2013 教育部-IBM 产学合作专业综合改革项目
厦门大学《数据挖掘原理及实践》课程习题

第 5 章 数据的相关和回归分析

1. 对于向量 X={2, -1, 0, 2, 0, -3}, Y={-1, 1, -1, 0, 0, -1}, 用 Pearson 方法测定相关性, 并讨论是正相关还是负相关?

2. 受某啤酒公司的委托, 尼尔森咨询公式就啤酒市场进行了详细的品牌调查。现在截获部分数据如下表, 现对啤酒品牌的相似度进行分析。

编号	啤酒品牌	热量/cal	钠含量/%	酒精含量/%	价格
1	Budweiser	144	19	4.7	0.43
2	Schlitz	181	19	4.9	0.43
3	Ionenbrau	157	15	4.9	0.48
4	Kronensourc	170	7	5.2	0.73
5	Heineken	152	11	5.0	0.77
6	Old-milaukee	145	23	4.6	0.26
7	Aucsberger	175	24	5.5	0.40
8	Strchs-bohemi	149	27	4.7	0.42

3. 经调查, 某地区住宅建筑面积和建筑成本的有关资料如下图, 求建筑面积与建筑成本的回归分析。

工地编号	建筑面积 (x)/万米^2	建造成本 (y)/万元
1	4	14.8
2	2	12.8
3	3	13.3
4	5	15.4
5	4	14.3
6	5	15.9

4. 某种水泥在凝固时放出的热量 y (卡/克)与水泥中的四种化学成分 x_1, x_2, x_3, x_4 有关, 具体观测得到的数据见下表, 试从中选出主要变量, 建立 y 关于它们的线性回归方程。

序号	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

5. 在一次关于公共交通的社会调查中, 一个调查项目是“乘坐公交汽车上下班, 还是骑自行车上下班”, 调查对象为工薪族群体, 数据见下表, 表中因变量“y=1”表示“主要乘坐公交汽车上下班”, “y=0”表示“主要骑自行车上下班”, 自变量为 x_1, x_2, x_3, 其中 “x_3=1” 表示“男性”, “x_3=0” 表示“女性”。对该数据建立 y 与自变量间的 Logistic 回归方程。

序号	年龄 x_1	月收入 x_2	性别 x_3	y
1	18	850	0	0
2	21	1200	0	0
3	23	850	0	1
4	23	950	0	1
5	28	1200	0	1
6	31	850	0	0
7	36	1500	0	1
8	42	1000	0	1
9	46	950	0	1
10	48	1200	0	0
11	55	1800	0	1
12	56	2100	0	1
13	58	1800	0	1
14	18	850	1	0
15	20	1000	1	0
16	25	1200	1	0
17	27	1300	1	0
18	28	1500	1	0
19	30	950	1	1
20	32	1000	1	0
21	33	1800	1	0
22	33	1000	1	0
23	38	1200	1	0
24	41	1500	1	0
25	45	1800	1	1
26	48	1000	1	0
27	52	1500	1	1
28	56	1800	1	1

6. 政治科学家们怀疑政治候选人的一些承诺和他们一旦当选之后兑现承诺两者之间存在一定的关系, 下表列出了 10 位政客在这方面的“追踪记录”, 请计算承诺和兑现之间的相关系数。

政客	承诺数	兑现数
1	21	7
2	40	5
3	31	6
4	62	1
5	28	5
6	50	3
7	55	2
8	43	6
9	61	3
10	30	5

7. 《1995 年美国统计摘要》列出了 1960-1994 年不同种类的消费价格指数。消费价格指数反映了所有城市消费者的购买模式。请就下表数据做一次回归分析, 确定运输与能源之间的关系, 并计算相关系数的置信区间。

年份	能源	运输
1960	22.4	29.8
1961	22.5	30.1
1962	22.6	30.8
1963	22.6	30.9
1964	22.5	31.4
1965	22.9	31.9
1966	23.3	32.3
1967	23.8	33.3
1968	24.2	34.3
1969	24.8	35.7
1970	25.5	37.5
1971	26.5	39.5
1972	27.2	39.9
1973	29.4	41.2
1974	38.1	45.8
1975	42.1	50.1
1976	45.1	55.1

8. 请阐述简单线性回归分析的基本假设有哪些。

9. 请简述回归分析的主要目的。

10. 简单相关分析有哪些典型的分析方法，它们各自的特点是什么？