**回归分析的基本思想及其初步应用易错点-高中数学选修1-2第一章**

**例1：**研究某灌溉渠道水的流速image161 与水深image204 之间的关系，测得一组数据如下：

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 水深image024 | 1.40 | 1.50 | 1.60 | 1.70 | 1.80 | 1.90 | 2.00 | 2.10 |
| 流速image026 | 1.70 | 1.79 | 1.88 | 1.95 | 2.03 | 2.10 | 2.16 | 2.21 |

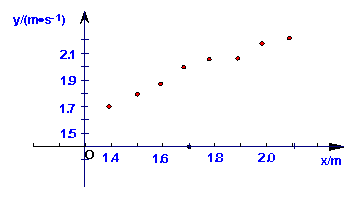
（1）求image161 对image204 的回归直线方程；

（2）预测水深为1.95image032 时水的流速是多少？

**分析：**本题考查如何求回归直线的方程，可先把有关数据用散点图表示出来，若这些点大致分布在通过散点图中心的一条直线附近，说明这两个变量线性相关，从而可利用我们学过的最小二乘估计思想及计算公式求得线性回归直线方程。

**解：**

（1）由于问题中要求根据水深预报水的流速，因此选取水深为解释变量，流速为预报变量，作散点图：



由图容易看出，image204 与image161 之间有近似的线性关系，或者说，可以用一个回归直线方程  
image198 来反映这种关系。由计算器求得image042 。

对image204 的回归直线方程为image048 。

（2）由（1）中求出的回归直线方程，把image050 代入，易得

image052 。

计算结果表示，当水深为image054 时可以预测渠水的流速为image056 。

**评注：**建立回归模型的一般步骤：

（1）确定研究对象，明确两个变量即解释变量和预报变量；

（2）画出散点图，观察它们之间的关系；

（3）由经验确定回归方程类型（若呈线性关系，选用线性回归方程）；

（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，或残差出现不随机的规律性，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**：1993年到2002年中国的国内生产总值(GDP)的数据如下：

|  |  |
| --- | --- |
| 年份 | GDP |
| 1993 | 34634.4 |
| 1994 | 46759.4 |
| 1995 | 58478.1 |
| 1996 | 67884.6 |
| 1997 | 74462.6 |
| 1998 | 78345.2 |
| 1999 | 82067.5 |
| 2000 | 89468.1 |
| 2001 | 97314.8 |
| 2002 | 104790.6 |

(1)作GDP和年份的散点图，根据该图猜想它们之间的关系应是什么。

(2)建立年份为解释变量，GDP为预报变量的回归模型，并计算残差。

(3)根据你得到的模型，预报2003年的GDP，并查阅资料，看看你的预报与实际GDP的误差是多少。

(4)你认为这个模型能较好地刻画GDP和年份的关系吗？请说明理由。

**解：**(1)由表中数据制作的散点图如下：



从散点图中可以看出GDP值与年份近线呈线性关系；

(2)用yt表示GDP值，t表示年份，根据截距和斜率的最小二乘计算公式，

得：image060

从而得线性回归方程：image062

残差计算结果见下表：

GDP值与年份线性拟合残差表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 年份 | 1993 | 1994 | 1995 | 1996 | 1997 |
| 残差 | -6422.269 | -1489.238 | 3037.493 | 5252.024 | 4638.055 |
| 年份 | 1998 | 1999 | 2000 | 2001 | 2002 |
| 残差 | 1328.685 | -2140.984 | -1932.353 | -1277.622 | -993.791 |

(3)2003年的GDP预报值为112976.360，根据国家统计局2004年统计，2003年实际GDP值为117251.9，所以预报与实际相-4275.540；

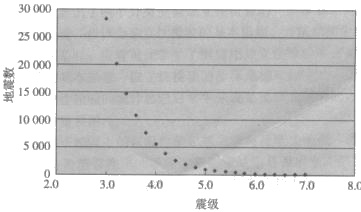
(4)上面建立的回归方程的R2=0.974，说明年份能够解释约97%的GDP值变化，因此所建立的模型能够很好地刻画GDP和年份的关系。

说明： 关于2003年的GDP的值来源，不同的渠道可能会有所不同。

**例3：**如下表所示，某地区一段时间内观察到的大于或等于某震级x的地震个数为N，试建立回归方程表述二者之间的关系。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 震级 | 3 | 3.2 | 3.4 | 3.6 | 3.8 | 4 | 4.2 | 4.4 | 4.6 | 4.8 | 5.0 |
| 地震数 | 28381 | 20380 | 14795 | 10695 | 7641 | 5502 | 3842 | 2698 | 1919 | 1356 | 973 |
| 震级 | 5.2 | 5.4 | 5.6 | 5.8 | 6 | 6.2 | 6.4 | 6.6 | 6.8 | 7 |  |
| 地震数 | 746 | 604 | 435 | 274 | 206 | 148 | 98 | 57 | 41 | 25 |  |

**解：**由表中数据得散点图如下：



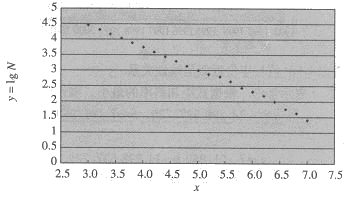
从散点图中可以看出，震级x与大于该震级的地震次数N之间不呈线性相关关系，随着x的减少，所考察的地震数N近似地以指数形式增长.

做变换y=lgN，

得到的数据如下表所示：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 3 | 3.2 | 3.4 | 3.6 | 3.8 | 4 | 4.2 | 4.4 | 4.6 | 4.8 | 5 |
| y | 4.453 | 4.309 | 4.170 | 4.029 | 3.883 | 3.741 | 3.585 | 3.431 | 3.283 | 3.132 | 2.988 |
| x | 5.2 | 5.4 | 5.6 | 5.8 | 6 | 6.2 | 6.4 | 6.6 | 6.8 | 7 |  |
| y | 2.873 | 2.781 | 2.638 | 2.438 | 2.314 | 2.170 | 1.991 | 1.756 | 1.613 | 1.398 |  |

x和y的散点图如下：



从这个散点图中可以看出x和y之间有很强的线性相差性，因此可以用线性回归模型拟合它们之间的关系。根据截距和斜率的最小二乘计算公式，得：image068

故线性回归方程为：image070

相关指数R2≈0.997，说明x可以解释y的99.7%的变化。因此，可以用回归方程  
image072 描述x和y之间的关系。

**例4：**电容器充电后，电压达到image074 ,然后开始放电，由经验知道，此后电压image094 随时间image114 变化的规律公式image080 表示，观测得时间image082 时的电压image084 如下表所示：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| image114 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| image094 | 100 | 75 | 55 | 40 | 30 | 20 | 15 | 10 | 10 | 5 | 5 |

试求电压image094 对时间image114 的回归方程。

**分析：**由于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，我们可通过对数变换把指数关系变为线性关系，通过线性回归模型来建立image094 与image114 之间的非线性回归方程。

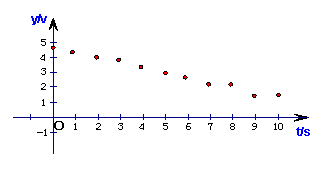
**解：**对image098 两边取自然对数得

image100 ，令image102 ,即image104 。

由所给数据可得

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| image114 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| image108 | 4.6 | 4.3 | 4.0 | 3.9 | 3.4 | 2.9 | 2.7 | 2.3 | 2.3 | 1.6 | 1.6 |

其散点图为：



由散点图可知image161 与image114 具有线性相关关系，可用image116 来表示。经计算得：image118 （最小二乘法），　　image120 ，　　即image122 。　　所以，image124 。

**评注：**一般地，有些非线性回归模型通过变换可以转化为线性回归模型，即借助于线性回归模型研究呈非线性回归关系的两个变量之间的关系：

（1）如果散点图中的点分布在一个直线状带形区域，可以选用线性回归模型来建模；

（2）如果散点图中的点的分布在一个曲线状带形区域，要先对变量作适当的变换，再利用线性回归模型来建模。

**本周练习：**

1.对具有相关关系的两个变量统计分析的一种常用的方法是（　　 ）

A．回归分析　　 B.相关系数分析　　 C.残差分析　　 D.相关指数分析

2.在画两个变量的散点图时，下面叙述正确的是（　　 ）

A．预报变量在image204 轴上，解释变量在image161 轴上

B.解释变量在image204 轴上，预报变量在image161 轴上

C.可以选择两个变量中任意一个变量在image204 轴上

D.可以选择两个变量中任意一个变量在image161 轴上

3.两个变量相关性越强，相关系数image202 （　　 ）

　　A．越接近于0　　　　 B.越接近于1　　 C.越接近于－1　　 D.绝对值越接近1

4.若散点图中所有样本点都在一条直线上，解释变量与预报变量的相关系数为（　　 ）　　A．0　　 B.1　　　　 C.－1　　　　 D.－1或1

5.一位母亲记录了她儿子3到9岁的身高，数据如下表：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年龄（岁） | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 身高（image140 | 94.8 | 104.2 | 108.7 | 117.8 | 124.3 | 130.8 | 139.0 |

由此她建立了身高与年龄的回归模型image142 ，她用这个模型预测儿子10岁时的身高，则下面的叙述正确的是（　　 ）

A.她儿子10岁时的身高一定是145.83image147

B.她儿子10岁时的身高在145.83image147 以上

C.她儿子10岁时的身高在145.83image147 左右

D.她儿子10岁时的身高在145.83image147 以下

6.两个变量有线性相关关系且正相关，则回归直线方程中，image198 的系数image151 （　　 ）　　A.image153 　　 B.image155 　　 C.image157 　　　　 D.image159

7.两个变量有线性相关关系且残差的平方和等于0，则（　　 ）

A.样本点都在回归直线上　　 B.样本点都集中在回归直线附近

C.样本点比较分散　　　　　 D.不存在规律

8.在建立两个变量image161 与image204 的回归模型中，分别选择了4个不同的模型，它们的相关指数image172 如下，其中拟合最好的模型是（　　 ）

A.模型1的相关指数image172 为0.98　　 B.模型2的相关指数image172 为0.80

C.模型3的相关指数image172 为0.50　　 D.模型4的相关指数image172 为0.25

9.相关指数image172 ＝　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　　 。

10.某农场对单位面积化肥用量image174 和水稻相应产量image176 的关系作了统计，得到数据如下：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| image204 | 15 | 20 | 25 | 30 | 35 | 40 | 45 |
| image202 | 330 | 345 | 365 | 405 | 445 | 450 | 455 |

如果image204 与image202 之间具有线性相关关系，求出回归直线方程，并预测当单位面积化肥用量为image212 时水稻的产量大约是多少？（精确到image188 ）

11.假设美国10家最大的工业公司提供了以下数据：

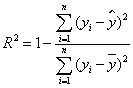
|  |  |  |
| --- | --- | --- |
| 公司 | 销售总额经x1/百万美元 | 利润x2/百万美元 |
| 通用汽车 | 126974 | 4224 |
| 福特 | 96933 | 3835 |
| 埃克森 | 86656 | 3510 |
| IBM | 63438 | 3758 |
| 通用电气 | 55264 | 3939 |
| 美孚 | 50976 | 1809 |
| 菲利普·莫利斯 | 39069 | 2946 |
| 克莱斯勒 | 36156 | 359 |
| 杜邦 | 35209 | 2480 |
| 德士古 | 32416 | 2413 |

（1）作销售总额和利润的散点图，根据该图猜想它们之间的关系应是什么形式；

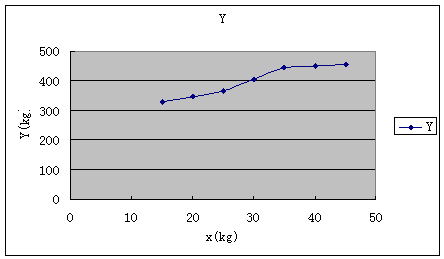
(2) 建立销售总额为解释变量，利润为预报变量的回归模型，并计算残差；

(3) 你认为这个模型能较好地刻画销售总额和利润之间的关系吗？请说明理由。

**参考答案：**  
　　A B D B　　 C A A A

9．

10.由于问题中要求根据单位面积化肥用量预报水稻相应的产量，因此选取单位面积的化肥用量为解释变量，相应水稻的产量为预报变量，作散点图：



由图容易看出，image204 与image202 之间有近似的线性关系，或者说，可以用一个回归直线方程

image198 来反映这种关系。由计算器求得image200 。

image202对image204 的回归直线方程为image206（ \*）。

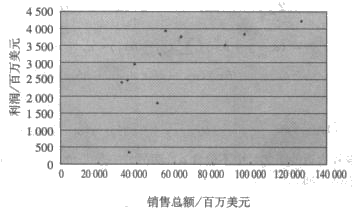
由（\*）中求出的回归直线方程，把image208 代入

易得image210。

计算结果表示，当单位面积化肥用量为image212 时水稻的产量大约是image214 .

11．

(1)将销售总额作为横轴，利润作为纵轴，根据表中数据绘制散点图如下：



由于散点图中的样本点基本上在一个带形区域分布，猜想销售总额与利润之间呈现线性相关关系；

(2)由最小二乘法的计算公式，得：image218

则线性回归方程为：image220

其残差值计算结果见下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 销售总额 | 126974 | 96933 | 86656 | 63438 | 55264 |
| 利润 | 4224 | 3835 | 3510 | 3758 | 3939 |
| 残差 | -361.034 | 19.015 | -42.894 | 799.487 | 1189.742 |
| 销售总额 | 50976 | 39069 | 36156 | 35209 | 32416 |
| 利润 | 1809 | 2946 | 359 | 2480 | 2413 |
| 残差 | -830.486 | 611.334 | -1901.09 | 244.150 | 248.650 |

(3)对于(2)中所建立的线性回归方程，相关指数为R2≈0.457，说明在线性回归模型中销售总额只能解释利润变化的46%，所以线性回归模型不能很好地刻画销售总额和利润之间的关系。

　　说明：此题也可以建立对数模型或二次回归模型等，只要计算和分析合理，就算正确。