第4课时　线性回归方程(2)



知识技能

1. 进一步掌握利用公式求线性回归方程的一般方法．

2. 利用相关系数解决一般回归分析的问题．

思想方法

进一步了解回归分析的基本思想、方法及应用，体会数形结合的思想．

核心素养

分别利用散点图和相关系数的方法，作线性回归分析，理解处理问题的方法，发展逻辑推理和数学运算素养．



重点：线性回归方程、相关系数．

难点：非线性回归方程．



问题导引

1. 用什么方法可以检验两个变量之间的线性相关程度？

提示　相关系数．

2. 求得的线性回归方程是否有实际意义？

提示　不一定．

即时体验

1. (多选)已知变量*x, y*的相关系数为*r*，线性回归方程为＝ *x*＋，下列说法中正确的有(ABD)

A. *r*是说明具有线性关系的两个变量间的相关密切程度和相关方向的统计指标

B. 表示*x*每增大(或减小)一个单位，*y*平均增大(或减小)个单位

C. 通过回归直线＝*x*＋及回归系数，可以精确反映变量的取值和变化趋势

D. 对同一资料，*r*与 的符号是相同的：*r*为正，说明*x*与*y*相关关系的方向是一致的；为正，说明*x*与*y*之间由回归方程所确定的变量关系是递增的

提示　通过回归直线＝ *x*＋及回归系数，可估计和预测变量的取值和变化趋势．

2. 某车间加工零件的数量*x*(单位：个)与加工时间*y*(单位：mi*n*)的统计数据如下表所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 零件的数量*x*/个 | 10 | 20 | 30 |
| 加工时间*y*/mi*n* | 21 | 30 | 39 |

现已求得上表数据的回归方程＝ *x*＋中的值为0.9，则据此回归模型可以预测，加工100个零件所需要的加工时间约为(C)

A. 84mi*n* B. 94mi*n*

C. 102mi*n* D. 112mi*n*

提示　＝20, ＝30，所以直线＝ *x*＋过点(20, 30)，可求得＝12.

3. 下面是两组数据的散点图，若求出相应的线性回归方程，则求出的线性回归方程可以用作预测和估计吗？

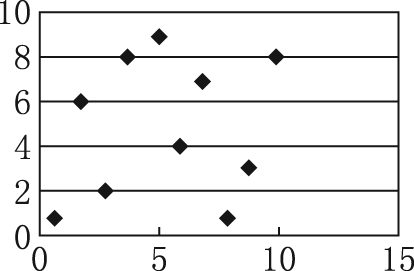


图1

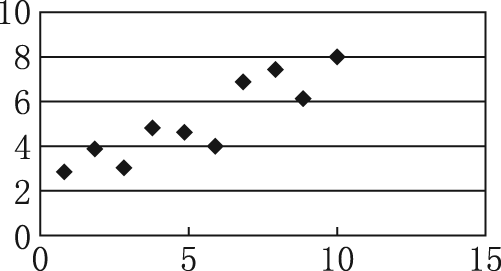


图2

第3题)

解　图1不可以，图2可以．(但是预测和估计时，也要注意范围和实际意义)



一、 问题情境

对任意给定的样本数据，由计算公式都可以求出相应的线性回归方程，但求得的线性回归方程未必有实际意义．

图1中的散点明显不在一条直线附近，不能进行线性拟合，求得的线性回归方程是没有实际意义的；图2中的散点基本上在一条直线附近，我们可以粗略地估计两个变量间有线性相关关系，但是它们线性相关的程度如何？如何较为精确地刻画线性相关关系呢？

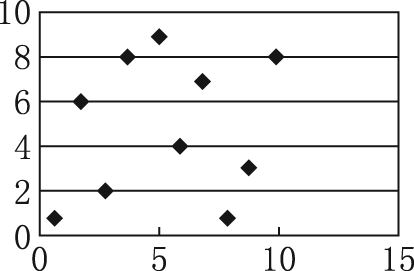


图1

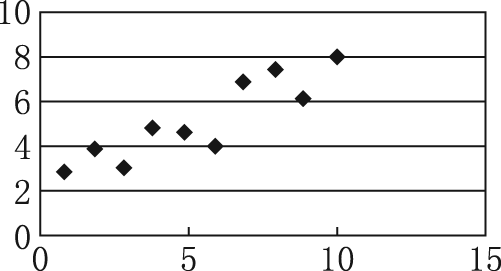


图2

这就是上节课提到的问题，即模型的合理性问题．为了回答这个问题，我们需要对变量*x*与*y*的线性相关性进行检验(简称相关性检验)．

二、 数学建构

1. 直线＝ *x*＋称为这*n*对数据的回归直线，此直线方程即为线性回归方程，而， 的计算公式如下：

或

2. 对于*x, y*随机取到的*n*对数据(*x*i*, y*i)(i＝1, 2, …， *n*)，样本相关系数*r*的计算公式为

r＝

＝

＝.

说明：(1) 线性相关系数*r*的绝对值很小，只是说明线性相关程度低，不一定不相关，可能是非线性相关的某种关系．

(2) *r*是对抽样数据而言的，有时即使|*r*|＝1，两者也不一定是线性相关的．故在统计分析时，不能就数据论数据，要结合实际情况进行合理解释．

三、 数学运用

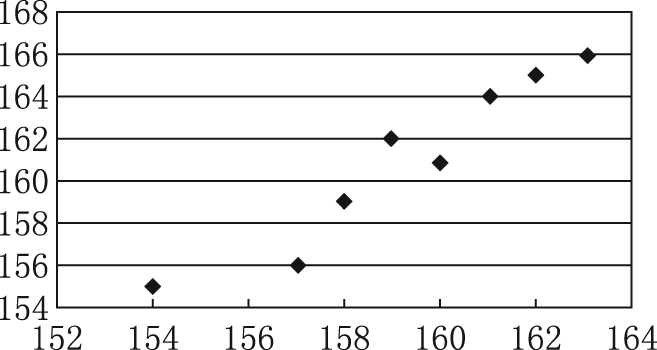
例1　下表是随机抽取的8对母女的身高数据，试根据这些数据探讨*y*与*x*之间的关系．(相关系数*r*的绝对值大于0.75时，认为两个变量有很强的线性相关性，精确到0.01)[1]

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 母亲的身高*x*/cm | 154 | 157 | 158 | 159 | 160 | 161 | 162 | 163 |
| 女儿的身高*y*/cm | 155 | 156 | 159 | 162 | 161 | 164 | 165 | 166 |

(见学生用书课堂本P95)

[处理建议]　画散点图，代入公式计算．

[规范板书]　解　所给数据的散点图如图所示：



(例1)

由图可以看出，这些点在一条直线附近，故*x*与*y*线性相关．

由表中数据计算得

*x*i＝1274, *y*i＝1288, *x*＝202944, *y*＝207484, *x*i*y*i＝205194，＝159.25, ＝161，

所以

r＝

≈0.963＞0.75，

从而可以认为*x*与*y*之间具有较强的线性相关关系．

又＝≈1.345，

＝－ ≈－53.191，

所以线性回归方程为＝－53.191＋1.345*x*.

[题后反思]　解决这类问题的解题步骤：

(1) 作出散点图，直观判断散点是否在一条直线附近；

(2) 求相关系数*r*；

(3) 计算， ，写出线性回归方程．

　研究机构对某校学生往返校时间的统计资料表明：该校学生居住地到学校的距离*x*(单位：km)和学生花费在上学路上的时间*y*(单位：mi*n*)有如下的统计数据：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 到学校的距离*x*/km | 1.8 | 2.6 | 3.1 | 4.3 | 5.5 | 6.1 |
| 花费的时间*y*/mi*n* | 17.8 | 19.6 | 27.5 | 31.3 | 36.0 | 43.2 |

由统计资料表明*y*与*x*具有线性相关关系．

(1) 判断*y*与*x*是否有很强的线性相关性．(相关系数*r*的绝对值大于0.75时，认为两个变量有很强的线性相关性，精确到0.01)

(2) 求线性回归方程＝ *x*＋.(精确到0.01)

(3) 将<27的时间数据i，称为美丽数据．现从这6个时间数据i中任取2个，求抽取的2个数据全部为美丽数据的概率．

参考数据：r＝， *y*i＝175.4, *x*i*y*i＝764.36, (*x*i－)(*y*i－)＝80.30, (*x*i－)2＝14.30, (*y*i－)2＝471.65, ＝82.13.

[规范板书]　解　(1) r＝

＝≈0.98>0.75，

所以*y*与*x*具有很强的线性相关性．

(2) ＝3.9, ＝*y*i≈29.23，

所以＝＝≈5.62, ＝－ ≈29.23－5.62×3.9≈7.31，

从而线性回归方程为＝5.62*x*＋7.31.

(3) 由(2)可知，当*x*＝3.1时，3＝24.732<27；当*x*＝4.3时，4＝31.476>27.所以满足<27的美丽数据共有3个．设3个美丽数据为*a, b, c*，另3个不是美丽数据的为*A, B, C*，则从这6个数据中任取2个共有15种情况，即*ab, ac, bc, aA, aB, aC, bA, bB, bC, cA, cB, cC, AB, AC, BC*，其中，抽取的2个数据全部为美丽数据的有3种情况，即*ab, ac, bc*，所以从这6个数据i中任取2个，抽取的2个数据全部为美丽数据的概率为.

[题后反思]　在求古典概型概率时，如果事件的总数不多，可以用列举法列出所有样本点，计数后可求概率．如果数据较大，样本点的个数较多，则可用排列组合思想求得样本点的个数，再计算概率.

例2　某地区对本地的企业进行了一次抽样调查，下表是这次抽查中所得到的各企业的人均资本*x*(单位：万元)与人均产值*y*(单位：万元)的数据：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 人均资本*x*/万元 | 3 | 4 | 5.5 | 6.5 | 7 |
| 人均产值*y*/万元 | 4.12 | 4.67 | 8.68 | 11.01 | 13.04 |
| 人均资本*x*/万元 | 8 | 9 | 10.5 | 11.5 | 14 |
| 人均产值*y*/万元 | 14.43 | 17.50 | 25.46 | 26.66 | 45.20 |

(1) 设*y*与*x*之间具有近似关系*y*≈*axb*(*a, b*为常数)，试根据表中数据估计*a*和*b*的值；

(2) 估计企业人均资本为16万元时的人均产值(精确到0.01)．[2]

(见学生用书课堂本P95)

[处理建议]　根据*x, y*所具有的关系可知，此问题不是线性回归问题，不能直接用线性回归方程处理．但由对数运算的性质可知，只要对*y*≈*axb*的两边取对数，就能将其转化为线性关系．

[规范板书]　解　(1) 在*y*≈*axb*的两边取常用对数，得lg*y*≈lg*a*＋*b*lg*x*.设lg*y*＝*z,* lg*a*＝*A,* lg*x*＝*X*，则*z*≈*A*＋*bX*.相关数据如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 人均资本*x*/万元 | 3 | 4 | 5.5 | 6.5 | 7 |
| 人均产值*y*/万元 | 4.12 | 4.67 | 8.68 | 11.01 | 13.04 |
| *X*＝lg*x* | 0.47712 | 0.60206 | 0.74036 | 0.81291 | 0.8451 |
| *z*＝lg*y* | 0.6149 | 0.66932 | 0.93852 | 1.04179 | 1.11528 |
| 人均资本*x*/万元 | 8 | 9 | 10.5 | 11.5 | 14 |
| 人均产值*y*/万元 | 14.43 | 17.50 | 25.46 | 26.66 | 45.20 |
| *X*＝lg*x* | 0.90309 | 0.95424 | 1.02119 | 1.0607 | 1.14613 |
| *z*＝lg*y* | 1.15927 | 1.24304 | 1.40586 | 1.42586 | 1.65514 |

易得

由lg＝－0.2155，得≈0.6088，即*a, b*的估计值分别为0.6088和1.5677.

(2) 由(1)知＝0.6088*x*1.567 7.

当*x*＝16时，＝0.6088×161.567 7≈47.01(万元)，

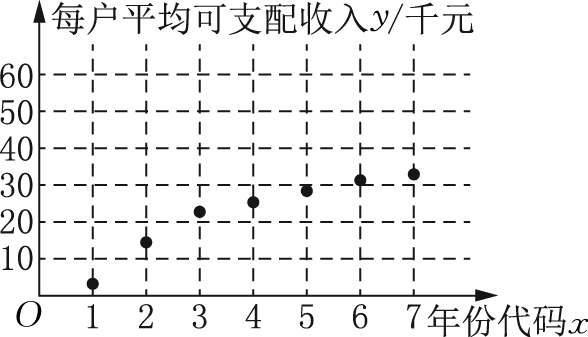
故当企业人均资本为16万元时，人均产值约为47.01万元．

[题后反思]　在解决非线性回归问题时，通过变量置换把非线性回归问题转化为线性回归问题来进行解答．

　发展扶贫产业，找准路子是关键．重庆市石柱土家族自治县中益乡华溪村不仅找准了路，还将当地打造成了种植中药材黄精的产业示范基地．通过种植黄精，华溪村村民的收入逐年递增．以下是2013年至2019年华溪村村民每户平均可支配收入的统计数据：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年　　份 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 |
| 年份代码*x* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 每户平均可支配收入*y*/千元 | 4 | 15 | 22 | 26 | 29 | 31 | 32 |

根据以上数据，绘制了如图所示的散点图．



(变式)

(1) 根据散点图判断＝＋*x*与＝＋l*nx*哪一个更适宜作为每户平均可支配收入*y*(单位：千元)关于年份代码*x*的回归方程模型(给出判断即可，不必说明理由)，并建立*y*关于*x*的回归方程(精确到0.1)．

(2) 根据(1)建立的回归方程，试预测要到哪一年华溪村的每户平均可支配收入才能超过35千元．

(3) 记“恰有一年的每户平均可支配收入超过22千元”为事件*A*，从2013年到2019年中任选两年，求事件*A*的概率．

参考数据：*u*i＝l*nx*i*,* ＝ui.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *x*i*y*i | ui*y*i | u | e2.1 |
| 22.7 | 1.2 | 759 | 235.1 | 13.2 | 8.2 |

参考公式：对于一组数据(*u*1, *v*1), (*u*2, *v*2), …， (*un, vn*)，其回归直线＝＋*u*的斜率和截距的最小二乘估计公式分别为＝， ＝－ .

[处理建议]　从散点图可知图象呈对数型，故选*y*＝*c*＋*d*l*nx*更适合，通过换元将模型转化为线性模型，然后利用表中的数据求解回归方程．

[规范板书]　解　(1) 选择＝＋l*nx*更适合．

＝＝≈14.2，

＝－ ＝22.7－14.2×1.2≈5.7，

所以回归方程为＝5.7＋14.2l*nx*.

(2) 令5.7＋14.2l*nx*>35，得*x*>e2.1≈8.2，

所以到2021年每户平均可支配收入能超过35千元．

(3) 由表中的数据可知，7年中有4年每户平均可支配收入超过22千元，3年每户平均可支配收入不超过22千元，

所以*P*(*A*)＝＝.

四、 课堂练习

1. 两个变量*x, y*的取值如下表所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x* | 4 | 5 | 6 | 7 |
| *y* | 2.5 | 3 | 4.5 | 6 |

若*x, y*具有线性相关关系，且得到的线性回归方程为＝ *x*＋，则(C)

A. ＞0, 4.5＝3.5＋

B. ＞0, 3＝4＋

C. ＞0, 4＝5.5＋

D. ＜0, 5.5＝4＋

提示　*y*随着*x*的增大而增大，所以＞0.

又＝5.5, ＝4，所以4＝5.5＋.

2. 某地区2014年至2020年农村居民家庭人均纯收入*y*(单位：千元)的数据如下表所示：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 年　份 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 |
| 年份代码*x* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 人均纯收入*y*/千元 | 2.9 | 3.3 | 3.6 | 4.4 | 4.8 | 5.2 | 5.9 |

(1) 求*y*关于*x*的线性回归方程；

(2) 利用(1)中的线性回归方程，分析2014年至2020年该地区农村居民家庭人均纯收入的变化情况，并预测该地区2022年农村居民家庭人均纯收入．

参考公式：＝， ＝－ .

解　(1) ＝＝4, ＝＝4.3，

*x*i*y*i＝1×2.9＋2×3.3＋3×3.6＋4×4.4＋

5×4.8＋6×5.2＋7×5.9＝134.4

*x*＝12＋22＋32＋42＋52＋62＋72＝140，

所以＝＝＝0.5，

＝－ ＝4.3－0.5×4＝2.3，

从而＝0.5*x*＋2.3.

(2) 因为＝0.5>0，所以2014年至2020年该地区农村居民家庭人均纯收入逐年增加，平均每年增加0.5千元左右．

将*x*＝9代入回归方程得＝6.8，

故预测该地区2022年农村居民家庭人均纯收入为6.8千元．

3. 一个国家的数学实力往往影响着国家的科技发展，几乎所有的重大科技进展都与数学息息相关，我国第五代通讯技术(5G)的进步就是源于数学算法的优化．华为公司所研发的Si*n*gleRA*N*算法在部署5G基站时可以把原来的4G、 3G基站利用起来以节省开支，华为创始人任正非将之归功于“数学的力量”．近年来，我国加大5G基站建设力度，基站已覆盖所有地级市，并逐步延伸到乡村．

某市2020年已建成的5G基站数*y*与月份*x*的数据如下表所示：

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *x* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| *y* | 283 | 340 | 428 | 547 | 701 | 905 | 1151 | 1423 | 1721 | 2109 | 2601 | 3381 |

探究上表中的数据发现，因年初受新冠疫情影响，5G基站建设进度比较慢，随着疫情得到有效控制，5G基站建设进度越来越快，根据散点图分析，已建成的5G基站数呈现先慢后快的非线性变化趋势，采用非线性回归模型*y*＝ e *x*拟合比较合理，请结合参考数据，求5G基站数*y*关于月份*x*的回归方程．(的值精确到0.01)

注：设*u*＝l*ny*，则*u*i＝l*ny*i(i＝1, 2, …， 12), 　≈1299.17, 　≈6.88, (*x*i－)2＝143, (*x*i－)·(*y*i－)＝37238, (*x*i－)(ui－)≈32.42.

对于样本(*x*i*, y*i)(i＝1, 2, …， *n*)的线性回归方程＝ *x*＋有＝， ＝－ .

解　对于指数模型*y*＝ e *x*，设*u*＝l*ny*，则*u*＝l*n*＋*x*，所以*u*与*x*具有线性相关关系．

因为＝＝6.5, ≈6.88,

(*x*i－)(ui－)≈32.42, (*x*i－)2＝143，

所以＝≈≈0.23，

*ln*＝－ ≈6.88－0.23×6.5≈5.39，

从而*u*＝5.39＋0.23*x*，即*y*＝e5.39＋0.23*x*.

五、 课堂小结

1. 相关系数*r*的计算公式与回归系数的计算公式的比较．

*r*的公式有下列形式：

r＝

＝.

的公式有下列形式：

＝＝

＝.

请根据题目具体情况灵活选用．

2. 求线性回归方程的步骤

(1) 用散点图或进行相关性检验判断两个变量是否具有线性相关关系；

(2) 求系数；

(3) 求： ＝－ ；

(4) 写出线性回归方程.



[1] 利用线性回归方程解决一些简单的线性模型的运算，利用相关系数作回归分析．

[2] 非线性模型问题．