北京农业科学

回归分析与回归设计
——在肥料与栽培试验中的应用

专 辑

1984

北京市农林科学院情报资料室

《北京农业科学》

（月　　刊）

1984年

编辑出版：《北京农业科学》编辑部
地　　址：北京市西郊板井村
印刷装订：北京市农林科学院情报室
发　　行：限国内发行
期刊登记：北京市期刊登记证第37号

目　　录

第一章　一元线性回归

第一节　回归分析能解决哪些问题……（ 1 ）
第二节　用选点法建立一元线性回归方程……（ 1 ）
第三节　用平均值法求一元线性回归方程……（ 3 ）
第四节　用最小二乘法求一元线性回归方程……（ 5 ）
第五节　回归方程的方差分析……（ 8 ）
第六节　一元线性回归方程的简化计算法……（ 10 ）
第七节　全部处理有重复的回归方程……（ 13 ）
第八节　部分处理有重复的回归方程……（ 17 ）
第九节　回归方程的预报与控制……（ 19 ）
第十节　两个回归方程的比较……（ 21 ）
第十一节　协方差分析……（ 23 ）

第二章　曲线回归

第一节　$Y=\frac{1}{a+bx}$型的曲线回归方程……（ 29 ）
第二节　$Y=\frac{x}{b+ax}$型的曲线回归方程……（ 31 ）
第三节　$Y=a+b\log X$型的曲线回归方程 ……（ 33 ）
第四节　$Y=dx^b$ 型的曲线回归方程……（ 35 ）
第五节　$Y=ab^x$ 型的曲线回归方程……（ 38 ）
第六节　$Y=ab^{\frac{1}{x}}$ 型的曲线回归方程 ……（ 40 ）
第七节　实验数据的修匀……（ 43 ）
第八节　$Y=c+\frac{x}{b+ax}$型的曲线回归方程……（ 50 ）
第九节　$Y=c+dx^b$ 型的曲线回归方程 ……（ 52 ）
第十节　$Y=c+ab^x$ 型的曲线回归方程……（ 55 ）

第三章　多元回归与多项式回归

第一节　二元一次线性回归方程……（ 56 ）
第二节　多元一次回归方程（解法一）……（ 60 ）
第三节　多元一次回归方程（解法二）……（ 64 ）

第四节　回归系数的显著性检验……（ 72 ）
第五节　各个自变量在多元回归中的作用……（ 74 ）
第六节　一元二次回归方程……（ 76 ）
第七节　二元二次回归方程……（ 80 ）
第八节　正交多项式回归方程……（ 83 ）
第九节　多元正交多项式回归方程……（ 89 ）

第四章　回归设计

第一节　正交设计基础知识……（ 94 ）
第二节　一次回归正交设计……（ 99 ）
第三节　快速登高试验计划……（ 105 ）
第四节　二次回归正交设计……（ 107 ）
第五节　三因素二次旋转设计……（ 116 ）
第六节　回归设计中采用其他编码尺度……（ 125 ）
第七节　二因素最优设计……（ 131 ）
第八节　六因素近似最优混合设计……（ 134 ）

附表：

附表一：F分布表
附表二：t分布表
附表三：正交多项式表
附表四：正交表
附表五：混合设计的X表与C表

第一章　一元线性回归

第一节　回归分析能解决哪些问题

1.确定二组或二组以上相对应的变量之间的相关关系，找出这些变量之间的定量关系式。

2.对这个定量关系式的可靠性进行统计检验。

3.根据可靠的变量之间的定量关系式作出预报和控制。

4.对多因素问题进行因素分析，确定各个因素之间的主次关系。

5.应用回归分析的原理，作出试验处理少、统计性质好的试验计划。

本章只涉及两组变量之间线性定量关系式的建立、统计检验与预报等问题。关于曲线回归、多因素之间的回归问题以及回归设计将在以后各章中叙述。

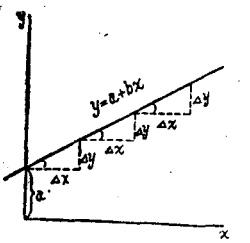
第二节　用选点法建立一元线性回归方程

1.所谓两个变量之间的关系，有相互影响的平行关系，如株高与干物重之间就是互为影响的因果关系。也有一个变量的取值，取决于另一个变量处于什么状态。例如产量与施肥量之间就只是施肥量对产量发生影响。把施肥量这一类的因素称为自变量或独立变量。把受独立变量影响的因素（如产量）称为倚变量或依靠变量。从独立变量来估测依靠变量就是所谓回归问题。对于具有平行关系的两个变量，也可以求出双方向的回归问题，即由因素x来估测因素y，也可以由因素y来估测因素x。

一元线性回归是最简单的回归问题。因素y的增量与因素x的增量的比例是一个固定的定量。也就是因素x增加多少，因素y也按一定比例增加若干，因此可以用一个直线方程来表达因素x与因素y之间的数量关系（如图1.1）

$$Y=a+bx$$

上式中a就是当x的取值为零时的y值。例如不施肥（即x＝0时）也有一定产量，这个产量就是a。在图1.1上a就是回归直线Y＝a＋bx与Y轴的交点，称为纵轴截距。很明显，b就是回归直线的斜率，称为回归系数。它反映了两个因素变化的比率。



（图1.1）

在一个回归方程中a值愈大，说明当x等于零时的y值愈大，譬如说，通过二个施肥量试验分别建立了两个线性回归方程，其中a值较大的方程的试验地地力可能要高些。如果b值较大，它对y的影响更为强烈些。因为当x的取值相同时，乘以较大的b值，再加到a值上去，也会得到更大的y值。由此可见b值的大小反映了因素x的作用强度。假如b为负值，x乘以一个负值再加到a值上去，就会使y值降低，因此b值的正负号决定了因素x的作用方向。

通过科学实验或调查研究取得一系列配对的数据后，就可以着手建立两个因素之间的回归方程。建立回归方程的方法有好多种，选点法是最简单的一种。

2.统计方法

以大豆试验叶片中含N量（mg/dm²）与碳水化合物（mg/dm²）测定数据为例（表1.1）

表1·1

样品号	X N(mg/dm²)	Y CH_2O(mg/dm²)	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	21.76	43.48	53.29	−9.81	96.20
2	18.68	46.08	46.43	0.65	0.42
3	16.35	38.19	38.49	−1.30	1.70
4	24.73	61.33	60.86	0.47	0.22
5	20.10	57.00	49.06	7.95	63.12
6	21.66	53.63	53.03	0.60	0.36
7	23.03	56.58	56.53	0.05	0.00
8	18.21	48.77	44.24	4.53	20.53
9	17.13	43.71	41.48	2.23	4.97
10	14.14	32.73	33.86	−1.10	1.20
11	15.05	35.84	36.18	−0.34	0.11
12	17.19	41.17	41.63	−0.46	0.22
13	14.59	41.25	35.00	6.25	39.01
14	16.76	41.23	40.54	0.69	0.48
15	16.48	41.16	39.82	1.34	1.78
16	13.50	31.92	32.23	−0.31	0.09
17	15.35	39.83	36.94	2.89	8.34
18	17.50	48.06	42.43	5.64	31.75
Σ	322.21	801.99	282.04	19.95	270.53

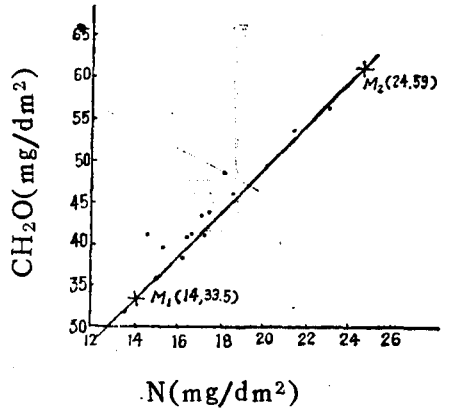


图 1.2

试求从叶片含N量估测叶片碳水化合物含量的回归方程。

先将数据标点在以x（含N量）为横轴，y（含CH_2O量）为纵轴的坐标纸上，可以看出这些点基本上自左下角向右上角比较集中地分布在一条直线的附近，可以用一根透明直尺在图上移动，尽量使尺的上下所分布的点数约各占一半，画出这根直线。（图1.2）

在直线两端各任取一个点M_1（x_1，y_1）与M_2(x_2,y_2)，读出M_1的坐标为（14,33.5），M_2的坐标为（24,59）。将M_1、M_2代入直线方程的两点式中：

$$\frac{\hat{y}-y_1}{x-x_1}=\frac{y_2-y_1}{x_2-x_1} \qquad (1.2)$$

得

$$\frac{\hat{y}-33.5}{x-14}=\frac{59-33.5}{24-14}=\frac{25.5}{10}$$

解得 $\hat{y}=2.55x-2.2$ (1.3)

即 $a=-2.2 \quad b=2.55$

上式 $\hat{y}$ 即为用回归方程计算出的y的估计值。将表1.1中18个点的x的数值代入式1.3，即可求出各个 $\hat{y}$ 值，填入表 1.1 的 $\hat{y}$ 一列之下。

每个观察值y与估计值 $\hat{y}$ 的差（$y-\hat{y}$）称为偏差或残差。它反映了回归方程的好坏。从图上直观地看出，这个回归直线大体上与各点分布趋势是一致的。

衡量 $\hat{y}$ 偏离y的偏差实际上用的是剩余标准差。表1.1最后一列的 $(y-\hat{y})^2$ 的总计值 $\Sigma(y-\hat{y})^2=270.53$ 称为剩余平方和，于是就可以从下式求出剩余标准差S：

$$S=\sqrt{\frac{\Sigma(y-\hat{y})^2}{N-2}}=\sqrt{\frac{270.53}{18-2}}=4.11 \qquad (1.4)$$

上式中N－2为剩余自由度，因全试验共18个叶片样品，即N＝18，故剩余自由度为 16。

3.注解与说明：

两个因素之间的回归分析，在科研中是经常遇到的。选点法是一个比较简单，又最粗放的方法，它具有很大的主观随意性，人为目测画的回归线会因人而异地发生很大偏差，所以只有在对回归关系分析精度要求不高时应用。当观察数据比较分散的情况下，所得回归方程的精度更低。

第三节　用平均值法求一元线性回归方程

1．为了克服选点法的主观随意性，可把全部数据分为两组，使两组所占的数据数目大体相等。将有关数据代入以下联立方程组。

$$\begin{cases}\Sigma\hat{y}_i'=ra+b\Sigma x_i' \\ \Sigma\hat{y}_i''=sa+b\Sigma x_i''\end{cases} \qquad (1.5)$$

式中r及s分别为两组配对数据的数目。

解这个联立方程组即可求出a与b。

2.统计方法：仍以上例将第1—9号样品的x数据相加得 $\Sigma x_i'$；第10—18号样品的x数据相加得 $\Sigma x_i''$；再分别将1—9号及10—18号样品的y数据相加得 $\Sigma y_i'$ 及 $\Sigma y_i''$。有关数据代入式（1.5），得

$$\begin{cases}448.77=9a+b\times 181.65 \\ 353.72=9a+b\times 140.56\end{cases}$$

解得 $a=2.93 \quad b=2.33$

回归方程为

$$\hat{y}=2.93+2.33x \qquad (1.6)$$

用式（1.6）也可以求出各个$\hat{y}$，列在表1.2中。

表1·2

样品号	X N(mg/bm²)	Y CH_2O(mg/bm²)	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	21.76	43.48	53.63	−10.15	103.04
2	18.68	46.08	46.45	− 0.37	0.14
3	16.35	38.19	41.03	− 2.84	8.04
4	24.73	61.33	60.55	0.78	0.61
5	20.10	57.00	49.76	7.24	12.37
6	21.66	53.63	53.40	0.23	0.05
7	23.03	56.58	56.59	− 0.01	0
8	18.21	48.77	45.36	3.41	11.63
9	17.13	43.71	42.84	0.87	0.75
10	14.14	32.76	35.88	− 3.12	9.71
11	15.05	35.84	38.00	− 2.16	4.65
12	17.19	41.17	42.98	− 1.81	3.29
13	14.59	41.25	36.92	4.33	18.71
14	16.76	41.23	41.98	− 0.75	0.56
15	16.48	41.16	41.33	− 0.17	0.03
16	13.50	31.92	34.39	− 2.47	6.08
17	15.35	39.83	38.70	1.13	1.29
18	17.50	48.06	43.71	4.36	18.97
Σ	322.21	801.99	803.49	− 1.5	239.92

同样可求出偏差（$y-\hat{y}$）及偏差平方和$\Sigma(y-\hat{y})^2$，并用式（1.4）求出剩余标准差

$$S=\sqrt{\frac{239.92}{18-2}}=3.87$$

可见平均值法的误差比选点法小。

3.注解与说明：如果把表1.2中的x，y原始数据，根据x由小到大的程序排个队再来计算如表1.3

再在样品9与12之间划为二组，则得

$$\begin{cases}345.89=9a+b\ (139.35)\\ 456.10=9a+b\ (182.86)\end{cases}$$

解得$a=-0.97 \quad b=2.53$

得$\hat{y}=2.53x-0.79$

表1·3

样品号	X N(mg/am²)	Y CH_2O(ma/am²)	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
16	13.50	31.92	33.37	−1.45	2.10
10	14.14	32.76	34.98	−2.22	4.93
13	14.59	41.25	36.12	5.13	26.32
11	15.05	35.84	37.29	−1.45	2.10
17	15.35	39.83	38.05	1.78	3.17
3	16.35	38.19	40.58	−2.39	5.71
15	16.48	41.16	40.90	0.26	0.07
14	16.76	41.23	41.61	−0.38	0.14
9	17.13	43.71	42.55	1.16	1.35
12	17.19	41.17	42.70	−1.53	2.34
18	17.50	48.06	43.49	4.57	20.98
8	18.21	48.77	45.28	3.49	12.18
2	18.68	46.08	46.47	−0.39	0.15
5	20.10	57.00	50.06	6.94	48.16
6	21.66	53.63	54.01	−0.38	0.14
1	21.76	43.48	54.26	−10.78	116.21
7	23.03	56.58	57.48	−0.90	0.81
4	24.73	61.33	61.78	−0.45	0.20
Σ	322.21	800.98	800.98	−1.02	247.06

用此式求出各个$\hat{y}$，$(y-\hat{y})$及$\Sigma(y-\hat{y})^2$填入表1·3中，也可以求出剩余标准差。

$$S=\sqrt{\frac{247.06}{16}}=3.93$$

当分组的样品发生变化时，可以求出略为不同的回归方程，可见本法优于选点法，但也不是很准确的。

第四节　用最小二乘法求一元线性回归方程

1.原理：要使回归方程$\hat{y}=a+bx$更好地反映出x与y两个因素在数量上的互变关系，应使y与$\hat{y}$的偏差$y-\hat{y}$尽可能小。一个试验中有若干对x与y的配对数据，而所有数据偏差之和等于零，即$\Sigma(y-\hat{y})=0$，根据最小二乘法必须在偏差平方和$\Sigma(y-\hat{y})^2$为最小值的前提下来求出纵轴截距a与回归系数b。因此必须使

$$Q=\sum_{1}^{N}(y-\hat{y})^2=\sum_{1}^{N}(y-a-bx)^2$$
$$=\Sigma y^2+Na^2+\Sigma(bx)^2-2a\Sigma y-2b\Sigma xy+2ab\Sigma x$$
$$=\text{最小} \qquad (1.7)$$

上式Q为偏差平方和，它是a、b的二元函数，为了求出Q的最小值，需要按微分公式

$$U^N=NU^{N-1} \qquad (1.8)$$

求出a、b的偏导数，并使其为零

$$\begin{cases}\dfrac{\partial Q}{\partial a}=-2\Sigma(y-a-bx)=0\\[2ex]\dfrac{\partial Q}{\partial b}=-2\Sigma(y-a-bx)x=0\end{cases} \qquad (1.9)$$

由式（1.9）得以下两个正规方程

$$\begin{cases}aN+b\Sigma x=\Sigma y & (A)\\ a\Sigma x+b\Sigma x^2=\Sigma xy & (B)\end{cases}$$

由（A）式：

$$a=\frac{\Sigma y-b\Sigma x}{N}=\frac{\Sigma y}{N}-\frac{b\Sigma x}{N}=\bar{y}-b\bar{x} \qquad (1.10)$$

将a代入（B）式得

$$\left(\frac{\Sigma y}{N}-\frac{b\Sigma x}{N}\right)\Sigma x+b\Sigma x^2=\Sigma xy$$

$$\frac{(\Sigma x)(\Sigma y)}{N}-\frac{b(\Sigma x)^2}{N}+b\Sigma x^2=\Sigma xy$$

移项：

$$b\Sigma x^2-\frac{b(\Sigma x)^2}{N}=\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{N}$$

$$b\left[\Sigma x^2-\frac{(\Sigma x)^2}{N}\right]=\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{N}$$

$$b=\frac{\Sigma xy-\dfrac{(\Sigma x)(\Sigma y)}{N}}{\Sigma x^2-\dfrac{(\Sigma x)^2}{N}}$$

令

$$l_{xy}=\Sigma xy-\frac{(\Sigma x)(\Sigma y)}{N}$$

$$l_{xx}=\Sigma x^2-\frac{(\Sigma x)^2}{N}$$

则

$$b=\frac{l_{xy}}{l_{xx}} \qquad (1.11)$$

式（1.10），（1.11）即为求a、b的常用公式

2、用最小二乘法求一元线性回归方程的方法：一般都列成表格计算。欲求出a与b，必先求出Σx、Σy、Σxy、Σx^2等总数，为了可靠性检验的需要，也需算出Σy^2，一般都在表格的下方进一步计算出lxx、lxy及lyy，其中

$$lyy=\Sigma y^2-\frac{(\Sigma y)^2}{N} \qquad (1.12)$$

表1·4

样品号	X N(mg/dm²)	Y CH_2O(mg/dm²)	X^2	Y^2	XY	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	21.76	43.48	473.50	1890.51	946.12	53.55	−10.07	101.42
2	18.68	46.08	348.94	2123.37	860.77	46.37	−0.29	0.09
3	16.35	38.19	267.32	1458.48	624.41	40.95	−2.76	7.59
4	24.73	61.33	611.57	3761.37	1516.69	60.47	0.86	0.74
5	20.10	57.00	404.01	3249.00	1145.70	49.68	7.32	53.54
6	21.66	53.63	469.16	2876.18	1161.63	53.32	0.31	0.10
7	23.03	56.58	530.38	3201.30	1303.04	56.51	0.07	0.00
8	18.21	48.77	331.60	2378.51	888.10	45.28	3.49	12.18
9	17.13	43.71	293.44	1910.56	748.75	42.76	0.95	0.90
10	14.14	32.76	199.94	1073.22	463.23	35.80	−3.04	9.22
11	15.05	35.84	226.50	1284.51	539.39	37.92	−2.08	4.31
12	17.19	41.17	295.50	1694.97	707.71	42.90	−1.73	3.00
13	14.59	41.25	212.89	1701.56	601.84	36.84	4.41	19.41
14	16.76	41.23	280.90	1699.91	691.01	41.90	−0.67	0.45
15	16.48	41.16	271.59	1694.15	678.32	41.25	−0.09	0.01
16	13.50	31.92	182.25	1018.89	430.92	34.31	−2.39	5.69
17	15.35	39.83	235.62	1586.43	611.39	38.62	1.21	1.48
18	17.50	48.06	306.25	2309.76	841.05	43.63	4.44	19.67
Σ	322.21	801.99	5941.34	36912.66	14760.07	802.05	−0.06	239.79

$\overline{X}=17.90$　　$\overline{Y}=44.56$　　$N=18$

$\Sigma X=322.21$　　$\Sigma Y=801.99$　　$\Sigma XY=14760.07$

$\Sigma X^2=5941.34$　　$\Sigma Y^2=36912.66$　　$(\Sigma X)(\Sigma Y)/N=14356.07$

$(\Sigma X)^2/N=5767.74$　　$(\Sigma Y)^2/N=35732.66$

$lxx=173.60$　　$lyy=1180$　　$lxy=404.01$

将以上各项数据代入式（1.11）及（1.10）

$$b=\frac{lxy}{lxx}=\frac{404.01}{173.60}=2.33$$

$$a=\bar{y}-b\bar{x}=44.56-2.33(17.90)=2.85$$

得回归方程为

$$\hat{y}=2.85+2.33x$$

以上式将各个x数据代入，求出理论估计值 $\hat{y}$，$y-\hat{y}$ 及 $\Sigma(y-\hat{y})^2$，均列入式1.4，并计算出剩余标准差 S

$$S=\sqrt{\frac{239.79}{16}}=3.87$$

剩余标准差比平均值法小。

第五节　回归方程的方差分析

任何两组对应的变量，不管其是否存在着线性回归关系，总是可以用最小二乘法求出它们的线性回归方程的。求出了回归方程并不等于就存在着显著的回归关系，因此对求出的回归方程需要进行显著性检验。

要对回归方程进行统计检验，先要了解数据波动的性质。每一个试验数据的取值，是由两方面原因引起的。第一，自变量x取不同数值（x_1、x_2、x_3…，x_u）对y产生的影响。第二，试验里未加控制的因素和随机误差共同的影响。进行方差分析必须将这两类造成数据波动的影响从试验数据y中分解开来。

试验数据的波动情况称为变差。变差的大小是通过与变差的平均值的比较来衡量的，即用y与它的平均值 $\bar{y}$ 的差（$y-\bar{y}$）来度量变差，（$y-\bar{y}$）称为离差。由于离差有正值与负值，所以 $\Sigma(y-\bar{y})=0$。为了克服这个障碍，改用离差平方和来表示总的变差，离差平方简称总平方和，并记作 l_{yy}

$$l_{yy}=\Sigma(y-\bar{y})^2 \qquad (1.13)$$

从下图可以看出，$y-\bar{y}$ 由二部分组成，

$$y-\bar{y}=(y-\hat{y})+(\hat{y}-\bar{y}) \qquad (1.14)$$

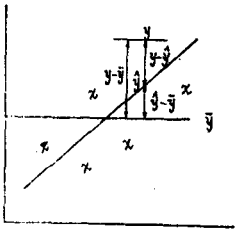


图.13

上式平方并对各点求和，得

$$\Sigma(y-\bar{y})^2=\Sigma[(y-\hat{y})+(\hat{y}-\bar{y})]^2$$
$$=\Sigma(y-\hat{y})^2+\Sigma(\hat{y}-\bar{y})^2$$
$$+2\Sigma(y-\hat{y})(\hat{y}-\bar{y})$$

根据式（1.9）的性质，$2\Sigma(y-\hat{y})(\hat{y}-\bar{y})=0$，所以

$$\Sigma(y-\bar{y})^2=\Sigma(y-\hat{y})^2+\Sigma(\hat{y}-\bar{y})^2 \qquad (1.15)$$

总平方和 $\Sigma(y-\bar{y})^2$ 分解为两项，其中 $\Sigma(\hat{y}-\bar{y})^2$ 是回归值 $\hat{y}$ 与平均值 $\bar{y}$ 之差的平方和，它反映了试验因素x不同取值时对试验数据y的影响程度，称作回归平方和，记作u

$\Sigma(y-\hat{y})^2$ 刻划了试验数据y偏离回归线的距离，这个偏距是由试验中其它未加控制的因素与随机误差共同引起的。于是又把 $y-\hat{y}$ 称作偏差或残差，把 $\Sigma(y-\hat{y})^2$ 称作偏差平方和或剩余平方和，记作Q

计算回归平方和时要将公式 $u=\Sigma(\hat{y}-\bar{y})^2$ 改变一下形式。

由于 $\hat{y}=a+bx$，　　$\bar{y}=a+b\bar{x}$

代入回归平方和式中，得

$$u=\Sigma(\hat{y}-\bar{y})^2=\Sigma(a+bx-a-b\bar{x})^2$$
$$=b^2\Sigma(x-\bar{x})^2 \qquad (1.16)$$

式(1.16)中的$\Sigma(x-\bar{x})^2$即自变量x的平方和，可证明如下

$$\Sigma(x-\bar{x})^2=\Sigma(x^2-2x\bar{x}+\bar{x}^2)$$
$$=\Sigma x^2-2\Sigma x\frac{\Sigma x}{N}+N\left(\frac{\Sigma x}{N}\right)^2$$
$$=\Sigma x^2-\frac{(\Sigma x)^2}{N}=l_{xx}$$

已知$b=\frac{l_{xy}}{l_{xx}}$

所以

$$b^2\Sigma(x-\bar{x})=b\frac{l_{xy}}{l_{xx}}l_{xx}=bl_{xy} \qquad (1.17)$$

式(1.17)是通常计算回归平方和的公式，l_{xy}已在计算过程中算出，所以用式(1.17)很便利地可求得回归平方和

同理：总平方和$\Sigma(y-\bar{y})^2$亦可仿照$\Sigma(x-\bar{x})^2$即为l_{yy}，亦已在前计算出来。

前已证明$\Sigma(y-\bar{y})^2=\Sigma(y-\hat{y})^2+\Sigma(\hat{y}-\bar{y})^2$
$$=Q+U$$

已知 $(y-\bar{y})^2=l_{yy}$

所以 $l_{yy}=Q+U$

$$Q=l_{yy}-U=l_{yy}-bl_{xy} \qquad (1.18)$$

在一元线性回归中，只有一个自变量，故回归自由度为1，于是回归平方和就是回归方差。总自由度为N－1，N是试验数据的数目，于是剩余自由度为N－1－1＝N－2。剩余平方和与剩余自由度之比即剩余方差。

于是

$$F=\frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\hat{y})^2/N-2}=\frac{U}{Q/N-2} \qquad (1.19)$$

它服从于自由度为1与N－2的F分布。不难理解在总平方和之中，回归平方和所占的比重愈大，回归方程的效果愈好。回归方程的方差分析可概括如表1.5

表1·5

变异来源	平方和	自由度	均方	F
回归	$U=\Sigma(\hat{Y}-\bar{Y})^2=bl_{xy}$	1	U	$\frac{U}{Q/(N-2)}$
剩余	$Q=\Sigma(Y-\hat{Y})^2=l_{yy}-bl_{xy}$	N－2	Q/(N－2)	
总计	$l_{yy}=\Sigma(Y-\bar{Y})^2$	N－1		

对大豆叶片中N与CH_2O的回归方程作方差分析如下：

表1·6

变异来源	平　方　和	自由度	均　方	F
回　归	$b1xy=2.326\times404.19=940.14594$	1	940.14594	61.99
剩　余	$Q=1182.8208-940.14594=242.67492$	$18-2=16$	15.167182	
总　计	$lyy=1182.8208$	$18-1=17$		

查F表，当信度$\alpha=0.01$时F的临界值为8.53，回归方程极显著。

第六节　一元线性回归方程的简化计算法

现在常用电子计算器一般为八位或十位数字。有时x与y的数值很大，x^2、y^2、xy的数值常超过八位或十位，因此有必要对数据进行简化。简化数据的方法常是将原始数据减去某数再除以一个常数，使数据的位数减少。用简化后数据计算回归方程与原始数据计算所得十分近似。上例数据为四位有效数字，本不需简化，为演示计算方法并检验与原始数据计算所得回归方程是否近似，故仍以上例来作简化计算。

原始数据有两位小数，故乘以100变成整数，再减去平均数附近的一个整数，并除以某数即可简化数据。简化公式如下：

$$x'=\frac{e_1x-c_1}{d_1}\quad ;\quad y'=\frac{e_2y-c_2}{d_2} \tag{1.20}$$

设　$c_1=1800$，$d_1=80$，$e_1=100$

$c_2=4456$，$d_2=80$，$e_2=100$

$$b=\frac{l_{xy}}{l_{xx}}=\frac{404.19}{173.77}=2.326$$

$$a=\bar{y}-b\,\bar{x}=44.57-2.326(17.9)=2.93$$

回归方程为

$$y=2.93+2.326x$$

与用原始数据求出的十分接近。

在表1.7下方有将$lx'x'$及$lx'y'$还原为l_{xx}及l_{xy}的计算公式，其来原如下.

已定义

$$x'=\frac{e_1x-c_1}{d_1}\text{ 及 }\bar{x}'=\frac{e_1\bar{x}-c_1}{d_1}$$

两式相减　$(x'-\bar{x}')=\dfrac{e_1x-c_1}{d_1}-\dfrac{e_1\bar{x}-c_1}{d_1}=\dfrac{e_1(x-\bar{x})}{d_1}$

左右平方　$(x'-\bar{x}')^2=\dfrac{e_1^2(x-\bar{x})_2}{d_1^2}$

表1·7

样品号	X N(mg/dm²)	Y CH_2O(mg/dm²)	$X'=\frac{100x-1800}{80}$	$Y'=\frac{100y-4456}{80}$	$(X')^2$	$(Y')^2$	$X'Y'$
1	21.76	43.48	4.7	− 1.4	22.09	1.96	−6.58
2	18.68	46.08	0.9	1.9	0.81	3.61	1.71
3	16.35	38.19	−2.1	− 8.0	4.41	64.00	−16.8
4	24.73	61.33	8.4	21.0	70.56	441.00	176.4
5	20.10	57.00	2.6	15.6	6.76	243.36	40.56
6	21.66	53.63	4.6	11.3	21.16	127.69	51.98
7	23.03	56.58	6.3	15.0	39.69	225.00	94.50
8	18.21	48.77	0.3	5.3	0.09	28.09	1.59
9	17.13	43.71	−1.1	− 1.1	1.21	1.21	1.21
10	14.14	32.76	−4.8	−14.8	23.04	219.04	71.04
11	15.05	35.84	−3.7	−10.9	13.69	118.81	40.33
12	17.19	41.17	−1.0	− 4.2	1.00	17.64	4.20
13	14.59	41.25	−4.3	− 4.1	18.43	16.81	17.63
14	16.76	41.23	−1.6	− 4.2	2.56	17.64	6.72
15	16.48	41.16	−1.9	− 4.3	3.61	18.49	8.17
16	13.50	31.92	−5.6	−15.8	31.36	249.64	88.48
17	15.35	39.83	−3.3	− 5.9	10.89	34.81	19.47
18	17.50	48.06	−0.6	4.4	0.36	19.36	−2.64
Σ			−2.2	− 0.2	271.78	1848.16	631.57

$\Sigma x' = -2.2$　　$\Sigma y' = -0.2$　　$N = 18$

$\overline{x'} = 271.78$　　$\overline{y'} = -0.0111$　　$\Sigma x'y' = 63157$

$\Sigma (x')^2 = 271.78$

$(\Sigma x')^2/N = 0.2689$　　$\Sigma (y')^2 = 1848.16$　　$(\Sigma x')(\Sigma y')N = 0.0244$

$lx'x' = 271.5111$　　$(\Sigma y')^2/N = 0.0022$　　$lx'y' = 631.5456$

$$lxx = \frac{d_1^2 lx'x'}{e_1^2} = 17377$$

$$lxy = \frac{d_1 d_2 lx'y'}{e_1 e_2} = 404.19$$

$$e_1^2 (x - \overline{x})^2 = (x' - \overline{x'})^2 d_1^2$$

$$(x - \overline{x})^2 = \frac{(x' - \overline{x}^1)^2 d_1^2}{e_1^2}$$

对各点求和

$$\Sigma (x - \overline{x})^2 = \frac{\Sigma (x' - \overline{x}')^2 d_1^2}{e_1^2}$$

$$l_{xx}=\frac{l_{x'x'}d_1^2}{e_1^2} \tag{1.21}$$

同理：

$$y'=\frac{e_2y-c_2}{d_2},\qquad \bar{y}'=\frac{e_2\bar{y}-c_2}{d_2}$$

两式相减 $y'-\bar{y}'=\frac{e_2y-c_2}{d_2}-\frac{e_2\bar{y}-c_2}{d_2}=\frac{e_2(y-\bar{y})}{d_2}$

左右都平方 $(y'-\bar{y}')=\frac{e_2^2(y-\bar{y})^2}{d_2}$

$$e_2^2(y-\bar{y})=(y'-\bar{y}')^2d_2^2$$

$$(y-\bar{y})^2=\frac{(y'-\bar{y}')^2d_2^2}{e_2^2}$$

对各点求和 $\Sigma(y-\bar{y})^2=\frac{\Sigma(y'-\bar{y}')^2d_2^2}{e_2^2}$

即 $$l_{yy}=\frac{l_{y'y'}d_2^2}{e_2^2} \tag{1.22}$$

由于 $(x'-\bar{x}')=\frac{e_1(x-\bar{x})}{d_1}$ ； $(y'-\bar{y}')=\frac{e_2(y-\bar{y})}{d_2}$

两式相乘 $e_1(x-\bar{x})=(x'-\bar{x}')d_1$ ； $e_2(y-\bar{y})=(y'-\bar{y}')d_2$

$$e_1e_2(x-\bar{x})(y-\bar{y})=(x'-\bar{x}')(y'-\bar{y}')d_1d_2$$

$$(x-\bar{x})(y-\bar{y})=\frac{(x'-\bar{x}')(y'-\bar{y}')d_1d_2}{e_1e_2}$$

对各点求和

$$\Sigma(x-\bar{x})(y-\bar{y})=\frac{\Sigma(x'-\bar{x}')(y'-\bar{y}')d_1d_2}{e_1e_2}$$

即 $$l_{xy}=\frac{l_{x'y'}d_1d_2}{e_1e_2} \tag{1.23}$$

还可以用其他方法对原始数据进行简化，例如：

$$x'=(x-c_1)d_1;\quad y'=(y-c_2)d_2$$

则l_{xx}、l_{yy}、l_{xy}的复原公式如下：

$$l_{xy}=\frac{1}{d_1^2}l_{x'x'} \tag{1.24}$$

$$l_{yy}=\frac{1}{d_2^2}l_{y'y'} \tag{1.25}$$

$$l_{xy}=\frac{1}{d_1d_2}l_{x'y'} \tag{1.26}$$

第七节　全部处理有重复的回归方程

在不设重复的回归方程中，总平方和可分解为回归平方和与剩余平方和两项。剩余平方和反映了自变量x以外的其他未控因素与随机误差两方面对试验数据y共同造成的数据波动。如果从中将随机误差的影响分离出去，剩下来的就是其他未加控制的因素对试验数据y的影响。这部分的影响愈大，势必使观察值y偏离回归值$\hat{y}$愈远，称为x与y的拟合较差，并把这部分平方和称为失拟平方和，记作SS_{lf}

为了求出误差平方和，试验要设置重复，相同处理不同重复之间的变差反映了真正的误差。

一个试验有N个处理，每个处理重复m次，于是有N·m个试验数据。为了书写方便，把一元线性回归方程中所用的符号添加下标来区别不同的处理和重复

$Y_{\alpha i}=a+bX_{\alpha}$，$\alpha=1,2,\cdots,N$；$i=1,2,\cdots,m$式中$\alpha$代表处理，i代表重复。处理有1到N个，重复有1到m个。

与不设重复的试验相似，求解回归系数与纵轴截距的公式相同，只是用$\overline{Y}_{\alpha}$代替Y_{α}

$$a=\overline{y}-b\overline{x}$$

$$b=\frac{\sum_{\alpha}X_{\alpha}Y_{\alpha}-\frac{(\sum_{\alpha}X_{\alpha})(\sum_{\alpha}\overline{y}_{\alpha})}{N}}{\sum_{\alpha}X_{\alpha}^{2}-\frac{(\sum_{\alpha}X_{\alpha}^{2})}{N}} \qquad (1.27)$$

上式中

$$\overline{x}=\frac{1}{N}\sum_{\alpha}X_{\alpha} \qquad (1.28)$$

$$\overline{Y}=\frac{1}{Nm}\sum_{\alpha}\sum_{i}Y_{\alpha i} \qquad (1.29)$$

$$\overline{Y}_{a}=\frac{1}{m}\sum_{i}Y_{\alpha i} \qquad (1.30)$$

式（1.28）为所有自变量的平均数。式（1.29）为包括重复试验在内的$Y_{\alpha i}$的平均数，式中所用$\sum_{\alpha}\sum_{i}Y_{\alpha i}$符号意即把不同处理各个重复的$Y_{\alpha i}$数据都总加到一起，Nm即所有$Y_{\alpha i}$数据一共有Nm个。式（1.30）是某一个处理不同重复的平均数。

从式（1.27）看出有重复试验求b值，是以各个配对的X与$\overline{Y}_{\alpha}$数据来计算的，即用各处理的平均值来代替各个重复的原始数据即可。

现仍用大豆叶片中含N量与碳水化合物量作为例子，只是CH_2O含量有两次重复数据，各样品的数据如表1.8