2019-2020(2) 《数据分析》期末考试
案例分析大作业部分

题目：根据2020.4.12的全球疫情况分析主要爆发的国家以及其未来增长趋势

姓名：李金哲

学号：201707010119

班级：信171

成绩：

日期：2020.06.13

目录

一、 问题探究 1：根据 4 月 12 日的新增死亡和新增感染可以将全球国家分为几类

1.1 观测数据提出假设

1.1.1 直方图结果展现

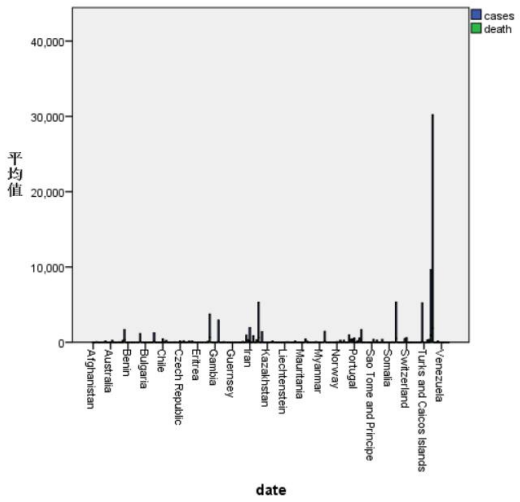

图 1-1-1 *分层直方图*

1.1.2 分析结果

根据图 1-1-1，能够明显的看出在 Venezuela 和 Turks and Caicos Islands 直接存在着两个国家的分层数据远高于其他的国家，根据原始数据（图 1-1-2）可以大致判断，United States 和 United Kingdom 两个国家在 4 月 12 日受新型冠状肺炎影响最为严重。

190	Tunisia	14	3
191	Turkey	5138	95
192	Turks and Caicos Islands	0	0
193	Uganda	0	0
194	Ukraine	308	4
195	United Arab Emirates	376	4
196	United Kingdom	8719	917
197	United States	28391	1831
198	United States Virgin Islands	1	0
199	Uruguay	7	0
200	Uzbekistan	172	1
201	Vatican	0	0
202	Venezuela	0	0
203	Vietnam	1	0
204	Yemen	0	0
205	Zambia	0	0
206	Zimbabwe	3	0

图 1-1-2 *原始数据*

1.2 通过主成分分析，得到各主成分的贡献率

1.2.1 主成分结果展现

表 1-2-1(1) 相关性矩阵

相关性矩阵			
		cases	death
相关性	cases	1.000	.945
	death	.945	1.000

表1-2-1(2) 巴特利特球度检验

KMO 和巴特利特检验		
KMO 取样适切性量数。		.500
巴特利特球形度检验	近似卡方	451.302
	自由度	1
	显著性	.000

表1-2-1(3) 总方差解释

总方差解释						
成分	初始特征值			提取载荷平方和		
	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	1.945	97.232	97.232	1.945	97.232	97.232
2	.055	2.768	100.000			
提取方法：主成分分析法。						

1.2.2 分析结果

表 1-2-1（1）是原有变量的相关系数矩阵。可以看到:两个变量的相关系数十分高且无限接近 1，因此可以得出两个变量呈现强线性关系，能够从中提取公共因子，**可以进行因子分析。**

由表 1-2-1（2），可知巴特利特球度检验统计量的观测值为 451.302，相应的概率 P-值接近 0。如果显著性水平 a 为 0.05，由于概率 P 值小于显著性水平 a，则应拒绝原假设，认为相关系数矩阵与单位阵有显著差异。同时，KMO 值为 0.500，根据 Kaiser 给出的 KMO 度量标准可知**原有变量可以进行但不太适合因子分析**，对上述结论起到补充说明。

根据表 1-2-1（3）可以得出:第 1 个因子的方差贡献为 1.945，解释原有 2 个变量总方差的 97.232%，累计方差贡献率为 97.232%;第 2 个因子的方差贡献为 0.055，解释原有 2 个变量总方差的 2.768%，累计方差贡献率为 100.000%。此时在初始解中由于提取了 2 个因子，原有变量的总方差均被解释，累计方差贡献率为 100%。**此时认定主**

成分之间应该是相互独立的，不存在包含不包含的关系。

1.3 通过聚类分析，得到具体分类结果

1.3.1 聚类分析结果展现

表1-3-1(1) 初始聚类中心

初始聚类中心			
	聚类		
	1	2	3
cases	5138	0	28391
death	95	0	1831

表1-3-1(2) 迭代历史

迭代历史记录[a]			
	聚类中心中的变动		
迭代	1	2	3
1	463.632	121.409	.000
2	.000	.000	.000
a. 由于聚类中心中不存在变动或者仅有小幅变动，因此实现了收敛。任何中心的最大绝对坐标变动为 .000。当前迭代为 2。初始中心之间的最小距离为 5138.878。			

表1-3-1(3) 最终聚类中心

最终聚类中心			
	聚类		
	1	2	3
Cases	4886	121	28391
Death	484	7	1831

表1-3-1(4) 聚类个案数

每个聚类中的个案数目		
聚类	1	6.000
	2	198.000
	3	1.000
有效		205.000
缺失		1.000

1.3.2 分析结果

表 1-3-1（1） 展示了 3 个类的初始类中心点的情况。3 个初始类中心点的数据分别是（5138，95）、（0，0）、（28391，1831）。 可得**初始中心中的第 3 类各指数均是最优的，第 1 类次之，第 2 类各指数最不理想。**

表 1-3-1（2）展示了 3 个类中心点每次送代时的偏移情况。由表 2-1-2 可知，第 1 次送代后，3 个类的中心点分别偏移了 463.632，121.409，0，第 1 类中心点偏移最大。迭代 1 次后，3 个类的中心点的偏移均小于指定的判定标准 (0.02)，聚类分析结束

表 1-3-1（4）显示了，聚类 2 包含的样本最多，聚类 3 包含的样本最少。通过 K 聚类分析可以根据不同国家的新型冠状肺炎的新增死亡人数和新增感染人数产生一个大致的分类情况。结合国家分类（图 1-3-2(a)~图 1-3-2(g)）以及 4 月 12 日新增死亡和感染人数的原始数据，能够分为 爆发程度严重，难以控制、爆发程度较为严重，不好控制、爆发程度轻微，容易控制，三个分类。**其中聚类结果中的第 1 类为爆发程度较为严重，不好控制，主要包含国家为：United Kingdom、Turkey、Spain、Italy、Germany、France，共 6 个国家；第 3 类为爆发程度严重，难以控制，主要包括国家为：United States，共 1 个国家；第 2 类为爆发程度轻微，容易控制，成员为剩余的 198 个国家。**

197	United States	3	.000

图 1-3-2(a) 国家分类

196	United Kingdom	1	3857.361

图 1-3-2(b) 国家分类

191	Turkey	1	463.632

图 1-3-2(c) 国家分类

176	Spain	1	61.671

图 1-3-2(d) 国家分类

73	Germany	1	2095.321

图 1-3-2(e) 国家分类

68	France	1	1778.408

图 1-3-2(f) 国家分类

97	Italy	1	234.615

图 1-3-2(g) 国家分类

二、 问题探究 2：根据 2019 年 12 月 31 日到 2020 年 4 月 12 日的新增死亡和新增感染人数变化趋势，预测 2020 年 4 月 13 日 United States 可能会有多少新增死亡和新增感染人数，对比实际数据进行分析

2.1 回归分析曲线

2.1.1 回归分析结果展现

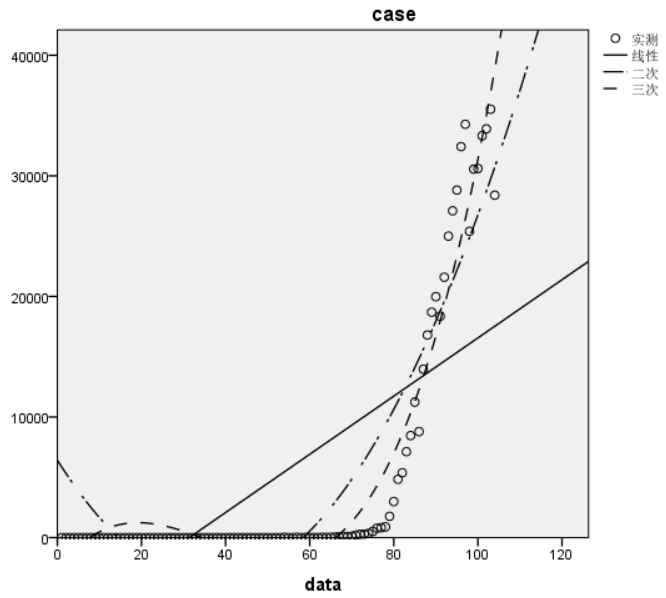

图2-1-1(a) 每日新增确诊回归曲线

表2-1-1-1(a) 每日新增确诊2次回归系数

系数					
	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
data	-550.231	53.831	-1.608	-10.222	.000
data ** 2	7.539	.497	2.387	15.178	.000
(常量)	6410.285	1224.466		5.235	.000

表2-1-1-1(b) 每日新增确诊3次回归系数

系数					
	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
data	465.041	80.994	1.359	5.742	.000
data ** 2	-16.519	1.788	-5.230	-9.240	.000
data ** 3	.153	.011	4.811	13.644	.000
(常量)	-2685.194	986.711		-2.721	.008

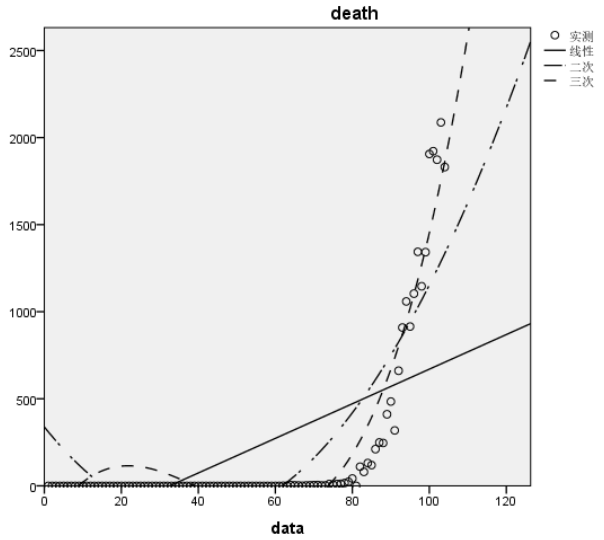

图2-1-1(b) 每日新增死亡回归曲线

表2-1-1-2(a) 每日新增死亡2次回归系数

系数					
	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
data	-27.512	3.396	-1.700	-8.102	.000
data ** 2	.357	.031	2.389	11.383	.000
(常量)	338.099	77.237		4.377	.000

表2-1-1-2(b) 每日新增死亡3次回归系数

系数					
	未标准化系数		标准化系数	t	显著性
	B	标准误差	Beta		
data	38.547	4.795	2.382	8.039	.000
data ** 2	-1.209	.106	-8.095	-11.420	.000
data ** 3	.010	.001	6.621	14.995	.000
(常量)	-253.701	58.416		-4.343	.000

2.1.2 分析结果

1）图2-1-1(a) 每日新增确诊回归曲线，可以确定二次项方程和三次项方程的变化趋势曲线较为贴近原数据，由于在回归方程显著性检验中，二次模型和三次模型的概率P均值近似为零(**故在此不粘贴回归方程显著性检验的表格了**)。当显著性水平a为0.05时，应拒绝回归系数检验的原假设，**在二次或三次模型每日新增确诊和日期呈现明显相关性，应保留在模型中**。

根据系数矩阵表2-1-1-1(a)和表2-1-1-1(b)，确定最佳的回归方程。若设显著性水平a为0.005时，3次项回归方程中的系数的概率P等于0.0008大于0.0005，接受原假设，不呈现显著性。而**2次项回归方程中的全部概率P均约等于0，小于0.0005，拒绝原假设，呈现相关性**。

回归方程为

$$每日新增确诊 = -550.231 * 爆发天数 + 7.539 * 爆发天数^2 + 6410.285$$

2）图2-1-1(b) 每日新增确诊回归曲线，可以确定二次项方程和三次项方程的变化趋势曲线较为贴近原数据，由于在回归方程显著性检验中，二次模型和三次模型的概率P均值近似为零(**故在此不粘贴回归方程显著性检验的表格了**)。当显著性水平a为0.05时，应拒绝回归系数检验的原假设，**在二次或三次模型每日新增确诊和日期呈现明显相关性，应保留在模型中**。

根据系数矩阵表2-1-1-1(a)和表2-1-1-1(b)，确定最佳的回归方程。若设显著性水平a为0.005时，**3次项回归方程和2次项回归方程中的全部概率P均约等于0，小于0.0005，拒绝原假设，呈现相关性**。

由于二次和三次回归方程的系数均符合假设，故二者之中选取最优构建方程，使用**后期更加贴合原数据的三次方程构建回归方程曲线**。

回归方程为

$$每日新增死亡 = 38.547 * 爆发天数 - 1.209 * 爆发天数^2 + 0.010 * 爆发天数^3 - 253.701$$

2.2 4 月 13 日 United States 新增死亡和新增感染人数预测

2.2.1 结果展示

data	case	death	FIT_1	LCL_1	UCL_1	FIT_4	LCL_4	UCL_4
100	30613	1906	26773.46771	18431.88101	35115.05440	1452.76921	1156.68032	1748.85809
101	33323	1922	27738.50148	19377.23565	36099.76731	1549.52076	1251.73751	1847.30401
102	33901	1873	28718.61252	20336.20420	37101.02084	1649.87774	1350.12733	1949.62816
103	35527	2087	29713.80082	21308.72977	38118.87187	1753.89980	1451.88291	2055.91668
104	28391	1831	30724.06640	22294.75554	39153.37725	1861.64654	1557.03665	2166.25643
105	.	.	31749.40923	23294.22479	40204.59367	1973.17762	1665.62047	2280.73477

图 2-2-1 预测结果

2.2.2 分析结果

根据图 2-2-1 可知，如果 **United States 不采用任何控制，在 4 月 13 日可能会产生 31749 名新增确诊患者(区间为 23294 ~ 40204)和 1973 名新增死亡的患者(区间为 1665 ~ 2280)**。

根据图 2-2-2 的 4 月 13 日美国新增患者人数，可以明显看出在 **4 月 13 日美国产生了 27176 名新增确诊患者和 1513 名新增死亡患者**，基本符合预测曲线，证明截止到 4 月 13 日，美国暂时没有采用任何新型冠状肺炎的防疫措施。

Country, Other	Total Cases	New Cases	Total Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical	Tot Cases/ 1M pop	Deaths/ 1M pop	Total Tests	Tests/ 1M pop
World	1,851,264	+71,422	114,160	+5,381	422,572	1,314,532	50,764	237	14.6		
USA	560,055	+27,176	22,090	+1,513	31,986	505,979	11,766	1,692	67	2,832,258	8,557

图 2-2-2　4 月 13 日美国疫情人数（MedSci 期刊数据）

来源（https://www.medsci.cn/article/show_article.do?id=57fb1921e573）