基于大数据的人寿保险公司续保数据分析

李青娟 杨 杉

四川大学锦城学院计算机与软件学院 四川 成都 611731

【摘 要】2020 年保险行业因疫情发展受阻，但随着疫情的消退，保险行业整体的发展一直呈稳步发展。本文针对婚姻状况与总保费之间的关系，总保费与年龄、过去三年平均年收入、缴费期限、保额之间的关系，总保费和保额之间是否具有相关性等问题，采用 SPSS 和 Excel 中的单因素方差分析、探索分析、相关性分析、线性回归模型等方法对续保相关数据进行数据分析，得出相关结论并提出建议。

【关键词】续保数据，数据分析，SPSS

1 引言

2020 年保险行业因疫情发展受阻，但随着疫情的消退，保险行业整体的发展一直呈稳步发展。在《中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要》中，全文提及“保险”关键词有 30 余次，强调“深化保险公司改革，提高商业保险保障能力”[1]。中国是全球最大的保险增量市场，在发达国家，保险市场已经处于缓慢发展或者基本饱和的状态，而我国的保险行业还有很大的发展空间，保险市场仍将长期处于快速发展阶段 。随着经济的高速发展，风险的产生概率也在不断的提升，对此，保险作为管控风险的重要手段，成为了经济社会发展的坚实后盾[2]。本文通过数据分析后利用所学知识，在 SPSS 和 Excel 中进行数据分析，从而得出保险行业及保险产品当前的发展状况，进一步规划出之后的发展方向。主要采用 SPSS 和 Excel 中的单因素方差分析、探索分析、相关性分析、线性回归模型、和简单相关分析的方法对续保相关数据进行数据分析，得出相关结论并提出建议。

2 研究思路

首先对保险公司的续保数据进行数据清洗，从而提高数据准确性主要是删去无效数据、会产生较大误差的数据以及对数据分析没有意义的数据，针对不同险种的变化对总保费的平均水平是否有显著性影响，婚姻状况与总保费之间的关系，总保费与年龄、过去三年平均年收入、缴费期限、保额之间的关系，总保费和保额是否具有相关性等问题。通过 SPSS 进行数据分析，主要采用单因素方差分析、探索分析、相关性分析、线性回归模型等方法对续保相关数据进行数据分析，根据相应的结果得出具体的数据分析报告，并得出相关结论和提出建议。

3 数据说明

3.1 数据准备

数据来源于人寿保险公司的续保数据，记录了不同机构不同险种的投保人信息。本数据共 218480 行，16 列，18.2M。主要包括客户的基本信息：客户号、性别、年龄、婚姻状况、过去三年平均年收入，教育程度，职业，家庭人口以及购买保险的基本信息：机构、险种、投保时间、投保时间、缴费方式、缴费期限、投保份数、总保费、保额。其中代表婚姻状况的数据中 M 指已婚，S 指未婚，D 指离异，R 指再婚，W 指丧偶，X 指未知。

3.2 数据清洗

将过去三年年收入为 0 的数据进行删除。教育程度和家庭人口数据是异常数据，没有显示正确的值，因此将对数据分析无用的数据列进行删除，例如家庭人口、机构、缴费方式、缴费期限、客户号、教育程度等。由于险种分组过多，无法做出单因素方差分析，通过筛选、替换和查找的方法对险种进行清洗，B 开头的险种用 1 表示，S 开头的险种用 2，Y 开头的险种用 3，然后是 4 和 6 开头的险种，得到 1,2,3,4,6，五个险种。其他险种由于数据过少不具有普遍性，会影响后续的分析，因此将这些数据删去。

4 数据分析

4.1 分析不同险种的变化对总保费的平均水平是否有显著性影响

首先在 SPSS 中用描述统计功能对数据中的险种进行频率分析，为后续的分析做准备。险种 1 的续保率最高，超过了 100%；险种 2 的续保率不足 40%；险种 3 的续保率接近一半 50%；险种 4 的续保率最低，只有大约 0.4%；险种 6 续保率也只有 2% 左右；总续保率约为 24%。还可以得到险种 2 数量明显高于其它险种，所以将险种 2 分为一组，险种 1 3 4 6 分为一组。对其进行方差齐次性检验，其显著性小于显著性水平 0.05，所以方差不具有齐次性。对险种和总保费进行单因素方差分析，其 ANOVA 表如表一所示，多重比较表如图一所示

由表 1 可知显著性 =0.000<0.05，df=4，至少有四种险种具有显著差异，显著性（P）= 0.00<005，所以险种 1 对总保费的平均水平的影响与其他四个险种是有显著差异的。由图 1 可知，因为方差不具有齐次性，所以只看比较表的下半部分。险种 1 与险种 4 对总保费的影响是没有显著差异的，其余险种之间是有显著差异的。险种 1 显著高于险种 2 6，低于险种 3；险种 2 显著高于险种 6，低于险种 3 4；险种 3 显著高于险种 4 6；险种 4 显著高于险种 6。最终得到险种对总保额的影响：

险种 3 > 险种 4 > 险种 1 > 险种 2 > 险种 6

表 1ANOVA

	平方和	df	均方	F	显著性
组间	8.174E9	4	2.043E9	197.301	0.000
组内	1.596E12	154118	10356651.53		
总数	1.604E12	154122			

Tamhane	1	2	134.5794728*	14.0548749	.000	95.223148	173.935797
		3	-1316.584602	91.1584720	.000	-1571.969485	-1061.199719
		4	-221.2615245	110.8384393	.377	-532.355193	89.832144
		6	966.9917510*	11.2199011	.000	935.564776	998.418726
	2	1	-134.5794728	14.0548749	.000	-173.935797	-95.223148
		3	-1451.164074	90.8752671	.000	-1705.757707	-1196.570443
		4	-355.8409974	110.6056363	.013	-666.287888	-45.394106
		6	832.4122782*	8.6216783	.000	808.274444	856.550112
	3	1	1316.584602	91.1584720	.000	1061.199719	1571.969485
		2	1451.164074	90.8752671	.000	1196.570443	1705.757707
		4	1095.323077	142.6391752	.000	695.572066	1495.074089
		6	2283.576353	90.4801683	.000	2030.086520	2537.066187
	4	1	221.2615245	110.8384393	.377	-89.832144	532.355193
		2	355.8409974*	110.6056363	.013	45.394106	666.287888
		3	-1095.323077	142.6391752	.000	-1495.074089	-695.572066
		6	1188.253275	110.2812470	.000	878.707516	1497.799035
	6	1	-966.9917510	11.2199011	.000	-998.418726	-935.564776
		2	-832.4122782	8.6216783	.000	-856.550112	-808.274444
		3	-2283.576353	90.4801683	.000	-2537.066187	-2030.086520
		4	-1188.253275	110.2812470	.000	-1497.799035	-878.707516

图 1

4.2 分析婚姻状况与总保费之间的关系

首先进行方差齐次性检验，由方差齐次性检验可知方差不具有齐次性。且由结果可得知离婚与结婚之间、结婚与单身之间总保费金额有显著性差异。离婚人士和单身人士的续保金额高于结婚人士。

选择分析栏中的描述统计中的探索分析，设置总保费为因变量列表，婚姻状况为因子列表，设置平均值的置信区间为 95%。由分析结果可知离异（D）人士的总保费金额置信区间为 1031-1189，已婚（M）人士的总保费金额置信区间为 885-912，未婚（S）人士的总保费金额置信区间为 1031-1086，丧偶（W）人士的总保费金额置信区间为 1010-1702，单身人士的续保率最高，对保费的投资也相应较高。婚姻状态为丧偶的人群方差最大，说明该类人群购买保险的能力参差不齐。

通过表 2 比较均值栏中的均值，的出每种婚姻状况下总保费的均值，可以分析得到：婚姻状态为未知的人群购买保险的能力最强，平均数排在第一名，而婚姻状态为已婚的人群购买保险的能力稍弱，中平均值排在最后一名。

表 2 均值

婚姻状况	总保费
D	1110.415149
M	898.756270
R	710.000000
S	1059.337522
W	1356.625193
X	1524.214267
总计	1045.304831

4.3 分析总保费与年龄、过去三年平均年收入、缴费期限、保额之间的关系

先对总保费与年龄、过去三年平均年收入、缴费期限、保额做相关性分析，在 SPSS 中采用相关性分析的方法，由于它们都是刻度集数据，所以选择 Pearson 相关系数，通过分析结果可以看出，总保费与年龄、过

系数a

模型		非标准化系数 B	非标准化系数 标准 误差	标准系数 试用版	t	Sig.
1	(常量)	428.247	7.640		56.051	.000
	保额	.043	.000	.271	131.777	.000
2	(常量)	2436.853	18.520		131.577	.000
	保额	.052	.000	.325	158.656	.000
	缴费期限	-132.486	1.120	-.242	-118.328	.000
3	(常量)	2275.612	18.200		125.032	.000
	保额	.047	.000	.298	147.452	.000
	缴费期限	-127.876	1.097	-.234	-116.589	.000
	过去三年平均年收入	.011	.000	.193	98.067	.000
4	(常量)	1272.774	33.290		38.233	.000
	保额	.040	.000	.300	140.774	.000
	缴费期限	-124.148	1.098	-.227	-113.017	.000
	过去三年平均年收入	.011	.000	.194	98.575	.000
	年龄	22.943	.639	.070	35.932	.000

a. 因变量: 总保费

图 2 系数

4.4 分析总保费和保额是否具有相关性

由于保额和总保费都是刻度级数据，所以用 Pearson 相关系数进行分析，由分析结果表 2 可以看出保额和总保费的相关系数为 0.281，结果具有两个 * 号表示保额与总保费之间有非常显著的正相关关系，第二行的 Sig 结果为 0.000 小于 0.01，所以拒绝原假设，即保额和总保费之间具有显著的相关性。所以保额和总

保费之间具有显著的正相关关系。

表 3 相关性

		总保费	保额
总保费	Pearson 相关性	1	.271
	显著性（双侧）N	218478	0.000
			218478
保额	Pearson 相关性	.271	1
	显著性 N	.000	218478
		218478	

5 结论及建议

5.1 结论

通过性关系分析得出保额和总保费之间具有显著的正相关关系。

通过对不同险种的变化对总保费的平均水平是否有显著性影响的分析得出任意两个险种对总保费的影响都是有显著差异的，险种 2 3 4 均显著高于险种 1，险种 1 显著高于险种 6。

通过分析婚姻状况与总保费之间的关系，得出婚姻状态为未知的人群购买保险的能力最强，平均数排在第一名,而婚姻状态为已婚的人群购买保险的能力稍弱，中平均值排在最后一名。

离婚和结婚之间、结婚和单身之间总保费金额有显著性差异。离婚人士和单身人士的续保金额高于结婚人士。

通过分析总保费与年龄、过去三年平均年收入、缴费期限、保额之间的关系得出总保费与年龄、过去三年平均年收入、缴费期限、保额之间都有显著的正相关性，可以建立线性回归模型。最后得出线性回归模型为 y（总保费）=1365.936+00.011* 过去三年平均年收入 +24.322* 年龄 +0.051* 保额 -137.54* 缴费期限，总保费都与投保人的年龄、过去三年平均年收入、保额与缴费数据之间有线性正相关性。总保费与保额之间也存在线性正相关关系。

5.2 建议

对于保险公司，通过结婚状况不同的人群与总保额相关性分析，可以明确的判断出婚姻状态为未知的人群是最为忠诚的。抓住这部分人群，就能够在稳住自己的市场份额。续保时可以着重向单身人士介绍。单身人士的续保率会较多且保费投资金额较多。对于保险的购买人群，应对收入较低的人群或者收入下降的人群，加大续保营销，从而获得更多的成交量。

续保目标也应放在收入高的人身上，因为他们有稳定的工资，对自己生命安全的重视程度更大

对于客户，应该慎重考虑后决定自己是否有必要续保，续保后是否有能力继续承担后续的保费。在收入可观的情况下，建议保值可适当买大。

【参考文献】

[1] 翟文博 . 把握保险行业发展机遇 [J]. 经济 , 2021,(04):103.

[2] 闻豪 . 中国保险业服务经济的绩效评价 [J]. 现代商业 , 2021,(09):50-52.