通过线性回归模型进行 IT 行业薪资变化分析

史杰

大连理工大学 辽宁省 大连市

摘 要：基于调查数据研究软件行业薪资水平的影响因素，可帮助求职者预估薪资情况，也可为软件公司制定薪资方案提供理论依据。数据包括某软件公司员工的工作年限、工作岗位、学历以及薪资。首先对数据进行适当处理，把工作岗位和学历等定性数据进行定量化。建立多元线性回归模型刻画工作年限、岗位、学历对薪资的具体影响作用。

关键词：软件行业薪资影响因素分析；多元线性回归模型；显著性检验；薪资预测；

0 引言

在 Internet 飞速发展的今天，互联网已成为人们快速获取、发布和传递信息的重要渠道，它在政治、经济、生活等各个方面发挥着重要的作用，已成为政府、企事业单位信息化建设中的重要组成部分，从而倍受人们的重视。随着中国软件业的迅猛发展，软件工程师的薪资也“水涨船高”，这是吸引大量毕业生进入软件行业的主要动力[1]。本文调查了各大招聘网上软件行业的招聘需求和薪资水平，并获取了某软件公司员工的具体薪资及相应的工作年限、学历、工作岗位等数据。基于实际调查数据，利用数学方法，建立了多元线性回归模型，并通过显著性检验，得到了软件行业员工的薪资与工作年限、学历、工作岗位等因素的依赖关系。

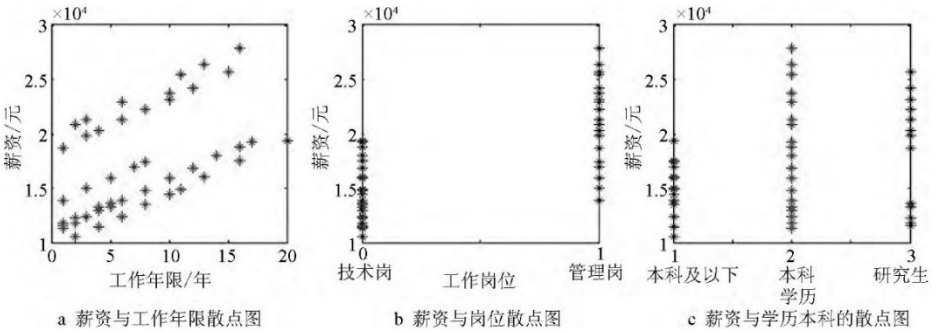
1 数据说明及分析

调查了某软件公司 57 个员工的薪资及相应的工作年限、学历和工作岗位，表 1 罗列了部分数据，其中学历包括大专、本科、研究生 3 个级别，分别对应数字 1，2，3，工作岗位分为管理岗和技术岗，对应数字 1，0，这样做的好处是把定性的级别定量化，统一为数据，便于分析处理。

表 1 某软件公司员工的薪资及对应的工作年限、学历、岗位数据

薪资/元	工作年限/年	学历(大专－1；本科－2；研究生－3)	岗位(管理－1；技术－0)
13 876	1	1	1
11 608	1	3	0
12 195	2	3	0
12 313	3	2	0
20 263	4	3	1
13 677	5	3	0
15 965	5	1	1
22 884	6	2	1
…	…	…	…
19 346	20	1	0

为了探究员工薪资与工作年限、学历和工作岗位这 3 种因素是否存在确定的依赖关系，以及如果存在该如何定量地表示，先对数据做粗略地描述性分析，考察薪资与每一种因素的相关性，记薪资为 y，工作年限为 $x1$，学历为 $x2$，岗位为 $x3$，分别画出 y 与 $x1$，$x2$，$x3$ 的散点图，见图 1 所示。



a 薪资与工作年限散点图　　b 薪资与岗位散点图　　c 薪资与学历本科的散点图

图 1 薪资与 3 个影响因素的散点图

从图 1 大致看出，薪资 y 与工作年限 $x1$ 呈现一定的线性关系，技术岗员工的平均工资比管理岗低，本科生和研究生的平均工资略高于大专生。然而，这仅仅是基于调查数据得到的粗略结果，样本数据未必能得到确定性关系，不同岗位和学历对薪资是否具有显著性差别以及具体的依赖关系如何，需要建立多元线性回归模型来进一步分析。

2 多元线性回归简介

回归分析（regression analysis）是确定 2 种或 2 种以上变量间相互依赖的定量关系的一种统计分析方法，其一般步骤如下[2]。

(1) 确定因变量与一个或多个自变量间的定量关系表达式，一般称之为回归方程；(2) 对求得的回归方程的可信度进行检验；(3) 判断自变量对因变量有无影响，对回归方程进行诊断和修正；(4) 利用所求得的回归方程进行预测和控制。

在回归分析中，多元线性回归是最基本的建模技术，要求因变量是连续的，自变量可以是连续的也可以是离散的，回归线的性质是线性的，其模型函数通常表示如下：

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n+\varepsilon \qquad (1)$$

式 (1) 中，y 是因变量；$x1$，$x2$，xn 为 n 个自变量；$\beta 1$，$\beta 2$，βn 代表 n 个回归系数；ε 是随机变量，代表随机误差，一般要求其服从正态分布。对于实际问题，将实测数据代入模型中，通过最小二乘原理拟合出回归系数，之后对模型进行诊断和修正，最终对实际问题进行控制或预测。

基于实测数据求回归系数时，可借助数学软件 MATLAB 来进行，省去复杂的计算，其命令如下[3]：

$$[b,bint,r,rint,stats]=regress(Y,X,alpha) \quad (2)$$

式 (2) 中，回归命令为 regress；Y 是因变量的实测值，X 是自变量的实测值，当自变量为多个时，X 为多个自变量实测值组成的矩阵。$alpha$ 代表显著性水平，缺省状态为 0.05，表明模型 95%成立。b 是由求出的回归系数组成的向量，$bint$ 为回归系数对应的置信区间，刻画了回归系数的有效范围。r 是模型结果与实测结果之间的误差向量，$rint$ 为其对应的置信区间。$stats$ 是检验回归模型的统计量，包含 4 个数据：决定系数 $R2$（越接近 1 越好）、检验量 F 值（较大的数）、与 F 值对应的概率 p（越接近 0 越好）、残差平方和 q（越小越好）。通过 regress 的回归命令，不仅很快求出了回归系数，更重要的是，通过观察统计量 $stats$的4个值，能够快速诊断回归模型是否成立并给出模型优度，因此在实际回归分析中应用十分广泛。

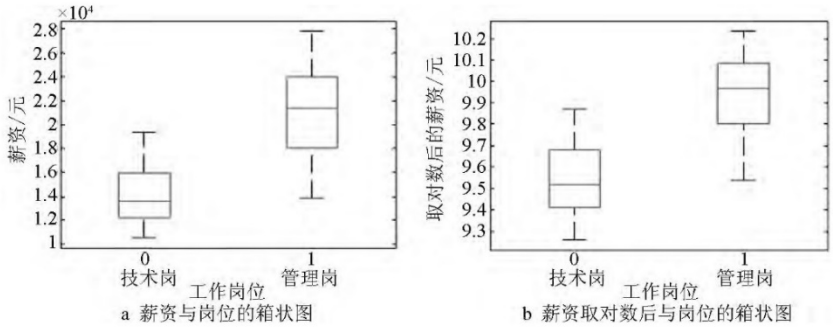
3 多元线性回归在软件行业薪资影响因素分析中的应用

为了探究软件行业员工的薪资与工作年限、学历和工作岗位这 3 个因素的关系，建立多元线性回归模型函数如下：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad (3)$$

式（3）中，y 代表员工薪资，是连续型变量；$x1$ 是工作年限（离散型），$x2$ 代表工作岗位（离散型），$x3$ 代表学历（离散型）。

式（3）中，做回归分析时，要求因变量与自变量的实测数据尽量同方差，避免由数据本身离散度不一致对模型造成不稳定的影响[4]。例如考察薪资 y 与工作岗位 $x2$（技术岗 0、管理岗 1）的方差，画箱状图（y, $x2$），观察矩形的上下边线间距（见图 2a），差别较大，表明两者的方差相差较多，常用的处理方法是取对数，对因变量 y 取对数，再次画箱状图（$log\ (y)$，$x2$），此时方差较为接近（见图 2b）。



a 薪资与岗位的箱状图　　b 薪资取对数后与岗位的箱状图

图 2 薪资本身及取对数后与工作岗位的箱状图

为此，将模型修正如下：

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad (4)$$

经过计算得到模型的检验统计量 $stats$ 的 4 个值如下：

$$stats = 0.7240 \quad 20.9862 \quad 0.0000 \quad 0.0225$$

最终得到模型的检验统计量 $stats$ 及回归系数 b 如下：

$$stats = 0.9871 \quad 496.2196 \quad 0.0000 \quad 0.0011$$

$$b = 9.2663 \ 0.0297 \ 0.2042 \ 0.0876 \ -0.0020$$

$$0.2132 \ 0.0561$$

此时 $stats$ 的第 3 个值接近 0，表明回归模型成立，第 1 个值 0.9871 接近 1，表明模型的拟合优度好，这说明软件行业员工的薪资与年限、岗位、学历显著相关，且工作岗位与学历有显著的交互影响。

据此模型可以分析和预测软件行业员工的薪资状况，具体来说其作用表现如下：

（1）帮助毕业生预估入职薪资。例如某本科毕业生应聘技术岗，将年限 $x1$=0，岗位 $x2$=0，学历对应逻辑变量 $x31$=1，$x32$=1，代入模型，可得其薪资期望值为 11521 元，即月薪期望值约为 1.15 万。

（2）帮助公司制定薪资方案。假设学历和岗位都相同时，考察不同工作年限薪资的差异，利用模型可

得逐年工资的平均涨幅约为5%，例如对新入职员工而言，每年月薪增长约为600，650，720等等;假设学历和年限都相同，以工作5年为例，管理岗的月薪比技术岗高约50%;假设年限和岗位都相同，考察不同学历，本科生月薪比大专生高约9%，研究生比本科生高约0.3%，表明不同学历间的工资差异不大。

4 结论

本文基于对某软件公司57名员工薪资的调查数据，建立了多元线性回归模型研究了员工薪资与工作年限、工作岗位、学历三者之间的关系，模型不仅通过了检验而且回归系数的拟合优度较高，能够帮助求职者预估薪资及涨幅情况，也可以为软件公司制定合理的薪资方案提供理论依据。

参考文献

[1] 天极数据调查中心. IT行情调查研究报告[M].北京:清华大学出版社，2015.

[2]何晓群，应用回归分析[M].北京:中国人民大学出版社，2015.

[3]王倩，加春燕.数学建模方法与应用[M].北京:北京师范大学出版社，2016.

[4]王松桂，线性回归与方差分析[M].北京:高等教育出版社，1999.

[5]姜启源，谢金星.实用数学建模[M].北京:高等教育出版社，2014.