**回归分析的基本思想及其初步应用公式-高中数学选修2-3第三章**

**探究**：对于一组具有线性相关关系的数据：

（) , () ，…， ()，

我们知道其回归方程的截距和斜率的最小二乘估计公式分别为：

 （1）

 （2）

其中，（）成为样本点的中心.

**注**：回归直线过样本中心.

你能推导出这两个计算公式吗？

从我们已经学过的知识知道，截距和斜率分别是使



取到最小值时的值.

由于







注意到







.





 在上式中，后两项和无关，而前两项为非负数，因此要使Q取得最小值，当且仅当前两项的值均为0，即有



这正是我们所要推导的公式．

下面我们从另一个角度来推导的公式．

人教A版选修2-2P37习题1.4A组第4题:

用测量工具测量某物体的长度，由于工具的精度以及测量技术的原因，测得n个数据

 .

**证明**：用这个数据的平均值

表示这个物体的长度，能使这n个数据的方差



最小．

**思考**：这个结果说明了什么？通过这个问题，你能说明最小二乘法的基本原理吗？

证明：由于，所以

，

令, 得。

可以得到， 是函数的极小值点，也是最小值点．

这个结果说明，用n个数据的平均值表示这个物体的长度是合理的，这就是最小二乘法的基本原理．

由最小二乘法的基本原理即得

**定理** 设,,则

 (\*)

当且仅当时取等号.

(\*)式说明, 是任何一个实数与的差的平方的平均数中最小的数.从而说明了方差具有最小性,也即定义标准差的合理性.

下面借助(\*)式求的最小值.



,

由(\*)式知,

















当且仅当,且时, 达到最小值

.

由此得到,其中是回归直线的斜率,是截距.

借助和配方法,我们给出了人教A版必修3的第二章统计第三节变量间的相关关系中回归直线方程的一个合理的解释.

**1、回归分析的基本步骤**：

(1) 画出两个变量的散点图.

(2) 求回归直线方程.

(3) 用回归直线方程进行预报.

下面我们通过案例，进一步学习回归分析的基本思想及其应用．

**2、举例**：

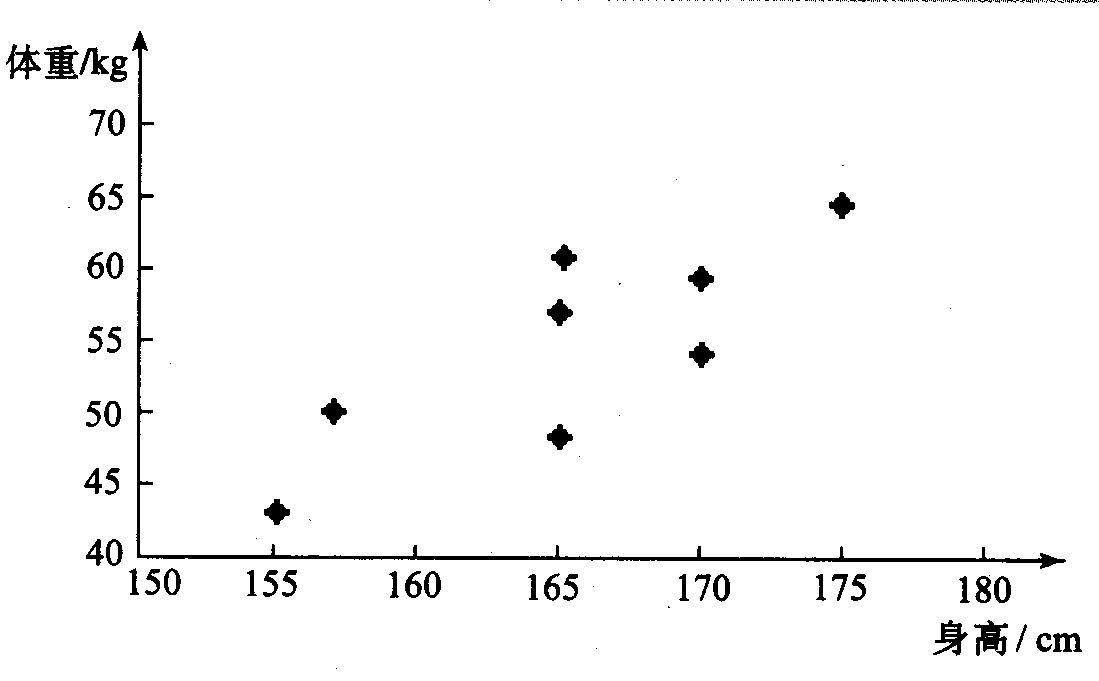
**例1**. 从某大学中随机选取 8 名女大学生，其身高和体重数据如表

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 身高/cm | 165 | 165 | 157 | 170 | 175 | 165 | 155 | 170 |
| 体重/kg | 48 | 57 | 50 | 54 | 64 | 61 | 43 | 59 |

求根据女大学生的身高预报体重的回归方程，并预报一名身高为 172 cm 的女大学生的体重．

解：由于问题中要求根据身高预报体重，因此选取身高为自变量 x ，体重为因变量 y .

作散点图(图3 . 1 一 1)



从图3. 1一1 中可以看出，样本点呈条状分布，身高和体重有比较好的线性相关关系，因此可以用线性回归方程来近似刻画它们之间的关系．

根据探究中的公式（1）和（2 ) ，可以得到.

于是得到回归方程

.

因此，对于身高172 cm 的女大学生，由回归方程可以预报其体重为

 ( kg ) .

是斜率的估计值，说明身高 x 每增加1个单位时，体重y就增加0.849 位，这表明体重与身高具有正的线性相关关系．如何描述它们之间线性相关关系的强弱？

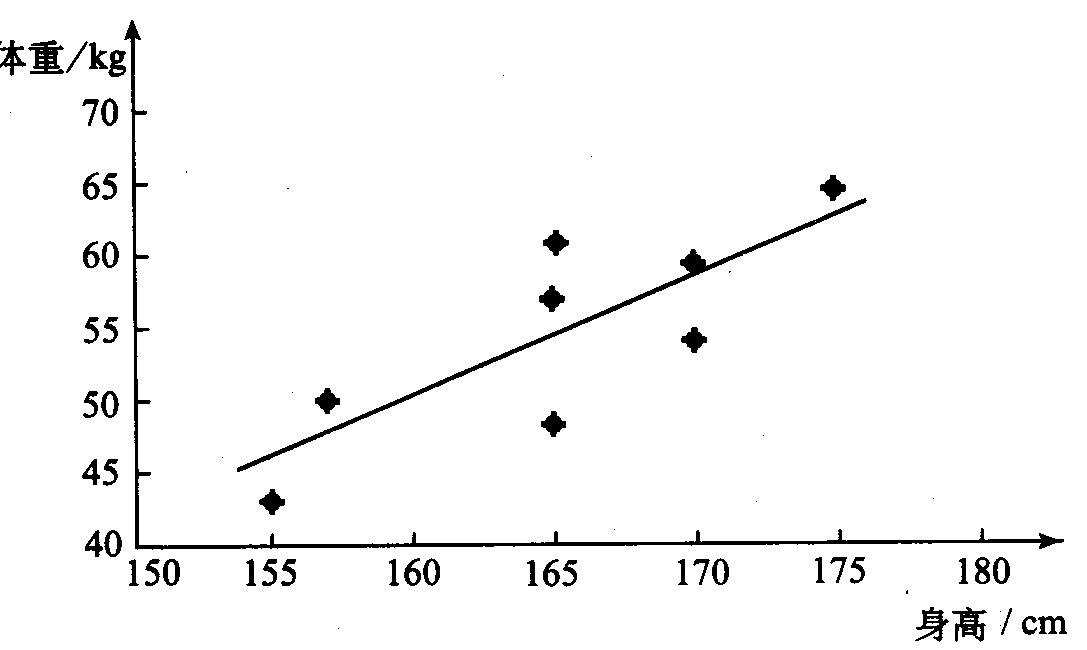
在必修 3 中，我们介绍了用相关系数；来衡量两个变量之间线性相关关系的方法．本相关系数的具体计算公式为



当r>0时，表明两个变量正相关；当r<0时，表明两个变量负相关．r的绝对值越接近1，表明两个变量的线性相关性越强；r的绝对值接近于0时，表明两个变量之间几乎不存在线性相关关系．通常，当r的绝对值大于0. 75 时认为两个变量有很强的线性相关关系．

在本例中，可以计算出r =0. 798．这表明体重与身高有很强的线性相关关系，从而也表明我们建立的回归模型是有意义的．

显然，身高172cm 的女大学生的体重不一定是60. 316 kg，但一般可以认为她的体重接近于60 . 316 kg .图3 . 1 一 2 中的样本点和回归直线的相互位置说明了这一点．



由于所有的样本点不共线，而只是散布在某一条直线的附近，所以身高和体重的关系可用下面的线性回归模型来表示：

, ( 3 )

这里 a 和 b 为模型的未知参数，e是 y 与之间的误差．通常e为随机变量，称为**随机误差**，它的均值 E (e）=0，方差D（e）=＞0 ．这样线性回归模型的完整表达式为：

 (4)

在线性回归模型（4）中，随机误差e的方差护越小，通过回归直线

 (5)

预报真实值y的精度越高．随机误差是引起预报值与真实值 y 之间的误差的原因之一，大小取决于随机误差的方差.

另一方面，由于公式（1）和（2）中 和为截距和斜率的估计值，它们与真实值a和b之间也存在误差，这种误差是引起预报值与真实值y之间误差的另一个原因．

**思考**:产生随机误差项e的原因是什么?

一个人的体重值除了受身高的影响外，还受许多其他因素的影响．例如饮食习惯、是否喜欢运动、度量误差等．事实上，我们无法知道身高和体重之间的确切关系是什么，这里只是利用线性回归方程来近似这种关系．这种近似以及上面提到的影响因素都是产生随机误差 e 的原因．

因为随机误差是随机变量，所以可以通过这个随机变量的数字特征来刻画它的一些总体特征．均值是反映随机变量取值平均水平的数字特征，方差是反映随机变量集中于均值程度的数字特征，而随机误差的均值为0，因此可以用方差来衡量随机误差的大小．

为了衡量预报的精度，需要估计护的值．一个自然的想法是通过样本方差来估计总体方差．如何得到随机变量的样本呢？由于模型（3）或（4）中的隐含在预报变量 y 中，我们无法精确地把它从 y 中分离出来，因此也就无法得到随机变量的样本．

解决问题的途径是通过样本的估计值来估计．根据截距和斜率的估计公式（1）和（2 ) , 可以建立回归方程

,

因此是（5）中的估计量．由于随机误差，所以是的估计量．对于样本点（) , () ，…， ()

而言，相应于它们的随机误差为

,

其估计值为

,

称为相应于点的**残差**（residual )．类比样本方差估计总体方差的思想，可以用



作为的估计量， 其中和由公式（1) (2）给出，Q（ ,）称为**残差平方和**（residual sum of squares )．可以用衡量回归方程的预报精度．通常，越小，预报精度越高．

在研究两个变量间的关系时，首先要根据散点图来粗略判断它们是否线性相关，是否可以用线性回归模型来拟合数据．然后，可以通过残差



来判断模型拟合的效果，判断原始数据中是否存在可疑数据．这方面的分析工作称为残差分析．表3一 2 列出了女大学生身高和体重的原始数据以及相应的残差数据．

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 身高/cm | 165 | 165 | 157 | 170 | 175 | 165 | 155 | 170 |
| 体重/kg | 48 | 57 | 50 | 54 | 64 | 61 | 43 | 59 |
| 残差 | -6.373 | 2.627 | 2.419 | -4.618 | 1.137 | 6.627 | -2.883 | 0.382 |

我们可以利用图形来分析残差特性作图时纵坐标为残差，横坐标可以选为样本编号，或身高数据，或体重的估计值等，这样作出的图形称为残差图．图 3 . 1 一 3 是以样本编号为横坐标的残差图．

从图3 . 1 一 3 中可以看出，第 1 个样本点和第 6 个样本点的残差比较大，需要确认在采集这两个样本点的过程中是否有人为的错误．如果数据采集有错误，就予以纠正，然后再重新利用线性回归模型拟合数据；如果数据采集没有错误，则需要寻找其他的原因．另外，残差点比较均匀地落在水平的带状区域中，说明选用的模型比较合适.这样的带状区域的宽度越窄，说明模型拟合精度越高，回归方程的预报精度越高．另外，我们还可以用相关指数来刻画回归的效果，其计算公式是：



显然，取值越大，意味着残差平方和越小，也就是说模型的拟合效果越好．在线性回归模型中，表示解释变量对于预报变量变化的贡献率． 越接近于1，表示回归的效果越好（因为越接近于1，表示解释变量和预报变量的线性相关性越强）．如果对某组数据可能采取几种不同的回归方程进行回归分析，也可以通过比较几个，选择大的模型作为这组数据的模型．

在例 1 中，=0. 64 ，表明“女大学生的身高解释了64 ％的体重变化”，或者说“女大学生的体重差异有 64 ％是由身高引起的”.

用身高预报体重时，需要注意下列问题：

1．回归方程只适用于我们所研究的样本的总体．例如，不能用女大学生的身高和体重之间的回归方程，描述女运动员的身高和体重之间的关系．同样，不能用生长在南方多雨地区的树木的高与直径之间的回归方程，描述北方干旱地区的树木的高与直径之间的关系．

2．我们所建立的回归方程一般都有时间性．例如，不能用 20 世纪 80 年代的身高体重数据所建立的回归方程，描述现在的身高和体重之间的关系．

3．样本取值的范围会影响回归方程的适用范围．例如，我们的回归方程是由女大学生身高和体重数据建立的，那么用它来描述一个人幼儿时期的身高和体重之间的关系就不恰当（即在回归方程中，解释变量 x 的样本的取值范围为［155cm,170cm〕 ，而用这个方程计算 x-70cm 时的y值，显然不合适．)

4．不能期望回归方程得到的预报值就是预报变量的精确值．事实上，它是预报变量的可能取值的平均值．

一般地，建立回归模型的基本步骤为：

(1）确定研究对象，明确哪个变量是解释变量，哪个变量是预报变量；

(2）画出确定好的解释变量和预报变量的散点图，观察它们之间的关系（如是否存在线性关系等） ;

(3）由经验确定回归方程的类型（如我们观察到数据呈线性关系，则选用线性回归方程 y=bx+a ) ;

(4）按一定规则估计回归方程中的参数（如最小二乘法）;

(5）得出结果后分析残差图是否有异常（个别数据对应残差过大，或残差呈现不随机的规律性等等），若存在异常，则检查数据是否有误，或模型是否合适等．

**例2**.现收集了一只红铃虫的产卵数*y*和温度*x*之间的7组观测数据列于下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 温度*x*o*C* | 21 | 23 | 25 | 27 | 29 | 32 | 35 |
| 产卵数*y*/个 | 7 | 11 | 21 | 24 | 66 | 115 | 325 |

(1)试建立*y*与*x*之间的回归方程；并预测温度为28o*C*时产卵数目。

(2)你所建立的模型中温度在多大程度上解释了产卵数的变化？

**探究：**

**方案1**（学生实施）：

（1）选择变量，画散点图。

（2）通过计算器求得线性回归方程:***=*19.87*x***-**463.73**

（3）进行回归分析和预测：

**R2=*r2*≈0.864*2*=0.7464**

预测当气温为28 时，产卵数为92个。这个线性回归模型中温度解释了74.64%产卵数的变化。

**困惑**：随着自变量的增加，因变量也随之增加，气温为28 时，估计产卵数应该低于66个，但是从推算的结果来看92个比66个却多了26个，是什么原因造成的呢？

**方案2：**

**(1)**找到变量t=x 2，将y=bx2+a转化成y=bt+a；

(2)利用计算器计算出y和t的线性回归方程：***y=*0.367*t*-202.54**

**(3)**转换回y和x的模型：

(4)***y=*0.367*x2* -202.54**

（5）计算相关指数**R2≈0.802**这个回归模型中温度解释了80.2%产卵数的变化。

预测：当气温为28 时，产卵数为85个。

**困惑**：比66还多19个，是否还有更适合的模型呢？

**方案3：**

**(1)**作变换z=lgy，将******转化成z=c2x+lgc1（线性模型）。

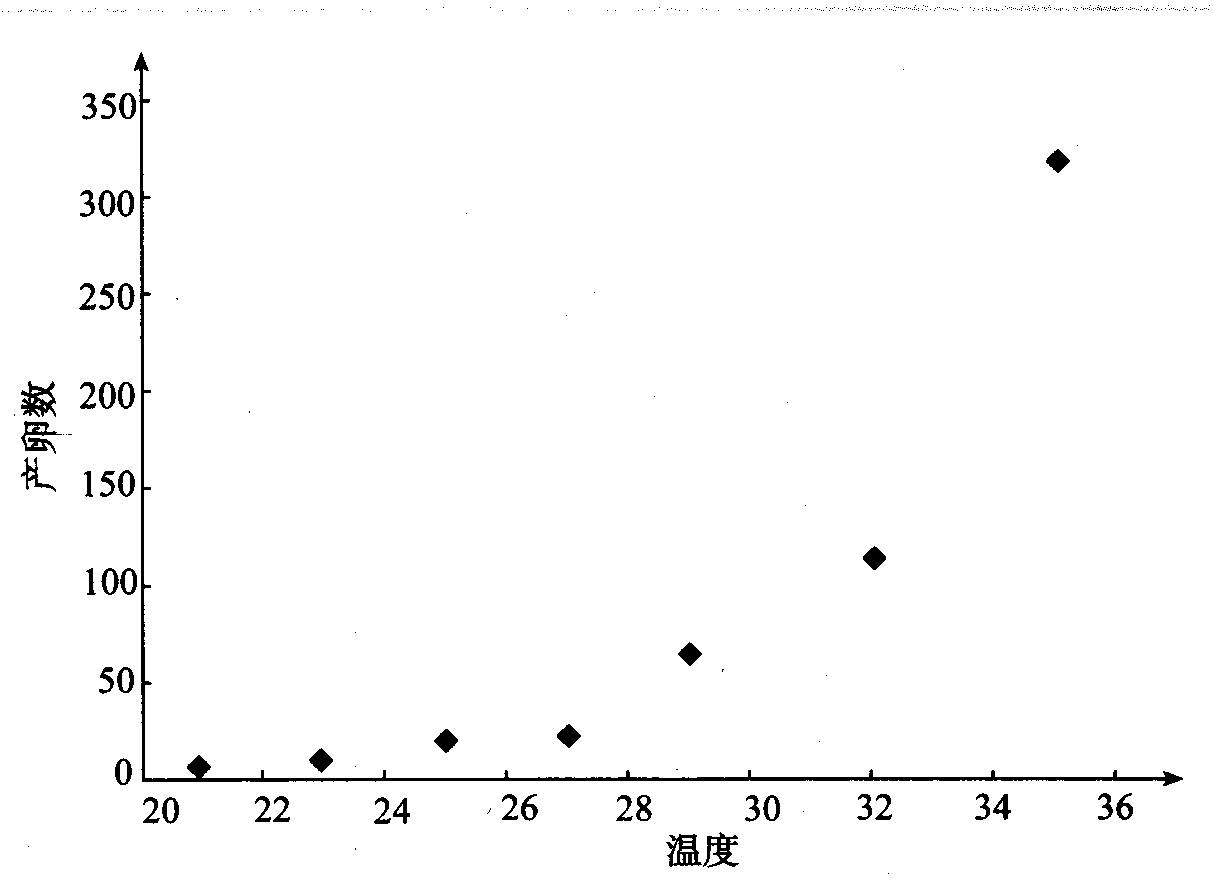
(2)利用计算器计算出z和x的线性回归方程： ***z=0.118x-1.672***

**(3)**转换回y和x的模型：

(4)计算相关指数**R2≈0.985**这个回归模型中温度解释了98.5%产卵数的变化。

预测：当气温为28 时，产卵数为4 2个。

解：根据收集的数据作散点图（图3. 1一4 ) .



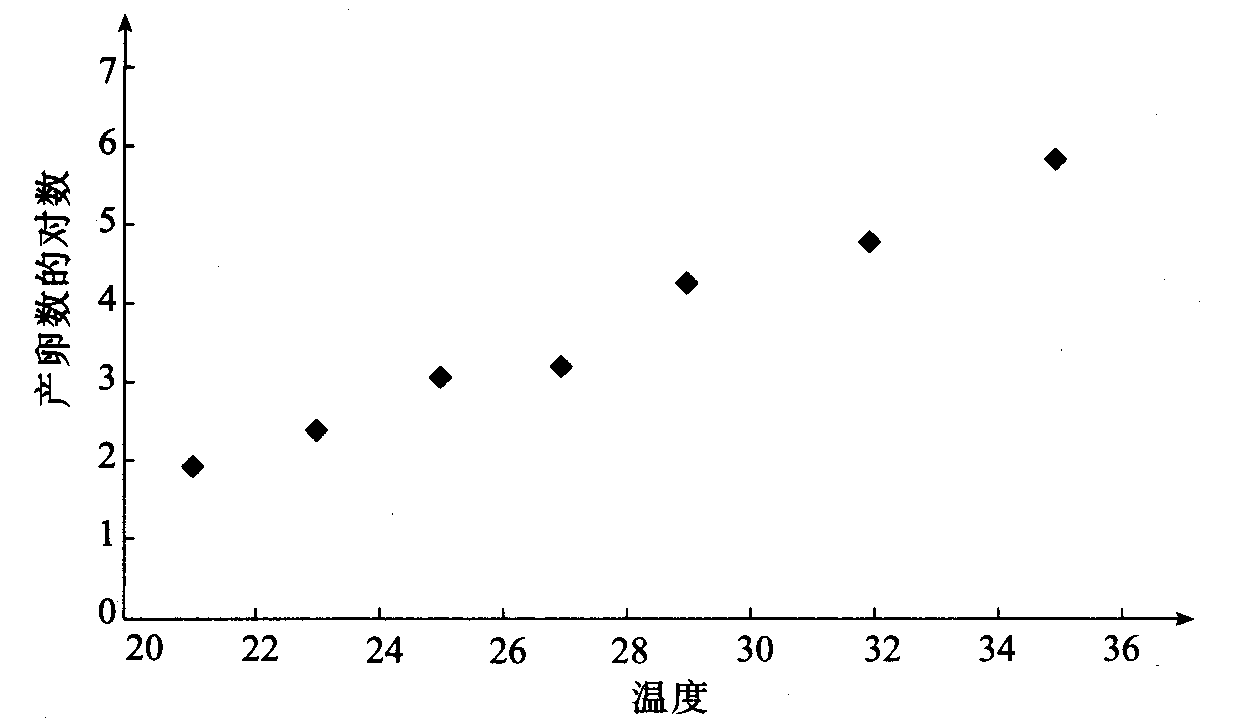
在散点图中，样本点并没有分布在某个带状区域内，因此两个变量不呈线性相关关系，所以不能直接利用线性回归方程来建立两个变量之间的关系．根据已有的函数知识，可以发现样本点分布在某一条指数函数曲线的周围，其中和是待定参数．现在，问题变为如何估计待定参数和．我们可以通过对数变换把指数关系变为线性关系．令，则变换后样本点应该分布在直线的周围．这样，就可以利用线性回归模型来建立 y 和 x 之间的非线性回归方程了．

由表3一3 的数据可以得到变换后的样本数据表 3一4 ，图3.1一5 给出了表 3 一 4 中数据的散点图．从图3.1一5 中可以看出，变换后的样本点分布在一条直线的附近，因此可以用线性回归方程来拟合．

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x | 21 | 23 | 25 | 27 | 29 | 32 | 35 |
| z | 1.946 | 3.398 | 3.045 | 3.178 | 4.190 | 4.745 | 5.784 |

由表 3 一 4 中的数据得到线性回归方程

.

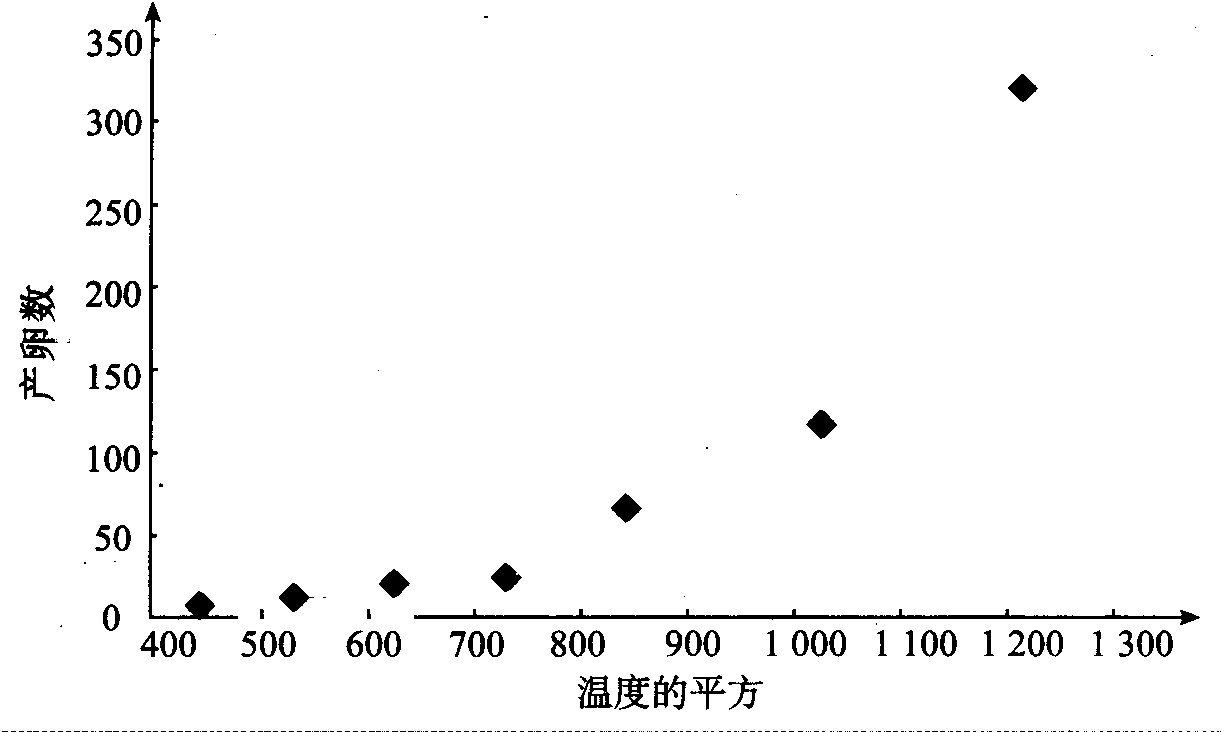


因此红铃虫的产卵数对温度的非线性回归方程为

. ( 6 )

另一方面，可以认为图3. 1一4 中样本点集中在某二次曲线的附近，其中和为待定参数．因此可以对温度变量做变换，即令，然后建立y与t之间的线性回归方程，从而得到y与x之间的非线性回归方程．表3一5 是红铃虫的产卵数和对应的温度的平方，图3 . 1一6 是相应的散点图．

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| t | 441 | 529 | 625 | 729 | 841 | 1024 | 1225 |
| x | 7 | 11 | 21 | 24 | 66 | 115 | 325 |



从图3.1一6 中可以看出，y与t的散点图并不分布在一条直线的周围，因此不宜用线性回归方程来拟合它，即不宜用二次曲线来拟合 y 和 x 之间的关系．这个结论还可以通过残差分析得到，下面介绍具体方法．

为比较两个不同模型的残差，需要建立两个相应的回归方程．前面我们已经建立了y

关于x 的指数回归方程，下面建立y关于x的二次回归方程．用线性回归模型拟合表 3 一 5 中的数据，得到 y 关于 t 的线性回归方程

,

即 y 关于 x 的二次回归方程为

 . ( 7 )

可以通过残差来比较两个回归方程（ 6 ）和（ 7 ）的拟合效果．用 xi表示表3一3 中第 1 行第 i 列的数据，则回归方程（ 6 ）和（ 7 ）的残差计算公式分别为

;

.

表3一6 给出了原始数据及相应的两个回归方程的残差．从表中的数据可以看出模型 ( 6 ）的残差的绝对值显然比模型（ 7 ）的残差的绝对值小，因此模型（ 6 ）的拟合效果比模型（ 7 ) 的拟合效果好．

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x | 21 | 23 | 25 | 27 | 29 | 32 | 35 |
| y | 7 | 11 | 21 | 24 | 66 | 115 | 325 |
|  | 0.557 | -0.101 | 1.875 | -8.950 | 9.230 | -13.381 | 34.675 |
|  | 47.696 | 19.400 | -5.832 | -41.000 | -40.1.4 | -58.265 | 77.968 |

在一般情况下，比较两个模型的残差比较困难．原因是在某些样本点上一个模型的残差的绝对值比另一个模型的小，而另一些样本点的情况则相反．这时可以通过比较两个模型的残差平方和的大小来判断模型的拟合效果．残差平方和越小的模型,拟合的效果越好．由表 3 一 6 容易算出模型（ 6 ）和（ 7 ）的残差平方和分别为

.

因此模型（6）的拟合效果远远优于模型（7）.

类似地，还可以用尸来比较两个模型的拟合效果，R2越大，拟合的效果越好．由表 3 一 6 容易算出模型（6）和（7）的R2分别约为 0 . 98 和 0 . 80 ，因此模型（ 6 ）的效果好于模型（7) 的效果．

对于给定的样本点（) , () ，…， ()，两个含有未知参数的模型

和,

其中 a 和 b 都是未知参数．可以按如下的步骤来比较它们的拟合效果：

(1)分别建立对应于两个模型的回归方程与, ，其中和分别是参数a和b的估计值；

(2)分别计算两个回归方程的残差平方和与;

( s ）若，则的效果比的好；反之，的效果不如的好．

**例2**：（提示后做练习、作业）研究某灌溉渠道水的流速y与水深x之间的关系，测得一组数据如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 水深xm | 1.40 | 1.50 | 1.60 | 1.70 | 1.80 | 1.90 | 2.00 | 2.10 |
| 流速ym/s | 1.70 | 1.79 | 1.88 | 1.95 | 2.03 | 2.10 | 2.16 | 2.21 |

（1）求y对x的回归直线方程；

（2）预测水深为1。95m 时水的流速是多少？

**解：**依题意，把温度作为解释变量x ，产卵个数y作为预报变量 , 作散点图，由观察知两个变量不呈线性相关关系。但样本点分布在某一条指数函数 y=c1ec2 x 周围.

令 z=lny , a=lnc1 , b=c2 则 z=bx+a

此时可用线性回归来拟合 z=0.272x-3.843

因此红铃虫的产卵数对温度的非线性回归方程为

Y=e0.272x-3.843.

3、从上节课的例1提出的问题引入线性回归模型：

Y=bx+a+e

解释变量x

预报变量y

随机误差 e

4、（1） 相关指数： 相关系数 r (公式) , r>0 正相关. R<0 负相关

R绝对值接近于1相关性强接 r绝对值 近于0 相关性几乎无



5、回忆建立模型的基本步骤 ① 例2 问题背景分析 画散点图。 ② 观察散点图，分析解释变量与预报变量更可能是什么函数关系。 ③ 学生讨论后建立自己的模型 ④ 引导学生探究如果不是线性回归模型如何估计参数。能否利用回归模型

通过探究体会有些不是线性的模型通过变换可以转化为线性模型 ⑤ 对数据进行变换后，对数据（新）建立线性模型 ⑥ 转化为原来的变量模型，并通过计算相关指数比较几个不同模型的拟合效果 ⑦ 总结建模的思想。鼓励学生大胆创新。 ⑧ 布置课后作业： 习题1.1 1、

6、复习与巩固：练习1：某班5名学生的数学和化学成绩如下表所示，对x与y进行回归分析，并预报某学生数学成绩为75分时，他的化学成绩。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | A | B | C | D | E |
| 数学x | 88 | 76 | 73 | 66 | 63 |
| 化学y | 78 | 65 | 71 | 64 | 61 |

**解略。**

练习2：某医院用光电比色计检验尿汞时，得尿汞含量 (mg/l) 与消光系数的结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 尿汞含量x | 2 | 4 | 6 | 8 | 10 |
| 消光系数y | 64 | 138 | 205 | 285 | 360 |

（1）求回归方程。（2）求相关指数R2。

**解：略**。