应用统计学

杜志渊

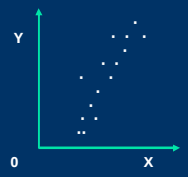
第四章 一元线性回归

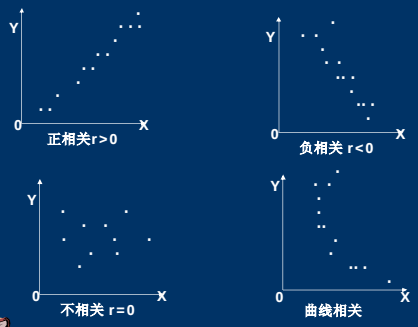
§4.1 一元线性回归模型

一、 变量之间的关系

函数关系 变量之间可以用数学公式表示。如圆的半径R与面积S等。

统计关系 两变量之间有一定依存关系，但没有严格的对应关系。如人的年龄和血压，身高和体重，储蓄额与居民收入等都是统计关系或相关关系。相关关系可以通过相关图表示出来。相关关系有线性相关和非线性相关（曲线相关）。





二、 两变量之间的线性相关系数

1、定义：变量X与y之间的线性相关程度可以用简单相关系数来度量。计算公式为：

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

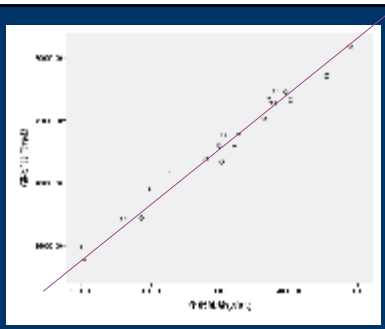
由样本数据得：

$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)(\sum_{i=1}^{n}y_i^2-n\bar{y}^2)}}$$

其中$-1\leq r\leq 1$，且$r>0$时称为正相关，$r<0$时为负相关，$r=0$时，两变量间不存在线性相关关系。

例4.1 讨论化肥施用量与粮食产量之间的关系.数据如下表:

化肥施用量x(万吨)	4541.05	3637.87	2287.49	3056.89	4883.7	3779.3	4021.09
粮食产量y(万吨)	48526.69	45110.87	40753.79	43824.58	50890.11	46370.88	46577.91
化肥施用量x(万吨)	2989.06	3021.9	3953.97	3212.13	3804.76	1598.28	1998.56
粮食产量y(万吨)	42947.44	41673.21	47244.34	43061.53	47336.78	37127.89	39515.07
化肥施用量x(万吨)	3710.56	3269.03	1017.12	1864.23	2797.24	1034.09	
粮食产量y(万吨)	46598.04	44020.92	34866.91	37184.14	41864.77	33717.78	



化肥施用量与粮食产量的关系图

2、相关系数的显著性检验

对样本相关系数r的相关性检验是确定两变量间的线性关系是否显著.

相关系数的检验过程：

r为两变量的相关系数的真值

1、假设 H_0：$\rho=0$；H_1：$\rho\neq0$

2、根据给出的样本数据，计算样本相关系数 r 和统计量 t；在原假设成立的条件下，检验统计量t服从t分布。

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad t\sim t(n-2)$$

3、根据给定的显著性水平α,查 临界值 $t_{\alpha/2}(n-2)$;

4、若 $t\geq t_{\alpha/2}(n-2)$, 说明r在统计上是显著的，即两个变量之间存在线性关系，若 $t<t_{\alpha/2}(n-2)$, 则两个变量之间不存在线性相关。

SPSS统计软件得到的化肥施用量与粮食产量的相关性输出表：

Correlations

		化肥施量(万吨)	粮食产量(万吨)
化肥施量(万吨)	Pearson Correlation	1	.989**
	Sig. (2-tailed)	.	.000
	N	20	20
粮食产量(万吨)	Pearson Correlation	.989**	1
	Sig. (2-tailed)	.000	.
	N	20	20

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

注意：在相关分析中两个变量的地位是平等的.

三、 一元线性回归的数学模型

如果两个变量之间存在线性关系，设一个变量为自变量，另一个是因变量，则两个变量之间的关系可以用一元线性回归模型表达。

设自变量X为确定性变量（解释变量），因变量Y（被解释变量）为随机变量，两者之间的数学结构式为

$$y=b_0+b_1X+e$$

式中：b_0、b_1是回归系数，e是随机项，表示除了变量X之外其他因素对变量Y的影响。

线性方程由两部分组成，一部分由X的变化引起，另一部分是由随机因素引起。

基本假设

假设1：随机项 e 服从正态分布，$N(0,\sigma^2)$，即：

$$E(e)=0,\quad Var(e)=s^2$$

假设2：随机项 e_i 之间是相互独立的，并具有相同的方差，即：

$$Cov(e_i,e_j)=\begin{cases}s^2 & i=j\\ 0 & i\neq j\end{cases} \qquad i=1,2,\mathrm{L},n;\ j=1,2,\mathrm{L},n$$

假设3：样本数据是相互独立的。

由假设知，随机变量y也服从正态分布，由公式 $y=b_0+b_1x+e$ 得出：

$$E(y)=b_0+b_1x,\quad Var(y)=s^2$$

在一般情况下，从研究的总体中抽取一个样本观察值（x_i,y_i），$i=1,2,\cdots,n$，对于样本X、y的每一组数，有

$$y_i=b_0+b_1x_i+e_i \qquad i=1,2,\mathrm{L},n$$

由假设条件知，$e_i\sim N(0,s^2)$，且$E(e_i)=0$, $Var(e_i)=s^2$,$i=1,2,\mathrm{L},n$

推导出观察值y_i (i=1,2,… n)也是相互独立的正态随机变量，且

$$E(y_i)=b_0+b_1x_i,\quad Var(y_i)=s^2$$

线性模型在平均意义上表达了变量Y与X的统计规律性.

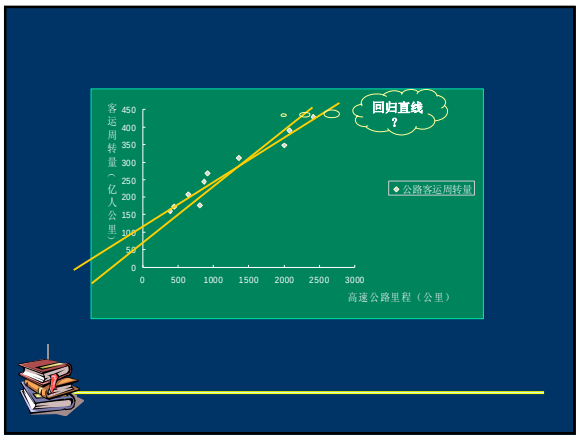
回归模型的矩阵表达式：

$$\begin{cases}Y=Xb+e\\ E(e)=0\\ Var(e)=s^2I_n\end{cases}$$

其中：$Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$ $b=\begin{bmatrix}b_0\\ b_1\end{bmatrix}$ $e=\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix}$ $X=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}$

线性回归的任务是：通过样本观察值对回归系数进行估计，求出 b_0,b_1的估计值 $\hat{b}_0,\hat{b}_1$，得出一元线性回归方程：

$$\hat{Y}=\hat{b}_0+\hat{b}_1X$$



§4.2 回归系数的最小二乘估计

一、普通最小二乘估计(OLSE)

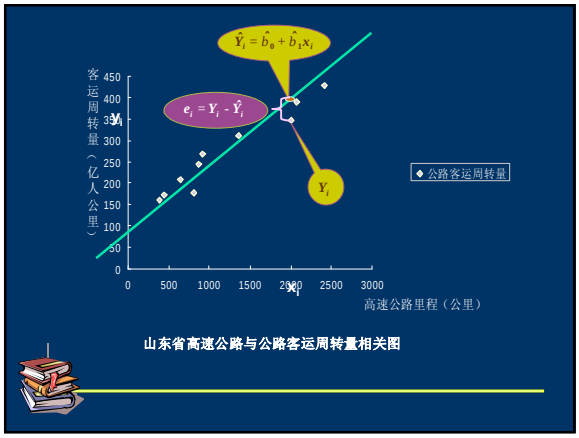
回归系数的估计是依赖于最小二乘法的基本思想，考虑观察值y_i与回归值$\hat{y}_i$的差(称为残差，记为e_i)．通过使残差平方和为最小来估计回归系数。通过样本得到残差平方和为：

$$Q(b_0,b_1)=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{b}_0-\hat{b}_1x_i)^2=\sum e_i^2$$

$\hat{y}_i=\hat{b}_0+\hat{b}_1X_i$

求极值得：

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum(y_i-\hat{b}_0-\hat{b}_1x_i)=0\\ \dfrac{\partial Q}{\partial b_1}=-2\sum x_i(y_i-\hat{b}_0-\hat{b}_1x_i)=0\end{cases}\leftrightarrow\begin{cases}\sum e_i=0\\ \sum e_ix_i=0\end{cases}$$



山东省高速公路与公路客运周转量相关图

由上面方程组可得：$\begin{cases}\sum y_i-n\hat{b}_0-\hat{b}_1\sum x_i=0\\ \sum x_iy_i-\hat{b}_0\sum x_i-\hat{b}_1\sum x_i^2=0\end{cases}$

$$\hat{b}_1=\frac{L_{xy}}{L_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

$$\hat{b}_0=\bar{y}-\hat{b}_1\bar{x}=\frac{1}{n}\sum_{i=1}^{n}y_i-\hat{b}_1\frac{1}{n}\sum_{i=1}^{n}x_i$$

这种方法称普通最小二乘估计(OLSE)，是线性回归方程中回归系数求解的基本方法。这样由样本估计得到的回归方程称为一元线性经验回归方程．记为：

$$\hat{Y}=\hat{b}_0+\hat{b}_1X$$

例： 根据恩格尔定律得出食品支出Y与实际收入X的关系的一元线性回归模型来反映.用表的形式列出原始数据X，Y的值及相应的计算数据。见表

解：根据计算表中的数据得出回归系数为：

$$\hat{b}_1=\frac{15\times44632-1516\times423}{15\times163634-1516^2}=0.1802$$

$$\hat{b}_0=\frac{423}{15}-0.1802\times\frac{1516}{15}=9.9872$$

所求的经验回归方程为：

$$\hat{y}_i=9.99+0.1802x_i$$

回归方程的实际意义是：当收入每增加一个单位时，食品支出会增加0.18单位，即使在收入为0的情况下，食品支出依然需要9.99单位。

续

编号	X	Y	XY	X^2	Y^2
1	102	27	2754	10404	729
2	96	26	2496	9216	676
3	97	25	2425	9409	625
4	102	28	2856	10404	784
5	91	27	2457	8281	729
6	158	36	5688	24964	1296
7	54	19	1026	2916	361
8	83	26	2158	6889	676
9	123	31	3813	15129	961
10	106	31	3286	11236	961
11	129	34	4386	16641	1156
12	138	38	5244	19044	1444
13	81	27	2187	6561	729
14	92	28	2576	8464	784
15	64	20	1280	4096	400
合计	1516	423	44632	163654	12311

返回

§4.3 回归系数的性质及统计意义

一、 $\hat{b}_0, \hat{b}_1$ 是y_i 的线性组合。

例如：
$$\hat{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{L_{xx}} = \frac{\sum (x_i - \bar{x})y_i}{L_{xx}} = \sum \frac{(x_i - \bar{x})}{L_{xx}} y_i$$

$$\hat{b}_1 = \sum C_i y_i \quad C_i = \frac{x_i - \bar{x}}{L_{xx}} \quad i = 1,2,\mathrm{L}\,,n$$

同理得：
$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = \sum (\frac{1}{n} - C_i \bar{x}) y_i$$

二、 $\hat{b}_0, \hat{b}_1$ 服从正态分布，且最小二乘估计的回归系数具有无偏性。

由于因变量y_i 服从正态分布，$\hat{b}_0$ 和 $\hat{b}_1$ 是y_i 的线性组合，故也服从正态分布。

即有：
$$E(\hat{b}_0) = b_0 \qquad s^2_{\hat{b}_0} = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right) = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)$$

$$E(\hat{b}_1) = b_1 \qquad s^2_{\hat{b}_1} = \frac{s^2}{L_{xx}} = \frac{s^2}{\sum (x_i - \bar{x})^2}$$

$$\hat{b}_0 \sim N\left(b_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)s^2\right); \qquad \hat{b}_1 \sim N\left(b_1, \frac{s^2}{L_{xx}}\right)$$

证明： $E(\hat{b}_1) = b_1$

因为有 $\hat{b}_1 = C_1 y_1 + C_2 y_2 + \mathrm{L} + C_n y_n$

由于 $\sum C_i = \sum \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2} = 0$

$$\sum C_i x_i = \sum C_i (x_i - \bar{x}) = 1$$

$$\begin{aligned} E(\hat{b}_1) &= \sum C_i E(y_i) \\ &= \sum C_i (b_0 + b_1 x_i) \\ &= b_0 \sum C_i + b_1 \sum C_i x_i \\ &= b_1 \end{aligned}$$

同理得： $Var(\hat{b}_1) = \sum Var(C_i y_i) = \sum C_i^2 Var(y_i) = \sum C_i^2 s^2$

由于 $\sum C_i^2 = \frac{\sum (x_i - \bar{x})^2}{(\sum (x_j - \bar{x})^2)^2} = \frac{1}{\sum (x_j - \bar{x})^2} = \frac{1}{L_{xx}}$

于是有 $Var(\hat{b}_1) = s^2_{\hat{b}_1} = \sum C_i^2 s^2 = \frac{s^2}{\sum (x_j - \bar{x})^2} = \frac{s^2}{L_{xx}}$

同理：可以证明关于 $\hat{b}_0$ 的相应的数字特征的结论。

注意：线性相关系数r与一元线性回归系数 β_1 符号相同。因为

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{L_{xy}}{L_{xx}} \sqrt{\frac{L_{xx}}{L_{yy}}} = \hat{b}_1 \sqrt{\frac{L_{xx}}{L_{yy}}}$$

在实际计算时，由于总体的方差 σ^2 常常是未知的，则用S^2近似地估计总体方差。故可得：

$$\hat{s}^2_{\hat{b}_0} = S^2_{\hat{b}_0} = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right) \qquad \hat{s}^2_{\hat{b}_1} = S^2_{\hat{b}_1} = \frac{S^2}{L_{xx}}$$

同理，还可以得到回归系数 $\hat{b}_0, \hat{b}_1$ 的协方差为：

$$Cov(\hat{b}_0, \hat{b}_1) = -\frac{\bar{x}}{L_{xx}} s^2$$

上式表明只有在 $\bar{x} = 0$ 时，两回归系数才是相互独立的。

Ø 在基本假设条件下，在各种线性无偏估计量中，由最小二乘估计得到的回归系数具有最小的方差界。

三、方差 σ^2 的估计S^2

在一元回归方程中，可以证明，未知参数 σ^2 的无偏估计 S^2 可以表示为：

$$\hat{s}^2 = S^2 = \frac{\sum e_i^2}{n - 2}$$

上式中，分子是残差平方和，
分母是自由度，
n是样本容量。

由于在估计回归系数时用到 $\sum e_i = 0, \sum e_i X_i = 0$，故可以得知自由度为n-2。

§4.4 回归模型的检验

一、回归系数的显著性检验

原假设$b_1=0$

首先，建立原假设： H_0: $b_1=0$

备选假设： H_1: $b_1\neq 0$

其次，计算回归系数的检验统计量t值 $t_{b_1}=\dfrac{\hat{b}_1-b_1}{s_{\hat{b}_1}}=\dfrac{\hat{b}_1}{S_{\hat{b}_1}}$

式中：$S_{\hat{b}_1}$是由样本计算的回归系数的估计标准差。

$S_{\hat{b}_1}$的计算公式： $S_{\hat{b}_1}=\sqrt{\hat{s}^2/L_{xx}}=\sqrt{S^2/L_{xx}}$

最后，根据给定的显著性水平a，查表得临界值 $t_{a/2}(n-2)$，

如果 $|t|<t_{a/2}(n-2)$，则接受原假设 H_0，即认为 $b_1=0$，说明回归系数与0的差异不显著，回归模型无效。

同理，可检验系数b_0。设原假设： H_0: $b_0=0$

备选假设： H_1: $b_0\neq 0$

再计算检验统计量t值 $t_{b_0}=\dfrac{\hat{b}_0-b_0}{s_{\hat{b}_0}}=\dfrac{\hat{b}_0}{S_{\hat{b}_0}}$

$S_{\hat{b}_0}$的计算公式： $S_{\hat{b}_0}=\sqrt{S^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{L_{xx}}\right)}$

最后，根据给定的显著性水平a，查表得临界值 $t_{a/2}(n-2)$，

如果 $|t|<t_{a/2}(n-2)$，则接受原假设 H_0，即认为 $b_0=0$，说明回归系数b_0与0的差异不显著.

三、回归方程的显著性检验。

对于线性回归方程整体进行检验判定两变量之间是否存在线性关系。

Ø 建立原假设： H_0: $b_1=0$

对立假设： H_1: $b_1\neq 0$

Ø构造检验统计量:

考虑每个y_i与平均值 $\bar{y}$ 之间的差异及总的偏差平方和。先将差异分解后再取平方得：

$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$

交叉项为0

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

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

$\sum(y_i-\bar{y})^2$ 称总偏差平方和，记为SST。

$\sum(\hat{y}_i-\bar{y})^2$ 称回归平方和,记作SSR。反映由X变化引起Y的波动。

$\sum(y_i-\hat{y}_i)^2$ 称残差平方和，记为SSE。反映了随机误差引起的波动。

$SST=\sum(y_i-\bar{y})^2=L_{yy}$

$SSR=\sum(\hat{y}_i-\bar{y})^2=\hat{b}_1^2L_{xx}$

$SSE=\sum(y_i-\hat{y}_i)^2=\sum e_i^2$

可以证明： $E(SSE)=(n-2)s^2$

$E(SSR)=s^2+b_1^2L_{xx}$

上式表明，当回归系数为0时，回归平方和只反映了随机误差引起的差异。故在原假设成立的条件下，有

$E(SSE/n-2)=s^2$； $E(SSR)=s^2$

在原假设成立的条件下还可以证明：

$\dfrac{SSR}{s^2}\sim c^2(1)$； $\dfrac{SSE}{s^2}\sim c^2(n-2)$

构造F统计量为： $F=\dfrac{SSR}{SSE/n-2}$

根据给定的显著性水平a和两个自由度，$df_1=1$，$df_2=n-2$，查临界值F_a（1，n-2），如果 $F>F_a$（1，n-2），则拒绝原假设 H_0，即回归效果是显著的；反之，回归效果是不显著的。

Ø 方差分析表

方差来源	平方和	自由度	均方	F值
回 归	SSR	1	$\overline{SSR}$ = SSR/1	$\dfrac{\overline{SSR}}{\overline{SSE}}$
残 差	SSE	n-2	$\overline{SSE}$ = SSE/n-2	
总 和	SST	n-1		

注意：一元线性回归方程的检验，回归系数的检验及相关系数的检验是等价的。

四、拟合优度检验（样本决定系数）

由于总离差平方和SST=SSER+ SSE，故在等式两边同除总离差平方和得：

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

则样本决定系数 r^2 定义为：

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}; \text{或} \quad r^2 = \frac{\hat{b}_1^2 L_{xx}}{L_{yy}} = \frac{L_{xy}^2}{L_{xx} L_{yy}}$$

在一元线性回归模型中，线性相关系数 r 是决定系数r^2的平方根。

由定义知，样本决系数是对回归方程拟合程度的综合测量，r^2 越接近1，拟合程度就越好。反之，样本决定系数越小，模型的拟合程度就越差。当$r^2=0$时，有SSR=0，变量X对Y没有关系。

五、估计标准误差（估计标准误）

变量Y的观察值y与回归值 $\hat{y}_i$ 的差异程度以下式表示：

$$S_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

上式为最小二乘残差 e_i 标准差。被称估计标准误。

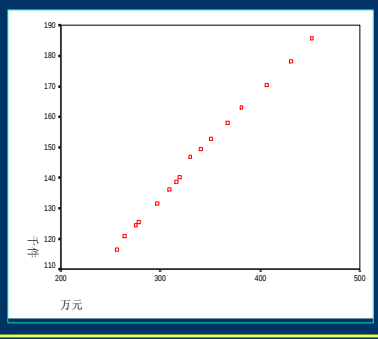
估计标准误是y值与回归直线变差的测度。可以用来判别回归方程的回归效果。

例： 根据表提供的数据，建立某地区居民对某产品的需求量与居民收入的回归方程。

解：令居民收入为自变量X，需求量为因变量Y，根据表中数据绘制散布图，根据已经给出的数据计算：

序号	Y	X	Y^2	X^2	XY
1	116.5	255.7	13572.25	65382.49	29789.05
2	120.8	263.3	14592.64	69326.89	31806.64
3	124.4	275.4	15475.36	75845.16	34259.76
4	125.5	278.3	15750.25	77450.89	34926.65
5	131.7	296.7	17344.89	88030.89	39075.39
6	136.2	309.3	18550.44	95666.49	42126.66
7	138.7	315.8	19237.69	99729.64	43801.46
8	140.2	318.8	19656.04	101633.44	44695.76
9	146.8	330	21550.24	108900	48444
10	149.6	340.2	22380.16	115736.04	50893.92
11	153	350.7	23409	122990.49	53657.1
12	158.2	367.3	25027.24	134909.29	58106.86
13	163.2	381.3	26634.24	145389.69	62228.16
14	170.5	406.5	29070.25	165242.25	69308.25
15	178.2	430.8	31755.24	185588.64	76768.56
16	185.9	451.5	34558.81	203852.25	83933.85
合计	2339.4	5371.6	348564.74	1855674.54	803822.07

根据数据得到相关图：



解：令居民收入为自变量X，需求量为因变量y，根据表中数据绘制散布图，根据已经给出的数据采用最小二乘法计算回归系数：

$$\hat{b}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 0.3524$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = 27.9123$$

得一元线性回归方程：

$$\hat{y} = 27.9123 + 0.3524X$$

并计算出各离差平方和为：

$$SST = \sum (y_i - \bar{y})^2 = 5191.35 \qquad SSE = \sum (y_i - \hat{y}_i)^2 = 17.57$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \hat{b}_1^2 L_{xx} = 5173.78$$

Ø 回归系数检验

建立假设： $H_0: b_1=0$； $H_1: b_1 \neq 0$

计算检验统计量 $t_{b_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}}$

$$\hat{S}^2 = S^2 = \frac{\sum e_i^2}{n-2} = \frac{17.57}{14} = 1.225$$

$$S_{\hat{b}_1} = \sqrt{\frac{S^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{1.225}{40495.87}} = 0.0055$$

$$t_{b_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}} = \frac{0.3524}{0.0055} = 64.2069$$

根据显著性水平a=0.05，查 $t_{0.025}(14)=2.1448$。显然有t值> $t_{0.025}(14)=2.1448$，表明回归系数是显著不等于0。

Ø 回归方程的显著性检验

建立假设： $H_0: \beta_1=0$ ； $H_1: b_1 \neq 0$

计算检验统计量 $F = \dfrac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \hat{y}_i)^2 / n-2} = \dfrac{SSR/1}{SSE/n-2}$

Q $SSR = 5173.78$

$SSE = 17.57$

$\therefore F = \dfrac{5173.78/1}{17.57/14} = 4122.53$

根据给定的显著性水平$\alpha=0.05$，两个自由度，$df_1=1, df_2=14$，查临界值$F_{0.05}(1, 14)=4.60$ 显然有

$F=4122.53>F_{0.05}(1,14)=4.60$

F检验通过，可以认为回归方程的回归效果是显著的；

Ø 拟合程度测定

计算样本可决系数：$r^2 = \dfrac{SSR}{SST} = \dfrac{5173.78}{5191.35} = 0.9966$

r^2接近于 1，表明回归直线与样本点的拟合程度很高。

六、回归系数的置信区间

根据区间估计的计算方法，给定置信水平1-α，则可以求出回归系数 b_0, b_1 的置信区间分别为：

$$(\hat{b}_0 - \hat{s}_{b_0} t_{\alpha/2}(n-2), \hat{b}_0 + \hat{s}_{b_0} t_{\alpha/2}(n-2))$$
$$(\hat{b}_1 - \hat{s}_{b_1} t_{\alpha/2}(n-2), \hat{b}_1 + \hat{s}_{b_1} t_{\alpha/2}(n-2))$$

由此得出例4.1中回归系数 b_1 的95%的置信区间为：

（0.3524-2.1448´0.0055，0.3524+2.1448´0.0055）

=（0.3406，0.3642）

§4.5 残差分析

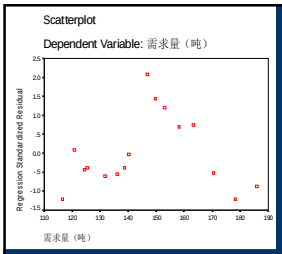
一、残差的概念与残差图

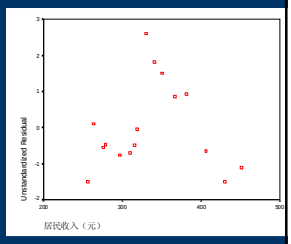
残差指实际观察值与回归值之差e_i，是随机项 ε_i 的估计。

$$e_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i$$
$$e_i = y_i - b_0 - b_1 x_i$$

以自变量X或因变量Y为横轴，以残差为纵轴可以得到残差图.

如果一个回归模型满足给定的基本假设条件，则残差应当在e=0 附近的带状区域内随机排列. 否则，则表明模型存在一定的问题. 书上给出了几种典型的残差图的实例.





例题的残差图

二、残差的性质

性质 1 $E(e_i)=0$

性质 2 $Var(e_i) = \left[1 - \dfrac{1}{n} - \dfrac{(x_i - \bar{x})^2}{L_{xx}}\right] s^2 = (1 - h_{ii}) s^2$

式中 $h_{ii} = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{L_{xx}}$ 称为杠杆值（Leverage Value），

有 $0 < h_{ii} < 1$。说明x_i远离平均值时杠杆值比较大，则方差较小。

性质 3 残差满足约束条件：$\sum e_i = 0; \sum e_i x_i = 0$

即残差项不是独立的。

三、标准化残差与学生化残差

残差分析中的异常值是指超过± 3 s的残差。由于普通的残差的方差不等，故引入标准化残差和学生化残差。

Ø 标准化残差 $ZRE_i = \dfrac{e_i}{\hat{s}}$

Ø 学生化残差 $SRE_i = \dfrac{e_i}{\hat{s}\sqrt{1 - h_{ii}}}$

标准化残差使残差具有可比性，一般|ZRE|>3的相应观测值判为异常值。但没有解决方差不等的问题。而学生化残差则进一步解决了标准化残差没有解决的方差不等的问题，在寻找异常值|SRE|>3方面有优越性。

§4.6 回归方程的应用——预测与控制

一、预测值及预测区间

点预测

预测某一点的值。以自变量的预测值$X=X_0$代入得到y_0的预测值 $\hat{y}_0$。

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 X_0$$

区间预测

由于是随机抽样，$\hat{y}$与y_0有一定的抽样误差，且不同的样本可以求出不同的回归系数值。因此，有必要对y_0值所在的区间作预测。可以证明：

$$\hat{y}_0 - y_0 \sim N[0, s^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2})]$$

因此，y_0的的概率为$(1-\alpha)$预测区间为：

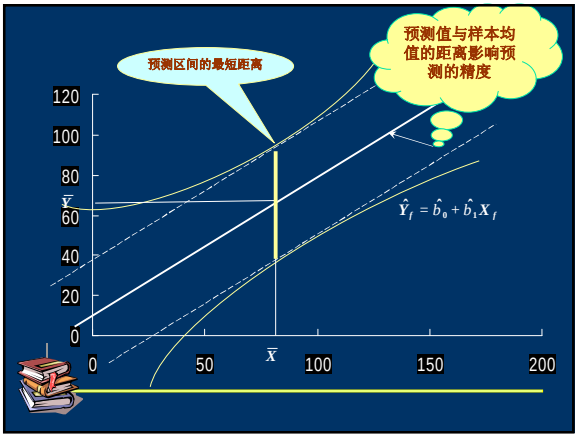
$$[\hat{y}_0 \pm t_{\alpha/2}(n-2) \times \hat{s}\sqrt{1 + h_{ii}}]$$

其中，h_{ii}为杠杆值，表明预测值远离平均值时，预测区间变大，预测精度降低。

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

当n较大时，t分布可以近似地用标准正态分布代替，根号下的值约等于1，则有y_0的预测区间近似为：

$$\hat{y}_0 \pm Z_{\alpha/2}\hat{s}$$



根据前面食品支出的例题，如果回归模型有效，可以预测当收入达到200单位时，相应的食品支出为：

$$\hat{y}_0 = 9.99 + 0.1802 \times 200 = 46.03$$

而预测值的95%的置信区间为：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \times \hat{s}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$= 46.03 \pm t_{0.025}(13) \times 3.334 \times \sqrt{1 + \frac{1}{15} + \frac{200 - 101.07}{20224.74}}$$

$$= (42.069, 49.988)$$

上式的实际意义：当收入达到200单位时，以95%的把握程度认为：食品支出大约在42.069至49.988之间。

二、控制问题

控制问题是预测的逆问题，与预测有着密切的关系，如果要求因变量y控制在一定范围内，即范围内以概率$1-\alpha$保证目标值y控制在（T_1，T_2）范围内，用数学表达式得：

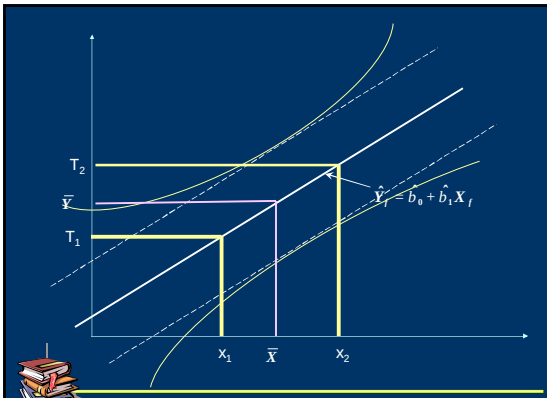
$$P\{T_1 < y < T_2\} = 1 - \alpha$$

从上式中反解出x值，就可以得到相应的x 的取值范围。近似解法为：

$$\begin{cases} \hat{y}(x) - 2\hat{s} = \hat{y}(x) - 2S > T_1 \\ \hat{y}(x) - 2\hat{s} = \hat{y}(x) + 2S < T_2 \end{cases}$$

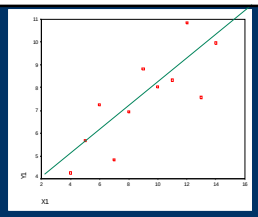
上式中：S为 σ 的估计值，

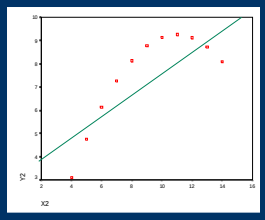
2近似等于标准正态分布95%的置信概率。

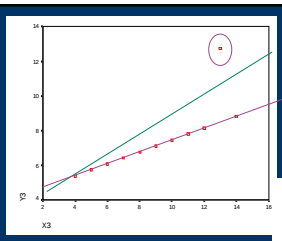


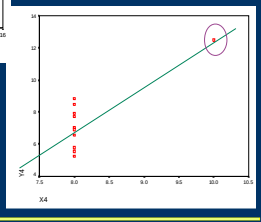
回归模型的进一步讨论：见例题P101。

第一组		第二组		第三组		第四组	
X	Y	X	Y	X	Y	X	Y
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.44	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	10	12.5









Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.667	.629	1.23660	2.788

a. Predictors: (Constant), X1

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.666	.629	1.23721	.385

a. Predictors: (Constant), X2

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.666	.629	1.23618	2.638

a. Predictors: (Constant), X3

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.817[a]	.667	.630	1.23570	1.296

a. Predictors: (Constant), X4

b. Dependent Variable: Y4

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.510	1	27.510	17.990	.002[a]
	Residual	13.763	9	1.529		
	Total	41.273	10			

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.500	1	27.500	17.966	.002[a]
	Residual	13.776	9	1.531		
	Total	41.276	10			

a. Predictors: (Constant), X2

b. Dependent Variable: Y2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.430	1	27.430	17.950	.002[a]
	Residual	13.753	9	1.528		
	Total	41.183	10			

a. Predictors: (Constant), X3

b. Dependent Variable: Y3

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.490	1	27.490	18.003	.002[a]
	Residual	13.742	9	1.527		
	Total	41.232	10			

a. Predictors: (Constant), X4

b. Dependent Variable: Y4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.000	1.125		2.667	.026
	X1	.500	.118	.816	4.241	.002

a. Dependent Variable: Y1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.001	1.125		2.667	.026
	X2	.500	.118	.816	4.239	.002

a. Dependent Variable: Y2

Coefficients[a]

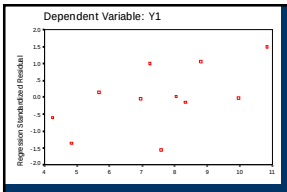
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.008	1.124		2.675	.025
	X3	.499	.118	.816	4.237	.002

a. Dependent Variable: Y3

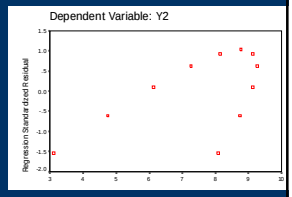
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14.995	5.315		-2.821	.020
	X4	2.750	.648	.817	4.243	.002

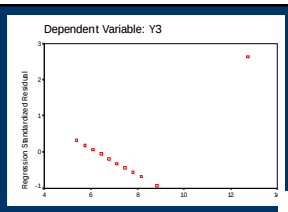
a. Dependent Variable: Y4



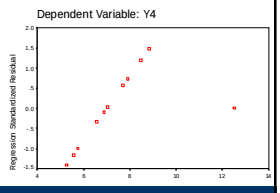
第一组数据的残差图有序列自相关的嫌疑.



第二组数据的残差图有曲线形式



第三组和第四组数据中都明显存在一个异常值.



这四个模型都需要进一步的分析.

一元线性回归应用步骤:

1．首先根据研究目的确定因变量y和自变量X；

2．收集数据后，首先进行相关分析，大致确定回归方程（线性或通过变换得到线性模型）；

3．建立一元线性回归模型，根据样本建立数据表；

4．用ＳＰＳＳ统计软件对样本数据进行分析，得到相应的输出结果。

5．对照输出结果对回归模型进行检验：包括拟合优度的检验（R平方）、回归系数的检验Ｐ值、回归方程的检验（方差分析表）、残差分析、寻找异常值等。

6．通过检验的回归模型称为有效方程，可以在实际中应用，也可以用于预测和控制问题。

§4.7 可化为线性的曲线函数

一、几种可划为线性的曲线函数

1. 抛物线函数（Quadratic 二次函数） $y=a+bX+cX^2$，经变换得：

$$y = a + bx_1 + cx_2$$

其中：$X_1 = X$，　$X_2 = X^2$，故方程变换后成为一个线性方程。

2. 双曲线函数（Inverse逆函数）$y=a+b/X$　做变换得:

$$y=a+bX_1 \qquad X_1=1/X$$

3．幂函数（Power）$y=aX^b$

令　$\ln y=\ln a+b\ln X$

即：　$y'=a'+bX'$;其中$y'=\ln y\ a'=\ln a\ X'=\ln X$

参数b是y对于自变量X的弹性，是X变动1%时, y变动的百分比。

4. 指数函数（Exponential）

$y = ae^{bx}$　　$\ln y = \ln a + bx$

5. 逆函数（Inverse）

$y = b_0 + (b_1 / x)$　　$y = b_0 + b_1 x'$　　$x' = 1/x$

6. 对数函数（Logarithmic）

$y = a + b\log x$

7. 逻辑函数（Logistic）

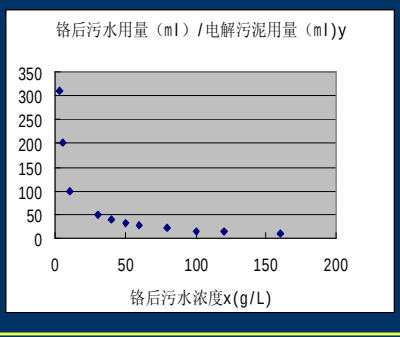
$$y = \frac{L}{1 + ae^{-bx}} \quad (L, a, b > 0)$$

$$ae^{-bx} = \frac{L}{y} - 1 \qquad \ln(\frac{L}{y} - 1) = \ln a - bx$$

实例分析 ： 某厂表面处理车间试验将铬后污水同电解质污泥混合生成无毒溶液。但实际排污的浓度不完全相同，且一定浓度的定量铬后污水只有同定量的电解污泥混合后，才能反应完全。现通过试验，找出铬后污水用量与电解污泥用量之比对于铬后污水浓度之间的关系，试验数据如下：

序号	铬后污水浓度x（g/L）	铬后污水用量（ml）/电解污泥用量（ml）y
1	3	310
2	5	200
3	10	100
4	30	49
5	40	40
6	50	32
7	60	28
8	80	23
9	100	16
10	120	14
11	160	10

解：首先根据数据得到散点图如下：



根据散点图的形状，试用幂函数进行曲线拟合。

假设曲线为幂函数，

由$y = aX^b$　对数变换 $\ln y = \ln a + b \ln X$

取$y' = \ln y$　　$x' = \ln x$　　$a' = \ln a$

以变换后计算得：

$L_{x'x'} = 17.462$，　$L_{y'y'} = 12.003$，$L_{x'y'} = -14.429$

由此得回归系数为：

$$\hat{b} = \frac{L_{x'y'}}{L_{x'x'}} = -0.8263, \quad a' = 6.6417$$

$\hat{a} = e^{a'} = 766.4$

回归方程为：　$Y = 766.4x^{-0.8263}$

对于线性函数应用相关系数表达其线性相关程度，对于非线性函数也可以用相关指数来反映一元非线性回归方程的优劣。

相关指数的定义为：

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{L_{y'y'}}$$

由例中的计算得相关指数为：

$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{L_{y'y'}} = 1 - \frac{0.080}{12.003} = 0.993$$

$L_{yy} = \sum (y_i - \bar{y})^2$

由此可见，曲线的拟合程度是很高的。可以用SPSS作多种函数的假设。

将变量X、Y进行对数变换，应用SPSS软件进行线性回归分析得：

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.997[a]	.993	.993	.09432	1.239

a. Predictors: (Constant), LNX

b. Dependent Variable: LNY

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.923	1	11.923	1340.206	.000[a]
	Residual	.080	9	.009		
	Total	12.003	10			

a. Predictors: (Constant), LNX

b. Dependent Variable: LNY

回归系数表

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.642	.085		78.280	.000
	LNX	-.826	.023	-.997	-36.609	.000

a. Dependent Variable: LNY

下面用SPSS曲线估计的方法寻找合适的拟合曲线。

Inverse 逆函数

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.999	.997	.997	5.392

The independent variable is 铬后污水浓度（g/L).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	91302.512	1	91302.512	3140.302	.000
Residual	261.670	9	29.074		
Total	91564.182	10			

The independent variable is 铬后污水浓度（g/L).

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 / 铬后污水浓度（g/L)	908.460	16.211	.999	56.038	.000
(Constant)	11.514	1.979		5.819	.000

逆函数的表达式为：

$$y = 11.51 + 908.46 / x$$

Power 幂函数

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.997	.993	.993	.094

The independent variable is 铬后污水浓度（g/L).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	11.923	1	11.923	1340.206	.000
Residual	.080	9	.009		
Total	12.003	10			

The independent variable is 铬后污水浓度（g/L).

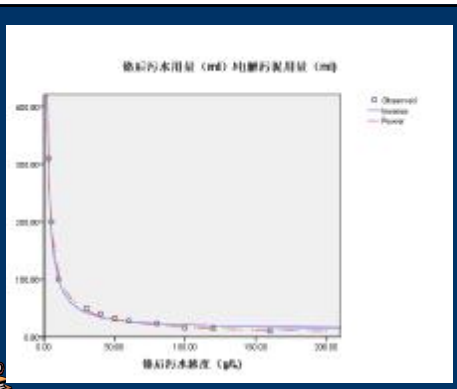
Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
ln(铬后污水浓度（g/L))	-.826	.023	-.997	-36.609	.000
(Constant)	766.418	65.027		11.786	.000

The dependent variable is ln(铬后污水用量（ml）/电解污泥用量（ml)).

幂函数的表达式为：

$$y = 766.418\, x^{-0.826}$$



一元线性回归小结：

1．一元线性回归模型的数学公式（总体与样本），要求掌握如何利用样本建立回归模型，注意理解高斯假设的意义。

2．理解线性回归的意义，明确回归分析与相关分析的关系。

3．掌握回归系数的求解方法——最小二乘法。

$$\hat{b}_1 = \frac{L_{xy}}{L_{xx}} = \frac{n\sum X_t Y_t - \sum X_t \sum Y_t}{n\sum X_t^2 - (\sum X_t)^2},\quad \hat{b}_0 = \frac{\sum Y_t}{n} - \hat{b}_1 \frac{\sum X_t}{n} = \bar{Y} - \hat{b}_1 \bar{X}$$

4．估计总体方差 σ^2 ：

计算中常用S^2估计 σ^2。

$$S^2 = \frac{\sum e_t^2}{n-2}$$

5．$\hat{b}_0, \hat{b}_1$ 的分布

由假设得出Y服从正态分布，且有： $y \sim N(b_0 + b_1 x, s^2)$

可以证明， $\hat{b}_0 \sim N(b_0, s^2_{\hat{b}_0})$； $\hat{b}_1 \sim N(b_1, s^2_{\hat{b}_1})$

其中：

$$s^2_{\hat{b}_0} = s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}\right) \qquad s^2_{\hat{b}_1} = \frac{s^2}{L_{xx}}$$

由样本得出回归系数方差的无偏估计为：

$$\hat{s}^2_{\hat{b}_0} = S^2_{\hat{b}_0} = S^2\left(\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}\right) \qquad \hat{s}^2_{\hat{b}_1} = S^2_{\hat{b}_1} = \frac{S^2}{L_{xx}}$$

6．模型检验

Ø **拟合程度的检验**：利用可决系数r^2对回归方程拟合程度的综合测量，可决系数越大，模型的拟合程度就越高，反之，可决系数越小，模型的拟合程度就越差:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Ø **回归系数的检验（t 检验）**

提出假设：H_0：b_1=0； H_1：$b_1 \neq 0$

计算检验统计量值

$$t_{\hat{b}_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}} = \frac{\hat{b}_1}{S/\sqrt{L_{xx}}}$$

对于给定的显著性水平a；查t分布表确定临界值(或用P值)。要注意当假设是双边检验比较临界值与t值后作出判断。

Ø **回归方程的检验（F检验）**

提出假设：H_0：b_1=0； H_1：$b_1 \neq 0$

计算离差平方和，列出方差分析表如下：

离差名称	离差平方和	自由度	均方差	F值
回归平方和	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	1	SSR	$F = \frac{SSR}{SSE/(n-2)}$
残差平方和	$SSE = \sum e_i^2$	n-2	SSE/(n-2)	
总离差平方和	$SST = \sum (Y_i - \bar{Y})^2$	n-1		

对于给定的显著性水平a，查F分布表确定临界值(或用P值)，当$F>F_a$时,拒绝原假设。

Ø **残差分析**

用因变量或者自变量做横轴，残差(标准化残差,学生化残差等)为纵轴，得到残差图,通过残差图是否随机排列和残差值的范围考察回归模型是否存在异方差性,是否存在异常值等.

7．利用回归方程进行预测

Ø **预测公式：**

由简单回归模型得出基本预测公式为：

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 X_0$$

·点预测：即给定X_0得到Y_0的预测值。

注意：内插与外推预测的区别。

· y_0的区间预测：

在总体标准差 σ^2未知的情况下，用其无偏估计S^2来代替，可以证明：

$$(y_0 - \hat{y}_0)/S_{e_0} \sim t(n-2)$$

由置信区间的确定方法得y_0的置信度为（1-a）的置信区间为

$$(\hat{y}_0 \pm t_{a/2}(n-2) \cdot S_{e_0})$$

· 预测的标准误S_{e0}：在标准假定下，有

$$S_{e_0} = s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \approx S\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

应用统计学

杜志渊

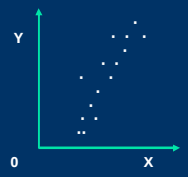
第四章 一元线性回归

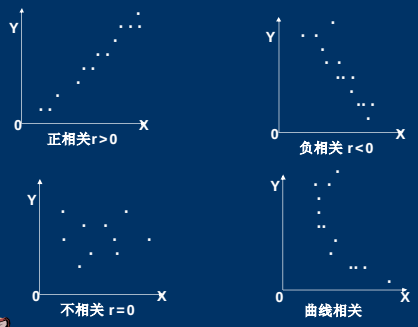
§4.1 一元线性回归模型

一、 变量之间的关系

函数关系 变量之间可以用数学公式表示。如圆的半径R与面积S等。

统计关系 两变量之间有一定依存关系，但没有严格的对应关系。如人的年龄和血压，身高和体重，储蓄额与居民收入等都是统计关系或相关关系。相关关系可以通过相关图表示出来。相关关系有线性相关和非线性相关（曲线相关）。





二、 两变量之间的线性相关系数

1、定义：变量X与y之间的线性相关程度可以用简单相关系数来度量。计算公式为：

$$r=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}$$

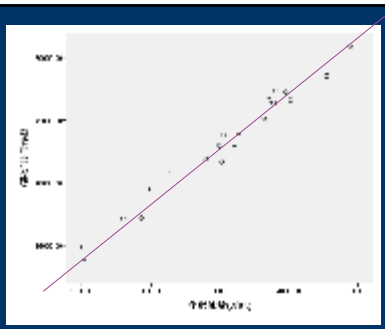
由样本数据得：

$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n}x_i^2-n\bar{x}^2)(\sum_{i=1}^{n}y_i^2-n\bar{y}^2)}}$$

其中$-1\leq r\leq 1$，且$r>0$时称为正相关，$r<0$时为负相关，$r=0$时，两变量间不存在线性相关关系。

例4.1 讨论化肥施用量与粮食产量之间的关系.数据如下表:

化肥施用量x(万吨)	4541.05	3637.87	2287.49	3056.89	4883.7	3779.3	4021.09
粮食产量y(万吨)	48526.69	45110.87	40753.79	43824.58	50890.11	46370.88	46577.91
化肥施用量x(万吨)	2989.06	3021.9	3953.97	3212.13	3804.76	1598.28	1998.56
粮食产量y(万吨)	42947.44	41673.21	47244.34	43061.53	47336.78	37127.89	39515.07
化肥施用量x(万吨)	3710.56	3269.03	1017.12	1864.23	2797.24	1034.09	
粮食产量y(万吨)	46598.04	44020.92	34866.91	37184.14	41864.77	33717.78	



化肥施用量与粮食产量的关系图

2、相关系数的显著性检验

对样本相关系数r的相关性检验是确定两变量间的线性关系是否显著.

相关系数的检验过程：

r为两变量的相关系数的真值

1、假设 H_0：$\rho=0$；H_1：$\rho\neq0$

2、根据给出的样本数据，计算样本相关系数 r 和统计量 t；在原假设成立的条件下，检验统计量t服从t分布。

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \qquad t\sim t(n-2)$$

3、根据给定的显著性水平α,查 临界值 $t_{\alpha/2}(n-2)$;

4、若 $t\geq t_{\alpha/2}(n-2)$, 说明r在统计上是显著的，即两个变量之间存在线性关系，若 $t<t_{\alpha/2}(n-2)$, 则两个变量之间不存在线性相关。

SPSS统计软件得到的化肥施用量与粮食产量的相关性输出表：

Correlations

		化肥施量(万吨)	粮食产量(万吨)
化肥施量(万吨)	Pearson Correlation	1	.989**
	Sig. (2-tailed)	.	.000
	N	20	20
粮食产量(万吨)	Pearson Correlation	.989**	1
	Sig. (2-tailed)	.000	.
	N	20	20

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

注意：在相关分析中两个变量的地位是平等的.

三、 一元线性回归的数学模型

如果两个变量之间存在线性关系，设一个变量为自变量，另一个是因变量，则两个变量之间的关系可以用一元线性回归模型表达。

设自变量X为确定性变量（解释变量），因变量Y（被解释变量）为随机变量，两者之间的数学结构式为

$$y=b_0+b_1X+e$$

式中：b_0、b_1是回归系数，e是随机项，表示除了变量X之外其他因素对变量Y的影响。

线性方程由两部分组成，一部分由X的变化引起，另一部分是由随机因素引起。

基本假设

假设1：随机项 e 服从正态分布，$N(0,\sigma^2)$，即：

$$E(e)=0,\quad Var(e)=s^2$$

假设2：随机项 e_i 之间是相互独立的，并具有相同的方差，即：

$$Cov(e_i,e_j)=\begin{cases}s^2 & i=j\\ 0 & i\neq j\end{cases} \qquad i=1,2,\mathrm{L},n;\ j=1,2,\mathrm{L},n$$

假设3：样本数据是相互独立的。

由假设知，随机变量y也服从正态分布，由公式 $y=b_0+b_1x+e$ 得出：

$$E(y)=b_0+b_1x,\quad Var(y)=s^2$$

在一般情况下，从研究的总体中抽取一个样本观察值（x_i,y_i），$i=1,2,\cdots,n$，对于样本X、y的每一组数，有

$$y_i=b_0+b_1x_i+e_i \qquad i=1,2,\mathrm{L},n$$

由假设条件知，$e_i\sim N(0,s^2)$，且$E(e_i)=0$, $Var(e_i)=s^2$,$i=1,2,\mathrm{L},n$

推导出观察值y_i (i=1,2,… n)也是相互独立的正态随机变量，且

$$E(y_i)=b_0+b_1x_i,\quad Var(y_i)=s^2$$

线性模型在平均意义上表达了变量Y与X的统计规律性.

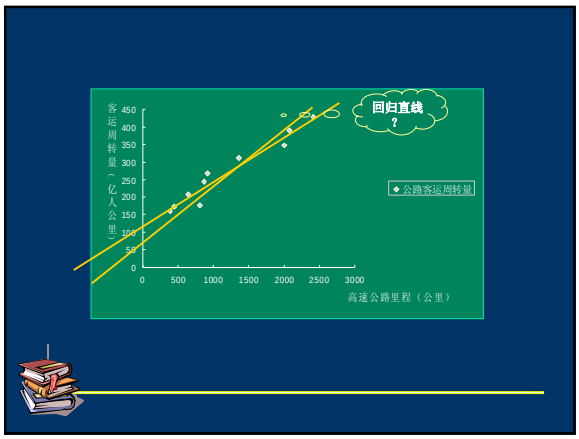
回归模型的矩阵表达式：

$$\begin{cases}Y=Xb+e\\ E(e)=0\\ Var(e)=s^2I_n\end{cases}$$

其中：$Y=\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix}$ $b=\begin{bmatrix}b_0\\ b_1\end{bmatrix}$ $e=\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix}$ $X=\begin{bmatrix}1 & x_1\\ 1 & x_2\\ \vdots & \vdots\\ 1 & x_n\end{bmatrix}$

线性回归的任务是：通过样本观察值对回归系数进行估计，求出 b_0,b_1的估计值 $\hat{b}_0,\hat{b}_1$，得出一元线性回归方程：

$$\hat{Y}=\hat{b}_0+\hat{b}_1X$$



§4.2 回归系数的最小二乘估计

一、普通最小二乘估计(OLSE)

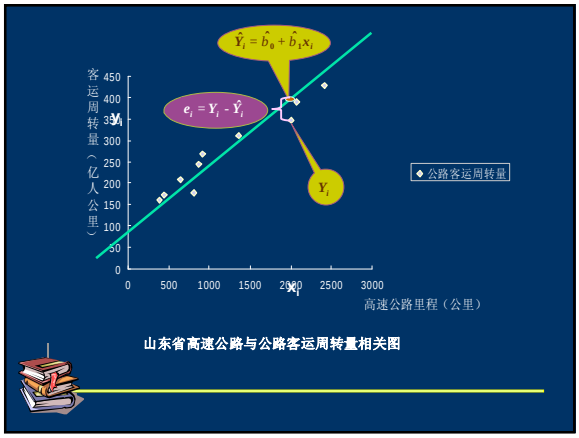
回归系数的估计是依赖于最小二乘法的基本思想，考虑观察值y_i与回归值$\hat{y}_i$的差(称为残差，记为e_i)．通过使残差平方和为最小来估计回归系数。通过样本得到残差平方和为：

$$Q(b_0,b_1)=\sum(y_i-\hat{y}_i)^2=\sum(y_i-\hat{b}_0-\hat{b}_1x_i)^2=\sum e_i^2$$

$\hat{y}_i=\hat{b}_0+\hat{b}_1X_i$

求极值得：

$$\begin{cases}\dfrac{\partial Q}{\partial b_0}=-2\sum(y_i-\hat{b}_0-\hat{b}_1x_i)=0\\ \dfrac{\partial Q}{\partial b_1}=-2\sum x_i(y_i-\hat{b}_0-\hat{b}_1x_i)=0\end{cases}\leftrightarrow\begin{cases}\sum e_i=0\\ \sum e_ix_i=0\end{cases}$$



山东省高速公路与公路客运周转量相关图

由上面方程组可得：$\begin{cases}\sum y_i-n\hat{b}_0-\hat{b}_1\sum x_i=0\\ \sum x_iy_i-\hat{b}_0\sum x_i-\hat{b}_1\sum x_i^2=0\end{cases}$

$$\hat{b}_1=\frac{L_{xy}}{L_{xx}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}$$

$$\hat{b}_0=\bar{y}-\hat{b}_1\bar{x}=\frac{1}{n}\sum_{i=1}^{n}y_i-\hat{b}_1\frac{1}{n}\sum_{i=1}^{n}x_i$$

这种方法称普通最小二乘估计(OLSE)，是线性回归方程中回归系数求解的基本方法。这样由样本估计得到的回归方程称为一元线性经验回归方程．记为：

$$\hat{Y}=\hat{b}_0+\hat{b}_1X$$

例： 根据恩格尔定律得出食品支出Y与实际收入X的关系的一元线性回归模型来反映.用表的形式列出原始数据X，Y的值及相应的计算数据。见表

解：根据计算表中的数据得出回归系数为：

$$\hat{b}_1=\frac{15\times44632-1516\times423}{15\times163634-1516^2}=0.1802$$

$$\hat{b}_0=\frac{423}{15}-0.1802\times\frac{1516}{15}=9.9872$$

所求的经验回归方程为：

$$\hat{y}_i=9.99+0.1802x_i$$

回归方程的实际意义是：当收入每增加一个单位时，食品支出会增加0.18单位，即使在收入为0的情况下，食品支出依然需要9.99单位。

续

编号	X	Y	XY	X^2	Y^2
1	102	27	2754	10404	729
2	96	26	2496	9216	676
3	97	25	2425	9409	625
4	102	28	2856	10404	784
5	91	27	2457	8281	729
6	158	36	5688	24964	1296
7	54	19	1026	2916	361
8	83	26	2158	6889	676
9	123	31	3813	15129	961
10	106	31	3286	11236	961
11	129	34	4386	16641	1156
12	138	38	5244	19044	1444
13	81	27	2187	6561	729
14	92	28	2576	8464	784
15	64	20	1280	4096	400
合计	1516	423	44632	163654	12311

返回

§4.3 回归系数的性质及统计意义

一、 $\hat{b}_0, \hat{b}_1$ 是y_i 的线性组合。

例如：
$$\hat{b}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{L_{xx}} = \frac{\sum (x_i - \bar{x})y_i}{L_{xx}} = \sum \frac{(x_i - \bar{x})}{L_{xx}} y_i$$

$$\hat{b}_1 = \sum C_i y_i \quad C_i = \frac{x_i - \bar{x}}{L_{xx}} \quad i = 1,2,\mathrm{L}\,,n$$

同理得：
$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = \sum (\frac{1}{n} - C_i \bar{x}) y_i$$

二、 $\hat{b}_0, \hat{b}_1$ 服从正态分布，且最小二乘估计的回归系数具有无偏性。

由于因变量y_i 服从正态分布，$\hat{b}_0$ 和 $\hat{b}_1$ 是y_i 的线性组合，故也服从正态分布。

即有：
$$E(\hat{b}_0) = b_0 \qquad s^2_{\hat{b}_0} = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right) = s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)$$

$$E(\hat{b}_1) = b_1 \qquad s^2_{\hat{b}_1} = \frac{s^2}{L_{xx}} = \frac{s^2}{\sum (x_i - \bar{x})^2}$$

$$\hat{b}_0 \sim N\left(b_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)s^2\right); \qquad \hat{b}_1 \sim N\left(b_1, \frac{s^2}{L_{xx}}\right)$$

证明： $E(\hat{b}_1) = b_1$

因为有 $\hat{b}_1 = C_1 y_1 + C_2 y_2 + \mathrm{L} + C_n y_n$

由于 $\sum C_i = \sum \frac{x_i - \bar{x}}{\sum (x_j - \bar{x})^2} = 0$

$$\sum C_i x_i = \sum C_i (x_i - \bar{x}) = 1$$

$$\begin{aligned} E(\hat{b}_1) &= \sum C_i E(y_i) \\ &= \sum C_i (b_0 + b_1 x_i) \\ &= b_0 \sum C_i + b_1 \sum C_i x_i \\ &= b_1 \end{aligned}$$

同理得： $Var(\hat{b}_1) = \sum Var(C_i y_i) = \sum C_i^2 Var(y_i) = \sum C_i^2 s^2$

由于 $\sum C_i^2 = \frac{\sum (x_i - \bar{x})^2}{(\sum (x_j - \bar{x})^2)^2} = \frac{1}{\sum (x_j - \bar{x})^2} = \frac{1}{L_{xx}}$

于是有 $Var(\hat{b}_1) = s^2_{\hat{b}_1} = \sum C_i^2 s^2 = \frac{s^2}{\sum (x_j - \bar{x})^2} = \frac{s^2}{L_{xx}}$

同理：可以证明关于 $\hat{b}_0$ 的相应的数字特征的结论。

注意：线性相关系数r与一元线性回归系数 β_1 符号相同。因为

$$r = \frac{L_{xy}}{\sqrt{L_{xx} L_{yy}}} = \frac{L_{xy}}{L_{xx}} \sqrt{\frac{L_{xx}}{L_{yy}}} = \hat{b}_1 \sqrt{\frac{L_{xx}}{L_{yy}}}$$

在实际计算时，由于总体的方差 σ^2 常常是未知的，则用S^2近似地估计总体方差。故可得：

$$\hat{s}^2_{\hat{b}_0} = S^2_{\hat{b}_0} = S^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right) \qquad \hat{s}^2_{\hat{b}_1} = S^2_{\hat{b}_1} = \frac{S^2}{L_{xx}}$$

同理，还可以得到回归系数 $\hat{b}_0, \hat{b}_1$ 的协方差为：

$$Cov(\hat{b}_0, \hat{b}_1) = -\frac{\bar{x}}{L_{xx}} s^2$$

上式表明只有在 $\bar{x} = 0$ 时，两回归系数才是相互独立的。

Ø 在基本假设条件下，在各种线性无偏估计量中，由最小二乘估计得到的回归系数具有最小的方差界。

三、方差 σ^2 的估计S^2

在一元回归方程中，可以证明，未知参数 σ^2 的无偏估计 S^2 可以表示为：

$$\hat{s}^2 = S^2 = \frac{\sum e_i^2}{n - 2}$$

上式中，分子是残差平方和，
分母是自由度，
n是样本容量。

由于在估计回归系数时用到 $\sum e_i = 0, \sum e_i X_i = 0$，故可以得知自由度为n-2。

§4.4 回归模型的检验

一、回归系数的显著性检验

原假设$b_1=0$

首先，建立原假设： H_0: $b_1=0$

备选假设： H_1: $b_1\neq 0$

其次，计算回归系数的检验统计量t值 $t_{b_1}=\dfrac{\hat{b}_1-b_1}{s_{\hat{b}_1}}=\dfrac{\hat{b}_1}{S_{\hat{b}_1}}$

式中：$S_{\hat{b}_1}$是由样本计算的回归系数的估计标准差。

$S_{\hat{b}_1}$的计算公式： $S_{\hat{b}_1}=\sqrt{\hat{s}^2/L_{xx}}=\sqrt{S^2/L_{xx}}$

最后，根据给定的显著性水平a，查表得临界值 $t_{a/2}(n-2)$，

如果 $|t|<t_{a/2}(n-2)$，则接受原假设 H_0，即认为 $b_1=0$，说明回归系数与0的差异不显著，回归模型无效。

同理，可检验系数b_0。设原假设： H_0: $b_0=0$

备选假设： H_1: $b_0\neq 0$

再计算检验统计量t值 $t_{b_0}=\dfrac{\hat{b}_0-b_0}{s_{\hat{b}_0}}=\dfrac{\hat{b}_0}{S_{\hat{b}_0}}$

$S_{\hat{b}_0}$的计算公式： $S_{\hat{b}_0}=\sqrt{S^2\left(\dfrac{1}{n}+\dfrac{\bar{x}^2}{L_{xx}}\right)}$

最后，根据给定的显著性水平a，查表得临界值 $t_{a/2}(n-2)$，

如果 $|t|<t_{a/2}(n-2)$，则接受原假设 H_0，即认为 $b_0=0$，说明回归系数b_0与0的差异不显著.

三、回归方程的显著性检验。

对于线性回归方程整体进行检验判定两变量之间是否存在线性关系。

Ø 建立原假设： H_0: $b_1=0$

对立假设： H_1: $b_1\neq 0$

Ø构造检验统计量:

考虑每个y_i与平均值 $\bar{y}$ 之间的差异及总的偏差平方和。先将差异分解后再取平方得：

$y_i-\bar{y}=(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})$

交叉项为0

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

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

$\sum(y_i-\bar{y})^2$ 称总偏差平方和，记为SST。

$\sum(\hat{y}_i-\bar{y})^2$ 称回归平方和,记作SSR。反映由X变化引起Y的波动。

$\sum(y_i-\hat{y}_i)^2$ 称残差平方和，记为SSE。反映了随机误差引起的波动。

$SST=\sum(y_i-\bar{y})^2=L_{yy}$

$SSR=\sum(\hat{y}_i-\bar{y})^2=\hat{b}_1^2L_{xx}$

$SSE=\sum(y_i-\hat{y}_i)^2=\sum e_i^2$

可以证明： $E(SSE)=(n-2)s^2$

$E(SSR)=s^2+b_1^2L_{xx}$

上式表明，当回归系数为0时，回归平方和只反映了随机误差引起的差异。故在原假设成立的条件下，有

$E(SSE/n-2)=s^2$； $E(SSR)=s^2$

在原假设成立的条件下还可以证明：

$\dfrac{SSR}{s^2}\sim c^2(1)$； $\dfrac{SSE}{s^2}\sim c^2(n-2)$

构造F统计量为： $F=\dfrac{SSR}{SSE/n-2}$

根据给定的显著性水平a和两个自由度，$df_1=1$，$df_2=n-2$，查临界值F_a（1，n-2），如果 $F>F_a$（1，n-2），则拒绝原假设 H_0，即回归效果是显著的；反之，回归效果是不显著的。

Ø 方差分析表

方差来源	平方和	自由度	均方	F值
回 归	SSR	1	$\overline{SSR}$ = SSR/1	$\dfrac{\overline{SSR}}{\overline{SSE}}$
残 差	SSE	n-2	$\overline{SSE}$ = SSE/n-2	
总 和	SST	n-1		

注意：一元线性回归方程的检验，回归系数的检验及相关系数的检验是等价的。

四、拟合优度检验（样本决定系数）

由于总离差平方和SST=SSER+ SSE，故在等式两边同除总离差平方和得：

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

则样本决定系数 r^2 定义为：

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}; \text{或} \quad r^2 = \frac{\hat{b}_1^2 L_{xx}}{L_{yy}} = \frac{L_{xy}^2}{L_{xx} L_{yy}}$$

在一元线性回归模型中，线性相关系数 r 是决定系数r^2的平方根。

由定义知，样本决系数是对回归方程拟合程度的综合测量，r^2 越接近1，拟合程度就越好。反之，样本决定系数越小，模型的拟合程度就越差。当$r^2=0$时，有SSR=0，变量X对Y没有关系。

五、估计标准误差（估计标准误）

变量Y的观察值y与回归值 $\hat{y}_i$ 的差异程度以下式表示：

$$S_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

上式为最小二乘残差 e_i 标准差。被称估计标准误。

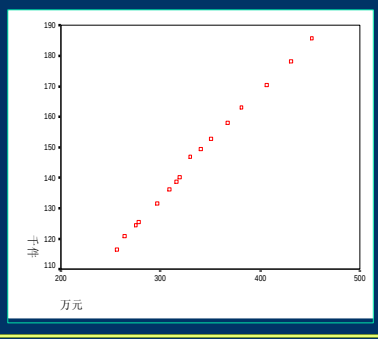
估计标准误是y值与回归直线变差的测度。可以用来判别回归方程的回归效果。

例： 根据表提供的数据，建立某地区居民对某产品的需求量与居民收入的回归方程。

解：令居民收入为自变量X，需求量为因变量Y，根据表中数据绘制散布图，根据已经给出的数据计算：

序号	Y	X	Y^2	X^2	XY
1	116.5	255.7	13572.25	65382.49	29789.05
2	120.8	263.3	14592.64	69326.89	31806.64
3	124.4	275.4	15475.36	75845.16	34259.76
4	125.5	278.3	15750.25	77450.89	34926.65
5	131.7	296.7	17344.89	88030.89	39075.39
6	136.2	309.3	18550.44	95666.49	42126.66
7	138.7	315.8	19237.69	99729.64	43801.46
8	140.2	318.8	19656.04	101633.44	44695.76
9	146.8	330	21550.24	108900	48444
10	149.6	340.2	22380.16	115736.04	50893.92
11	153	350.7	23409	122990.49	53657.1
12	158.2	367.3	25027.24	134909.29	58106.86
13	163.2	381.3	26634.24	145389.69	62228.16
14	170.5	406.5	29070.25	165242.25	69308.25
15	178.2	430.8	31755.24	185588.64	76768.56
16	185.9	451.5	34558.81	203852.25	83933.85
合计	2339.4	5371.6	348564.74	1855674.54	803822.07

根据数据得到相关图：



解：令居民收入为自变量X，需求量为因变量y，根据表中数据绘制散布图，根据已经给出的数据采用最小二乘法计算回归系数：

$$\hat{b}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = 0.3524$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} = 27.9123$$

得一元线性回归方程：

$$\hat{y} = 27.9123 + 0.3524X$$

并计算出各离差平方和为：

$$SST = \sum (y_i - \bar{y})^2 = 5191.35 \qquad SSE = \sum (y_i - \hat{y}_i)^2 = 17.57$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = \hat{b}_1^2 L_{xx} = 5173.78$$

Ø 回归系数检验

建立假设： $H_0: b_1=0$； $H_1: b_1 \neq 0$

计算检验统计量 $t_{b_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}}$

$$\hat{S}^2 = S^2 = \frac{\sum e_i^2}{n-2} = \frac{17.57}{14} = 1.225$$

$$S_{\hat{b}_1} = \sqrt{\frac{S^2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{1.225}{40495.87}} = 0.0055$$

$$t_{b_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}} = \frac{0.3524}{0.0055} = 64.2069$$

根据显著性水平a=0.05，查 $t_{0.025}(14)=2.1448$。显然有t值> $t_{0.025}(14)=2.1448$，表明回归系数是显著不等于0。

Ø 回归方程的显著性检验

建立假设： $H_0: \beta_1=0$ ； $H_1: b_1 \neq 0$

计算检验统计量 $F = \dfrac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \hat{y}_i)^2 / n-2} = \dfrac{SSR/1}{SSE/n-2}$

Q $SSR = 5173.78$

$SSE = 17.57$

$\therefore F = \dfrac{5173.78/1}{17.57/14} = 4122.53$

根据给定的显著性水平$\alpha=0.05$，两个自由度，$df_1=1, df_2=14$，查临界值$F_{0.05}(1, 14)=4.60$ 显然有

$F=4122.53>F_{0.05}(1,14)=4.60$

F检验通过，可以认为回归方程的回归效果是显著的；

Ø 拟合程度测定

计算样本可决系数：$r^2 = \dfrac{SSR}{SST} = \dfrac{5173.78}{5191.35} = 0.9966$

r^2接近于 1，表明回归直线与样本点的拟合程度很高。

六、回归系数的置信区间

根据区间估计的计算方法，给定置信水平1-α，则可以求出回归系数 b_0, b_1 的置信区间分别为：

$$(\hat{b}_0 - \hat{s}_{b_0} t_{\alpha/2}(n-2), \hat{b}_0 + \hat{s}_{b_0} t_{\alpha/2}(n-2))$$
$$(\hat{b}_1 - \hat{s}_{b_1} t_{\alpha/2}(n-2), \hat{b}_1 + \hat{s}_{b_1} t_{\alpha/2}(n-2))$$

由此得出例4.1中回归系数 b_1 的95%的置信区间为：

（0.3524-2.1448´0.0055，0.3524+2.1448´0.0055）

=（0.3406，0.3642）

§4.5 残差分析

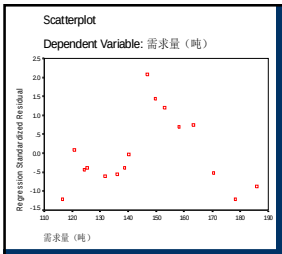
一、残差的概念与残差图

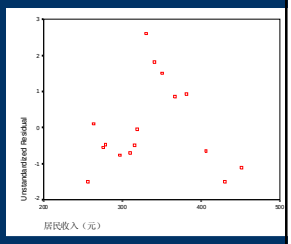
残差指实际观察值与回归值之差e_i，是随机项 ε_i 的估计。

$$e_i = y_i - \hat{y}_i = y_i - \hat{b}_0 - \hat{b}_1 x_i$$
$$e_i = y_i - b_0 - b_1 x_i$$

以自变量X或因变量Y为横轴，以残差为纵轴可以得到残差图.

如果一个回归模型满足给定的基本假设条件，则残差应当在e=0 附近的带状区域内随机排列. 否则，则表明模型存在一定的问题. 书上给出了几种典型的残差图的实例.





例题的残差图

二、残差的性质

性质 1 $E(e_i)=0$

性质 2 $Var(e_i) = \left[1 - \dfrac{1}{n} - \dfrac{(x_i - \bar{x})^2}{L_{xx}}\right] s^2 = (1 - h_{ii}) s^2$

式中 $h_{ii} = \dfrac{1}{n} + \dfrac{(x_i - \bar{x})^2}{L_{xx}}$ 称为杠杆值（Leverage Value），

有 $0 < h_{ii} < 1$。说明x_i远离平均值时杠杆值比较大，则方差较小。

性质 3 残差满足约束条件：$\sum e_i = 0; \sum e_i x_i = 0$

即残差项不是独立的。

三、标准化残差与学生化残差

残差分析中的异常值是指超过± 3 s的残差。由于普通的残差的方差不等，故引入标准化残差和学生化残差。

Ø 标准化残差 $ZRE_i = \dfrac{e_i}{\hat{s}}$

Ø 学生化残差 $SRE_i = \dfrac{e_i}{\hat{s}\sqrt{1 - h_{ii}}}$

标准化残差使残差具有可比性，一般|ZRE|>3的相应观测值判为异常值。但没有解决方差不等的问题。而学生化残差则进一步解决了标准化残差没有解决的方差不等的问题，在寻找异常值|SRE|>3方面有优越性。

§4.6 回归方程的应用——预测与控制

一、预测值及预测区间

点预测

预测某一点的值。以自变量的预测值$X=X_0$代入得到y_0的预测值 $\hat{y}_0$。

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 X_0$$

区间预测

由于是随机抽样，$\hat{y}$与y_0有一定的抽样误差，且不同的样本可以求出不同的回归系数值。因此，有必要对y_0值所在的区间作预测。可以证明：

$$\hat{y}_0 - y_0 \sim N[0, s^2(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2})]$$

因此，y_0的的概率为$(1-\alpha)$预测区间为：

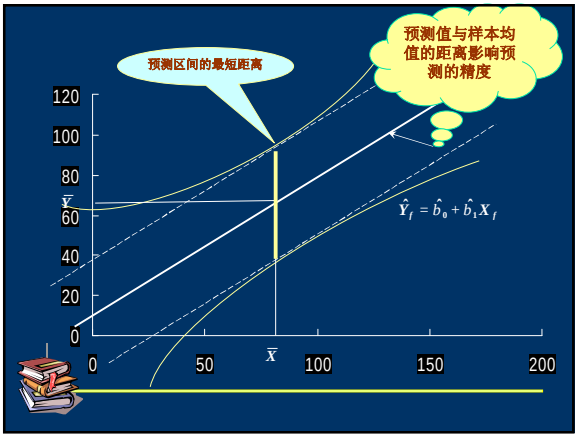
$$[\hat{y}_0 \pm t_{\alpha/2}(n-2) \times \hat{s}\sqrt{1 + h_{ii}}]$$

其中，h_{ii}为杠杆值，表明预测值远离平均值时，预测区间变大，预测精度降低。

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

当n较大时，t分布可以近似地用标准正态分布代替，根号下的值约等于1，则有y_0的预测区间近似为：

$$\hat{y}_0 \pm Z_{\alpha/2}\hat{s}$$



根据前面食品支出的例题，如果回归模型有效，可以预测当收入达到200单位时，相应的食品支出为：

$$\hat{y}_0 = 9.99 + 0.1802 \times 200 = 46.03$$

而预测值的95%的置信区间为：

$$\hat{y}_0 \pm t_{\alpha/2}(n-2) \times \hat{s}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$= 46.03 \pm t_{0.025}(13) \times 3.334 \times \sqrt{1 + \frac{1}{15} + \frac{200 - 101.07}{20224.74}}$$

$$= (42.069, 49.988)$$

上式的实际意义：当收入达到200单位时，以95%的把握程度认为：食品支出大约在42.069至49.988之间。

二、控制问题

控制问题是预测的逆问题，与预测有着密切的关系，如果要求因变量y控制在一定范围内，即范围内以概率$1-\alpha$保证目标值y控制在（T_1，T_2）范围内，用数学表达式得：

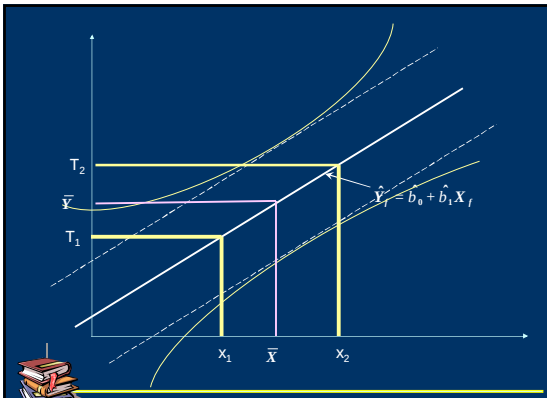
$$P\{T_1 < y < T_2\} = 1 - \alpha$$

从上式中反解出x值，就可以得到相应的x 的取值范围。近似解法为：

$$\begin{cases} \hat{y}(x) - 2\hat{s} = \hat{y}(x) - 2S > T_1 \\ \hat{y}(x) - 2\hat{s} = \hat{y}(x) + 2S < T_2 \end{cases}$$

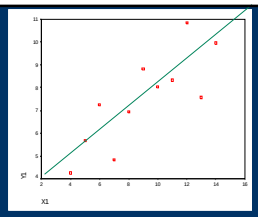
上式中：S为 σ 的估计值，

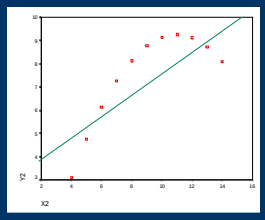
2近似等于标准正态分布95%的置信概率。

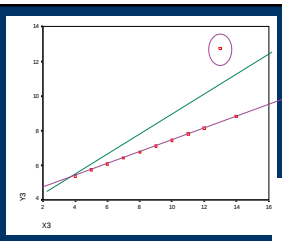


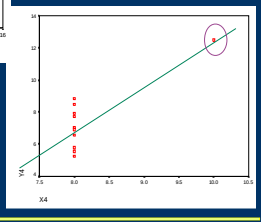
回归模型的进一步讨论：见例题P101。

第一组		第二组		第三组		第四组	
X	Y	X	Y	X	Y	X	Y
4	4.26	4	3.1	4	5.39	8	6.58
5	5.68	5	4.74	5	5.73	8	5.76
6	7.24	6	6.13	6	6.08	8	7.71
7	4.82	7	7.26	7	6.44	8	8.84
8	6.95	8	8.14	8	6.77	8	8.47
9	8.81	9	8.77	9	7.11	8	7.04
10	8.04	10	9.14	10	7.46	8	5.25
11	8.33	11	9.26	11	7.81	8	5.56
12	10.84	12	9.13	12	8.15	8	7.91
13	7.58	13	8.74	13	12.74	8	6.89
14	9.96	14	8.1	14	8.84	10	12.5









Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.667	.629	1.23660	2.788

a. Predictors: (Constant), X1

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.666	.629	1.23721	.385

a. Predictors: (Constant), X2

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.816[a]	.666	.629	1.23618	2.638

a. Predictors: (Constant), X3

Model Summary [b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.817[a]	.667	.630	1.23570	1.296

a. Predictors: (Constant), X4

b. Dependent Variable: Y4

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.510	1	27.510	17.990	.002[a]
	Residual	13.763	9	1.529		
	Total	41.273	10			

a. Predictors: (Constant), X1

b. Dependent Variable: Y1

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.500	1	27.500	17.966	.002[a]
	Residual	13.776	9	1.531		
	Total	41.276	10			

a. Predictors: (Constant), X2

b. Dependent Variable: Y2

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.430	1	27.430	17.950	.002[a]
	Residual	13.753	9	1.528		
	Total	41.183	10			

a. Predictors: (Constant), X3

b. Dependent Variable: Y3

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27.490	1	27.490	18.003	.002[a]
	Residual	13.742	9	1.527		
	Total	41.232	10			

a. Predictors: (Constant), X4

b. Dependent Variable: Y4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.000	1.125		2.667	.026
	X1	.500	.118	.816	4.241	.002

a. Dependent Variable: Y1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.001	1.125		2.667	.026
	X2	.500	.118	.816	4.239	.002

a. Dependent Variable: Y2

Coefficients[a]

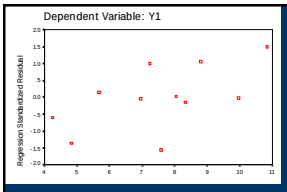
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.008	1.124		2.675	.025
	X3	.499	.118	.816	4.237	.002

a. Dependent Variable: Y3

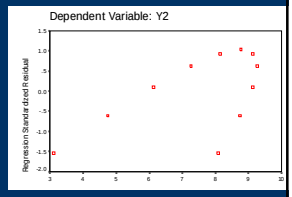
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-14.995	5.315		-2.821	.020
	X4	2.750	.648	.817	4.243	.002

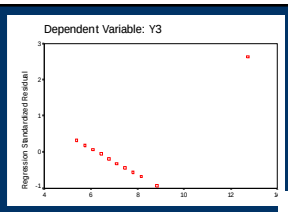
a. Dependent Variable: Y4



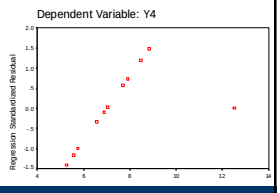
第一组数据的残差图有序列自相关的嫌疑.



第二组数据的残差图有曲线形式



第三组和第四组数据中都明显存在一个异常值.



这四个模型都需要进一步的分析.

一元线性回归应用步骤:

1．首先根据研究目的确定因变量y和自变量X；

2．收集数据后，首先进行相关分析，大致确定回归方程（线性或通过变换得到线性模型）；

3．建立一元线性回归模型，根据样本建立数据表；

4．用ＳＰＳＳ统计软件对样本数据进行分析，得到相应的输出结果。

5．对照输出结果对回归模型进行检验：包括拟合优度的检验（R平方）、回归系数的检验Ｐ值、回归方程的检验（方差分析表）、残差分析、寻找异常值等。

6．通过检验的回归模型称为有效方程，可以在实际中应用，也可以用于预测和控制问题。

§4.7 可化为线性的曲线函数

一、几种可划为线性的曲线函数

1. 抛物线函数（Quadratic 二次函数） $y=a+bX+cX^2$，经变换得：

$$y = a + bx_1 + cx_2$$

其中：$X_1 = X$，　$X_2 = X^2$，故方程变换后成为一个线性方程。

2. 双曲线函数（Inverse逆函数）$y=a+b/X$　做变换得:

$$y=a+bX_1 \qquad X_1=1/X$$

3．幂函数（Power）$y=aX^b$

令　$\ln y=\ln a+b\ln X$

即：　$y'=a'+bX'$;其中$y'=\ln y\ a'=\ln a\ X'=\ln X$

参数b是y对于自变量X的弹性，是X变动1%时, y变动的百分比。

4. 指数函数（Exponential）

$y = ae^{bx}$　　$\ln y = \ln a + bx$

5. 逆函数（Inverse）

$y = b_0 + (b_1 / x)$　　$y = b_0 + b_1 x'$　　$x' = 1/x$

6. 对数函数（Logarithmic）

$y = a + b\log x$

7. 逻辑函数（Logistic）

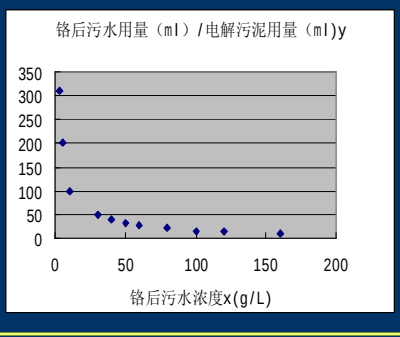
$$y = \frac{L}{1 + ae^{-bx}} \quad (L, a, b > 0)$$

$$ae^{-bx} = \frac{L}{y} - 1 \qquad \ln(\frac{L}{y} - 1) = \ln a - bx$$

实例分析 ： 某厂表面处理车间试验将铬后污水同电解质污泥混合生成无毒溶液。但实际排污的浓度不完全相同，且一定浓度的定量铬后污水只有同定量的电解污泥混合后，才能反应完全。现通过试验，找出铬后污水用量与电解污泥用量之比对于铬后污水浓度之间的关系，试验数据如下：

序号	铬后污水浓度x（g/L）	铬后污水用量（ml）/电解污泥用量（ml）y
1	3	310
2	5	200
3	10	100
4	30	49
5	40	40
6	50	32
7	60	28
8	80	23
9	100	16
10	120	14
11	160	10

解：首先根据数据得到散点图如下：



根据散点图的形状，试用幂函数进行曲线拟合。

假设曲线为幂函数，

由$y = aX^b$　对数变换 $\ln y = \ln a + b \ln X$

取$y' = \ln y$　　$x' = \ln x$　　$a' = \ln a$

以变换后计算得：

$L_{x'x'} = 17.462$，　$L_{y'y'} = 12.003$，$L_{x'y'} = -14.429$

由此得回归系数为：

$$\hat{b} = \frac{L_{x'y'}}{L_{x'x'}} = -0.8263, \quad a' = 6.6417$$

$\hat{a} = e^{a'} = 766.4$

回归方程为：　$Y = 766.4x^{-0.8263}$

对于线性函数应用相关系数表达其线性相关程度，对于非线性函数也可以用相关指数来反映一元非线性回归方程的优劣。

相关指数的定义为：

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{L_{y'y'}}$$

由例中的计算得相关指数为：

$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{L_{y'y'}} = 1 - \frac{0.080}{12.003} = 0.993$$

$L_{yy} = \sum (y_i - \bar{y})^2$

由此可见，曲线的拟合程度是很高的。可以用SPSS作多种函数的假设。

将变量X、Y进行对数变换，应用SPSS软件进行线性回归分析得：

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.997[a]	.993	.993	.09432	1.239

a. Predictors: (Constant), LNX

b. Dependent Variable: LNY

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11.923	1	11.923	1340.206	.000[a]
	Residual	.080	9	.009		
	Total	12.003	10			

a. Predictors: (Constant), LNX

b. Dependent Variable: LNY

回归系数表

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.642	.085		78.280	.000
	LNX	-.826	.023	-.997	-36.609	.000

a. Dependent Variable: LNY

下面用SPSS曲线估计的方法寻找合适的拟合曲线。

Inverse 逆函数

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.999	.997	.997	5.392

The independent variable is 铬后污水浓度（g/L).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	91302.512	1	91302.512	3140.302	.000
Residual	261.670	9	29.074		
Total	91564.182	10			

The independent variable is 铬后污水浓度（g/L).

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 / 铬后污水浓度（g/L)	908.460	16.211	.999	56.038	.000
(Constant)	11.514	1.979		5.819	.000

逆函数的表达式为：

$$y = 11.51 + 908.46 / x$$

Power 幂函数

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.997	.993	.993	.094

The independent variable is 铬后污水浓度（g/L).

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	11.923	1	11.923	1340.206	.000
Residual	.080	9	.009		
Total	12.003	10			

The independent variable is 铬后污水浓度（g/L).

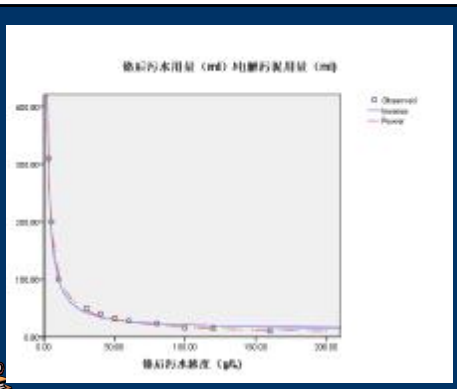
Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
ln(铬后污水浓度（g/L))	-.826	.023	-.997	-36.609	.000
(Constant)	766.418	65.027		11.786	.000

The dependent variable is ln(铬后污水用量（ml）/电解污泥用量（ml)).

幂函数的表达式为：

$$y = 766.418\, x^{-0.826}$$



一元线性回归小结：

1．一元线性回归模型的数学公式（总体与样本），要求掌握如何利用样本建立回归模型，注意理解高斯假设的意义。

2．理解线性回归的意义，明确回归分析与相关分析的关系。

3．掌握回归系数的求解方法——最小二乘法。

$$\hat{b}_1 = \frac{L_{xy}}{L_{xx}} = \frac{n\sum X_t Y_t - \sum X_t \sum Y_t}{n\sum X_t^2 - (\sum X_t)^2},\quad \hat{b}_0 = \frac{\sum Y_t}{n} - \hat{b}_1 \frac{\sum X_t}{n} = \bar{Y} - \hat{b}_1 \bar{X}$$

4．估计总体方差 σ^2 ：

计算中常用S^2估计 σ^2。

$$S^2 = \frac{\sum e_t^2}{n-2}$$

5．$\hat{b}_0, \hat{b}_1$ 的分布

由假设得出Y服从正态分布，且有： $y \sim N(b_0 + b_1 x, s^2)$

可以证明， $\hat{b}_0 \sim N(b_0, s^2_{\hat{b}_0})$； $\hat{b}_1 \sim N(b_1, s^2_{\hat{b}_1})$

其中：

$$s^2_{\hat{b}_0} = s^2\left(\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}\right) \qquad s^2_{\hat{b}_1} = \frac{s^2}{L_{xx}}$$

由样本得出回归系数方差的无偏估计为：

$$\hat{s}^2_{\hat{b}_0} = S^2_{\hat{b}_0} = S^2\left(\frac{1}{n} + \frac{\bar{X}^2}{L_{xx}}\right) \qquad \hat{s}^2_{\hat{b}_1} = S^2_{\hat{b}_1} = \frac{S^2}{L_{xx}}$$

6．模型检验

Ø **拟合程度的检验**：利用可决系数r^2对回归方程拟合程度的综合测量，可决系数越大，模型的拟合程度就越高，反之，可决系数越小，模型的拟合程度就越差:

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Ø **回归系数的检验（t 检验）**

提出假设：H_0：b_1=0； H_1：$b_1 \neq 0$

计算检验统计量值

$$t_{\hat{b}_1} = \frac{\hat{b}_1}{S_{\hat{b}_1}} = \frac{\hat{b}_1}{S/\sqrt{L_{xx}}}$$

对于给定的显著性水平a；查t分布表确定临界值(或用P值)。要注意当假设是双边检验比较临界值与t值后作出判断。

Ø **回归方程的检验（F检验）**

提出假设：H_0：b_1=0； H_1：$b_1 \neq 0$

计算离差平方和，列出方差分析表如下：

离差名称	离差平方和	自由度	均方差	F值
回归平方和	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	1	SSR	$F = \frac{SSR}{SSE/(n-2)}$
残差平方和	$SSE = \sum e_i^2$	n-2	SSE/(n-2)	
总离差平方和	$SST = \sum (Y_i - \bar{Y})^2$	n-1		

对于给定的显著性水平a，查F分布表确定临界值(或用P值)，当$F>F_a$时,拒绝原假设。

Ø **残差分析**

用因变量或者自变量做横轴，残差(标准化残差,学生化残差等)为纵轴，得到残差图,通过残差图是否随机排列和残差值的范围考察回归模型是否存在异方差性,是否存在异常值等.

7．利用回归方程进行预测

Ø **预测公式：**

由简单回归模型得出基本预测公式为：

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 X_0$$

·点预测：即给定X_0得到Y_0的预测值。

注意：内插与外推预测的区别。

· y_0的区间预测：

在总体标准差 σ^2未知的情况下，用其无偏估计S^2来代替，可以证明：

$$(y_0 - \hat{y}_0)/S_{e_0} \sim t(n-2)$$

由置信区间的确定方法得y_0的置信度为（1-a）的置信区间为

$$(\hat{y}_0 \pm t_{a/2}(n-2) \cdot S_{e_0})$$

· 预测的标准误S_{e0}：在标准假定下，有

$$S_{e_0} = s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}} \approx S\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$