数据岗位招聘分析

——基于岗位招聘数据线性回归分析

strength library and muscle laboratory（第七小组）

樊可 韩思越 黄永晟 杨远琨 俞铖昊

一.背景

随着物联网、社交网络、云计算等技术普及，计算能力、存储空间、网络带宽的高速发展，如何对大规模数据进行分析，提炼其背后的实际价值，已然成为时代话题。数据分析岗位由此应运而生。作为新兴产业，公司经营者，应当如何制定合理薪资，招聘真正有价值有潜力的数据分析人才，以有限的薪资成本换取更高更理想的企业收益；求职应聘者，如何科学理性地评估自我职业技能水平，合理准确定位就业方向，选择适合个人的职位并取得理想薪资；高校教育者，又该如何建设选择合适的学生培养方向，以期培养高校生毕业后事业有成，能够找到高薪资工作，享受美好舒适生活，都有待数据的解答。

我组先通过对数据分析岗位招聘的相关信息进行分词提取等预处理工作，筛选有价值信息。再以薪资为因变量，其余因素作为自变量建立线性回归模型。最后，由所得模型结果，分别针对企业经营机构，求职应聘者，高校教育机构，以数据说理，提出合理建议。

二.数据预处理及变量说明

本案例收集了 2016 年 9 月各大招聘网站发布的数据分析岗位招聘的相关信息，对大数据相关行业岗位薪资相关影响因素展开研究。我们首先对数据进行预处理，筛除薪资项缺失数据异常数据，并筛除重复数据信息，最终保留 7111 条数据。

便于后续分析，我们对数据按照变量类别，分为“岗位信息”、“公司信息”、“求职者要求”三个部分，对 “地区”,“公司类别”等离散变量分别制成哑变量形式,“公司类别”,“行业类别”等变量中数量较少的值均进行合并，并用最高薪资与最低薪资的平均值作为平均薪资。

为从 “描述”,“职位”两段自然语言文本栏中提取可供分析的信息，我们利用分词处理方法，去除标点、中文停止词等无效信息后进行词频统计，并分别筛选出词频较高且具有数据价值

的“技能”,“岗位名称”作为因变量用于后续分析。经预处理后，用于下文描述分析与回归分析讨论的变量情况，如表 1 所示。

表一　变量表

变量类型		变量名		详细说明	取值范围	备注
因变量		平均薪资		单位：千元/月	1.5 —— 400	
自变量	岗位信息	岗位名称		文本数据	如：数据分析师、产业运营经理	用于描述分析
		所属行业		定性变量 共 7 个水平	如：互联网/电子商务、金融/投资/证券、计算机软件等	
	公司信息	公司规模		定性变量 共 7 个水平	少于 50 人、50-150 人、150-500 人、500-1000 人、1000-5000 人、5000-10000 人、10000 人以上	非显著性变量，用于描述分析
		所在地区		定性变量 共 6 个水平	北京、上海、深圳、河北、山西、陕西	
		公司类别		定性变量 共 6 个水平	民营公司、上市公司、国企、合资、外资、其他	用于描述分析
	求职者要求	学历背景		定性变量 共 7 个水平	无、中专、高中、大专、本科、硕士、博士	
		从业经验		定性变量 共 4 个水平	1 年以下、1-3 年、3-5 年、5-10 年	连续变量 分组处理
		技能	软件技能	定性变量 共 11 个水平	Excel、SQL、SEO、PPT、SPSS、SAS、R、WORD、PYTHON、JAVA、HADOOP	热门技能包 高频关键词
			其他技能	定性变量 共 5 个水平	数据分析、数据挖掘、统计、数学、金融	

三. 描述分析

在分析前，先了解下此次分析的数据收集范围。如图 1 所示，本次收集数据集中于上海、北京、深圳三大一线城市与河北、山西、陕西三大内陆省份，能够基本代表国内数据岗位招聘就业情况。并且由图 1 可以看出，一线城市数据分析岗位平均薪资普遍明显高于内陆省份。

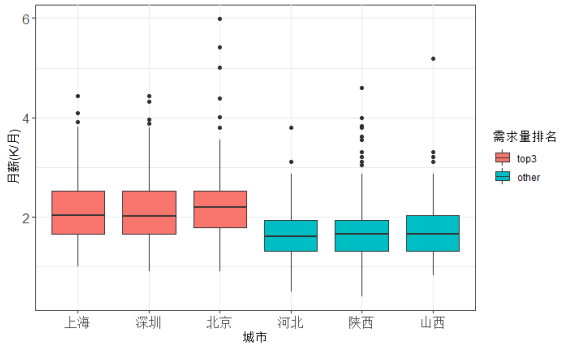

图 1 所在地区与月薪均值箱线图

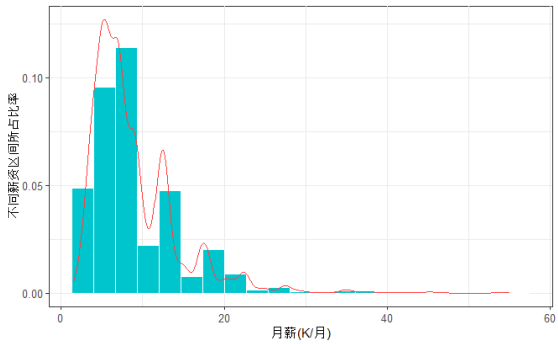

图 2 月薪均值分布直方图

本案例所关注的因变量是平均薪资（单位：千元/月）。从图 2 薪资分布直方图中可以看出，数据分析岗位薪资情况整体呈现正偏态分布，即分布高峰偏向低薪资，长尾向高薪资方向延伸。由此表明，绝大部分数据分析岗位薪资仍然在平均水平以下，约为每月 7000 元左右，而每月薪资高于 2 万元的高收入人群仍不在少数，其中月薪最高值来自汇鼎财富（北京）投资有限公司，月薪高达 40 万。

再考察各类自变量对于薪资的影响。首先是求职者能力需求方面。如图 3 所示，尽管工作经验三年以内，平均薪资水平没有较大提升，但超过三年以后，随着工作经验的增长，企业制定平均薪资水平逐步开始提高。

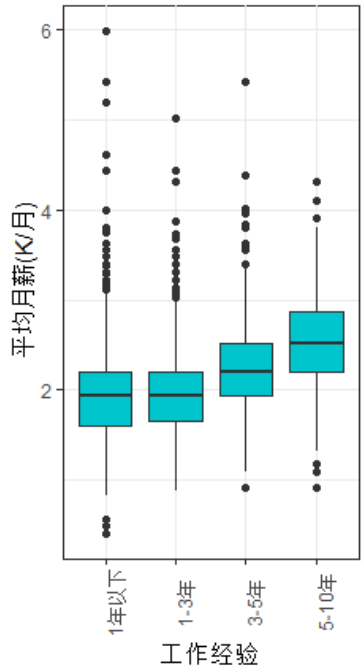

图 3 工作经验与月薪箱线图

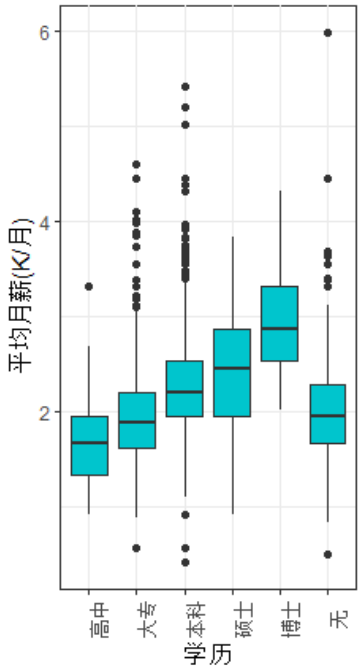

图 4 学历背景与月薪箱线图

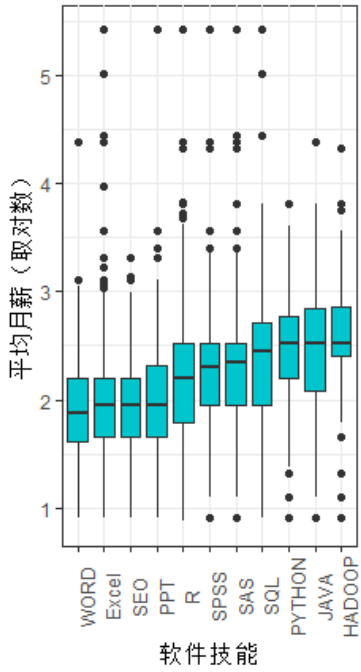

图 5 软件技能与月薪箱线图

于此同时，随着学历的逐步提升，薪资水平也得到明显增长，如图 4 所示。在软件技能方面，图 5 也说明，Excel，Word，PPT 作为最为基础的办公软件，即便再精通也难以依次拿到较高薪资，而驾驭 SQL,PYTHON,JAVA,HADOOP 等高级编程语言，往往能够拿到相对较高的薪资。

再看公司信息方面，企业规模与公司类别（图 6，图 7）两项因素对于月薪的影响并不明显。可见无论是选择一家上万人的国企，还是选择一家一百人不到的小型民营公司，对个人的薪资影响并不明显，与之相比更为重要的还是学历经验等个人因素。

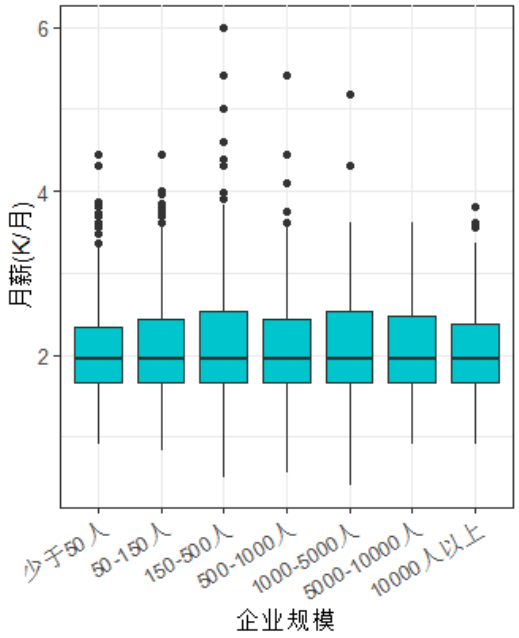

图 6 企业规模与月薪均值箱线图

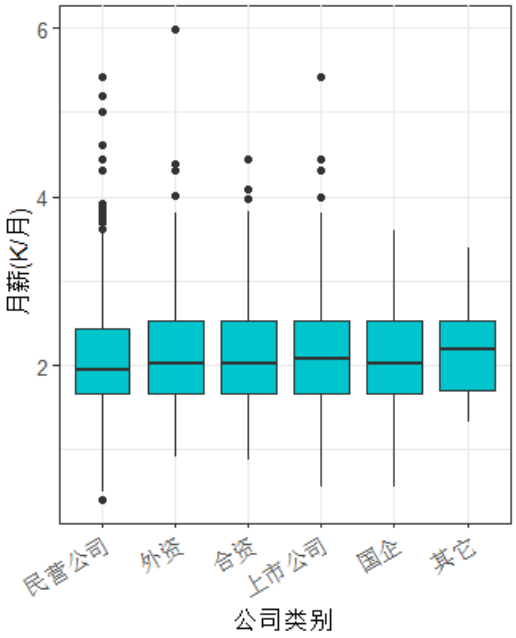

图 7 公司类别与月薪均值箱线图

四.线性回归模型

为了更直观地描述比较各因素对于最终数据分析岗位薪资的影响，我们尝试以薪资为因变量，由描述分析结果选取对薪资影响较大的因素作为自变量建立线性回归模型。通过模型结果，对各个因素影响进行比较解释分析。由于线性回归效果不甚理想，对线性回归模型做如下操作：

步骤一：由月薪均值分布直方图（图 2）可见，高于 10 万的平均薪资数据样本量过少，该类数据不仅不适合线性回归拟合，也破坏了数据集中性，严重影响拟合效果。我们推测该类岗位仅针对少数突出数据人才,对绝大多数就业群体参考价值不大,因此为得到较为理想的拟合效果，对该类数据进行舍弃。舍弃后模型诊断图如图 8 所示。

步骤二：易见，图 8 内 Q-Q 图与正态分布有较大差异。猜测是由于少数离群点对于数据产生了较大的影响，可能原因是市场本身配置不均与公司企业未合理考虑市场实际情况制定不合理薪资所致，该类离群点属于极少数现象，不具备普适参考价值，因此将离群点删除后重新进行线性回归，模型诊断图如图 9 所示。

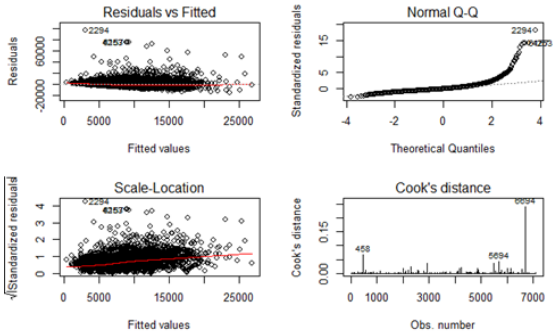

图 8 线性回归模型诊断图一

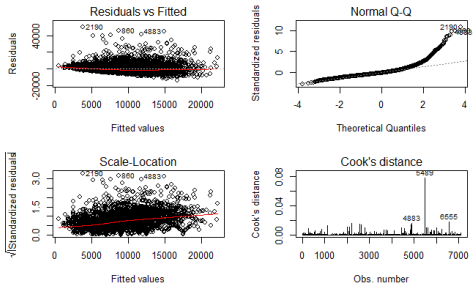

图 9 线性回归模型诊断图二

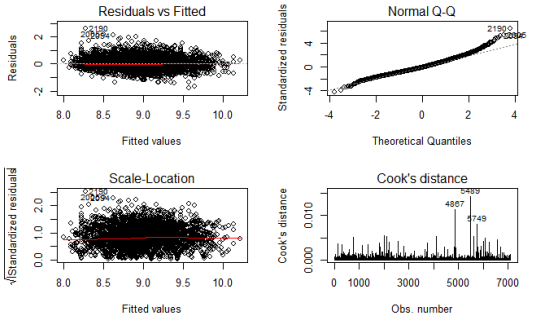

图 10 线性回归模型诊断图三

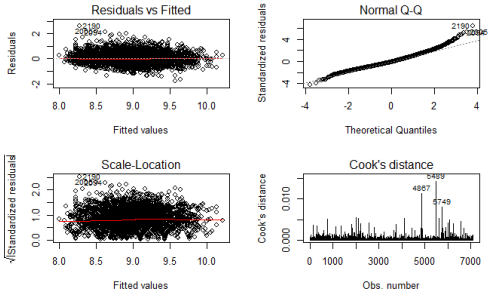

图 11 线性回归模型诊断图四

步骤三：图 9 表明数据本身即偏离正态，因此考虑以对数处理后的平均薪资为因变量，对地区、行业类别、公司类别、经验、学历和掌握的技术为自变量进行回归，其模型诊断图如图 10 所示。从正态 Q-Q 上可知正态性改善良好，同时 R 值也有较大的提升，因此采用对数线性回归比线性回归更加合理。

步骤四：在对数线性模型的拟合过程中由于因变量过多，易发生过度拟合。因此通过逐步回归，由贝叶斯信息准则进行变量筛选，防止模型复杂度过高，过度拟合造成的模型高精度假象。尽管拟合效果略有所降低，但有效规避过拟合问题，模型诊断图造成影响不大，如图 11 所示。

如果进一步考虑交互效应，由于各因素变量本身对线性模型拟合影响均较低，交互项对于线性模型拟合影响更低。交互项将在贝叶斯信息准则筛选中同样被筛除，因此在该模型内考虑交互项尽管能够略微提升拟合精度，但权衡过度拟合后仍然不合适。

表二 线性回归效果

	R^2	R_a^2	F	p
直接拟合	0.1406	0.1364	33.1	< 0.001
步骤一：删除极端数据	0.2798	0.2762	78.5	< 0.001
步骤二： 删除离群点	0.3259	0.3226	97.5	< 0.001
步骤三：对数线性回归	0.4046	0.4016	137.1	< 0.001
步骤四：筛除过拟合项	0.4019	0.3998	190.0	< 0.001

如上各步线性回归效果如表二所示。经过模型诊断后得到 p 值远小于 0.001，说明模型结果能够被接受。拟合优度R_a^2近似为 0.4，由此说明线性拟合效果并不理想，自变量与因变量间并无明显线性关系，仅能够作为各自变量影响的初步判断。

我们了计算修正后的模型方差膨胀因子，余下变量的 VIF 都在 1 附近，由此排除了多重共线性对于模型的影响。并使用五折交叉检验，在修正后模型的基础上进行交叉验证，多次计算 RMSE 结果均在 0.4 左右，由此表明模型的预测能力良好，并未存在严重过拟合现象。最终保留变量估计值、标准误和 t 值如表三所示。

表三 线性回归参数估计

变量	哑变量层次	回归系数 $\times 10^1$	标准误差 $\times 10^2$	t 值	P 值
截距		87.691	3.006	291.713	<0.001
所在地区	河北	-3.800	2.205	-17.234	<0.001
	山西	-3.190	3.239	-9.851	<0.001
	陕西	-3.285	2.308	-14.235	<0.001
	上海	0.296	1.495	1.978	0.048
	深圳	0.261	1.675	1.557	0.119
所属行业	互联网/电子商务	2.190	2.648	8.272	<0.001
	计算机软件	2.050	3.041	6.742	<0.001
	金融/投资/证券	2.565	2.894	8.862	<0.001
	快速消费品（食品、饮	1.078	3.332	3.235	0.001

	料、化妆品)				
	贸易/进出口	0.479	3.641	1.315	0.188
	其他	0.683	2.564	2.663	0.008
	专业服务(咨询、人力资源、财会)	1.989	3.812	5.218	<0.001
学历背景	博士	7.443	11.917	6.246	<0.001
	大专	-2.300	1.165	-19.749	<0.001
	高中	-3.381	4.051	-8.347	<0.001
	硕士	1.699	3.306	5.139	<0.001
	无	-0.942	1.630	-5.779	<0.001
	中专	-3.060	3.429	-8.923	<0.001
从业经验		1.019	0.280	36.407	<0.001
软件技能	Excel	-1.306	1.394	-9.370	<0.001
	SQL	1.293	1.773	7.292	<0.001
	SAS	0.621	2.054	3.024	0.003
	R	0.964	1.235	7.810	<0.001
	WORD	-1.093	2.431	-4.500	<0.001
	HADOOP	2.032	3.009	6.754	<0.001

对照系数估计的结果，控制其他因素不变时，得出以下结论：

(1)所在地区：一线城市的平均薪资明显高于内陆地区，比如北京的平均薪资比河北高了 38%，可见求职者选择一线城市往往能够得到更高薪资，而公司企业则需要权衡一线城市带来的区位优势与更为高额的薪资成本，以选择企业区位。

(2)所属行业：互联网公司，计算机软件公司和金融类和专业服务类公司的平均薪资高于其他公司。求职应聘者选择以上行业更易取得高收入，公司企业选择该类热门方向也可带来高业绩。

(3)学历背景：在有学历要求的情况下，学历越高的平均薪资越高，博士学历比本科学历平均薪资高了 74%。无论对于求职者还是高校教育者，重视高校教育，提升学历都是正确的选择。

(4)从业经验：工作经验每多一年，平均薪资高出 10%。可见工作经验的逐步积累也能够作为加薪筹码。

(5)软件技能：掌握 SQL、SAS、R、RHADOOP 比没有任何编程要求工作的平均薪资分别高出 13%、6%、9.6%和 20%。可见掌握几门基本高级编程语言，能够大幅加薪，这也为高校教育者，求职者提供了发展方向。

五.总结

为分析数据分析岗位招聘数据，我组通过文本分词提炼有效信息，并通过线性回归模型进行建模，描述比较各因素对于最终数据分析岗位薪资的影响。由建模结果可以得出，数据分析岗位薪资主要与个人学历与工作经验有关，与企业规模，公司类别几乎无关。相应地，我们分别针对公司经营者，求职应聘者，高校教育者，提出以下建议：

1.公司经营者：制订薪资时应当着重衡量所需学历，经验与所需软件高低，对于较高学历，具有较长工作经验，掌握高级语言就业者应制定更为高额的薪资以挽留该类人才，对学历、经验与技能较为不足的人才也可收缩薪资，以增加企业最终受益。

2.求职应聘者：由于数据分析岗位招聘相当看重学业与经验，因此建议有志于数据分析的求职者更多重视自我深造，提升学业层次，积累就业经验才能够取得较高薪资；行业方向方面更推荐互联网，计算机和金融方向，该类热门方向数据分析工作普遍薪资较高；于此同时，求职应聘者也应当更多注重自我能力提升，无论是驾驭高级的软件技能，作为一份加薪条件，还是自我能力的全方面提升，以应对工作与生活当中的一切挑战。

3.高校教育者：针对数据分析岗位教育，高校应当注重培养学生各方面高级编程语言的能力，以应对各行业对于高级编程语言的需求。于此同时，一线高校也不应仅仅注重于对口就业岗位的能力灌输，也应当尝试数据科学理论体系的培养与教育，以期为数据科学行业带来真正的科研质变。

附：小组分工情况

数据清洗：全员讨论各变量处理方式，韩思越负责代码

分词处理：全员讨论分词后有价值信息提取筛选，黄永晟负责代码

描述统计：全员讨论各变量统计可视化方式与呈现信息价值，俞铖昊负责代码

回归分析：全员讨论自变量与交互项选择以及数据诊断方法，樊可负责代码

项目报告：全员讨论报告呈现内容筛选，杨远琨负责报告最终整合