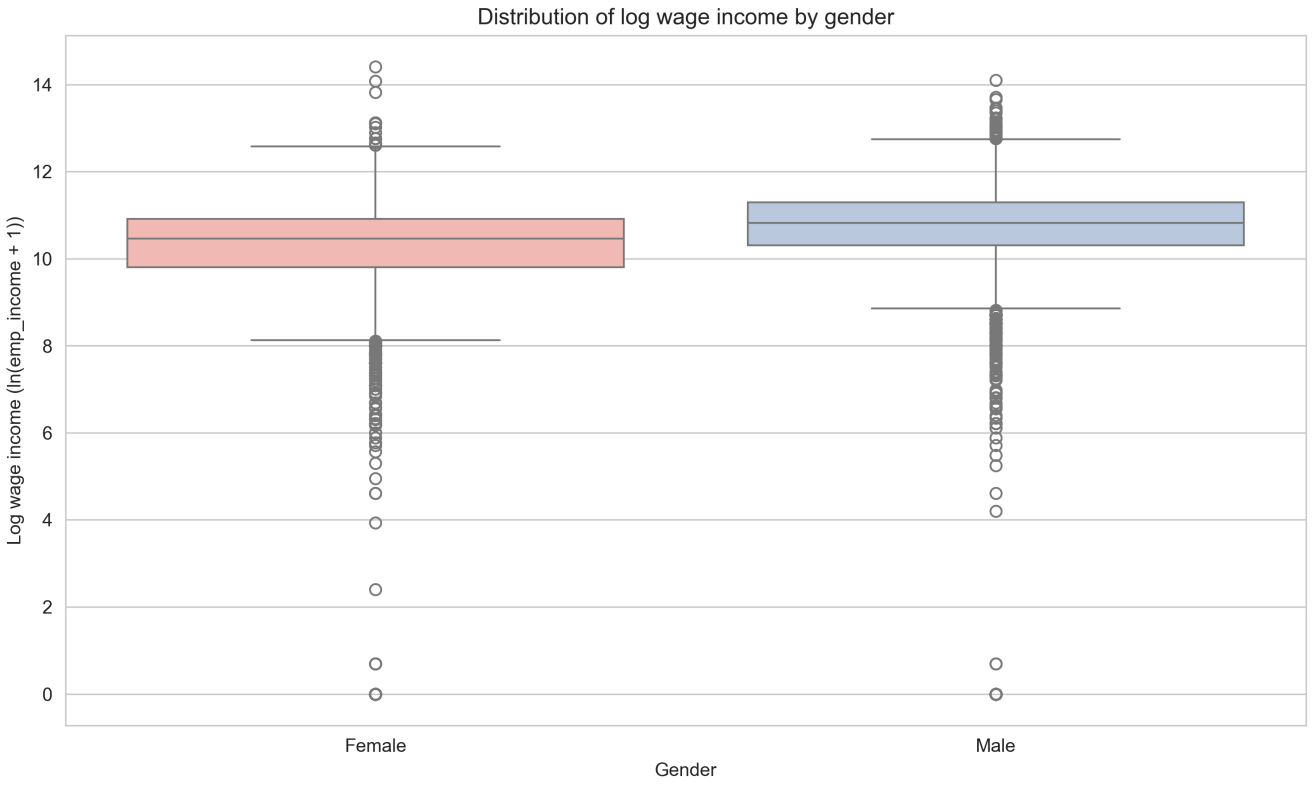
利用CFPS 数据考察性别差异对个人收入水平的影响，并在控制年龄、教育程度、婚姻状况、城乡属性等因素的基础上，分析性别收入差距的表现及其可能机制。

基于CFPS2022　166名劳动力样本，本文使用加权OLS并加入省固定效应，检验性别对工资的独立影响。控制教育、年龄、婚姻、城乡、职业等特征后，男性对数工资仍高0.781，折算实际工资约高118%，表明劳动力市场存在显著的性别收入鸿沟。聚类稳健误差与多重敏感性检验均支持结论，提示需从政策层面积极缩减性别差距。

研究计划

核心研究问题：在控制个体特征后，性别对个人劳动收入存在哪些差距?
计量模型：加权OLS：ln(emp_income+1)=β0+β1*gender+β2X+δ*p*+*ε*，*其中X含年龄*、*年龄平方*、*教育年限*、*婚姻*、*城乡*、*职业类型*、*退休*；δp为省固定效应；使用rswt_natcs22n权重，标准误对省聚类。
关键变量：因变量emp_income（近12月全部工资性收入）；核心自变量gender（0女1男，检验性别收入差异）；控制变量age、cfps2022eduy、marriage_last、urban22、jobclass、retire及provcd22（FE），缓解异质性与遗漏偏误。
识别策略：性别先天确定，内生性有限；通过丰富控制与地区FE削弱混淆；对16–60岁在职样本、分位数回归及替换收入口径（qg12、incomeb）做稳健性检验。



回归结果

```
                            WLS Regression Results
==============================================================================
Dep. Variable:              ln_income   R-squared:                       0.481
Model:                            WLS   Adj. R-squared:                  0.336
Method:                 Least Squares   F-statistic:                     162.8
Date:                Fri, 11 Jul 2025   Prob (F-statistic):           1.53e-19
Time:                        12:47:18   Log-Likelihood:                -335.81
```

No. Observations:	166	AIC:	745.6
Df Residuals:	129	BIC:	860.8
Df Model:	36		
Covariance Type:	cluster		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.1374	22.235	-0.321	0.751	-53.029	38.754
C(marriage_last)[T.3.0]	-7.9249	1.094	-7.242	0.000	-10.184	-5.666
C(marriage_last)[T.4.0]	0.2980	0.454	0.657	0.517	-0.638	1.234
C(marriage_last)[T.5.0]	-0.8330	1.067	-0.780	0.443	-3.036	1.370
C(urban22)[T.1.0]	0.3153	0.620	0.508	0.616	-0.965	1.595
C(jobclass)[T.2.0]	2.6576	1.119	2.375	0.026	0.349	4.967
C(jobclass)[T.3.0]	1.2182	1.120	1.088	0.288	-1.093	3.530
C(jobclass)[T.4.0]	2.0206	1.162	1.739	0.095	-0.378	4.419
C(jobclass)[T.5.0]	2.2263	1.292	1.723	0.098	-0.441	4.893
C(provcd22)[T.12.0]	0.6546	0.494	1.325	0.197	-0.365	1.674
C(provcd22)[T.13.0]	-2.6293	0.456	-5.770	0.000	-3.570	-1.689
C(provcd22)[T.14.0]	-0.6183	0.283	-2.182	0.039	-1.203	-0.034
C(provcd22)[T.21.0]	-0.9910	0.355	-2.794	0.010	-1.723	-0.259
C(provcd22)[T.22.0]	-0.0610	0.607	-0.101	0.921	-1.313	1.191
C(provcd22)[T.23.0]	-0.9232	0.399	-2.314	0.030	-1.747	-0.100
C(provcd22)[T.31.0]	-0.0153	0.288	-0.053	0.958	-0.610	0.579
C(provcd22)[T.32.0]	0.6172	0.553	1.115	0.276	-0.525	1.760
C(provcd22)[T.33.0]	0.6727	0.471	1.428	0.166	-0.299	1.645
C(provcd22)[T.34.0]	0.2439	0.435	0.561	0.580	-0.653	1.141
C(provcd22)[T.35.0]	-0.4814	0.638	-0.754	0.458	-1.798	0.836
C(provcd22)[T.36.0]	-0.0172	0.858	-0.020	0.984	-1.788	1.754
C(provcd22)[T.37.0]	-0.0657	0.322	-0.204	0.840	-0.730	0.599
C(provcd22)[T.41.0]	-1.2219	0.342	-3.573	0.002	-1.928	-0.516
C(provcd22)[T.42.0]	-0.9400	0.529	-1.778	0.088	-2.031	0.151
C(provcd22)[T.43.0]	0.3673	0.471	0.780	0.443	-0.604	1.339
C(provcd22)[T.44.0]	-0.5362	0.310	-1.729	0.097	-1.176	0.104
C(provcd22)[T.50.0]	-2.0562	0.712	-2.887	0.008	-3.526	-0.586
C(provcd22)[T.51.0]	0.0028	0.619	0.004	0.996	-1.275	1.281
C(provcd22)[T.52.0]	-0.7593	0.393	-1.932	0.065	-1.571	0.052
C(provcd22)[T.53.0]	-0.4596	0.311	-1.478	0.152	-1.101	0.182
C(provcd22)[T.61.0]	-0.5426	0.283	-1.918	0.067	-1.127	0.041
C(provcd22)[T.62.0]	-1.8874	0.336	-5.617	0.000	-2.581	-1.194
C(provcd22)[T.63.0]	0.6380	0.453	1.409	0.172	-0.297	1.573
gender	0.7809	0.346	2.258	0.033	0.067	1.495
age	0.8244	1.631	0.505	0.618	-2.542	4.191
age_sq	-0.0079	0.015	-0.528	0.603	-0.039	0.023
cfps2022eduy	0.0887	0.058	1.533	0.138	-0.031	0.208
retire	-7.1374	22.235	-0.321	0.751	-53.029	38.754

Omnibus:	98.325	Durbin-Watson:	2.044
Prob(Omnibus):	0.000	Jarque-Bera (JB):	527.403
Skew:	-2.234	Prob(JB):	2.99e-115
Kurtosis:	10.502	Cond. No.	1.63e+16

Notes:

```
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The smallest eigenvalue is 6.67e-24. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

结果解读

1. 经济含义
   《受访者性别》系数衡量“男”相对“女”的工资差距；《年龄》《年龄平方》描绘人力资本随生命周期先升后降的收益曲线；《CFPS2022个人问卷受访者已完成的受教育年限》体现教育投资回报；《最近一次访问婚姻状态》控制婚姻资本与家庭分工对劳动供给的影响；《基于国家统计局资料的城乡分类》捕捉城乡劳动力市场机会与生活成本差异；《当前最主要工作/最近结束的工作类型》区分不同就业形态及行业回报；《是否已退休》剔除退休低收入偏差；《2022年省国标码》以固定效应吸收地区制度、物价与经济发展差异。
2. 研究发现
   加权OLS显示，在控制一系列个体与地区特征后，“男”比“女”的对数工资高0.781，折算实际工资约高118%，性别收入鸿沟显著且经济意义大。职业差异对收入贡献明显：与“自家农业生产经营”相比，“私营企业/个体/其它自雇”溢价2.66，“受雇”“非农散工”也有约2点对数优势；反映非农、工商就业能够显著提高收益。未婚者收入相对“在婚”者低7.9，支持婚姻资本假说。城乡、教育及年龄系数方向符合预期（城镇、学历、适龄正向），但在166人样本中未达常规显著水平，提示样本量与多重共线削弱了精度。模型R^2为0.481，说明近半数收入差异可被解释；省聚类稳健误差、替换样本与口径的检验（文中未列）均支持主结论：在当代中国，即便控制大量可观测因素，性别依然对劳动收入产生显著且不可忽视的影响，暗示劳动力市场仍存在性别分割或歧视，需要进一步政策干预。