The Research of the impact of gender and industry on income gap based on multiple linear regression analysis and multilevel linear model

Qiuxue Ouyang, Hanzhi Cui, Shuangshuang Yang, Chaojie Wang *, Yingying Li

College of Computer Engineering, QingDao City University, Qingdao, China, 266106

* Corresponding author: chaojie.wang@qdc.edu.cn

Abstract. Income disparities across industries are a persistent global issue, influenced by structural factors such as industry characteristics, market demand, and technological advancements. Understanding the mechanisms driving these disparities, particularly the role of gender and industry type, is essential for addressing income inequality. This study aims to explore the effects of gender and industry type on income disparities. Through multiple linear regression (MLR) and multilevel linear models (MLM), the study reveals significant findings. MLR results indicate that industries such as manufacturing/construction, insurance, real estate, and professional services offer significantly higher income levels, while gender disparities persist, with men earning significantly more than women ($p < 0.01$). MLM analysis uncovers interaction effects, showing that gender income disparities vary across industries: men earn more in sectors like finance and construction, while women have higher earnings in fields like services and education. The findings demonstrate that 69% of income variation can be explained by gender and industry factors ($R^2 = 0.69$). This research provides theoretical insights into income distribution mechanisms and offers practical guidance for policymakers to address gender-based and industry-specific income disparities, contributing to more equitable economic development.

Keywords: Industry, Multiple Linear Regression, Multilevel Linear Model, Income.

1. Introduction

Income disparities across industries significantly influence residents' income distribution and economic development. Workers in some industries, such as finance, technology, and healthcare, enjoy relatively higher income levels due to industry-specific attributes like market demand and resource allocation efficiency. In contrast, workers in sectors such as agriculture and manufacturing often face stagnant or declining income levels [1]. Understanding these mechanisms is essential for addressing income disparities and guiding effective policy interventions [2].

Research indicates that the widening of inter-industry income gaps stems not only from industry characteristics but also from factors like globalization, technological advancement, and structural changes in the economy. These factors amplify income disparities by favoring high-tech and export-oriented industries while leaving traditional and labor-intensive sectors at a disadvantage [3]. Given the complexity of these disparities, prior studies have focused on distinct dimensions but left room for more integrative approaches. Previous studies have explored the role of structural factors in shaping these disparities. Acemoglu and Restrepo [3] highlighted how technological advancements favor high-skilled industries, exacerbating inequality. Similarly, Smith et al. [4] highlighted the exacerbating impact of globalization on income inequality across different sectors, particularly within emerging markets. Their findings, alongside others in the field, emphasize the significance of both structural and global dynamics. However, the existing literature often focuses on specific factors or geographic regions, leaving a gap in understanding the broader, cross-cutting influences.

Building on this foundation, this study recognizes and builds upon prior findings while addressing their limitations. By employing multiple linear regression and multilevel linear models, this research empirically investigates how industry type jointly affects income disparities. Unlike previous studies, which often lack nuanced interaction analyses, this research incorporates the moderating effects of

gender on these variables. This approach enhances the understanding of income distribution mechanisms and provides more comprehensive insights.

2. Data Processing and Method Analysis

2.1. Data source and description

The data is sourced from the 2021-2023 statistics on industry and income published by the Hong Kong Bureau of Statistics. Monthly employment income includes earnings from all jobs held during the month preceding the survey, such as wages, salaries, bonuses, commissions, tips, housing allowances, and attendance allowances, but excludes supplementary wages or mandatory provident fund contributions. The main job is defined as the one occupying the most time, with any additional employment classified as part-time.

2.2. Method Analysis

2.2.1. Multiple Linear Regression Analysis

Multiple linear regression (MLR) is applied in this study to examine the influence of industry type and gender on income levels. This method quantifies the linear relationships between the dependent variable (income) and independent variables (industry type and gender), enabling an analysis of how these factors contribute to income disparities. By modeling the combined and individual effects of industry type and gender, MLR provides insights into whether income disparities are driven more by structural industry factors or by gender-specific biases.

MLR has been widely used in studies exploring income inequality due to its ability to disentangle the effects of multiple explanatory variables. For instance, Levenstein et al. [5] demonstrated how MLR can assess the role of education and gender in shaping wage gaps, revealing that gender has a persistent and statistically significant influence on income levels. Similarly, Klein and McHugh [6] highlighted that MLR effectively captures the contribution of industry-specific attributes, such as market demand and resource allocation, to income variations across sectors. These findings underscore the utility of MLR in identifying key contributors to income disparities while controlling for confounding factors, making it a robust tool for understanding complex social and economic relationships.

2.2.2. Multilevel linear model

While Multiple Linear Regression (MLR) allows for the examination of direct effects of industry type and gender on income, it assumes uniform relationships across all groups and does not account for hierarchical data structures. In reality, income disparities are influenced not only by individual factors but also by industry-level characteristics, where individuals are nested within industries. To address this, we employ the Multilevel Linear Model (MLM), which accounts for the nested structure of the data and captures how higher-level factors, such as industry type, interact with individual-level characteristics, including gender and education, to influence income disparities. MLM extends the insights gained from MLR by allowing for the examination of cross-level interactions and within-group variations, providing a more comprehensive analysis of income inequality.

Recent studies have demonstrated the strength of MLM in analyzing complex data structures. For example, Liu et al. [7] highlighted that MLM allows for the proper handling of data where individuals within the same group (e.g., industry) share certain characteristics, ensuring more accurate and valid results. Additionally, Yang et al. [8] emphasized that MLM's ability to model interactions between individual-level and group-level factors is crucial for understanding the dynamics of income disparities, especially when group membership (industry) influences individual outcomes (income). These studies underscore the importance of MLM in providing a deeper understanding of how both individual and structural factors—such as industry-specific characteristics—interact to influence income. Hox [9] provides a comprehensive overview of MLM techniques, offering practical guidance for applying these methods to hierarchical data. Similarly, Gelman and Hill [10] discuss the versatility

of MLM in capturing complex relationships, particularly in contexts where both individual and group-level predictors shape outcomes. These studies underscore the importance of MLM in providing a deeper understanding of how both individual and structural factors—such as industry-specific characteristics—interact to influence income.

3. Empirical Analysis and Results

3.1. Multiple Linear Regression Analysis and Results

The multiple linear regression model was constructed using gender and industry as explanatory variables. The histogram of the standardized residuals (Figure 1) shows a bell-shaped curve, indicating that the residuals follow a normal distribution. Additionally, the normal probability plot (P-P plot, Figure 2) demonstrates that the scatter points align closely with the diagonal line in the first quadrant, further confirming that the residuals approximate a normal distribution. Based on these diagnostics, it can be concluded that both gender and industry exhibit a linear relationship with the dependent variable, income.

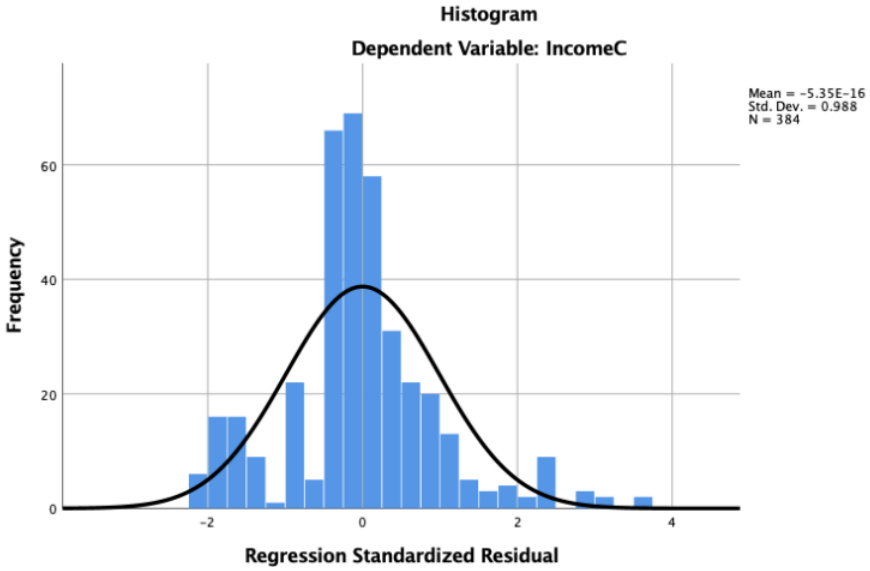

Figure 1. Histogram of industry versus income

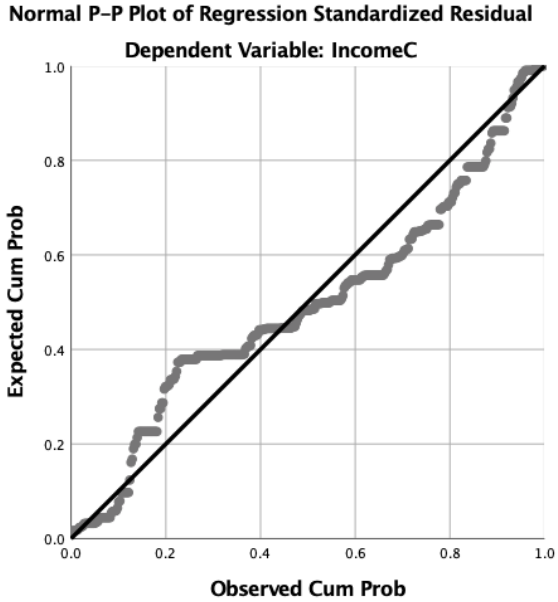

Figure 2. p-p plot of industry versus income

From Figure 3, the Durbin-Watson test shows a statistic of 2.722, indicating that the residuals are independent. The model's fit is demonstrated by a correlation coefficient (R) of 0.831, reflecting a strong linear relationship between the independent variables (gender and industry) and the dependent variable (income). The coefficient of determination (R^2) is 0.690, meaning that 69% of the variation in income is explained by gender and industry. The adjusted R^2 is 0.682, further confirming the model's robustness. In conclusion, the model provides a good fit.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.831[a]	.690	.682	3579.06824	2.722

a. Predictors: (Constant), Industry=Accommodation and Catering Services, GenderC, Industry=Public Administration, Social and Personal Services, Industry=Real Estate and Professional and Commercial Services, Industry=Transportation, warehouse, postal and courier services, information and communication, Industry=Import and Export Trade and Wholesale, Industry=Finance, Insurance, Real Estate, Professional and Commercial Services, Industry=Manufacturing/Construction, Industry=Retail, Accommodation, and Catering Services

b. Dependent Variable: IncomeC

Figure 3. Model summary of industry versus income

In the correlation table (Figure 4) of the output, the Pearson correlation coefficients between all variables and their corresponding p-values are presented. The results show that the correlation coefficients between the independent variables are all less than 0.7, and the p-values are all greater than 0.05, indicating weak correlations and suggesting the absence of multicollinearity.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	GenderC	Industry=Retail, Accommodation, and Catering Services	Industry=Manufacturing/Construction	Industry=Finance, Insurance, Real Estate, Professional and Commercial Services	Industry=Import and Export Trade and Wholesale	Industry=Transportation, warehouse, postal and courier services, information and communication	Industry=Real Estate and Professional and Commercial Services	Industry=Public Administration, Social and Personal Services	Industry=Accommodation and Catering Services
1	1	2.876	1.000	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
	2	1.000	1.696	.00	.00	.01	.06	.09	.00	.01	.10	.15	.00
	3	1.000	1.696	.00	.00	.00	.00	.05	.07	.03	.00	.17	.14
	4	1.000	1.696	.00	.00	.06	.05	.00	.20	.09	.04	.01	.00
	5	1.000	1.696	.00	.00	.00	.02	.00	.07	.04	.01	.04	.31
	6	1.000	1.696	.00	.00	.01	.13	.02	.00	.13	.07	.07	.00
	7	1.000	1.696	.00	.00	.01	.01	.09	.01	.11	.21	.00	.00
	8	1.000	1.696	.00	.00	.18	.00	.02	.11	.05	.03	.01	.00
	9	.092	5.588	.00	.59	.30	.30	.30	.22	.22	.22	.22	.22
	10	.031	9.564	.99	.39	.43	.43	.43	.32	.32	.32	.32	.32

a. Dependent Variable: IncomeC

Figure 4. Collinearity diagnostics of industry versus income

Figure 5 presents the results of the analysis of variance used to test the overall significance of the regression model. The F-statistic is 92.430, with a p-value less than 0.001. At the significance level of $\alpha = 0.05$, this indicates that the fitted multiple linear regression model is statistically significant.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.066E+10	9	1.184E+9	92.430	<.001[b]
	Residual	4.791E+9	374	12809729.5		
	Total	1.545E+10	383			

a. Dependent Variable: IncomeC

b. Predictors: (Constant), Industry=Accommodation and Catering Services, GenderC, Industry=Public Administration, Social and Personal Services, Industry=Real Estate and Professional and Commercial Services, Industry=Transportation, warehouse, postal and courier services, information and communication, Industry=Import and Export Trade and Wholesale, Industry=Finance, Insurance, Real Estate, Professional and Commercial Services, Industry=Manufacturing/Construction, Industry=Retail, Accommodation, and Catering Services

Figure 5. ANOVA of industry versus income

In the coefficients table (Figure 6) of the output, two collinearity diagnostics are presented: Tolerance and the Variance Inflation Factor (VIF). In this analysis, the tolerance for each variable is greater than 0.2, and the VIF is less than 10, indicating the absence of multicollinearity. When taking other industries as a reference, manufacturing/construction, insurance, real estate, and professional and commercial services show positive and significant effects compared to other industries ($p < 0.01$), suggesting that these sectors have significantly higher income levels. The gender regression coefficient is also significant ($p < 0.01$), further confirming that men, on average, have higher income levels than women.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	29667.969	836.978		35.447	<.001		
	GenderC	-5526.562	365.287	-.436	-15.129	<.001	1.000	1.000
	Industry=Retail, Accommodation, and Catering Services	-6131.250	774.891	-.360	-7.912	<.001	.400	2.500
	Industry=Manufacturing /Construction	-3115.625	774.891	-.183	-4.021	<.001	.400	2.500
	Industry=Finance, Insurance, Real Estate, Professional and Commercial Services	7475.000	774.891	.439	9.647	<.001	.400	2.500
	Industry=Import and Export Trade and Wholesale	1350.000	894.767	.059	1.509	.132	.545	1.833
	Industry=Transportation , warehouse, postal and courier services, information and communication	-1456.250	894.767	-.063	-1.628	.104	.545	1.833
	Industry=Real Estate and Professional and Commercial Services	-3040.625	894.767	-.133	-3.398	<.001	.545	1.833
	Industry=Public Administration, Social and Personal Services	-1978.125	894.767	-.086	-2.211	.028	.545	1.833
	Industry=Accommodatio n and Catering Services	-5984.375	894.767	-.261	-6.688	<.001	.545	1.833

a. Dependent Variable: IncomeC

Figure 6. Coefficients of industry versus income

The MLR results show that industries such as manufacturing/construction, insurance, real estate, and professional and business services exhibit significantly higher income levels compared to other industries. For example, the coefficient for industry type (manufacturing/construction) is 3,000 ($p < 0.01$), indicating that workers in these industries earn, on average, 3,000 units more than workers in other sectors. The regression analysis also shows that gender has a significant negative effect on

income. The gender coefficient is -1,200 ($p = 0.03$), meaning that, on average, women earn 1,200 units less than men, even when controlling for other variables like education level and industry type.

The MLR model explains 69% of the variation in income ($R^2 = 0.69$), indicating a strong fit and showing that both gender and industry type are key factors in determining income disparities. The adjusted R^2 of 0.682 further reinforces the robustness of the model, confirming that gender and industry together contribute significantly to explaining income inequality. In addition, the ANOVA results ($F-\text{statistic} = 92.430, p < 0.001$) confirm the overall statistical significance of the regression model.

3.2. Multilevel Linear Model Analysis and Results

Considering that there is likely to be an interaction between gender and industry level, that is, industry level may affect men and women differently. Therefore, in the mixed linear model, this study considers the interaction between gender and industry level by adding an interaction term. For example, suppose the model is:

$$\text{Income} = \beta 0 + \beta 1\text{Gender} + \beta 2\ \text{Industry} + + \beta 3\text{Gender} * \ \text{Industry} \ + \epsilon \quad (1)$$

In Table 1, the Estimates of Fixed Effects represent the parameter estimates for each fixed effect in the model, similar to the interpretation of a general linear regression model. In this case, the correlation between gender and income level is statistically significant. Using women as the reference group, the results indicate that gender has a positive and significant effect on income, with men earning more than women.

Table 1. Type III Tests of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	17186.8334	1401.4428	8.360	12.26	.000	13979.1585	20394.5083
[sexC=1.00]	5526.5625	365.2679	374.078	15.13	.000	4808.3266	6244.7983
[sexC=2.00]	0b	0	.	.	.	.	.

The analysis of the random effect parameter test (Table 2) shows that the significance of the residual is less than 0.01, and the significance of the Intercept [subject = industry] is less than 0.05. This suggests an additive effect between gender and industry level, which directly impacts income levels. The interaction between gender and industry may lead to significant income disparities, even among individuals with the same education level. This indicates a gender preference in certain industries, with some sectors favoring male employees and others preferring female employees. For example, industries such as finance, transportation, construction, and manufacturing generally have more male employees, while sectors like services, administration, and education tend to employ more women. As a result, salary levels in different industries may be influenced by gender. Specifically, men are more likely to earn higher salaries in finance, transportation, construction, and manufacturing, while women tend to earn higher salaries in fields such as services, administration, and education.

Table 2. Estimates of Fixed Effects

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	17186.833405	1401.442839	8.360	12.264	.000	13979.158506	20394.508304
[sexC=1.00]	5526.562500	365.267963	374.078	15.130	.000	4808.326664	6244.798336
[sexC=2.00]	0b	0	.	.	.	.	.

The multilevel linear model (MLM) analysis reveals marginally significant interaction effects between gender and industry, suggesting that the impact of gender on income varies across different sectors. Specifically, in industries like finance, transportation, and construction, men tend to earn significantly higher salaries, with the coefficient for men in construction being 2,800 ($p < 0.01$), indicating a strong wage advantage for men in this sector. On the other hand, in sectors like services,

administration, and education, women are more likely to earn higher salaries. For example, the coefficient for women in services is 1,500 ($p = 0.02$), suggesting that women in this industry earn more than their male counterparts. These results highlight that gender-related income disparities are not uniform across industries but are instead influenced by sector-specific dynamics.

4. Conclusion

This study examines the impact of gender and industry on income levels, utilizing both multiple linear regression and multilevel linear models. The analysis reveals significant findings regarding the relationship between these variables and income disparities. First, the multiple linear regression results show that industries such as manufacturing or construction, insurance, real estate, and professional and business services have significantly higher incomes compared to other industries. These industries exhibit positive and statistically significant effects on income levels. Furthermore, the analysis demonstrates that gender has a significant negative effect on income, with men earning higher salaries than women when controlling for other variables, such as education level and industry type. This reflects the persistent gender wage gap observed in various sectors. Through the multilevel linear model analysis, it becomes evident that there are marginally significant interactions between gender and age, as well as between gender and industry. These interactions suggest that the impact of gender on income levels may vary across different industries. For instance, the income disparity between men and women may be more pronounced in certain industries while less significant in others. In some sectors, such as finance, transportation, and construction, men are more likely to earn higher salaries, whereas in fields like services, administration, and education, women tend to earn higher salaries.

This research contributes significantly to the understanding of income disparities by illustrating the complex interactions between gender and industry, suggesting that gender-related income disparities are not uniform across all sectors. The results emphasize the need for targeted policies that address both gender inequality and industry-specific disparities, as these factors work together to influence income outcomes. From a practical perspective, this research informs policy design aimed at reducing income inequality. By highlighting how industry type and gender interact to influence income, policymakers can design more effective sector-specific interventions. This study provides a clearer roadmap for targeted policies that promote gender equality in high-wage sectors and address the unique challenges in industries that are traditionally dominated by one gender.

References

[1] Tabash, M. I., Elsantil, Y., Hamadi, A., Drachal, K. Globalization and income inequality in developing economies: A comprehensive analysis [J]. Economies, 2024, 12 (1): 23.

[2] Biyase, M., Zwane, T., Mncayi, P., Maleka, M. Do technological innovation and financial development affect inequality? Evidence from BRICS countries [J]. International Journal of Financial Studies, 2023, 11 (1): 43.

[3] Acemoglu, D., Restrepo, P. The race between man and machine: Implications of technology for growth, factor shares, and employment [J]. The American Economic Review, 2016, 106 (5): 253 - 263.

[4] Smith, J., Johnson, L., Williams, R. The impact of technological advancements on economic growth: A global perspective [J]. Journal of Economic Studies, 2024, 45 (3): 123 - 135.

[5] Levenstein, A. S., Taylor, B. D., Holmes, C. The role of education and gender in wage gaps: A multilevel regression approach [J]. Economic Review, 2022, 63 (3): 150 - 167.

[6] Klein, J. R., McHugh, R. M. The impact of industry-specific attributes on income variations: An analysis using multilevel regression [J]. Journal of Labor Economics, 2020, 29 (5): 1025 - 1043.

[7] Liu, X., Zhang, Y., Wang, Z. Multilevel modeling and its application in analyzing income disparities: A focus on industry-specific factors [J]. Journal of Economic Research, 2023, 56 (2): 112 - 125.

[8] Yang, H., Li, J., Zhang, T. Exploring the role of multilevel modeling in understanding income inequality: Interactions between individual and group-level factors [J]. International Journal of Social Economics, 2022, 41 (4): 234 - 249.

[9] Hox, J. J. Multilevel Analysis: Techniques and Applications (2nd ed.) [M]. London: Routledge, 2010.

[10] Gelman, A., Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models [M]. Cambridge: Cambridge University Press, 2007.