

个体薪资水平析因研究

——基于分类型数据分析方法

致理-数理 1 刘苏青* 2021013371

工 02 余丹[†] 2019030022

致理-数理 0 孟灏[‡] 2020012560

2024 / 01 / 16

摘 要

本文采用包含薪资等个人信息的数据集，基于列联表分析、逻辑回归、累积比数模型等分类型数据研究方法，考察对个体薪资水平有明确影响的因素，并就变量间的定量关系给出解释。

关键词：薪资水平，卡方检验，比值比，逻辑回归，累积比数模型。

*liu-sq21@mails.tsinghua.edu.cn

[†]shed19@mails.tsinghua.edu.cn

[‡]mengh20@mails.tsinghua.edu.cn

目录

1	研究背景	3
2	探索性数据分析	3
2.1	基本统计描述	3
2.2	连续变量分布分析	4
2.3	不同特征下的薪资分布分析	4
3	比值比分析：学历、工作经验和薪资水平之间的关系	6
3.1	薪资与工作经验	6
3.2	薪资与学历	7
4	逻辑回归模型：薪资水平与各项特征的定量关系	7
4.1	初步尝试——可加性模型	8
4.2	模型改进	8
5	累积比数模型：不同工作类别和薪资的具体关系	9
6	总结与反思	12
6.1	总结	12
6.2	反思	12
7	附录	13
7.1	小组分工	13
7.2	图表 & R output 补充	13

1 研究背景

在全球化加速的今天，经济的快速增长引领了工业结构的转型和劳动市场的扩展。这种转变促使我们重新审视那些影响职业满意度和薪酬水平的关键因素。职场多元化不仅表现在工作性质的变化上，也体现在员工背景、职业发展路径及薪酬结构的日益复杂化。了解和量化这些变量之间的相互作用对于政策制定者、企业管理者以及个人职业生涯规划均具有重要意义。为此，我们在 kaggle 平台上选取共含有 6704 条样本的 [Salary_Data.csv](#) 数据集，其数据主要来自印度国内的调查问卷、招聘网站等公开信息来源。该数据集包括年龄、性别、教育水平、工作名称、工作经验、薪资和工作类别等特征。通过对这些数据的综合分析，我们旨在揭示职场薪资和员工背景背后的逻辑关系，以及对比不同工作类别下各影响因素的具体效应。

2 探索性数据分析

2.1 基本统计描述

该数据集一共 7 列，列名分别为 Age, Gender, Education Level, Job Title, Years of Experience, Salary 和 Job Category。其中 Age, Years of Experience 和 Salary 这 3 列为连续型变量，其他 4 列为分类型变量，其基本的统计描述如下：

- *Age*: 平均年龄约为 33.62 岁，年龄范围从 21 岁到 62 岁。
- *Gender*: 数据包含 “Male” 和 “Female” 两种性别，其中男性 3674 次出现，女性 3013 次出现，男女比例接近 1:1。
- *Education Level*: 教育水平有 4 种不同的分类，分别为 “High School”, “Bachelor’s”, “Master’s” 和 “PhD”，其中 “Bachelor’s”（学士学位）最常见，共有 3021 次出现，“High School” 最少，共 436 次出现。
- *Job Title*: 共有 193 种不同的职位名称，其中 “Software Engineer”（软件工程师）最常见，共有 518 次出现。
- *Years of Experience*: 平均工作经验约为 8.09 年，范围从 0 到 34 年。
- *Salary*: 平均薪资约为 115,327 卢布，薪资范围从 350 卢布到 250,000 卢布。

- *Job Category*: 有 5 种不同的职位类别，其中 “Technology and Engineering”（技术与工程）最常见，共有 2644 次出现。

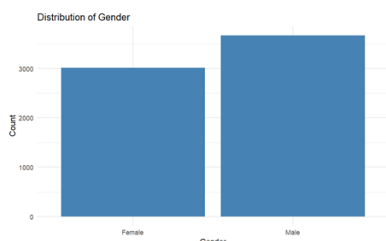


Figure 1: Distribution of Gender

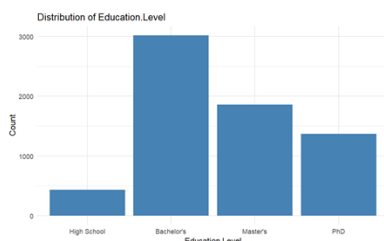


Figure 2: Distribution of Education Level

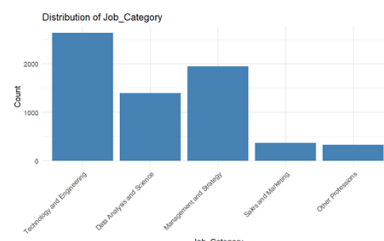


Figure 3: Distribution of Job Category

2.2 连续变量分布分析

对于 Age, Years of Experience 和 Salary 这 3 个连续型变量，我们分别做出了它们的直方图以初步分析其分布，可以看到 Age 的分布是右偏的，且大部分集中在 30 岁左右；Years of Experience 的分布明显右偏，且大部分集中在 0-10 年；Salary 的分布存在多峰情况，在薪资为 50000 卢布和 180000 卢布附近存在两个小峰值。

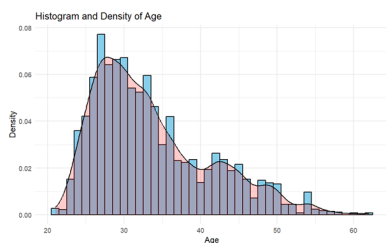


Figure 4: Distribution of Age

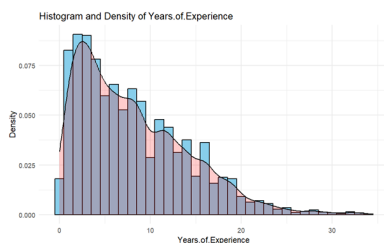


Figure 5: Distribution of Experience

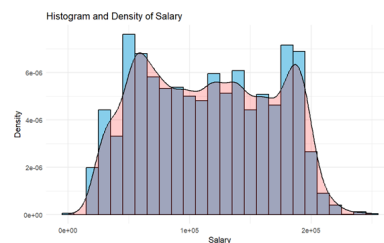


Figure 6: Distribution of Salary

2.3 不同特征下的薪资分布分析

本研究尤为关注 Salary 薪资的分布，故进一步我们考察了不同特征下的 Salary 薪资的分布。

- 薪资按性别分布：在性别与薪资的分布上，男性和女性的薪资中位数有所差异，其中男性的中位数薪资似乎高于女性。男性的薪资分布也显示出更大的分散性，尤其是在较高薪资范围内。

- 薪资按教育水平分布：从教育水平的箱线图中，我们可以看到随着教育水平的提高，薪资的中位数和四分位数范围通常也在提高。具有博士学位的人群薪资中位数最高，而高中学历的人群薪资中位数最低，薪资分布更集中。
- 薪资按职位类别分布：不同的职位类别薪资分布有显著差异。例如，“Technology and Engineering”（技术与工程）和“Management and Strategy”（管理与策略）类别的薪资中位数和上四分位数较高，而“Sales and Marketing”（销售与市场）和“Other Professions”（其他职业）的薪资分布相对较低。
- 工作经验与薪资的关系：从工作经验与薪资的散点图中可以看出，随着工作经验年数的增加，薪资也呈现出上升趋势。这表明工作经验是影响薪资的一个重要因素。不过，即使在相同的经验年数，薪资也存在相当大的变异性，这可能受到其他因素如行业、地区、职位级别等的影响。

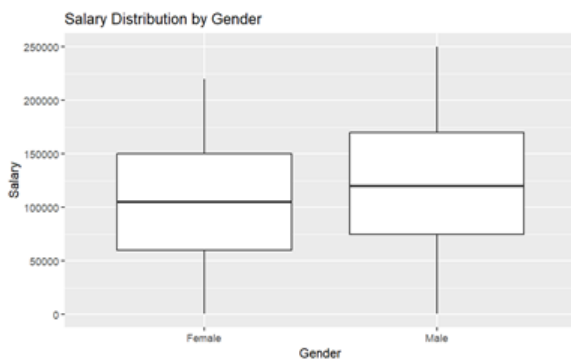


Figure 7: Salary Distribution by Gender

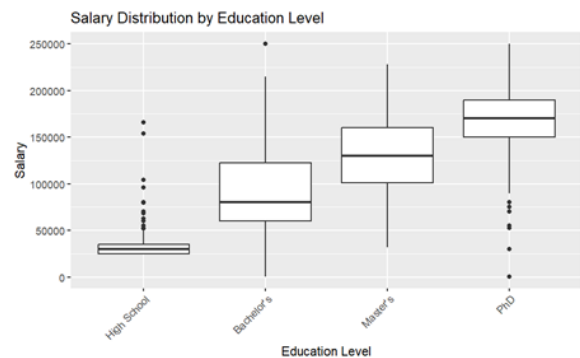


Figure 8: Salary Distribution by Education Level

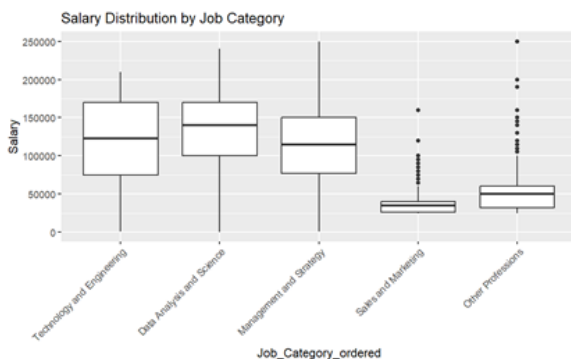


Figure 9: Salary Distribution by Job Category



Figure 10: Relationship between Years of Experience and Salary

3 比值比分析：学历、工作经验和薪资水平之间的关系

将 Salary 薪资按照“150000 卢布”为界，分为“High Salary”与“Low Salary”；将工作经验年限要求以“5 年”为界，分为“Experienced”和“Inexperienced”；学历分为“High School”，“Bachelor’s”，“Master’s”和“PhD”。

具体分完类别后，按照学历、工作经验和薪资水平的分类构造三维列联表：

	<i>High School</i>	<i>Bachelor's</i>	<i>Master's</i>	<i>PhD</i>
Experienced				
High Salary	4	426	626	1040
Low Salary	5	958	943	279
Inexperienced				
High Salary	0	29	3	0
Low Salary	427	1608	286	50

下面对此三向列联表做边缘比值比分析。

3.1 薪资与工作经验

	<i>Experienced</i>	<i>Inexperienced</i>
High Salary	2096	32
Low Salary	2185	2371

$$\theta_{\text{薪资-经验}} = \frac{2096 \times 2371}{2185 \times 32} = 71.12$$

卡方检验的结果同样显示了极其显著的统计差异（p-value < 2.2e-16），表明工作经验与薪资水平之间存在显著的相关性。比值比的结果为 71.11907，这表示有工作经验的人群比没有工作经验的人群有更高的几率获得高薪。这个比值比的大小指示经验在职场中的重要性，尤其是对于薪资水平。

3.2 薪资与学历

	<i>High School</i>	<i>Bachelor's</i>	<i>Master's</i>	<i>PhD</i>
High Salary	4	455	629	1040
Low Salary	432	2566	1229	329

$$\theta_{\text{高中-本科}} = \frac{4 \times 2566}{432 \times 455} = 0.0522$$

$$\theta_{\text{本科-硕士}} = \frac{455 \times 1229}{2566 \times 629} = 0.3465$$

$$\theta_{\text{硕士-博士}} = \frac{629 \times 329}{1229 \times 1040} = 0.1619$$

卡方检验的结果显示了极其显著的统计差异 ($p\text{-value} < 2.2e-16$), 这意味着在统计学上, 不同教育水平与薪资水平之间存在显著关联。比值比的结果显示, 对于 “High School” 和 “Bachelor’s” 学历的比值比是 0.0522, 这表明与拥有学士学位的人相比, 仅有高中学历的人获得高薪工作的几率要低得多; 对于 “Bachelor’s” 和 “Master’s” 学历的比值比是 0.3465, 这意味着拥有硕士学位的个体比仅有学士学位的个体有更高的几率获得高薪工作, 但差异没有高中与学士那么大; 对于 “Master’s” 和 “PhD” 学历的比值比是 0.1619, 这表明拥有博士学位的个体相比于硕士学位的个体来说, 获得高薪工作的可能性更大。这些比值比表明随着教育水平的提高, 获得高薪工作的几率也随之增加。这是符合普遍认知的, 因为较高的教育水平通常会提供更多的就业机会和更高的收入潜力。特别地, 高中学历与学士学历之间的巨大比值比差异强调了高等教育在职业发展中的重要性。

4 逻辑回归模型：薪资水平与各项特征的定量关系

根据之前的探索性分析, 我们发现收入与工作年限之间并不是简单的线性关系, 因此若想利用包括工作年限在内的诸多解释变量去解释并预测收入水平, 一般的线性模型大概率不适用。为此, 我们考虑引入逻辑回归模型。

将 Salary 薪资按照 “150000 卢布” 为界, 分为 “High Salary” 与 “Low Salary”, 作为 binary 响应变量; 以 Age, Gender, Education Level, Years of Experience 和 Job Category 为

解释变量，其中 Gender, Education Level 和 Job Category 是分类型变量，而 Age 和 Years of Experience 为连续型变量。

4.1 初步尝试——可加性模型

首先不考虑解释变量之间的交互效应，基于各解释变量的一阶效应，尝试构建可加性 (additive) 逻辑回归模型：

$$\text{logit } P(\text{High Salary}) = \beta_0 + \beta_1 \text{Age} + \beta_2^T \text{Gender} + \beta_3^T \text{Edu} + \beta_4 \text{Exp} + \beta_5^T \text{Job}$$

模型参数估计结果见附录图 18。该逻辑回归模型表明年龄、学历、工作经验和工作类别是高收入的显著预测因子。性别在这个模型中不是一个显著预测因子。

注意到这里年龄的系数的估计值为负，即在其他因素不变的情况下，年龄越大，获得高收入的概率越低。该结论显然与普遍的认知以及探索性数据分析部分显示的趋势相悖，这很可能是由工作经验和年龄两个解释变量间较强的多重共线性引起的。

因此，当前的模型并不能很好地解释或预测收入情况，需加以改进。

4.2 模型改进

考虑到年龄和工作经验具有较强的多重共线性（原数据集中 Age 和 Exp 二者的相关系数约为 0.94），同时年龄与学历也有一定的相关性，我们接下来不再将 Age 变量放入模型当中。

此外，某些二阶交互效应可能对收入水平有明确的影响。例如，工作类型和工作经验的二阶效应也许是显著的：一些工种可能“越老越吃香”，而另一些则是“青春饭”。由此，我们需要尝试向模型中加入二阶交互项。

Gender, Edu, Exp 和 Job 这四个解释变量可能构成的二阶模型有很多。我们对其选择的标准是：在保证模型中所有效应同时显著的前提下，使模型的 AIC 尽可能小。据此，我们最终选出模型：

$$\text{logit } P(\text{High Salary}) = \beta_0 + \beta_1^T \text{Edu} + \beta_2^T \text{Gender} + (\beta_3 + \beta_{23}^T \text{Gender} + \beta_{34}^T \text{Job}) \cdot \text{Exp} + \beta_4^T \text{Job}$$

模型的 AIC 为 4118.3，参数估计如下：


```

Coefficients:
(Intercept)
Education.LevelBachelor's
Education.LevelMaster's
Education.LevelPhD
GenderMale
Years.of.Experience
Job_CategoryData Analysis and Science
Job_CategoryManagement and Strategy
Job_CategorySales and Marketing
Job_CategoryOther Professions
GenderMale:Years.of.Experience
Years.of.Experience:Job_CategoryData Analysis and Science
Years.of.Experience:Job_CategoryManagement and Strategy
Years.of.Experience:Job_CategorySales and Marketing
Years.of.Experience:Job_CategoryOther Professions

Estimate Std. Error z value
-6.00061 1.33583 -4.492
2.59342 1.31505 1.972
2.40150 1.31266 1.829
3.10757 1.31342 2.366
-1.14176 0.20939 -5.453
0.32825 0.01940 16.921
0.78646 0.23476 3.350
-2.39331 0.31894 -7.504
-41.66008 774.70258 -0.054
-10.29759 5.28196 -1.950
0.12170 0.02069 5.882
-0.05946 0.02604 -2.283
0.11454 0.02884 3.972
2.36216 50.18312 0.047
0.52634 0.35521 1.482

Pr(>|z|)
7.05e-06 ***
0.048598 *
0.067326 .
0.017981 *
4.96e-08 ***
< 2e-16 ***
0.000808 ***
6.19e-14 ***
0.957114
0.051226 .
4.06e-09 ***
0.022419 *
7.12e-05 ***
0.962457
0.138398

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 11: 二阶逻辑回归模型参数估计

改进后的模型当中，所有一、二阶效应都是显著的。具体来说：

1. 其他因素相同时，本科、硕士、博士学历者明显比高中学历者更可能获得高收入，其中博士学历者在收入方面优势最大；此外，工作经验的增加显著有利于收入的提高；
2. 相同学历和工种下，工作经验较少时，女性获得高收入的几率比男性更大；而随着工龄的增加，男性获得高收入的几率增长得比女性更快；
3. 相同学历和性别下，工作经验较少时，数据科学岗位的高收入比例比工程与技术岗位更高，而管理与策略岗位的高收入比例则相对较低；随着工龄的增加，在收入方面，相比工程与技术岗位，管理与策略岗位具有更好的成长性，而数据科学岗位成长性较差。

5 累积比数模型：不同工作类别和薪资的具体关系

由上一部分可知，不同工作类别可能对 Salary 有不同的影响，因此我们选择 Job_Category 里面样本数量较多的三类：Technology and Engineering、Data and Science 以及 Management and Strategy 分别进行分析，剩余两类样本数量较少，单独使用 vglm 函数时会出

现迭代不收敛的情况。我们考虑根据总体 Salary 的分位数进行切分（此处我们选择将 Salary 平分为 8 个类别），便于建立 cumulative logit model 来进一步分析。

参考逻辑回归的模型，我们需要考虑性别和工作经验的二阶交互效应。因此我们对三种工作类别均选择如下的 cumulative logit model：

$$\begin{aligned} \text{logit } P(\text{Salary} \leq j) = & \alpha_j + \beta_1 \cdot \text{Gender} + \beta_2 \cdot \text{Edu}(B) + \beta_3 \cdot \text{Edu}(M) \\ & + \beta_4 \cdot \text{Edu}(P) + \beta_5 \cdot \text{Exp} + \beta_6 \cdot \text{Gender} * \text{Exp} \end{aligned}$$

Technology and Engineering, Data and Science, Management and Strategy 三种工作类别各自对应的结果分别见附录图 19, 20, 21。我们可以横向比较不同工作类别下各影响因素的具体效应：

- **GenderMale:** Technology and Engineering 和 Management and Strategy 的性别项为负 (-0.74347 & -1.49981)，意味着在其他条件不变的情况下，男性相比女性处于较低收入水平的 odds 更小。这一现象说明男性可能比女性更适合技术岗和管理岗的工作，或者说技术岗和管理岗的工作可能存在一些关于性别的刻板印象。

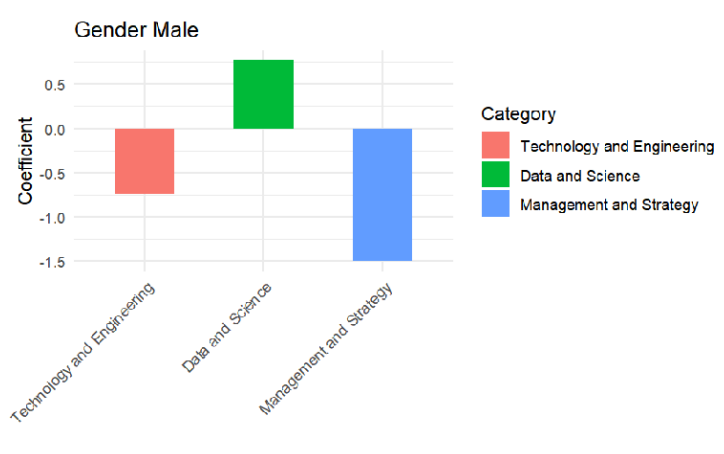


Figure 12: Gender

- **Education.LevelBachelor's, Master's, PhD:** Technology and Engineering 的教育水平项均为负且数值逐渐减小 (-1.12395 & -1.83793 & -3.80215)，意味着在其他条件不变的情况下，学历越高，处于较低收入水平的 odds 越小。Data and Science 类别的工作缺少 High School 的样本，因此选择 Education.LevelBachelor's 作为 baseline，其硕士项和博士项均为负且数值逐渐减小 (-0.57880 & -0.75094)，也意味着在其他条件不变的情况下，学历越高，处于较低收入水平的 odds 越小。Management and Strategy 的教育水平项均为负 (-4.78334 & -5.24996 & -4.42738)，意味着在其他条件

不变的情况下，取得本科及以上学历相较高中学历处于较低收入水平的 odds 均更小。但是三者并未呈递减关系，可能是由于本科学位的人已经足以胜任管理层的工作，取得博士学位对薪资的边际效用递减。

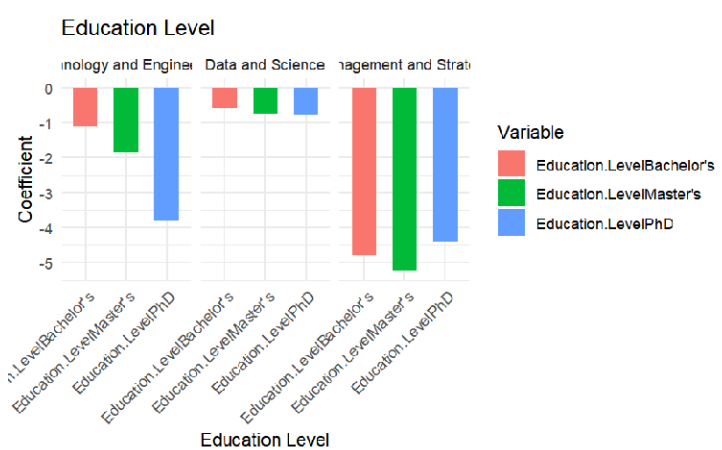


Figure 13: Education

- *Years.of.Experience*: 三者的工作经验项 (-0.42316&-0.30482&-0.63859) 均为负，表现出一致性。表明工作经验越多，处于较低收入水平的 odds 减小。这个结果与我们的直觉相符，即无论哪个类别的工作，更多的工作经验通常与更高的收入相关。

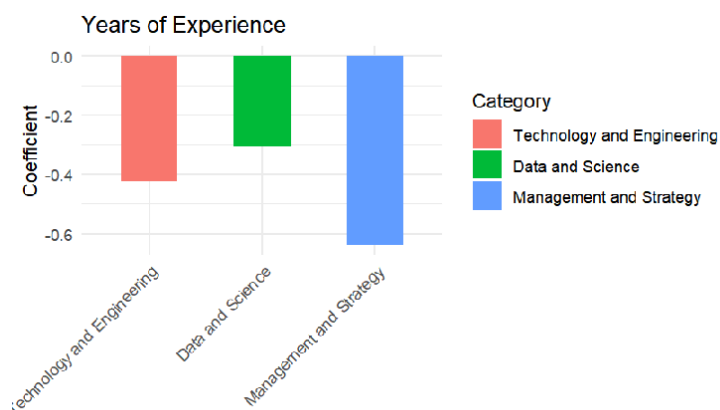


Figure 14: Experience

- *Gender*Years.of.Experience*: Technology and Engineering 和 Data and Science 类别的交互效应均不显著。而 Management and Strategy 类别的交互效应为正，意味着在其他条件不变的情况下，工作年限越长，不同性别处于较低收入水平的 odds 的差距逐渐减小。这一现象说明随着工作经验的增长，工作能力一定程度上消除了关于性别的刻板印象。

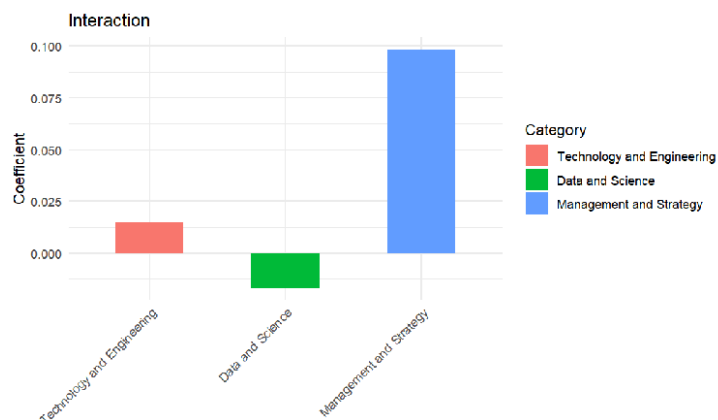


Figure 15: Gender*Experience

6 总结与反思

6.1 总结

综上所述，我们的研究可以大致总结出以下几条结论：

- 薪资水平确实存在性别差异，且部分符合人们的刻板印象，不过性别差异会随着工作经验的增加而缓和；
- 本科以上的学历都能显著提升薪资水平，但部分类别的工作未必学历越高薪资越高；
- 无论哪种类别的工作，丰富的工作经验都能显著提升薪资水平。

6.2 反思

- 当前数据集所包含的数据多集中在 IT 行业，可以考虑增加数据的多样性和代表性，例如收集更广泛的地区或不同行业的数据；
- 当前研究仅基于性别、学历和工作经验等基本信息，可以额外探索其他相关因素，如地理位置或社会经济状态对薪资水平的影响，或许能提供更全面的视角。

7 附录

7.1 小组分工

- 刘苏青：数据集获取、累积比数模型、总结与反思、报告撰写；
- 余丹：数据集获取、探索性数据分析、比值比分析、报告撰写；
- 孟灏：数据集获取、逻辑回归模型、报告撰写。

7.2 图表 & R output 补充

Table 1: 不同收入水平对应的样本个数

<i>High Salary</i>	<i>Low Salary</i>
2128	4556

Table 2: 不同工作经验水平对应的样本个数

<i>Experienced</i>	<i>Inexperienced</i>
4281	2403

Table 3: 不同学历对应的样本个数

<i>High School</i>	<i>Bachelor's</i>	<i>Master's</i>	<i>PhD</i>
436	3021	1858	1369

```
Pearson's Chi-squared test with Yates' continuity correction  
  
data: income_experience_table  
X-squared = 1606.7, df = 1, p-value < 2.2e-16
```

Figure 16: 薪资水平与工作经验的相关性检验

Pearson's Chi-squared test

```
data: income_education_table
X-squared = 1815.9, df = 3, p-value < 2.2e-16
```

Figure 17: 薪资水平与学历的相关性检验

```
Call:
glm(formula = Income_Level ~ Age + Gender + Education.Level +
  Years.of.Experience + Job_Category, family = binomial(),
  data = data_clean)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5629  -0.4309  -0.2027   0.2356   2.7182

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.37196    1.65486  -2.038  0.04159
Age              -0.23508    0.01656 -14.193 < 2e-16
GenderMale         0.01901    0.08574   0.222  0.82456
Education.LevelBachelor's  4.64369    1.64146   2.829  0.00467
Education.LevelMaster's   4.50572    1.63959   2.748  0.00599
Education.LevelPhD        5.25638    1.64056   3.204  0.00136
Years.of.Experience    0.73083    0.02768  26.401 < 2e-16
Job_CategoryData Analysis and Science -0.20371    0.11526  -1.767  0.07717
Job_CategoryManagement and Strategy -1.51689    0.10745 -14.117 < 2e-16
Job_CategorySales and Marketing    -3.86102    1.60706  -2.403  0.01628
Job_CategoryOther Professions    -3.10872    0.46957  -6.620 3.59e-11

(Intercept)      *
Age              ***
GenderMale
Education.LevelBachelor's  **
Education.LevelMaster's   **
Education.LevelPhD        **
Years.of.Experience  ***
Job_CategoryData Analysis and Science .
Job_CategoryManagement and Strategy ***
Job_CategorySales and Marketing  *
Job_CategoryOther Professions  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 8363.5 on 6683 degrees of freedom
Residual deviance: 3944.5 on 6673 degrees of freedom
AIC: 3966.5
```

Figure 18: 一阶逻辑回归模型参数估计

```

Call:
vglm(formula = Income_Level ~ Gender + Education.Level + Years.of.Experience +
      Gender:Years.of.Experience, family = cumulative(parallel = TRUE),
      data = subset_data_tech)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1      0.73011    0.37048   1.971  0.04876 *
(Intercept):2      2.76393    0.37340   7.402 1.34e-13 ***
(Intercept):3      3.97994    0.37718  10.552 < 2e-16 ***
(Intercept):4      5.04988    0.38154  13.235 < 2e-16 ***
(Intercept):5      6.47782    0.39012  16.605 < 2e-16 ***
(Intercept):6      8.29608    0.40718  20.375 < 2e-16 ***
(Intercept):7     11.24352    0.43482  25.858 < 2e-16 ***
GenderMale        -0.74347    0.12498  -5.949 2.70e-09 ***
Education.LevelBachelor's -1.12395    0.36312  -3.095 0.00197 **
Education.LevelMaster's -1.83793    0.36845  -4.988 6.09e-07 ***
Education.LevelPhD  -3.80215    0.38506  -9.874 < 2e-16 ***
Years.of.Experience -0.42316    0.01389 -30.461 < 2e-16 ***
GenderMale:Years.of.Experience 0.01518    0.01361   1.115 0.26465
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors: 7

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3]),
logitlink(P[Y<=4]), logitlink(P[Y<=5]), logitlink(P[Y<=6]), logitlink(P[Y<=7])

Residual deviance: 7305.305 on 18495 degrees of freedom

Log-likelihood: -3652.653 on 18495 degrees of freedom

Number of Fisher scoring iterations: 8

No Hauck-Donner effect found in any of the estimates

```

Figure 19: Technology and Engineering 的累积比数模型

```

Call:
vglm(formula = Income_Level ~ Gender + Education.Level + Years.of.Experience +
      Gender:Years.of.Experience, family = cumulative(parallel = TRUE),
      data = subset_data_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1     -1.70798    0.20519  -8.324 < 2e-16 ***
(Intercept):2     -0.34603    0.17567  -1.970 0.048867 *
(Intercept):3      0.28326    0.17082   1.658 0.097263 .
(Intercept):4      1.44943    0.17230   8.412 < 2e-16 ***
(Intercept):5      2.95596    0.18747  15.768 < 2e-16 ***
(Intercept):6      4.09679    0.20286  20.195 < 2e-16 ***
(Intercept):7      5.74015    0.23075  24.876 < 2e-16 ***
GenderMale         0.76989    0.19953   3.858 0.000114 ***
Education.LevelMaster's -0.57880    0.15086  -3.837 0.000125 ***
Education.LevelPhD   -0.75094    0.16400  -4.579 4.67e-06 ***
Years.of.Experience -0.30482    0.02300 -13.252 < 2e-16 ***
GenderMale:Years.of.Experience -0.01662    0.02164  -0.768 0.442483
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors: 7

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3]),
logitlink(P[Y<=4]), logitlink(P[Y<=5]), logitlink(P[Y<=6]), logitlink(P[Y<=7])

Residual deviance: 4443.648 on 9760 degrees of freedom

Log-likelihood: -2221.824 on 9760 degrees of freedom

Number of Fisher scoring iterations: 7

No Hauck-Donner effect found in any of the estimates

```

Figure 20: Data and Science 的累积比数模型

```

Call:
vglm(formula = Income_Level ~ Gender + Education.Level + Years.of.Experience +
      Gender:Years.of.Experience, family = cumulative(parallel = TRUE),
      data = subset_data_mgmt)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1      5.58016    0.75117   7.429 1.10e-13 ***
(Intercept):2      7.90913    0.76566  10.330 < 2e-16 ***
(Intercept):3      9.45483    0.77238  12.241 < 2e-16 ***
(Intercept):4     10.94384    0.77761  14.074 < 2e-16 ***
(Intercept):5     13.30927    0.78794  16.891 < 2e-16 ***
(Intercept):6     14.13975    0.79183  17.857 < 2e-16 ***
(Intercept):7     15.19451    0.79758  19.051 < 2e-16 ***
GenderMale        -1.49981    0.20351  -7.370 1.71e-13 ***
Education.LevelBachelor's -4.78334    0.74713  -6.402 1.53e-10 ***
Education.LevelMaster's -5.24996    0.75345  -6.968 3.22e-12 ***
Education.LevelPhD   -4.42738    0.75993  -5.826 5.68e-09 ***
Years.of.Experience -0.63859    0.01920 -33.253 < 2e-16 ***
GenderMale:Years.of.Experience 0.09859    0.02057   4.793 1.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors: 7

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2]), logitlink(P[Y<=3]),
logitlink(P[Y<=4]), logitlink(P[Y<=5]), logitlink(P[Y<=6]), logitlink(P[Y<=7])

Residual deviance: 5309.396 on 13644 degrees of freedom

Log-likelihood: -2654.698 on 13644 degrees of freedom

Number of Fisher scoring iterations: 8

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1'

```

Figure 21: Management and Strategy 的累积比数模型