	.257	.295	.284	.272	.277	.245	.255	.284	.265	.302	.241	.310	.239
y	1.35	1.50	1.33	1.39	1.28	1.06	1.32	1.23	1.15	1.35	1.08	1.20	1.15

(1) 样本相关系数为何？

(2) 以估计的回归直线，解释 y 变异的百分比为何？

(3) 因变量与自变量的相关性是否显着？

4. 某份早报刊载过国外进口中古车的售价(y ,千美元)与其车龄(x, 年)，数据如下：

X	1	2	2	3	3	4	6	7	8	10
Y	5.45	4.80	5.00	4.00	3.70	3.20	3.15	2.69	1.90	1.47

(1) 绘出散布图。

(2) 求最小平方的直线回归方程式，并在散布图中绘出此直线。

(3) 求回归直线之斜率的 95% 置信区间。

(4) 以回归直线，求车龄为 5 年的平均售价之估计值，并求其 95% 置信区间。

(5) 求车龄为 5 年之售价的 90% 预测区间。

(6) 是否可由上面的数据，依回归直线预测车龄为 20 年的售价，并说明理由。

5. 已知因变量 x 与自变量 y 的样本数据如下：

$$n = 20,\ \sum x = 160,\ \sum y = 240,\ \sum x^2 = 1536,\ \sum xy = 1832,\ \sum y^2 = 2965$$

(1) 求最小平方回归直线方程式。

(2) x 与 y 之间的样本相关系数为何？

(3) 因变量 x 与自变量 y 的相关性是否显着？

6. 已知下列 10 组(x,y) 值 ：

x	3	3	4	5	6	6	7	8	8	9
y	9	5	12	9	14	16	22	18	24	22

(1) 绘出散布图。

(2) 计算 $\bar{x}$，$\bar{y}$，S_{xx}，S_{yy}，S_{xy}

(3) 求最小平方的直线回归的方程式，并在散布图中绘出此直线。

(4) 求对应于 x = 3 时，预测值 Y。

(5) 计算判定系数 r^2。

(6) 求残差平方和，绘出残差散布图。

(7) 估计误差方差，计算 SS_E

(8) 求回归直线之斜率的 90% 置信区间。

(9) 求对应于 x = 4 时，因变量 Y 的期望值的 90% 置信区间。

(10) 求对应于 x = 6 时，因变量 Y 的 90% 预测区间。

(11) b_0 与 b_1 之估计标准误

(12) 检验 $H_0 : \beta = 0$。

(13) 将总平方和分解为解释的平方和与残差平方和

7. 已知下列 (x,y) 的资料：

X	170	147	166	125	182	133	146	125	136	179	174	128	152
Y	698	518	725	485	745	538	485	625	471	798	945	578	625
X	157	174	185	171	102	150	192						
Y	558	698	745	611	458	538	778						

(1) 求最小平方回归直线的方程式。

(2) 求回归直线之斜率的 95% 置信区间。

(3) 求 $x = 150$ 时，预测 y 的 95% 预测区间。

(4) 令 $x = 175$ 与 $x = 195$ ，重作上小题(3)。

8. 科学家认为蟋蟀的鸣叫频率(x)与气温(y)有关。观察 15 只蟋蟀，得下列(x,y)的资料：

音调频率(每秒) x	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1
气温 (F) y	88.6	71.6	93.3	84.3	80.6	75.2	69.7	82.0
音调频率(每秒) x	15.4	16.3	15.0	17.2	16.0	17.0	14.4	
气温 (F) y	69.4	83.3	79.6	82.6	80.6	83.5	76.3	

(1) 求最小平方的直线回归方程式。

(2) 求直线斜率的 95% 置信区间。

(3) 求 $x = 15$ 时，预测的气温。

9. 某连锁家俱销售公司的总经理认为家俱销售员的销售状况与经验有很大的关系，从旗下的连锁店中随机抽样 10 名家俱销售员，得到以下销售员的经验年资(以年计)与前一个月的销售(以 NT$1000 元计)：

经验年资	0	2	10	3	8	5	12	7	20	15
销售	7	9	20	15	18	14	20	17	30	25

假设经验年资与前一个月销售均为正态分布，想知道此连锁家俱销售公司总经理的想法是否正确，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释经验年资与销售是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般年资为 16 年经验年资销售员每月销售的 95%置信区间。

(7) 请估计某位年资为 16 年经验年资销售员每月销售的 95%置信区间。

10. 某求职网站的营销专员研究教育水平与薪资是否有关系，从网站的求职者中随机抽样 15 名，得到以下求职者的教育年资(以年计)与前一个工作每月的薪资(以 NT$1000 元计)：

教育年资	16	11	15	8	12	10	13	14	15	12	16	7	13	9	14
每月薪资	58	40	55	35	43	41	52	49	52	45	62	30	49	38	56

假设教育年资与前一个工作每月薪资均为正态分布，想知道此求职网站营销专员的想法是否正确，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释教育年资与前一个工作每月薪资是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般教育年资为 12 年的求职者每月薪资的 95%置信区间。

(7) 请估计某位教育年资为 12 年的求职者每月薪资的 95%置信区间。

11. 某电影制作公司的财务分析师发现电影是否卖座与是否有大明星参与演出有关系，一般而言参与演出的大明星越多电影越卖座，想知道此财务分析师的想法是否正确，从过去投资的电影中随机抽样 10 部，得到以下电影收入信息(以 NT$100,000 元计)与参与演出的最高薪水两名大明星的薪资(以 NT$100,000 元计)：

电影编号	最高薪水两名大明星的薪资	电影收入
1	9.2	48
2	13.1	65
3	2.5	18
4	3.2	20
5	6.2	31
6	5.5	26
7	14.7	73
8	4.3	23
9	8.8	39
10	11.2	58

假设明星薪资与电影收入均为正态分布，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释明星薪资与电影收入是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般最高薪水两名大明星薪资为 NT$600,000 元年电影收入的 95%置信区间。

(7) 请估计某部最高薪水两名大明星薪资为 NT$600,000 元年电影收入的 95%置信区间。

12. 某保险公司的业务员发现车祸的发生与下雨大小很有关系，一般而言雨下得越大车祸越容易发生，想知道此保险公司业务员的想法是否正确，从过去一年中随机抽样 10 个雨天，得到下雨多寡信息(以公厘计)与车祸次数：

下雨多寡	1.25	3.05	1.20	2.30	2.54	8.90	3.81	7.62	2.52	5.08
车祸次数	5	6	2	4	8	14	7	13	7	10

假设下雨多寡与车祸次数均为正态分布，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释下雨多寡与车祸次数是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般下雨为 4 公厘日子车祸次数的 95%置信区间。

(7) 请估计某下雨为 4 公厘日子车祸次数的 95%置信区间。

13. 某爱车人杂志的车辆效能分析师想知道车速与每公升里程的关系，一般而言车开得越快每公升里程越低，想知道此爱车人杂志车辆效能分析师的想法是否正确，特别做一实验测试 10 种不同车速，得到以下车速信息(以公里计)与每公升里程：

车速	25	30	35	40	45	50	55	60	65	70
每公升里程	44	41	38	37	35	34	32	30	27	25

假设每公升里程为正态分布，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释下车速信息与每公升里程是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般车速为 38 公里每公升里程的 95%置信区间。

(7) 请估计某车速为 38 公里每公升里程的 95%置信区间。

14. 某出版社教科书的编辑认为书的成本与书的页数有直接的关系，因此书的卖价也会跟着相关，一般而言书越厚成本越高卖价越高，想知道此出版社教科书编辑的想法是否正确，特别从书局卖的教科书中随机抽样 12 本书，得到以下页数信息与每本书的价格：

页数	844	727	360	915	295	706	410	905	1058	865	677	912
书价	1100	1000	700	1200	600	1000	800	1060	1210	1080	840	1160

假设书的卖价为正态分布，请根据这个样本回答以下的问题。

(1) 请画出散布图。并从图中解释页数信息与每本书的价格是否有线性关系？

(2) 请找出上述问题的简单线性模型。

(3) 请计算与解释请上述简单线性模型的决策系数与相关系数。

(4) 请用 5%的显着性水平检验上述问题的简单线性模型。

(5) 请解释上述问题的简单线性模型的斜率。

(6) 请估计一般页数为 850 页教科书价格的 95%置信区间。

(7) 请估计某页数为 850 页教科书价格的 95%置信区间。