**回归分析的基本知识点及习题**

　　 **本周难点：**
　　（1）求回归直线方程，会用所学的知识对实际问题进行回归分析.
　　（2）掌握回归分析的实际价值与基本思想.
　　（3）能运用自己所学的知识对具体案例进行检验与说明.
　　（4）残差变量的解释；
　　（5）偏差平方和分解的思想；
　　　**１．回归直线：**
　　如果散点图中点的分布从整体上看大致在一条直线附近，我们就称这两个变量之间具有线性相关关系，这条直线叫作回归直线。
　　求回归直线方程的一般步骤：
　　①作出散点图（由样本点是否呈条状分布来判断两个量是否具有线性相关关系），若存在线性相关关系→②求回归系数 →③写出回归直线方程 ，并利用回归直线方程进行预测说明.
　　**2.回归分析：**
　　对具有相关关系的两个变量进行统计分析的一种常用方法。
　　建立回归模型的基本步骤是：
　　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画好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（线性关系）.
　　③由经验确定回归方程的类型.
　　④按一定规则估计回归方程中的参数 （最小二乘法）；
　　⑤得出结论后在分析残差图是否异常，若存在异常，则检验数据是否有误，后模型是否合适等.
　**4.残差变量 的主要来源：**
　　（1）用线性回归模型近似真实模型（真实模型是客观存在的，通常我们并不知道真实模型到底是什么）所引起的误差。可能存在非线性的函数能够更好地描述 与 之间的关系，但是现在却用线性函数来表述这种关系，结果就会产生误差。这种由于模型近似所引起的误差包含在 中。
　　（2）忽略了某些因素的影响。影响变量 的因素不只变量 一个，可能还包含其他许多因素（例如在描述身高和体重关系的模型中，体重不仅受身高的影响，还会受遗传基因、饮食习惯、生长环境等其他因素的影响），但通常它们每一个因素的影响可能都是比较小的，它们的影响都体现在 中。
　　（3）观测误差。由于测量工具等原因，得到的 的观测值一般是有误差的（比如一个人的体重是确定的数，不同的秤可能会得到不同的观测值，它们与真实值之间存在误差），这样的误差也包含在 中。
上面三项误差越小，说明我们的回归模型的拟合效果越好。
**二、例题选讲**
　　1为研究某市家庭平均收入与月平均生活支出的关系，该市统计部门随机调查了10个家庭，得数据如下：

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 家庭编号 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| xi（收入）千元 | 0.8 | 1.1 | 1.3 | 1.5 | 1.5 | 1.8 | 2.0 | 2.2 | 2.4 | 2.8 |
| yi（支出）千元 | 0.7 | 1.0 | 1.2 | 1.0 | 1.3 | 1.5 | 1.3 | 1.7 | 2.0 | 2.5 |

（1）判断家庭平均收入与月平均生活支出是否相关？

（2）若二者线性相关，求回归直线方程.

**解** （1）作出散点图：



观察发现各个数据对应的点都在一条直线附近，所以二者呈线性相关关系.

（2）= (0.8+1.1+1.3+1.5+1.5+1.8+2.0+2.2+2.4+2.8)=1.74,

=（0.7+1.0+1.2+1.0+1.3+1.5+1.3+1.7+2.0+2.5）=1.42，

=≈0.813 6，=1.42-1.74×0.813 6≈0.004 3， ∴回归方程=0.813 6x+0.004 3. 2下表提供了某厂节能降耗技术改造后生产甲产品过程中记录的产量x(吨）与相应的生产能耗y（吨）标准煤的几组对照数据.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| x | 3 | 4 | 5 | 6 |
| y | 2.5 | 3 | 4 | 4.5 |

（1）请画出上表数据的散点图；

（2）请根据上表提供的数据，用最小二乘法求出y关于x的线性回归方程=x+；

（3）已知该厂技改前100吨甲产品的生产能耗为90吨标准煤.试根据（2）求出的线性回归方程，预测生产100吨甲产品的生产能耗比技改前降低多少吨标准煤？

**解** （1）散点图如下图:



 （2）==4.5,==3.5

=3×2.5+4×3+4×5+6×4.5=66.5.=32+42+52+62=86

∴===0.7

 =-=3.5-0.7×4.5=0.35.

∴所求的线性回归方程为=0.7x+0.35.

（3）现在生产100吨甲产品用煤

y=0.7×100+0.35=70.35，

∴降低90-70.35=19.65(吨)标准煤.

3科研人员为了全面掌握棉花新品种的生产情况，查看了气象局对该地区年降雨量与年平均气温的统计数据（单位分别是mm,℃)，并作了统计.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年平均气温 | 12.51 | 12.84 | 12.84 | 13.69 | 13.33 | 12.74 | 13.05 |
| 年降雨量 | 748 | 542 | 507 | 813 | 574 | 701 | 432 |

（1）试画出散点图；

（2）判断两个变量是否具有相关关系.

**解** （1）作出散点图如图所示，



(2)由散点图可知，各点并不在一条直线附近，所以两个变量是非线性相关关系.

4在研究硝酸钠的可溶性程度时，对于不同的温度观测它在水中的溶解度，得观测结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 温度(x) | 0 | 10 | 20 | 50 | 70 |
| 溶解度（y) | 66.7 | 76.0 | 85.0 | 112.3 | 128.0 |

由资料看y与x呈线性相关，试求回归方程.

**解** =30,==93.6.

=≈0.880 9.=-=93.6-0.880 9×30=67.173.

∴回归方程为=0.880 9x+67.173.

5.某企业上半年产品产量与单位成本资料如下：

|  |  |  |
| --- | --- | --- |
| 月份 | 产量（千件） | 单位成本（元） |
| 1 | 2 | 73 |
| 2 | 3 | 72 |
| 3 | 4 | 71 |
| 4 | 3 | 73 |
| 5 | 4 | 69 |
| 6 | 5 | 68 |

（1）求出线性回归方程；

（2）指出产量每增加1 000件时，单位成本平均变动多少？

（3）假定产量为6 000件时，单位成本为多少元？

**解** （1）n=6，=21，=426，=3.5,=71,=79，=1 481，

===-1.82.

=-=71+1.82×3.5=77.37.

回归方程为=+x=77.37-1.82x.

（2）因为单位成本平均变动=-1.82＜0,且产量x的计量单位是千件,所以根据回归系数b的意义有:

产量每增加一个单位即1 000件时,单位成本平均减少1.82元.

(3)当产量为6 000件时,即x=6,代入回归方程:

=77.37-1.82×6=66.45（元）

当产量为6 000件时，单位成本为66.45元.

1.观察下列散点图，则①正相关；②负相关；③不相关.它们的排列顺序与图形对应顺序是 .



**答案** a,c,b

2.回归方程=1.5x-15,则下列说法正确的有 个.

①=1.5-15②15是回归系数a③1.5是回归系数a④x=10时，y=0

**答案** 1

3.**（2009.湛江模拟）**某地区调查了2～9岁儿童的身高，由此建立的身高y(cm)与年龄x(岁)的回归模型为=8.25x+60.13,下列叙述正确的是 .

①该地区一个10岁儿童的身高为142.63 cm

②该地区2～9岁的儿童每年身高约增加8.25 cm

③该地区9岁儿童的平均身高是134.38 cm

④利用这个模型可以准确地预算该地区每个2～9岁儿童的身高

**答案** ②

4.某人对一地区人均工资x(千元）与该地区人均消费y(千元）进行统计调查，y与x有相关关系，得到回归直线方程=0.66x+1.562.若该地区的人均消费水平为7.675千元，估计该地区的人均消费额占人均工资收入的百分比约为 .

**答案** 83%

5.某化工厂为预测产品的回收率y,需要研究它和原料有效成分含量x之间的相关关系,现取8对观测值,计算,得=52, =228, =478, =1 849,则其线性回归方程为 .

**答案 **=11.47+2.62x

6.有下列关系：①人的年龄与他（她）拥有的财富之间的关系；②曲线上的点与该点的坐标之间的关系；③苹果的产量与气候之间的关系；④森林中的同一种树木，其断面直径与高度之间的关系.其中，具有相关关系的是 .

**答案** ①③④

7.已知关于某设备的使用年限x与所支出的维修费用y(万元），有如下统计资料：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 使用年限x | 2 | 3 | 4 | 5 | 6 |
| 维修费用y | 2.2 | 3.8 | 5.5 | 6.5 | 7.0 |

若y对x呈线性相关关系，则回归直线方程=x+表示的直线一定过定点 .

**答案** （4，5）

**二、解答题**

8.期中考试结束后，记录了5名同学的数学和物理成绩，如下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  学生学科 | A | B | C | D | E |
| 数学 | 80 | 75 | 70 | 65 | 60 |
| 物理 | 70 | 66 | 68 | 64 | 62 |

（1）数学成绩和物理成绩具有相关关系吗？

（2）请你画出两科成绩的散点图，结合散点图，认识（1）的结论的特点.

**解** （1）数学成绩和物理成绩具有相关关系.

（2）以x轴表示数学成绩，y轴表示物理成绩，可得相应的散点图如下：



由散点图可以看出，物理成绩和数学成绩对应的点不分散，大致分布在一条直线附近.

9.以下是某地搜集到的新房屋的销售价格y和房屋的面积x的数据：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 房屋面积x（m2) | 115 | 110 | 80 | 135 | 105 |
| 销售价格y(万元） | 24.8 | 21.6 | 18.4 | 29.2 | 22 |

（1）画出数据对应的散点图；

（2）求线性回归方程，并在散点图中加上回归直线.

**解** （1）数据对应的散点图如图所示：



（2）=109,=23.2,=60 975,=12 952,=≈0.196 2

=-≈1.814 2

∴=0.196 2x+1.814 2.

10.某公司利润y与销售总额x(单位：千万元）之间有如下对应数据：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| x | 10 | 15 | 17 | 20 | 25 | 28 | 32 |
| y | 1 | 1.3 | 1.8 | 2 | 2.6 | 2.7 | 3.3 |

（1）画出散点图；（2）求回归直线方程；

（3）估计销售总额为24千万元时的利润.

**解** （1）散点图如图所示：



（2)=(10+15+17+20+25+28+32)=21,

=(1+1.3+1.8+2+2.6+2.7+3.3)=2.1,=102+152+172+202+252+282+322=3 447,

=10×1+15×1.3+17×1.8+20×2+25×2.6+28×2.7+32×3.3=346.3,

==≈0.104,

=-=2.1-0.104×21=-0.084,

∴=0.104x-0.084.

(3)把x=24(千万元）代入方程得，

=2.412（千万元）.

∴估计销售总额为24千万元时，利润为2.412千万元.

11某种产品的广告费支出x与销售额y(单位：百万元）之间有如下对应数据:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | 2 | 4 | 5 | 6 | 8 |
| y | 30 | 40 | 60 | 50 | 70 |

（1）画出散点图；

（2）求回归直线方程；

（3）试预测广告费支出为10百万元时，销售额多大？

**解** （1）根据表中所列数据可得散点图如下：



（2）列出下表，并用科学计算器进行有关计算：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 |
| xi | 2 | 4 | 5 | 6 | 8 |
| yi | 30 | 40 | 60 | 50 | 70 |
| xiyi | 60 | 160 | 300 | 300 | 560 |

因此，==5,= =50,

=145, =13 500, =1 380.

于是可得：===6.5;

=-=50-6.5×5=17.5.

因此，所求回归直线方程为：=6.5x+17.5.

（3）根据上面求得的回归直线方程，当广告费支出为10百万元时，=6.5×10+17.5=82.5(百万元)，即这种产品的销售收入大约为82.5百万元.

**独立性检验的基本知识点及习题**

　　　**本周内容：**
　　 **一、基础知识梳理**
　　**1.独立性检验**
　　利用随机变量 来确定在多大程度上可以认为“两个分类变量有关系”的方法称为两个分类变量的独立性检验。
　　**2.判断结论成立的可能性的步骤：**
　　（1）通过三维柱形图和二维条形图，可以粗略地判断两个分类变量是否有关系，但是这种判断无法精确地给出所得结论的可靠程度。
　　（2）可以利用独立性检验来考察两个分类变量是否有关系，并且能较精确地给出这种判断的可靠程度。
　　**二、例题选讲**
　　**例1.**为了探究患慢性气管炎是否与吸烟有关，调查了339名50岁以上的人，调查结果如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 　　  | 患病 | 不患病 | 合计 |
| 吸烟 | 43 | 162 | 205 |
| 不吸烟 | 13 | 121 | 134 |
| 合计 | 56 | 283 | 339 |

　　试问：50岁以上的人患慢性气管炎与吸烟习惯有关吗？
　　**分析：**最理想的解决办法是向所有50岁以上的人作调查，然后对所得到的数据进行统计处理，但这花费的代价太大，实际上是行不通的，339人相对于全体50岁以上的人，只是一个小部分，已学过总体和样本的关系，当用样本平均数，样本方差去估计总体相应的数字特征时，由于抽样的随机性，结果并不唯一。现在情况类似，我们用部分对全体作推断，推断可能正确，也可能错误。如果抽取的339个调查对象中很多人是吸烟但没患慢性气管炎，而虽不吸烟因身体体质差而患慢性气管炎，能够得出什么结论呢？我们有95%（或99%）的把握说事件 与事件 有关，是指推断犯错误的可能性为5%（或1%），这也常常说成是“以95%（或99%）的概率”是一样的。
　　**解：**根据列联表中的数据，得
　　 。
　　因为 ，所以我们有99%的把握说：50岁以上的人患慢性气管炎与吸烟习惯有关。

**例2．**甲乙两个班级进行一门考试，按照学生考试成绩优秀和不优秀统计成绩后，得到如下的列联表：

班级与成绩列联表

|  |  |  |  |
| --- | --- | --- | --- |
| 　　  | 优秀 | 不优秀 | 总计 |
| 甲班 | 10 | 35 | 45 |
| 乙班 | 7 | 38 | 45 |
| 总计 | 17 | 73 | 90 |

　　画出列联表的条形图，并通过图形判断成绩与班级是否有关；利用列联表的独立性检验估计，认为“成绩与班级有关系”犯错误的概率是多少。
　　**解：**列联表的条形图如图所示：



　　由表中数据计算得K2的观察值为k≈0.653>0.455。
　　由下表中数据

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P（K2≥k） | 0.50 | 0.40 | 0.25 | 0.15 | 0.10 | 0.05 | 0.025 | 0.010 | 0.005 | 0.001 |
| k | 0.455 | 0.708 | 1.323 | 2.072 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 | 10.828 |

得：P(K2≥0.455)≈0.50，
　　从而有50%的把握认为“成绩与班级有关系”，即断言“成绩优秀与班级有关系”犯错误的概率为0.5。
　　**评注：**（1）画出条形图后，从图形上判断两个分类变量之间是否有关系。这里通过图形的直观感觉的结果可能会出错。
　　（2）计算得到K2的观测值比较小，所以没有理由说明“成绩优秀与班级有关系”。这与反证法也有类似的地方，在使用反证法证明结论时，假设结论不成立的条件下如果没有推出矛盾，并不能说明结论成立也不能说明结论不成立。在独立性检验中，在假设“成绩优秀与班级没有关系”的情况下，计算得到的K2的值比较小，且P(K2≥0.653)≈0.42，说明事件(K2≥0.653)不是一个小概率事件，这个事件的发生不足以说明“成绩优秀与班级没有关系”，即没有理由说明“成绩优秀与班级有关系”。这里没有推出小概率事件发生类似于反证法中没有推出矛盾。
　　**例3．**为考察某种药物预防疾病的效果，进行动物试验，得到如下的列联列表：

药物效果与动物试验列联表

|  |  |  |  |
| --- | --- | --- | --- |
| 　　  | 患病 | 未患病 | 总计 |
| 服用药 | 10 | 45 | 55 |
| 没服用药 | 20 | 30 | 50 |
| 总计 | 30 | 75 | 105 |

　　请问能有多大把握认为药物有效？
　　**解：** 假设“服药情况与是否患病之间没有关系”，则K2的值应比较小；如果K2的值很大，则说明很可能“服药情况与是否患病之间有关系”。由题目中所给数据计算，得K2的观测值为k≈6.110，而P(K2≥5.024)≈0.025，所以有97.5%的把握认为“服药情况与是否患病之间有关系”，即大约有97.5%的把握认为药物有效。

　　**例4．**在一次恶劣气候的飞行航程中调查男女乘客在机上晕机的情况如下表所示，根据此资料你是否认为在恶劣气候中男人比女人更容易晕机？

|  |  |  |  |
| --- | --- | --- | --- |
| 　　  | 晕机 | 不晕机 | 合计 |
| 男人 | 24 | 31 | 55 |
| 女人 | 8 | 26 | 34 |
| 合计 | 32 | 57 | 89 |

　　**分析：**这是一个 列联表的独立性检验问题，根据列联表的数据求解。
　　**解：**由条件中数据，计算得：　 ，
　　因为 ，所以我们没有理由说晕机是否跟男女性别有关，尽管这次航班中男人晕机的比例 比女人晕机的比例 高，但我们不能认为在恶劣的气候飞行中男人比女人更容易晕机。
　　**评注：**在使用 统计量作 列联表的独立性检验时，要求表中的4个数据大于等于5，为此，在选取样本的容量时一定要注意这一点，本例中的4个数据都大于5，且满足这一要求的。
例5在对人们休闲的一次调查中，共调查了124人，其中女性70人，男性54人。女性中有43人主要的休闲方式是看电视，另外27人主要的休闲方式是运动；男性中有21人主要的休闲方式是看电视，另外33人主要的休闲方式是运动。
　　（1）根据以上数据建立一个 的列联表；
　　（2）检验性别与休闲方式是否有关系。
（1） 的列联表：

|  |  |  |  |
| --- | --- | --- | --- |
| 　　  | 看电视 | 运动 | 合计 |
| 女 | 43 | 27 | 70 |
| 男 | 21 | 33 | 54 |
| 合计 | 64 | 60 | 124 |

　　（2）假设休闲方式与性别无关，计算
　　 ；
　　因为 ，所以有理由认为假设休闲方式与性别无关是不合理的，即我们有97.5%的把握认为休闲方式与性别无关。
例6调查某医院某段时间内婴儿出生的时间与性别的关系，得到下面的数据表。试问能以多大把握认为婴儿的性别与出生的时间有关系。

|  |  |  |  |
| --- | --- | --- | --- |
| 出生时间 性别 | 晚上 | 白天 | 合计 |
| 男婴 | 24 | 31 | 55 |
| 女婴 | 8 | 26 | 34 |
| 合计 | 32 | 57 | 89 |

　　　7．由所给数据计算得K2的观测值为k≈3.689，而由

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| P（K2≥k） | 0.50 | 0.40 | 0.25 | 0.15 | 0.10 | 0.05 | 0.025 | 0.010 | 0.005 | 0.001 |
| k | 0.455 | 0.708 | 1.323 | 2.072 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 | 10.828 |

　　知P(K2≥2.706)=0.10
　　所以有90%的把握认为“婴儿的性别与出生的时间有关系”。