ISSN: 0976-3376

Available Online at http://www.journalajst.com

ASIAN JOURNAL OF SCIENCE AND TECHNOLOGY

Asian Journal of Science and Technology
Vol. 17, Issue, 09, pp. 14530-14542, September, 2026

RESEARCH ARTICLE

A COMPARATIVE EVALUATION OF OLS, ROBUST REGRESSION, AND QUANTILE REGRESSION UNDER DISTRIBUTIONAL HETEROGENEITY AND OUTLIER CONTAMINATION

[1]Manali De, [2]Tiyasha Samanta, [3]Manaswini Mondal and [4,*]Argha Nath Bhattacharyya

[1]Department of Statistics, Aliah University, Kolkata; [2]Department of Statistics, Sister Nivedita University, Kolkata; [3]ICMR - National Institute for Research in Bacterial Infections, Kolkata; [4]Department of Statistics, Sister Nivedita University, Kolkata

ARTICLE INFO

Article History:
Received 29th June, 2026
Received in revised form
25th July, 2026
Accepted 20th August, 2026
Published online 30th September, 2026

Key words:
Income Inequality, Ordinary Least Squares, Robust Regression, Quantile Regression, K-fold cross-validation, Socio-economic Determinants.

***Corresponding author:**
Argha Nath Bhattacharyya

ABSTRACT

This study examines the impact of key socio-economic factors—education, work experience, gender, and age—on income, with a particular focus on capturing heterogeneity across the income distribution. A structured preprocessing framework is employed, including missing value treatment, categorical variable transformation, and outlier detection using the Interquartile Range (IQR) method. The dataset is divided into training (80%) and test (20%) subsets to assess model generalizability, and K-fold cross-validation is applied to evaluate predictive stability. Ordinary Least Squares (OLS) regression is initially used to estimate average effects, followed by diagnostic tests to assess assumptions such as normality, homoscedasticity, and the presence of influential observations. To address these limitations, Robust Regression is employed to obtain stable estimates in the presence of outliers, while Quantile Regression is implemented at the 25th, 50th, and 75th percentiles to capture distributional heterogeneity. The results show that OLS estimates are sensitive to extreme observations and assumption violations, whereas Robust Regression improves stability. Quantile Regression reveals that the returns to education increase across higher income quantiles, indicating significant heterogeneity. Model evaluation using test data and cross-validation suggests that OLS achieves the lowest RMSE, while alternative methods provide deeper insights into the income structure. Overall, the study highlights the limitations of mean-based approaches and demonstrates that combining robust and quantile-based methods offers a more comprehensive framework for analyzing income inequality.

Citation: ***Manali De, Tiyasha Samanta, Manaswini Mondal and Argha Nath Bhattacharyya.* 2026.** "A comparative evaluation of ols, robust regression, and quantile regression under distributional heterogeneity and outlier contamination.". *Asian Journal of Science and Technology*, 17, (09), 14530-14542.

Copyright©2026, Manali De et al. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

INTRODUCTION

"The welfare of a nation can scarcely be inferred from a measure of national income." — Simon Kuznets

There are many socio-economic issues worldwide, and one of the most frequent and most important issues that are faced by an individual is income inequality. It adversely affects the social mobility and the economic development of a society (1). Therefore, income inequality has a considerable effect on the overall quality of everyday life . It has been seen that the unequal income distribution influences access to higher education, good healthcare facilities, suitable employment opportunities and the living standards of the society (2, 3). Hence, it is essential for individuals to understand the structure as well as the factors of income distribution to create any necessary change in the economy and society. Previous studies have highlighted that there are a high association between income inequality and wide economic and social outcomes. To formulate an evidence-based policy, it is obvious to analysed the wage and salary differences (4). Hence, the researchers are more focused to identifying the factors causing variations in in personal earnings and income distribution between different groups of the population (3). Some socio-economic characteristics are identified as the key factors of Income distri-bution, such as education, work experience, age and gender (1, 3). Education is widely recognized as an important factor of earnings (5). In general, higher levels of educational attainment are linked to better skills, productivity, and job opportunities. Similarly, working experience raises higher income by gaining practical knowledge and professional skills. Age is also linked with career advancement and participation in the labor market, while gender wage gaps remain an important issue in labor economics and studies of in-equality (3). The differences in income levels between individuals and groups are heavily influenced by these socioeconomic determinants.

Different regression techniques have been widely used to analysis the income determi-nants. Generally, income distributions are skewed and distinguished by heteroscedasticity, outliers, and influential observations (6). Hence, there are several statistical challenges with income-related data. In these cases, average -based models can be inadequate in capturing the complexity of the underlying distribution. Ordinary Least Squares (OLS) regression is the fundamental statistical method that used to estimates the mean rela-tionship between predictors and outcome variable (2). But the drawback of OLS is it is highly sensitive to the outliers and it may cause the biasedness to interpret the socio-economic relationships (7). Hence, using only mean-based estimation may hide important differences across various segments of the income distribution. Robust regression is used to fulfil the limitations of OLS. It reduces the influence of outliers. In the presence of outliers, robust regression provides more stable parameter estimates than the OLS (7). Quantile regression allows studying the impact of predictor effects across different points of the conditional income distribution rather than focus-ing only on the mean. This is especially useful in the field of inequality research where the effects of socio-economic variables may vary significantly between lower-income and higher-income groups (8). Earlier studies have shown that quantile regression provides more insight regarding wage inequality, heterogeneous socio-economic effects, and distri-butional dynamics than OLS and robust regression. This study approaches examining the impact of socio-economic factors on income through multi-regression models. This study accomplished Ordinary Least Squares (OLS) regression, robust regression, and quantile regression techniques to examine the effect of education, work experience, age, and gender on salary distribution. Additionally, di-agnostic tests and k-fold cross-validation will be performed to assess the assumptions, predictive validity, and generalizability of the model used. The study aims to provide a more comprehensive understanding of income determination and distributional hetero-geneity across different income groups through the application of mean modeling, robust modeling, and distribution modeling of income.

METHODOLOGY

In order to comprehend how many socioeconomic factors affect income, this study employs a structured analytical technique. Multiple regression techniques are employed to capture both average effects and variances across various income levels rather than depending on a single model (1). The average association between income and the explanatory variables is estimated in the first step of the study using OLS regression (2). This model offers a foundational understanding of the average salary impact of variables such as education and experience and gender. However, OLS depends on a number of presumptions, including the absence of extreme observations and constant error variance. Statistical tests and residual plots are diagnostic techniques used to verify these assumptions (2, 6). Relying just on OLS can produce inaccurate findings since income data frequently contains outliers. Robust regression is used to solve this problem. By giving extreme values less weight during estimation, this technique lessens their impact (7). Consequently, robust regression yields coefficients that are more stable and more accurately represent the normal pattern in the data. Because the Salary distribution in this study exhibits significant variability and the existence of abnormally high values, robust regression is very helpful. While OLS focuses on the mean, it does not capture how relationships may differ across income groups. Quantile regression is used to get around this restriction (9). The 25th, 50th, and 75th percentiles are used to estimate the models. This makes it possible to investigate how the effects of age, gender, education, and experience differ for those with low, middle, and high incomes. This method helps determine whether particular factors have greater influence at different income levels and gives a more comprehensive picture of income inequality (3, 10, 11, 12). A 5-fold cross-validation method is used to assess the models' dependability (13). To guarantee consistency, this technique splits the training data into smaller sections and tests the model several times. The model's performance is evaluated using:

- Mean Squared Error (MSE), which calculates the average squared difference be-tween actual and forecasted values.
- Mean Absolute Error (MAE), which is the average of the absolute discrepancies between the actual and anticipated values.
- Root Mean Squared Error (RMSE), which is squared root of MSE.

These metrics aid in comparing the accuracy of each model's income prediction. Three distinct regression algorithms are combined in the methodology to produce a fair analysis:

- Ordinary Least Squares (OLS) explains the average effect.
- Robust regression enhances reliability in the presence of outliers.
- Quantile regression captures variation across different income levels.

The study avoids the drawbacks of a single strategy and provides a more thorough un-derstanding of income determination by combining all three approaches.

Data Availability: The dataset used in this study is generated and maintained in Microsoft Excel for the pur-pose of empirical analysis. The original dataset consists of individual-level observations with the following variables:

- **Age:** Age of the individual (in years)
- **Gender:** Categorical variable indicating gender (Male/Female)
- **Education. Level:** Educational attainment recorded as *Bachelor*, *Master*, and *PhD*
- **Years. of. Experience:** Total years of professional work experience

- **Salary:** Monthly income in Rs. (response variable)

The dataset is structured to reflect realistic socio-economic patterns and is used ex-clusively for methodological and academic purposes. It does not contain any personally identifiable information. The dataset is publicly available and can be accessed at: https://docs.google.com/spreadsheets/d/1wIW-_AxHsYsAuKMgmtql32hptHy37BUonvmLtB2Bv0k/edit?gid=1695379265#gid=1695379265 Researchers may use the dataset for replication and further academic research.

Data Preprocessing: Prior to model estimation, the dataset was preprocessed to ensure quality and consis-tency. Missing values were removed using listwise deletion (na.omit) to obtain a complete dataset (14). The categorical variable Education. Level (Bachelor, Master, PhD) was converted into an ordinal numerical variable Education Level Num using the mapping (15):

Bachelor = 1, Master = 2, PhD = 3, while Gender was treated as a factor variable.

Descriptive Statistics: Table 1 presents the descriptive statistics of the principal variables considered in the study.

Table 1: Summary Statistics of Variables

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Age	22	32	40	40.92	50	60
Years of Experience	0	8	17	16.86	25	35
Salary (Rs./Monthly)	10,568	57,709	85,770	154,802	121,875	698,885

The dataset consists of 335 observations with a nearly balanced gender distribution (169 females and 166 males). The age of individuals ranges from 22 to 60 years, with an average age of 40.92 years, indicating that the sample primarily represents mid-career in-dividuals. Similarly, the average years of experience is 16.86 years, suggesting a workforce with moderate to substantial professional exposure. Table 2 shows the distribution of individuals across different education levels.

Table 2. Distribution of Education Levels

Education Level	Frequency
Bachelor (1)	105
Master (2)	108
PhD (3)	122

The distribution indicates a relatively balanced representation across educational cat-egories, with a slightly higher proportion of individuals holding a PhD degree. This balanced structure enhances the reliability of comparative analysis across education lev-els. Table 3 reports additional descriptive measures for continuous variables.

Table 3. Numerical Summary Statistics

Variable	Mean	Standard Deviation	Minimum	Maximum
Salary	154,802.5	175,437.7	10,568	698,885
Age	40.92	10.78	22	60
Years of Experience	16.86	10.12	0	35

The descriptive analysis reveals substantial heterogeneity in the dataset, particularly in salary distribution. The mean salary (154,802.5) is considerably higher than the median salary (85,770) (16), indicating a pronounced right-skewed distribution. This suggests that a relatively small number of high-income individuals significantly elevate the average income level. Furthermore, the large standard deviation of salary (175,437.7), which exceeds the mean, highlights extreme variability and the presence of high-value observations. The maximum salary (698,885) is substantially larger than the third quartile (121,875), re-inforcing the existence of upper-tail outliers. Such characteristics are typical of income data and reflect underlying income inequality. In contrast, age and years of experience exhibit more symmetric distributions with moderate variability, as indicated by their relatively lower standard deviations. The median values (40 years for age and 17 years for experience) suggest that the sample is concentrated around mid-career individuals. To understand the underlying structure and distribution of the data, several graphical analyses were conducted, focusing on salary distribution and its relationship with key explanatory variables.

Salary Distribution: The histogram in Figure 1 reveals a pronounced right-skewed distribution of salary. A majority of individuals fall within lower to middle income ranges, while a small proportion earns significantly higher salaries, forming a long right tail. This skewness indicates the presence of income inequality and suggests that mean-based models may be influenced by extreme values (16, 17).

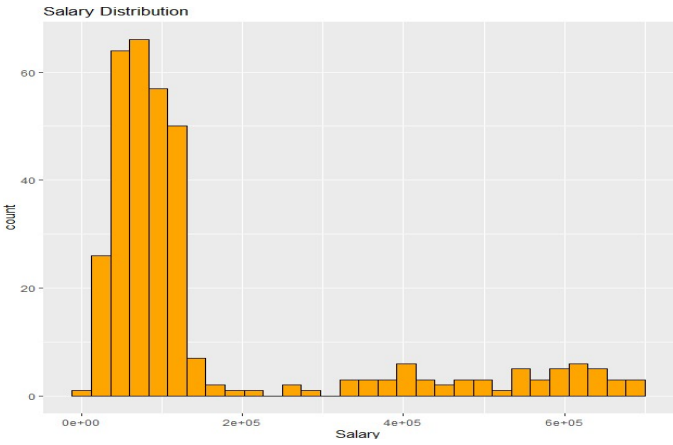

Figure 1. Histogram of Salary Distribution

Detection of Outliers

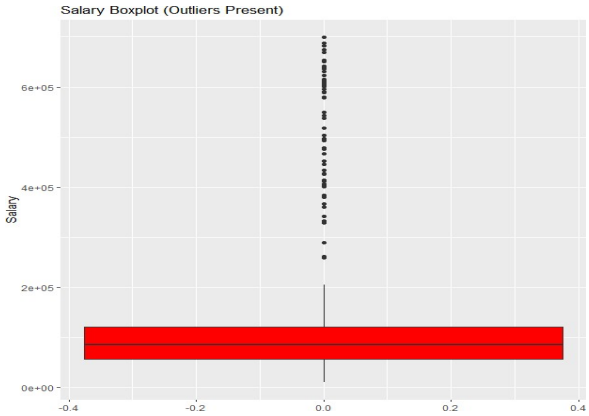

Figure 2 illustrates the presence of several extreme observations (outliers) in the upper tail of the salary distribution. These outliers may exert undue influence on regression estimates, particularly in Ordinary Least Squares (OLS), thereby motivating the use of robust regression techniques (18, 19).

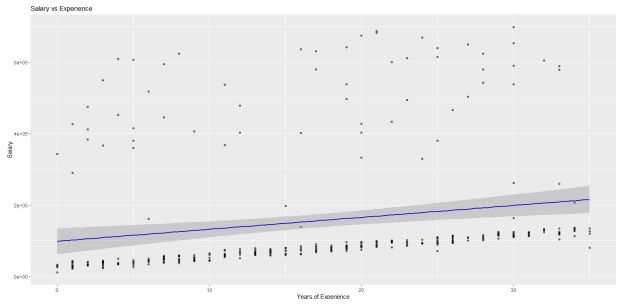

Figure 2. Boxplot of Salary Showing Outliers

Relationship Between Salary and Experience

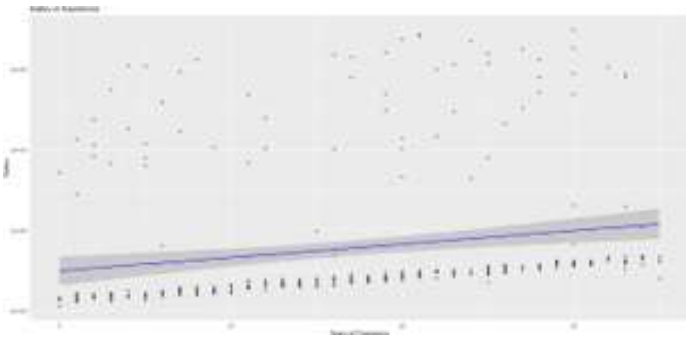

Figure 3. Scatter Plot of Salary vs. Years of Experience with Fitted Line

The scatter plot in Figure 3 shows a positive relationship between salary and years of experience. The fitted regression line suggests that, on average, salary increases with experience. However, the wide dispersion of points around the line indicates substantial variability, implying that experience alone does not fully explain salary differences. This further highlights the need for multivariate modeling.

Salary Across Education Levels

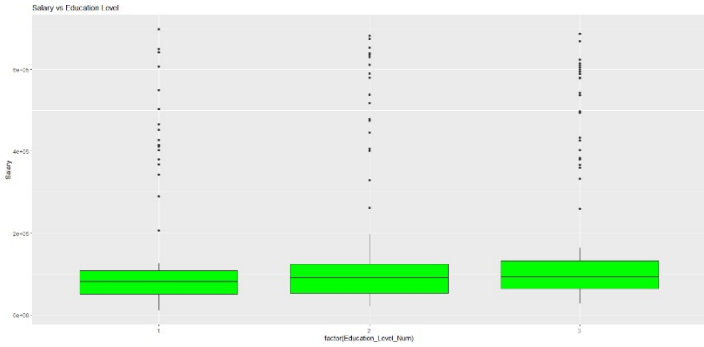

Figure 4. Boxplot of Salary Across Education Levels

Figure 4 presents the distribution of salary across different education levels. It is evident that higher education levels are generally associated with higher median salaries. However, the spread within each category is considerable, and overlapping distributions suggest heterogeneity in returns to education. This observation supports the application of quantile regression to capture differential effects across the income distribution.

Ordinary Least Squares (OLS) Regression: The dataset is randomly partitioned into two subsets, with 80% of the observations allocated to the training set and the remaining 20% reserved for the test set (13). All model estimation, diagnostic testing, and parameter tuning are performed exclusively on the training data to ensure unbiased learning of underlying relationships. The test dataset, which remains completely unseen during model fitting, is subsequently used to evaluate the predictive performance and generalizability of the estimated models (2). This approach enables an objective comparison between models based on their ability to replicate observed outcomes in new data, thereby providing a robust assessment of their out-of-sample predictive accuracy.

Model Specification; To examine the relationship between salary and key socio-economic determinants, an Ordinary Least Squares (OLS) regression model was estimated using the training dataset (80% of the sample). The model is specified as follows:

$$Salary_i = \beta_0 + \beta_1 Education_i + \beta_2 Experience_i + \beta_3 Gender_i + \beta_4 Age_i + \epsilon_i$$

where $i = 1, 2, 3, \ldots, n$. (20)

Table 4 presents the estimated coefficients along with their standard errors, t-values, and significance levels.

Table 4. OLS Regression Results

Variable	Estimate	Std. Error	t-value	p-value
Intercept	-267.0	52,442.6	-0.005	0.9959
Education Level	26,074.5	12,726.7	2.049	0.0415*
Years of Experience	2,939.0	1,035.7	2.838	0.0049**
Gender (Male)	-28,290.6	20,950.2	-1.350	0.1780
Age	1,637.2	971.4	1.685	0.0931

Significance levels: * $p < 0.05$, ** $p < 0.01$

and Model Fit Statistics

Table 5. Model Fit Statistics

Statistic	Value
Residual Standard Error	171,400
R^2	0.0565
Adjusted R^2	0.0423
F-statistic	3.981
p-value (F-test)	0.0037

The results indicate that both education and years of experience have statistically significant positive effects on salary. Specifically, a one-unit increase in the education level (e.g., from Bachelor to Master) is associated with an average increase of approximately 26,074.5 Rs. in salary, holding other factors constant. Similarly, each additional year of experience increases salary by approximately 2,939 Rs., which is statistically significant at the 1% level. The coefficient for gender (Male) is negative, suggesting that males earn, on average, 28,290.6 Rs. less than females in this sample. However, this effect is not statistically significant, indicating insufficient evidence to conclude a gender-based difference in salary within this model. Age has a positive coefficient, implying that salary tends to increase with age, but its effect is only marginally significant at the 10% level. This suggests that age may play a role in income determination, though its influence is weaker compared to education and experience.) Despite some statistically significant predictors, the overall explanatory power of the model is relatively low, as indicated by an R^2 value of 0.0565. This implies that only about 5.65% of the variation in salary is explained by the included variables. The adjusted R^2 further confirms this limited explanatory capacity. However, the F-test is statistically significant (p-value = 0.0037), indicating that the model is jointly significant and that at least one predictor contributes meaningfully to explaining salary variation. The large residual standard error (171,400) reflects substantial unexplained variability in salary, which may be attributed to omitted variables, non-linear relationships, or the presence of outliers, as identified in the exploratory analysis (20). The weak model fit and high residual variability are consistent with earlier exploratory findings, which indicated a highly skewed salary distribution and the presence of extreme values. These characteristics violate key OLS assumptions, particularly homoscedasticity and sensitivity to outliers. Therefore, while OLS provides a baseline understanding of average effects, its limi-tations in the presence of heterogeneity and outliers necessitate the use of more robust and distribution-sensitive methods, such as Robust Regression and Quantile Regression, which are explored in subsequent sections.

Diagnostic Analysis of OLS Model: To assess the validity of the classical linear regression assumptions, a series of diagnostic plots were generated, as shown in Figure 5 (2).

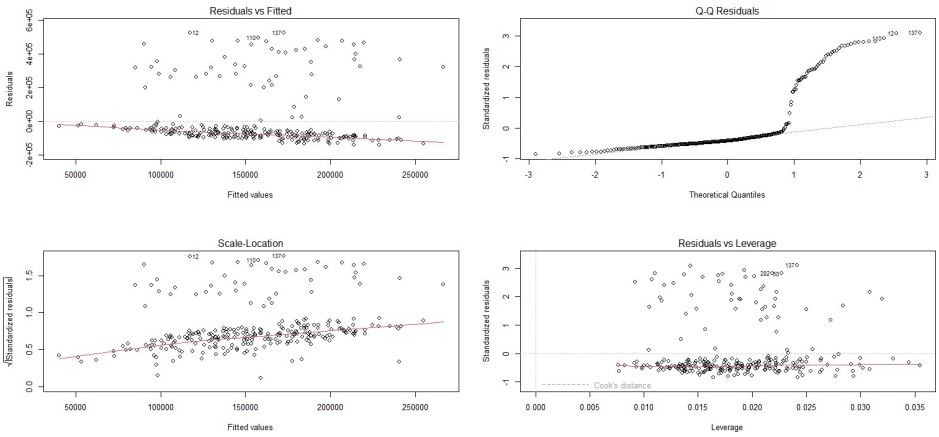

Figure 5. Diagnostic Plots for OLS Model

The diagnostic plots provide important insights into the adequacy of the OLS model. The Residuals vs Fitted plot indicates a non-random pattern in the residuals, sug-gesting potential violation of the homoscedasticity assumption. Ideally, residuals should be randomly scattered around zero (21); however, the observed structure implies that the variance of errors may not be constant across fitted values. The Normal Q-Q plot reveals deviations from the straight reference line, particu-larly in the tails. This suggests that the residuals are not perfectly normally distributed, which is consistent with the earlier finding of a right-skewed salary distribution. The presence of heavy tails indicates the influence of extreme observations. The Scale-Location plot further supports the presence of heteroscedasticity, as the spread of residuals appears to increase with fitted values. This implies that the variability of salary predictions is not uniform across different levels of income. The Residuals vs Leverage plot highlights the presence of influential observations. Points with high leverage and large residuals may disproportionately affect the estimated regression coefficients, leading to instability in OLS estimates. Overall, the diagnostic analysis suggests that key assumptions of the OLS model, particularly homoscedasticity and normality of residuals—are likely violated. These find-ings reinforce the need for robust estimation techniques and distribution-sensitive models, such as Robust Regression and Quantile Regression, which are better suited to handle outliers and heterogeneity in the data.

Residual Distribution Analysis: To further examine the distributional properties of the error terms, a histogram of the residuals from the OLS model is presented in Figure 6. The residual histogram indicates a deviation from the normality assumption of the OLS model, as the distribution is asymmetric and exhibits skewness rather than a bell-shaped form. The presence of long tails suggests the influence of extreme observations, consistent with detected outliers in the salary variable. This violation of normality, also supported by the Q-Q plot, may affect the reliability of inference. Consequently, it justifies the use of alternative approaches such as robust and quantile regression, which are less sensitive to such deviations (7, 10).

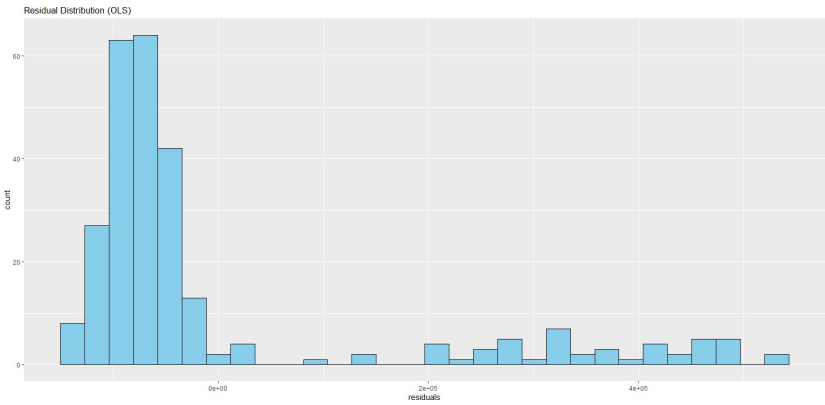

Figure 6. Histogram of Residuals from OLS Model

Influence Diagnostics Using Cook's Distance: To identify influential observations that may disproportionately affect the regression es-timates, Cook's Distance was computed for each observation (19). The results are illus-trated in Figure 7.

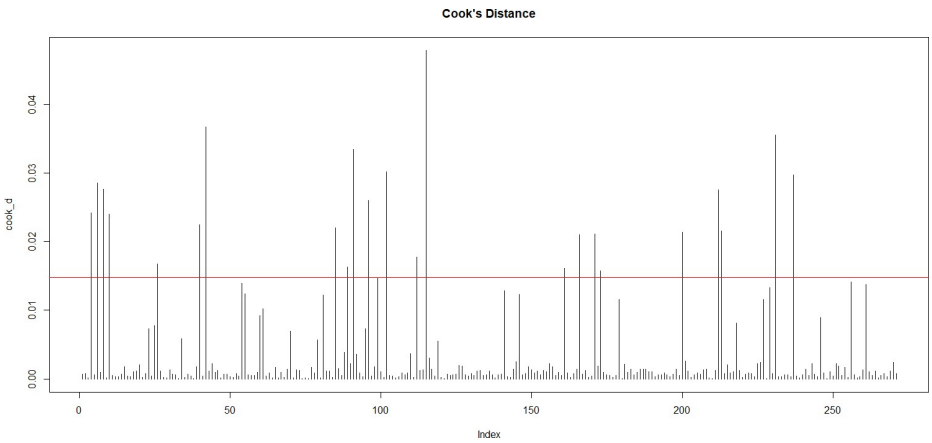

Figure 7. Cook's Distance for OLS Model

Cook's Distance measures the influence of individual observations on the regression es *n* timates. Observations exceeding the threshold are considered influential (19). The plot shows several points above this cutoff, indicating the presence of influential observations that can distort OLS estimates. This finding aligns with earlier evidence of outliers in the salary distribution. Since OLS is sensitive to such points, the results may be unstable or biased. Therefore, the presence of influential observations justifies the use of robust and quantile regression methods, which are less affected by extreme values and provide more reliable estimates.

Test for Heteroscedasticity: To examine whether the variance of the error terms remains constant, the Breusch–Pagan test was conducted (22). The results are reported in Table 6. The Breusch–Pagan test is used to examine whether the variance of the error terms is constant across observations.

•Null Hypothesis (H_0): The error terms are homoscedastic, i.e., they have con-stant variance.
•Alternative Hypothesis (H_1): The error terms are heteroscedastic, i.e., the variance of the errors is not constant.

Table 6. Breusch–Pagan Test for Heteroscedasticity

Statistic	Value
BP Statistic	8.29
Degrees of Freedom	4
p-value	0.0815

The null hypothesis of the Breusch–Pagan test assumes homoscedasticity, i.e., constant variance of the error terms. The obtained p-value (0.0815) is greater than the conventional 5% significance level but less than 10%, indicating weak evidence against the null hypothesis. This suggests that while heteroscedasticity is not strongly significant at the 5% level, there is marginal evidence of non-constant variance in the residuals. This finding aligns with the earlier diagnostic plots, which indicated some degree of heteroscedasticity. There-fore, caution is warranted when interpreting OLS estimates, and robust methods may still be preferable (6, 7).

Multicollinearity Diagnostics: To assess multicollinearity among the explanatory variables, both the Variance Inflation Factor (VIF) and the Condition Index (CI) were computed (23). The Variance Inflation Factor (VIF) values are presented in Table 7.

Table 7. Variance Inflation Factor (VIF)

Variable	VIF
Education Level	1.009
Years of Experience	1.005
Gender	1.012
Age	1.007

All VIF values are close to 1, which is well below the commonly accepted threshold of 5 (or 10) (21, 23). This indicates that there is no significant multicollinearity among the explanatory variables. Therefore, the estimated coefficients are stable and not adversely affected by linear dependence among predictors. The Condition Index (CI) values are reported in Table 8.

Table 8. Condition Index Values

Index	Value
CI_1	1.00
CI_2	4.45
CI_3	48.02
CI_4	91.93

The Condition Index provides a more sensitive measure of multicollinearity. Values below 10 indicate weak collinearity, values between 10 and 30 suggest moderate collinearity, and values above 30 indicate strong multicollinearity. In this case, the third and fourth indices (48.02 and 91.93) exceed the threshold of 30, indicating the presence of moderate to strong multicollinearity in certain linear combinations of predictors. However, since the VIF values are very low, this suggests that the multicollinearity is not arising from simple pairwise correlations but rather from more complex linear dependencies. Overall, while multicollinearity does not pose a major concern, the presence of heteroscedasticity and influential observations (as identified earlier) indicates that the OLS model may not fully satisfy classical assumptions. These findings further justify the use of robust and quantile regression techniques in subsequent analysis to obtain more reliable and distribution-sensitive estimates.

Robust Regression Results: To address the influence of outliers and violations of classical OLS assumptions, a robust regression model was estimated using the M-estimation approach (7, 18, 24). The results are presented in Table 9.

Table 9. Robust Regression Results

Variable	Estimate	Std. Error	t-value
Intercept	12,362.06	2,752.85	4.49
Education Level	6,267.92	668.06	9.38
Years of Experience	2,997.07	54.37	55.13
Gender (Male)	-869.97	1,099.73	-0.79
Age	98.44	50.99	1.93

The residual standard error of the robust regression model is 7,587, which is substantially lower than that observed in the OLS model. This indicates a significantly improved fit when the influence of extreme observations is reduced. The robust regression results reveal notable differences compared to the OLS estimates. The effect of education remains positive and highly significant, but the magnitude is considerably smaller (6,267.92) compared to the OLS estimate (26,074.5). This suggests that OLS may have overestimated the returns to education due to the influence of high-income outliers. Similarly, the coefficient for years of experience remains positive and highly significant, with an estimated increase of approximately 2,997 Rs. per additional year of experience. This estimate is very stable and closely aligned with the OLS result, indicating that the effect of experience is robust across different estimation methods. The coefficient for gender (Male) remains negative but is not statistically significant, consistent with the OLS findings. This suggests that gender does not have a strong independent effect on salary within this dataset after controlling for other factors. The effect of age is positive and marginally significant, indicating a slight increase in salary with age. However, the magnitude is relatively small compared to other predictors.

Comparison with OLS Results

A key observation is the substantial reduction in coefficient magnitudes for education and the intercept when moving from OLS to robust regression. This indicates that OLS estimates were influenced by extreme observations in the upper tail of the salary distribution. Additionally, the drastic reduction in residual variability demonstrates that the robust model provides a better representation of the central tendency of the data by down-weighting outliers.

Quantile Regression Analysis: To capture heterogeneity in the effects of socio-economic determinants across the income distribution, quantile regression models were estimated at the 25th, 50th (median), and 75th percentiles (9, 10, 25, 26). The results are summarized in Table 10.

Table 10. Quantile Regression Estimates at Different Income Levels

Variable	Q(0.25)	Q(0.50)	Q(0.75)
Intercept	8750.92	12105.00	9262.78
Education Level	6088.70	6325.25	8162.76
Years of Experience	3046.95	2993.50	2955.43
Gender (Male)	-448.21	-742.10	-2518.36
Age	53.41	70.05	279.93

The quantile regression results reveal substantial heterogeneity in the effects of explanatory variables across different points of the income distribution.

Education: The impact of education increases significantly as we move from lower to higher income quantiles. At the 25th percentile, the return to education is approximately 6,088, which increases slightly at the median (6,325) and rises sharply to 8,163 at the 75th percentile. This indicates that higher levels of education yield disproportionately larger benefits for individuals in the upper income group (3, 5, 27, 28).

This finding highlights the presence of *increasing returns to education*, suggesting that education plays a more critical role in determining high-income outcomes compared to lower-income levels.

Years of Experience: The effect of experience remains positive and remarkably stable across all quantiles, with estimates around 3,000. This suggests that work experience contributes consistently to income regardless of the income level, indicating a uniform return to experience across the distribution.

Gender: The coefficient for gender (Male) is negative across all quantiles, implying that males earn less than females in this dataset after controlling for other factors. However, the effect becomes more pronounced at higher quantiles, with a larger negative coefficient at the 75th percentile. Despite this pattern, the confidence intervals include zero, indicating that the gender effect is not statistically significant across quantiles (29).

Age: The effect of age is relatively small at lower and median quantiles but increases substantially at the 75th percentile. This suggests that age plays a more important role in determining higher incomes, possibly reflecting accumulated experience or seniority effects in higher-paying positions.

Evidence of Distributional Heterogeneity

The variation in coefficient magnitudes across quantiles clearly demonstrates that the relationship between predictors and income is not uniform. In particular:

- Education exhibits increasing returns at higher income levels.
- Experience has a stable and consistent effect across all quantiles.
- Age shows a stronger influence in the upper tail of the distribution.
- Gender effects remain weak and statistically insignificant.

These findings confirm that mean-based models such as OLS are insufficient to capture the full complexity of income determination (4, 8, 30), as they mask important variations across the distribution. To visualize the variation in coefficients across quantiles, Figure 8 presents the esti-mated coefficients for each predictor at different quantiles.

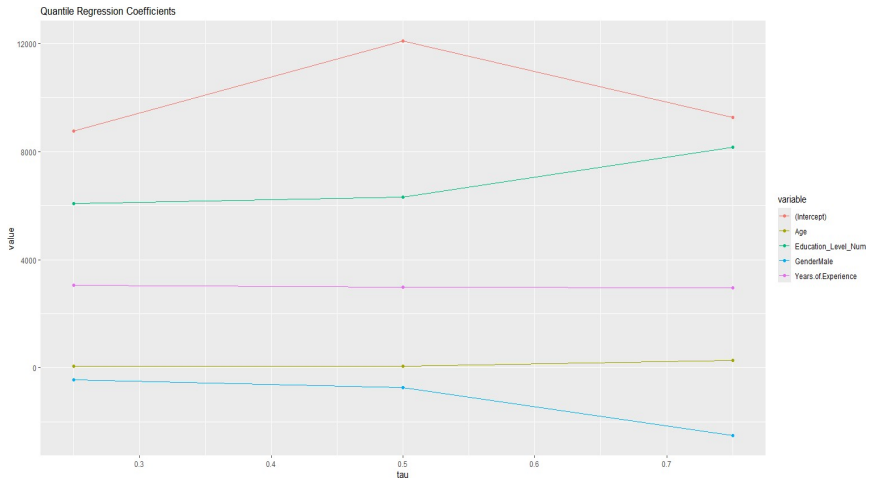

Figure 8. Quantile Regression Coefficients Across Different Quantiles

The graphical representation clearly illustrates the changing impact of predictors across the income distribution. The upward trend in the education coefficient confirms increasing returns, while the relatively flat line for experience indicates stability. The widening variation for age and gender at higher quantiles reflects increased uncertainty and heterogeneity in high-income groups. The quantile regression analysis provides strong evidence of heterogeneous effects of socio-economic determinants across the income distribution. Unlike OLS and robust regression, which focus on central tendencies, quantile regression reveals that the impact of key variables varies significantly between lower and higher income groups. This highlights the importance of adopting distribution-sensitive modeling approaches when analyzing income inequality, as they provide deeper insights into the underlying structural relationships.

Model Validation and Performance Comparison

K-Fold Cross-Validation: To assess the stability and predictive performance of the model, 5-fold cross-validation was performed on the training dataset (13, 31). The results are presented in Table 11.

Table 11. Cross-Validation Performance (5-Fold)

Metric	Value
RMSE	173,055.6
R^2	0.0535
MAE	126,412.2

The cross-validation results indicate relatively high prediction errors and a low R^2 value (0.0535), suggesting that the model explains only a small proportion of the variability in salary. This reflects the inherent complexity and variability in income data, where unobserved factors may play a significant role. The consistency of error metrics across folds, however, suggests that the model is stable and does not suffer from severe over fitting (13).

Model Performance Comparison: To evaluate predictive accuracy, multiple models—OLS, Robust Regression, and Quantile Regression (at 25%, 50%, and 75% quantiles)—were compared using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The results for both training and test datasets are presented in Tables 12 and 13.

Table 12. Training Data Performance

Model	MSE	RMSE	MAE
OLS	28,824,936,088	169,779.1	123,770.55
Robust	35,230,063,279	187,696.7	81,782.29
QR (0.25)	36,004,201,685	189,747.7	82,564.90
QR (0.50)	35,439,234,641	188,253.1	81,650.09
QR (0.75)	34,144,016,565	184,781.0	84,339.09

Table 13. Test Data Performance

Model	MSE	RMSE	MAE
OLS	29,436,968,833	171,572.1	128,682.55
Robust	34,511,102,458	185,771.6	81,537.78
QR (0.25)	35,232,168,637	187,702.3	82,644.77
QR (0.50)	34,720,123,010	186,333.4	81,602.12
QR (0.75)	33,516,692,561	183,075.6	84,282.87

The comparative analysis reveals important differences in model performance:

- OLS Model: The OLS model achieves the lowest RMSE in both training and test datasets (2), indicating the best overall predictive accuracy in terms of squared error minimization. However, it exhibits the highest MAE, suggesting sensitivity to extreme values.
- Robust Regression: The robust model significantly reduces MAE compared to OLS, indicating improved performance in terms of typical prediction error (7). This confirms its effectiveness in handling outliers, although its RMSE is higher due to reduced emphasis on extreme values.
- Quantile Regression: The quantile models provide intermediate performance. The median regression (Q50) shows balanced performance, while Q75 performs slightly better in RMSE than other quantiles. These models are particularly useful for capturing distributional heterogeneity rather than minimizing overall prediction error (8, 10).

The performance metrics between training and test datasets are relatively close across all models, indicating good generalization and absence of severe overfitting. The small differences in RMSE suggest that the models maintain predictive stability when applied to unseen data. Based on the lowest RMSE on the test dataset, the OLS model emerges as the best-performing model in terms of overall predictive accuracy. Although the OLS model provides the best predictive performance according to RMSE, it fails to account for heteroscedasticity, outliers, and distributional heterogeneity. In contrast, robust regression and quantile regression offer more reliable and interpretable insights into the underlying data structure. Therefore, while OLS may be preferred for prediction, quantile regression is more appropriate for understanding income inequality and heterogeneous effects across different segments of the population.

Actual vs Predicted Comparison: To visually assess the predictive performance of different models, scatter plots of ac-tual versus predicted salary values were generated for the test dataset. The results are presented in Figure 9.

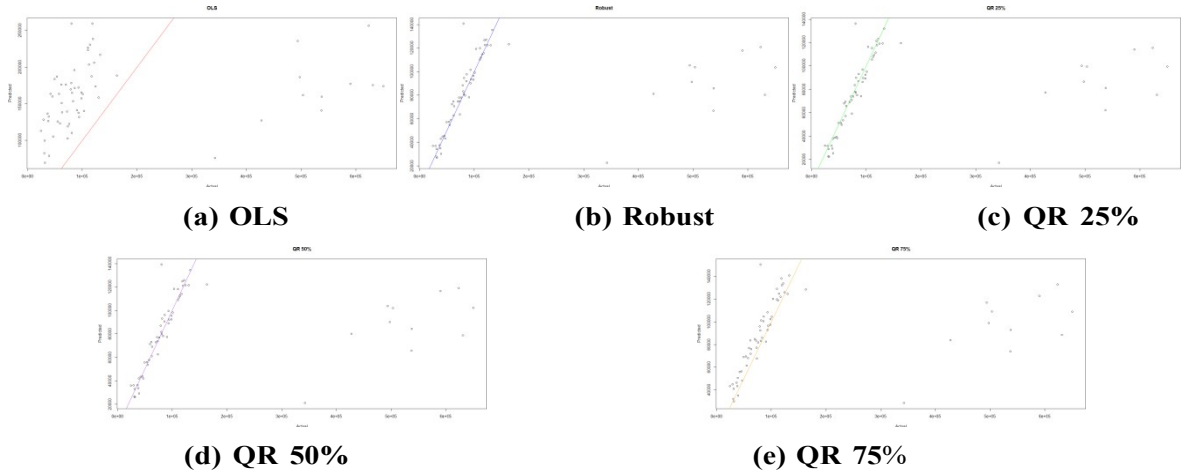

Figure 9. Actual vs Predicted Salary for Different Models

The scatter plots compare the predicted salary values against the actual observed values for each model. The red diagonal line represents the ideal scenario where predicted values perfectly match the actual values.

OLS Model: The OLS model shows predictions that are relatively closer to the 45-degree reference line compared to other models. This indicates better overall predictive accuracy in terms of minimizing squared errors. However, noticeable deviations at higher salary levels suggest that OLS struggles to accurately capture extreme values, likely due to the influence of outliers and heteroscedasticity.

Robust Regression: The robust regression model exhibits tighter clustering of pre-dictions around the central region, indicating improved performance for typical observa-tions. However, predictions tend to deviate more from the reference line for higher salary values, reflecting the model's reduced sensitivity to extreme observations.

Quantile Regression (25%, 50%, 75%): The quantile regression models display distinct patterns across different parts of the distribution:

- The 25th quantile model tends to underestimate higher salary values, as it focuses on the lower tail of the distribution.
- The 50th quantile (median) model provides balanced predictions around the central region but does not fully capture extreme values.
- The 75th quantile model performs relatively better for higher salary values, show-ing improved alignment with the upper tail of the distribution.

DISCUSSION

This study provides a comprehensive analysis of the determinants of income by integrating classical regression techniques with robust and distribution-sensitive methods. The findings highlight several important methodological and empirical insights. First, the Ordinary Least Squares (OLS) results indicate that education and work experience are statistically significant determinants of income, while gender and age exhibit weaker effects. However, diagnostic analysis reveals that key OLS assumptions are not fully satisfied. The presence of influential observations, mild heteroscedasticity, and non-normal residuals suggests that OLS estimates may be sensitive to extreme values and may not adequately capture the underlying data structure (6, 19, 22). Second, the robust regression results demonstrate substantial improvements in coefficient stability and residual variability. In particular, the magnitude of the education effect is considerably reduced compared to OLS, indicating that the latter overestimates returns due to the influence of high-income outliers. The consistent effect of experience across both models suggests that it is a stable and reliable predictor of income. These findings confirm the importance of accounting for outliers in empirical income analysis. Third, the quantile regression analysis provides the most significant contribution of the study by revealing heterogeneous effects across the income distribution. The results show that returns to education increase substantially at higher income quantiles, indicating that education plays a more prominent role in determining high-income outcomes. In contrast, the effect of work experience remains relatively constant across quantiles, suggesting uniform returns. The impact of age becomes more pronounced in the upper tail, while gender effects remain statistically weak throughout. These findings demonstrate that mean-based models such as OLS are insufficient for capturing the complexity of income dynamics. By focusing only on average effects, they obscure important variations across different segments of the population. Quantile regression, on the other hand, provides a more nuanced understanding of inequality by uncovering distribution-specific relationships (4, 10, 30). Fourth, the model validation results indicate that all models exhibit reasonable generalization performance, with relatively small differences between training and test errors. The OLS model achieves the lowest RMSE on the test dataset, indicating superior predictive accuracy in terms of squared error minimization.

However, robust regression achieves lower MAE, reflecting better performance in capturing typical observations. Quantile regression models offer intermediate predictive performance but provide valuable insights into different parts of the income distribution. Overall, the comparative analysis highlights a fundamental trade-off between predictive accuracy and interpretability (32). While OLS performs well in terms of aggregate prediction, robust and quantile regression models provide more reliable and informative insights into the structure of the data.

CONCLUSION

This study develops a comprehensive framework for analyzing income determination by combining data preprocessing, diagnostic testing, and multiple regression approaches. The results demonstrate that income is influenced by key socio-economic factors such as education and work experience, but their effects vary significantly across the income distribution. The findings emphasize that reliance on a single modeling approach, particularly OLS, may lead to incomplete or potentially misleading conclusions. While OLS provides strong predictive performance, it is sensitive to outliers and fails to capture distributional het-erogeneity. Robust regression offers improved stability in the presence of extreme observations, while quantile regression reveals important variations in the effects of predictors across different income levels. From a methodological perspective, the study underscores the importance of integrating multiple analytical techniques to obtain a more accurate and comprehensive understanding of economic relationships. From a substantive perspective, the results suggest that policies aimed at improving educational attainment may have a greater impact on higher-income groups, thereby potentially widening income inequality if not complemented by inclusive measures. Future research may extend this framework by incorporating additional variables such as industry, occupation, and regional factors, as well as exploring nonlinear and dynamic modeling approaches. Moreover, the use of larger and more diverse datasets could further enhance the robustness and generalizability of the findings. In conclusion, this study highlights the necessity of moving beyond traditional mean-based models toward more flexible and robust approaches that capture the full distributional characteristics of income. Such an approach is essential for developing more effective and equitable economic policies.

Author Contributions: All authors contributed equally to the conception, design, analysis, interpretation of data, manuscript preparation, and final approval of the submitted version of the manuscript.

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

1. Joshua Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics. *Princeton University Press*, 2009.
2. Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2012.
3. Krishna Muniyoor et al. Wage differentials and inequality: A quantile regression analysis. *Vision: The Journal of Business Perspective*, 2024.
4. Vytaras Brazauskas, Francesca Greselin, and Ricardas Zitikis. Measuring income inequality via percentile relativities. *Quality & Quantity*, 58:4859–4896, 2024.
5. Ishmael Amartey. Public education spending and income inequality. *Working Paper*, 2026.
6. Halbert White. A heteroskedasticity-consistent covariance matrix estimator. *Econo-metrica*, 48(4):817–838, 1980.
7. Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, 2009.
8. Alessio Farcomeni and Marco Geraci. Quantile ratio regression. *Statistics and Com-puting*, 34(94), 2024.
9. Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.
10. Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.
11. Zhongsheng He and Jing Li. Income inequality and environmental pollution: Quan-tile regression approach. *Heliyon*, 10:e33397, 2024.
12. Olumide O. Olaoye and Mamdouh A. Al-Faryan. Fiscal policy and income inequality: A quantile regression analysis. *International Journal of Social Economics*, 2024.
13. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduc-tion to Statistical Learning: With Applications in R*. Springer, 2 edition, 2021.
14. Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 3 edition, 2019.
15. Alan Agresti. *Analysis of Ordinal Categorical Data*. Wiley, 2 edition, 2010.
16. Frank A. Cowell. *Measuring Inequality*. Oxford University Press, 3 edition, 2011.
17. Anthony B. Atkinson. *Inequality: What Can Be Done?* Harvard University Press, 2015.
18. Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. Wiley, 1987.
19. R. Dennis Cook. Detection of influential observation in linear regression. *Techno-metrics*, 19(1):15–18, 1977.
20. Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2 edition, 2010.
21. John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. SAGE, 3 edition, 2019.
22. Trevor S. Breusch and Adrian R. Pagan. A simple test for heteroscedasticity. *Econo-metrica*, 47(5):1287–1294, 1979.
23. David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identi-fying Influential Data and Sources of Collinearity*. Wiley, 1980.
24. Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel.
25. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, 1986.

26. Moshe Buchinsky. Recent advances in quantile regression models: A practical guide-line for empirical research. *Journal of Human Resources*, 33(1):88–126, 1998.
27. Roger Koenker. Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176, 2017.
28. Jacob Mincer. *Schooling, Experience, and Earnings*. National Bureau of Economic Research, 1974.
29. David Card. The causal effect of education on earnings. *Handbook of Labor Eco-nomics*, 3:1801–1863, 1999.
30. Francine D. Blau and Lawrence M. Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3):789–865, 2017.
31. Xiaobo Shen and Pingsheng Dai. A regression method for estimating gini index by decile. *Humanities and Social Sciences Communications*, 11:1235, 2024.
32. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2 edition, 2009.
33. Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

****