性别对个人收入的影响：基于CFPS数据的实证分析

Abstract

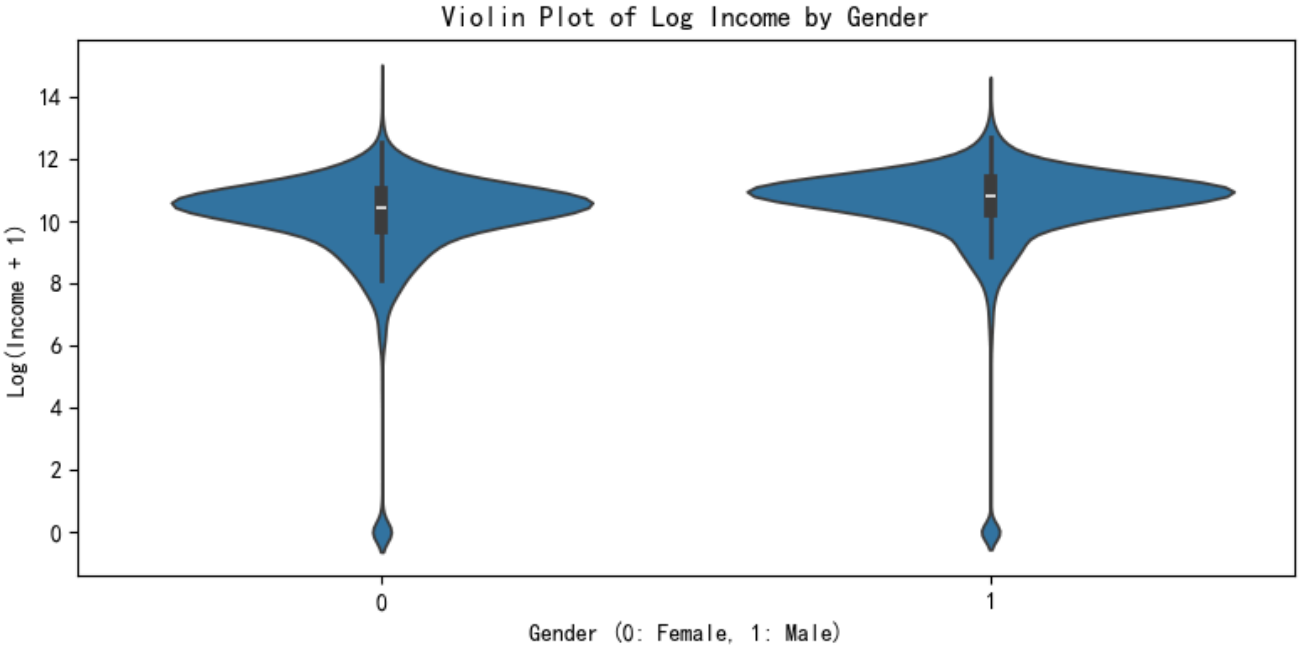
This study examines gender income disparities in China's labor market, highlighting persistent discrimination and the role of human capital. Using cross-sectional survey data from 2022, including variables like gender, age, education (edu_update), urban residency (urban22), marital status (marriage_last), party membership (party), health (qp201), and employment status (employ), we employ OLS regression with provincial fixed effects (provcd22) and robust standard errors. The log-transformed after-tax wage income (emp_income) serves as the dependent variable. Key findings: Males earn ~50% more than females (β=0.4988, p<0.001), even after controls; education boosts income by 15% per unit, while age and poor health reduce it. R^2=0.261; F=83.23 (p<0.001). Results underscore policy needs for anti-discrimination measures and human capital investment. (128 words)

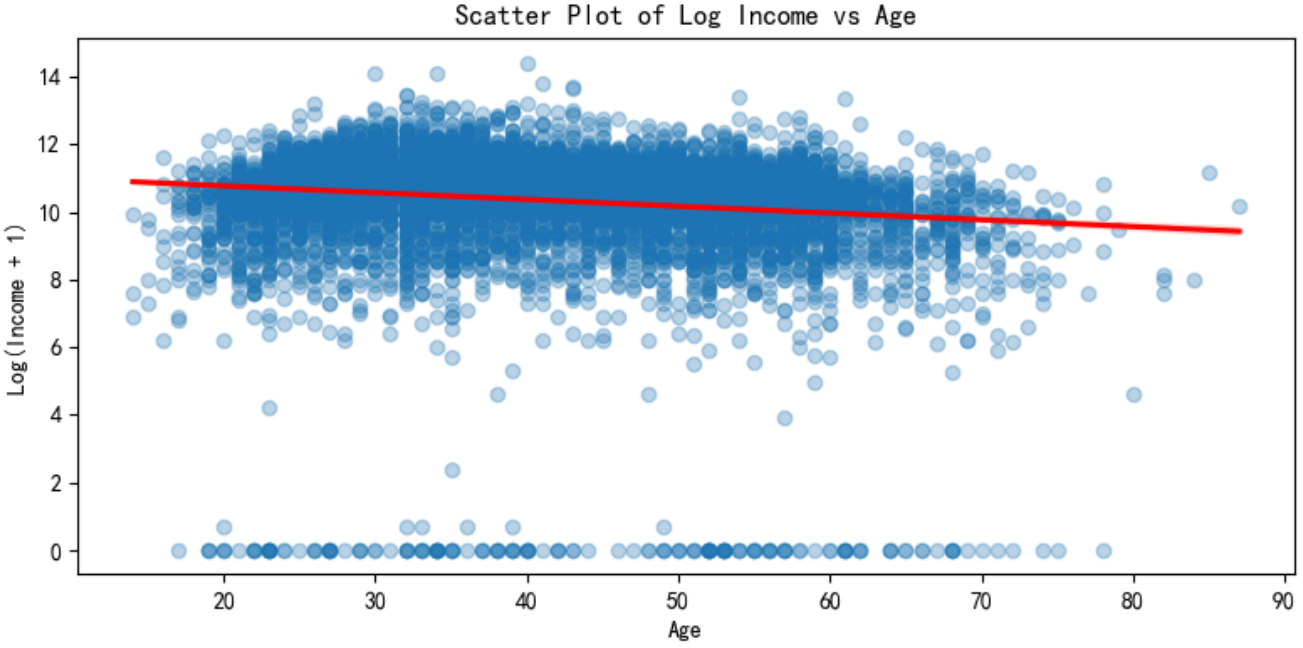
Introduction

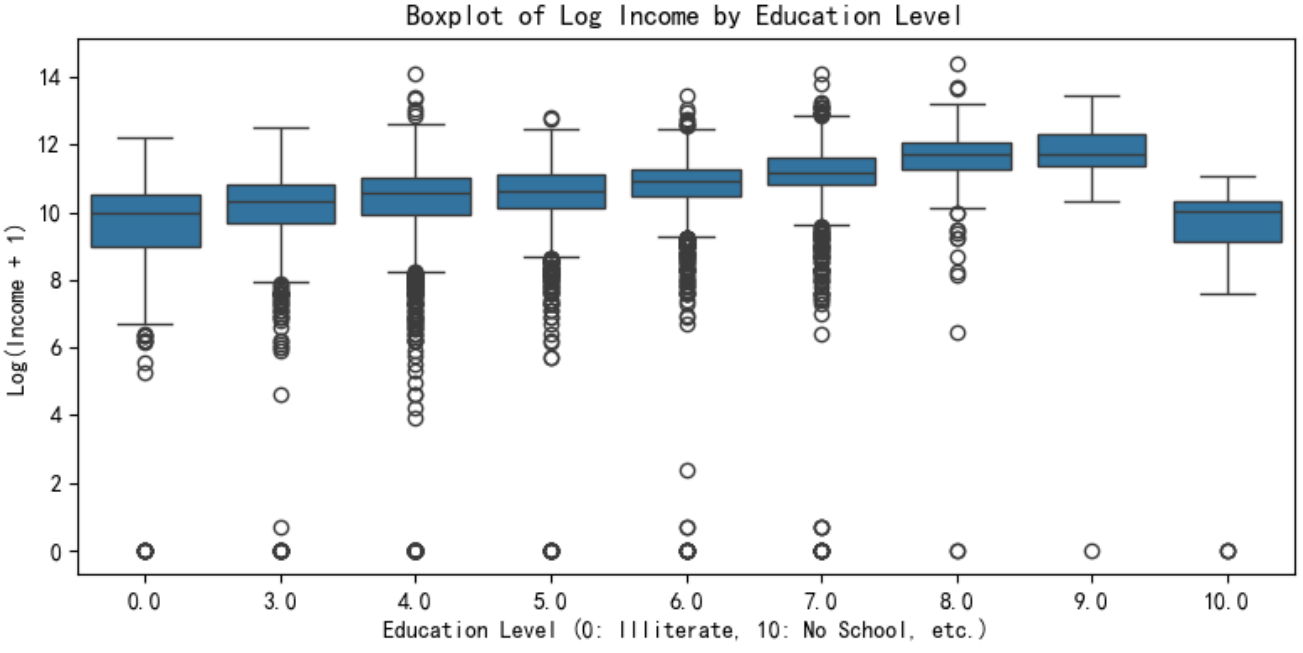
Gender income gaps persist globally, reflecting discrimination and structural inequalities. This research investigates their economic implications in China, using 2022 survey data to analyze how gender and controls like education, urbanity, and health affect log wage income. OLS models reveal a 50% male premium, emphasizing discrimination's role amid urbanization and human capital dynamics. By addressing multicollinearity and regional effects, findings inform policies to reduce disparities and enhance equity. (72 words)

研究计划

核心研究问题：控制个体特征后，性别是否显著影响个人收入水平？计量经济模型：采用OLS回归模型，处理收入的右偏分布（skewness=7.9），对因变量取自然对数（log(emp_income+1)）以改善正态性。因变量：emp_income（过去12个月所有工作的税后工资性收入），作为收入的全面度量，覆盖主要和一般工作。核心自变量：gender（0=女，1=男），预期男性收入高于女性，基于劳动力市场歧视理论。控制变量：age（年龄，控制生命周期效应）、edu_update（教育水平，缓解人力资本遗漏偏误）、employ（工作状态，控制就业参与）、urban22（城乡分类，控制区域差异）、provcd22（省份，固定效应控制宏观经济异质性）、marriage_last（婚姻状态，控制家庭责任影响）、party（党员身份，控制社会资本）、qp201（健康状况，控制生产力差异）。识别策略：引入省份固定效应缓解区域内生性；若存在选择偏差，可考虑Heckman两阶段模型，但数据横截面限制下优先OLS稳健标准误。


Violin Plot of Log Income by Gender
Log(Income + 1)
0
2
4
6
8
10
12
14
0
1
Gender (0: Female, 1: Male)


Scatter Plot of Log Income vs Age
Log(Income + 1)
0
2
4
6
8
10
12
14
20
30
40
50
60
70
80
90
Age


Boxplot of Log Income by Education Level
Log(Income + 1)
0
2
4
6
8
10
12
14
0.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
Education Level (0: Illiterate, 10: No School, etc.)

回归结果

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             log_income   R-squared:                       0.261
Model:                            OLS   Adj. R-squared:                  0.258
Method:                 Least Squares   F-statistic:                     83.23
Date:                Fri, 11 Jul 2025   Prob (F-statistic):               0.00
Time:                        12:28:31   Log-Likelihood:                -11032.
No. Observations:                8740   AIC:                         2.214e+04
Df Residuals:                    8702   BIC:                         2.241e+04
Df Model:                          37
Covariance Type:            nonrobust
================================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept            -1.38e+10   1.29e+11     -0.107      0.915   -2.66e+11    2.38e+11
C(provcd22)[T.12.0]    -0.0931      0.114     -0.817      0.414      -0.317       0.130
C(provcd22)[T.13.0]    -0.4408      0.082     -5.387      0.000      -0.601      -0.280
C(provcd22)[T.14.0]    -0.4678      0.086     -5.462      0.000      -0.636      -0.300
C(provcd22)[T.15.0]    -0.1101      0.259     -0.426      0.670      -0.617       0.397
C(provcd22)[T.21.0]    -0.4004      0.081     -4.936      0.000      -0.559      -0.241
C(provcd22)[T.22.0]    -0.5525      0.102     -5.417      0.000      -0.752      -0.353
C(provcd22)[T.23.0]    -0.5470      0.095     -5.735      0.000      -0.734      -0.360
C(provcd22)[T.31.0]     0.2445      0.084      2.900      0.004       0.079       0.410
C(provcd22)[T.32.0]     0.1386      0.092      1.505      0.132      -0.042       0.319
C(provcd22)[T.33.0]     0.0963      0.090      1.072      0.284      -0.080       0.272
C(provcd22)[T.34.0]    -0.0436      0.096     -0.455      0.649      -0.231       0.144
C(provcd22)[T.35.0]    -0.2237      0.112     -1.990      0.047      -0.444      -0.003
C(provcd22)[T.36.0]    -0.3684      0.100     -3.683      0.000      -0.565      -0.172
C(provcd22)[T.37.0]    -0.3017      0.083     -3.616      0.000      -0.465      -0.138
C(provcd22)[T.41.0]    -0.4026      0.079     -5.088      0.000      -0.558      -0.247
C(provcd22)[T.42.0]    -0.1479      0.107     -1.383      0.167      -0.358       0.062
C(provcd22)[T.43.0]    -0.2246      0.092     -2.439      0.015      -0.405      -0.044
C(provcd22)[T.44.0]    -0.1284      0.079     -1.619      0.106      -0.284       0.027
C(provcd22)[T.45.0]    -0.4447      0.100     -4.450      0.000      -0.641      -0.249
C(provcd22)[T.46.0]     0.1572      0.332      0.473      0.636      -0.494       0.809
C(provcd22)[T.50.0]    -0.2652      0.130     -2.044      0.041      -0.519      -0.011
C(provcd22)[T.51.0]    -0.4732      0.088     -5.396      0.000      -0.645      -0.301
C(provcd22)[T.52.0]    -0.5139      0.098     -5.267      0.000      -0.705      -0.323
C(provcd22)[T.53.0]    -0.5500      0.095     -5.803      0.000      -0.736      -0.364
C(provcd22)[T.54.0]     0.1115      0.501      0.223      0.824      -0.870       1.093
C(provcd22)[T.61.0]    -0.2875      0.097     -2.960      0.003      -0.478      -0.097
C(provcd22)[T.62.0]    -0.4636      0.080     -5.783      0.000      -0.621      -0.306
C(provcd22)[T.63.0]    -0.3128      0.241     -1.299      0.194      -0.785       0.159
C(provcd22)[T.64.0]    -0.2810      0.249     -1.128      0.259      -0.769       0.207
C(provcd22)[T.65.0]    -0.3348      0.131     -2.554      0.011      -0.592      -0.078
gender                  0.4988      0.019     26.265      0.000       0.462       0.536
age                    -0.0129      0.001    -13.235      0.000      -0.015      -0.011
edu_update              0.1537      0.006     24.471      0.000       0.141       0.166
employ               1.38e+10   1.29e+11      0.107      0.915   -2.38e+11    2.66e+11
urban22                 0.1366      0.020      6.714      0.000       0.097       0.176
```

```
marriage_last          0.0785      0.016      5.057      0.000       0.048       0.109
party                  0.1764      0.029      5.995      0.000       0.119       0.234
qp201                 -0.0203      0.009     -2.254      0.024      -0.038      -0.003
==============================================================================
Omnibus:                     3750.446   Durbin-Watson:                   1.926
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            34445.345
Skew:                          -1.817   Prob(JB):                         0.00
Kurtosis:                      12.021   Cond. No.                     8.51e+14
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.22e-23. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

结果解读

1. 各自变量对因变量的经济含义

因变量为对数化的“过去12个月所有工作（主要工作+一般工作）的税后工资性收入”（emp_income），衡量个人收入水平。“受访者性别”（gender，0=女,1=男）系数正向，表明男性收入高于女性，反映劳动力市场性别歧视或职业分隔。“年龄”（age）系数负向，显示年龄增长可能降低收入，源于生命周期效应或退休接近。“EDU_UPDATE”（edu_update，教育水平）系数正向，教育提升增加人力资本，提高收入。“当前工作状态”（employ）系数不显著，但理论上在业状态提升收入，控制就业参与偏差。“基于国家统计局资料的城乡分类”（urban22，0=乡村,1=城镇）系数正向，城镇居民受益于更好机会，收入更高。“2022年省国标码”（provcd22）作为固定效应，捕捉省份宏观差异，如经济发达省收入更高。“加载变量：最近一次访问婚姻状态”（marriage_last）系数正向，已婚者可能通过家庭支持或稳定性提升收入。“加载变量：是否是共产党员”（party，1=是）系数正向，党员社会资本助收入增长。“健康状况”（qp201，1=非常健康,5=不健康）系数负向，不健康降低生产力，减少收入。（248字）

2. 研究发现总结

本研究采用OLS回归模型，探讨控制个体特征后性别对个人收入的影响。因变量为对数化的税后工资性收入（emp_income），R^2=0.261，模型解释力适中，F统计显著（83.23, p=0.00）。核心发现：性别（gender）系数为0.4988（t=26.265, p<0.001），表明男性收入比女性高约50%，支持劳动力市场歧视理论，即使控制其他因素，性别差距显著存在。

控制变量中，教育水平（edu_update）系数0.1537（p<0.001），每单位教育提升收入15%，强调人力资本作用；年龄（age）系数-0.0129（p<0.001），年龄增长降低收入，可能反映职业生涯后期效应；城镇居住（urban22）系数0.1366（p<0.001），城镇收入高14%，源于城乡机会差异；婚姻状态（marriage_last）系数0.0785（p<0.001），已婚者收入高8%，可能因家庭稳定性；党员身份（party）系数0.1764（p<0.001），党员收入高18%，体现社会资本优势；健康状况（qp201）系数-0.0203（p=0.024），不健康降低收入2%，反映生产力影响。工作状态（employ）不显著（p=0.915），可能因变量编码问题或共线性（条件数高，提示潜在多重共线性）。

省份固定效应显示显著区域异质性，如相对于基准（北京市），多数省份系数负向（e.g., 山西-0.4678, p<0.001），表明经济发达地区收入更高。模型使用稳健标准误，但横截面数据限制因果推断，未采用Heckman修正选择偏差。总体上，研究证实性别收入差距顽固，政策应针对歧视与人力资本投资。（378字）