What factors explain wages?

Group 4

Imane el Bouzrouti, Máté Horgász, Simona Pavlovičová, Duy Phan

(2766193), (2813063), (2817540), (2802376)

January 2025

1 Abstract

This study examines whether parental education is a primary determinant of children's wages or if other factors, such as individual education, work experience and cognitive ability have a greater influence. We also compare three predictive approaches, linear regression, the K-nearest neighbors method, and regression trees to determine which better predicts wage outcomes. Furthermore, we conduct a simulation study investigating the effect of non-normal residuals on model performance. Through a detailed analysis, we highlight the relationship between parent's education, other key variables, and overall wage determination.

Keywords: wage determination, parents' education, linear regression, K-nearest neighbor regression, regression trees, simulation study

2 Introduction

In many labor economics studies, people investigate the factors that determine wages. While in most of them, a person's own education and work experience play a significant role (Heckman et al., 2006; Becker, 1975), other factors such as cognitive ability or family background can also affect earnings. In fact, since parents generally care about their children's well-being, they invest in their education. Because we expect a positive correlation between education and wage, we assume that higher-income or better-educated parents are more likely to invest heavily in their children's education (Borjas, 2020). Consequently, we expect a positive correlation between parents' and children's education, which can ultimately affect children's wages.

Many research articles also investigate this relationship. Early studies by Becker (1975) and Mincer (1974) emphasize that parents who remain in school longer can create a better learning environment for their children, leading to greater educational opportunities and potentially higher earnings later. More analyses by Borjas (2020) and Card (1999) also highlight that well-educated parents invest more in their children, supporting them to achieve more schooling and, in turn, better wages. These studies suggest that parents' education can affect a child's future income both directly through the parents' level of education and indirectly by encouraging children to pursue additional years of education themselves. Overall, the interaction between a family's educational background and an individual's education is vital to understanding wage explanation.

In our analysis, we aim to examine whether this relationship between parents' education and a child's wage persists once other factors are considered. Although parental education may significantly shape our future earnings, we hope that it is not the only determinant. Therefore, we hypothesize that parental education could have a notable impact, but we include other predictors in the model to assess the potential influence of other variables from the data set. Our first research question is "Are there variables that are more influential in explaining wages or are our future earnings only dependent on our parents' education?".

Moreover, we compare the performance of multiple predictive methods. Hence, our second research question is "Does the K-nearest neighbors approach predict the wages more accurately than linear regression or regression trees?". Given the previous assumptions about the positive correlation between wages and education, which might be further extended through parental education, we suspect the relationship between the wage and the parents' education to be linear. Therefore, we hypothesized that linear regression would predict wages more accurately than the K-nearest neighbors method.

To answer these research questions, we build a thorough analysis of modeling and testing using different approaches.

3 Data

3.1 Data set

Our dataset consists of 935 observations, capturing monthly wage information along with 17 educational, demographic, and family background variables (summarized in Table 1). In our analysis, we aim to investigate whether the parents' education directly influences wages and if this is the only factor that explains them. Therefore, we focus particularly on the mother's and father's years of education, *meduc* and *feduc*. We combine these measures into a single parental education variable, *peduc*. However, since the data set does not directly contain this variable, we have to make an assumption about the equal linear effect on the wage of both parents.

Table 1: List of variables and their definitions.

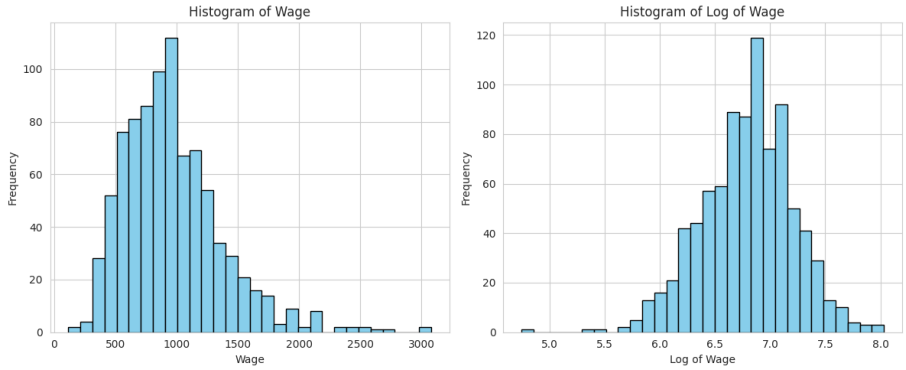
Variable	Description
wage	Monthly earnings
hours	Average weekly hours
IQ	IQ score
KWW	Knowledge of World of Work score
educ	Years of education
exper	Years of work experience
tenure	Years with current employer
age	Age in years
married	=1 if married
black	=1 if Black
south	=1 if living in the South
urban	=1 if living in an SMSA (urban)
sibs	Number of siblings
brthord	Birth order
meduc	Mother's education
feduc	Father's education
lwage	Natural log of wage

Before starting to build a model, we clean the data set and analyze its descriptive statistics and plots, which allows us to identify any necessary data transformations.

3.2 Descriptive statistics

We first examine the data using test statistics, as summarized in Table 2. In particular, the *wage* ranges from 115 to 3,078, with a mean of approximately 977.10 and a median of around 923, indicating a pronounced right skew. In contrast, the natural logarithm of wages (*lwage*) spans from 4.75 to 8.03, with a mean and median of 6.80 and 6.83, respectively. Figure 1 further illustrates that *lwage* exhibits less skewness and more symmetry than the *wage*. Moreover, it resembles the normal distribution more closely, and hence we chose to use *lwage* instead of *wage* for our study. This is a common practice in labor economics. Researchers often log-transform wages to better satisfy the assumptions of linear models, namely homoskedasticity and normality of residuals, which leads to more reliable inference (Wooldridge, 2019).

Figure 1: Histogram of *wage* and *lwage*



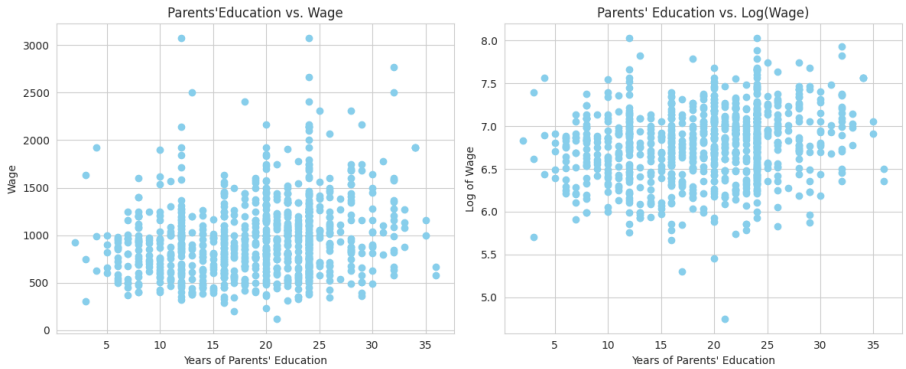
3.3 Outliers

As indicated by the descriptive statistics in Table 2 and the scatter plots in Figure 2, a small number of observations have wages exceeding 3,000 (with a maximum of 3,078), which is above the mean and hence they are defined as outliers. Since these values represent the high earners rather than data-entry errors, we opt to keep them. To mitigate their impact, we apply the log-transformation to wages, as previously noted (Mincer, 1974). We can see in Figure 2 that we no longer exhibit these outliers associated with high-earners after this transformation.

Table 2: Descriptive statistics for selected variables.

	wage	**lwage**	**IQ**	**educ**	**exper**	**peduc**
count	722.00	722.00	722.00	722.00	722.00	722.00
mean	977.10	6.80	102.12	13.66	11.33	21.06
min	115.00	4.75	54.00	9.00	1.00	2.00
median	923.00	6.83	103.00	13.00	11.00	22.00
max	3078.00	8.03	145.00	18.00	22.00	36.00

Figure 2: Scatter plots of *peduc* vs *wage* and *lwage*



3.4 Missing values

We identify a total of 213 missing values in the variable *peduc*. To proceed with our analysis, we decide to drop these values. However, it is important to note that the missingness might be correlated with other variables (i.e *black*), which could potentially cause biased results (Bennett, 2001). We do not take this into account for our analysis.

4 Methods

To begin with, we divide the data into a training and testing set. Moreover, we use a parametric method to determine the relationship between *wage* and *peduc*. Finally, we compare the parametric method with non-parametric methods in terms of predicting wages.

4.1 Training and Testing set

We randomly split the data into two sets, training and testing set, using a 80%-20% ratio. This means that the training set contains 80% of the observations, while the test set holds the remaining 20%. We fit the models on the training set, and then apply that fitted model to predict the outcome for observations in the test set. Splitting the data in this way helps prevent overfitting and ensures that the performance of the model can be evaluated on unseen data. By keeping the test set separate from the modeling process, we obtain an unbiased assessment of how well the final model generalizes to new observations (James et al., 2013).

4.2 Selecting variables

To determine which predictors to include in our model, we create a correlation matrix and conduct a preliminary hypothesis testing. This approach helps us to understand the linear relationships between the variables. After observing low correlations and high p-values for certain attributes (*brthord*, *south*, *sibs*, *KWW*, *black*), we decide to drop them. In addition, we group the remaining variables into four categories, which helps us to create a model that is statistically significant and interpretable.

(a) Main focus variables: *lwage*, *peduc*

(b) Educational and cognitive variables: *educ*, *IQ*

(c) Variables related to job experience: *exper*

(d) Personal variables: *married*, *urban*

4.3 Inference

For an accurate prediction, it is crucial to build a reliable regression model and focus on inference. This step involves understanding relationships, testing hypotheses about these relationships, and interpreting the coefficients.

Linear regression is a standard technique for analyzing how explanatory variables influence an outcome, which is usually defined as a 'true' model. By specifying the model 1 and fitting it to our sample data, we aim to approximate this 'true' model as closely as possible (Wooldridge, 2019).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad j = 0, 1, 2, .., p \tag{1}$$

where, y_i is the dependent variable, β_0 is the intercept, x_{ij} is the j-th regressor, β_j is the slope coefficient of the j-th regressor, ε_i is the error term.

We estimate this model using the Ordinary Least Squares (OLS) and we get a new model, see Equation 2. This estimation method minimizes the sum of squared residuals (SSR) and provides the Best Linear Unbiased Estimates (BLUE) when the following key assumptions are met (Stock and Watson, 2015):

1. The model is linear in its parameters.
2. The data is drawn from a random sample.
3. No multicollinearity.
4. Homoskedasticity.
5. Normality of the residuals.

We begin with a simple linear regression model that has only one regressor. In our study, we link (*lwage*) to (*peduc*), which takes the form of model 2 with *p = 1*.

To further extend this model, we incorporate additional variables. At each step, we perform the t-test or the partial F-test to see whether the additional regressors improve the fit. If these tests are rejected, we add the variables to our final model. Note that we use 5% significance level for the testing. Furthermore, we inspect the residual plots to diagnose heteroskedasticity or non-linearity, and finally we monitor the adjusted R-squared ($\bar{R}^2$), BIC and AIC, which account for unnecessary complexity (James et al., 2013). We add the variables to our final model when $\bar{R}^2$ increases and both BIC and AIC decrease.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}, \qquad j = 0, 1, 2, .., p \tag{2}$$

where $\hat{y}_i$ is the estimated dependent variable (estimated *lwage*), $\hat{\beta}_0$ is the estimated intercept, x_{ij} is the j-th regressor (i.e *peduc*), $\hat{\beta}_j$ is the estimated slope coefficient of the j-th regressor.

4.4 Assumptions

4.4.1 Linearity

Firstly, we check the linearity assumption using the Ramsey Reset test, which adds polynomial terms of the fitted values (i.e. $\hat{y}^2$) to the model 2 (Volkova and Pankina, 2013). Under the null hypotheses, these additional terms have no effect, suggesting that the original linear specification is correct. This test computes the F-statistics with the corresponding p-value, hence a small p-value implies a violation of this assumption. In addition, we plot the residuals against each explanatory variable to look for any nonlinear patterns.

4.4.2 Random sample

Since the data set is given to us and we do not have more information about the sample, we assume that this assumption holds.

4.4.3 Multicollinearity

We examine multicollinearity by computing the Variance Inflation Factor (VIF) for each explanatory variable. It measures how much the variance of a coefficient is magnified due to linear relationships among predictors. A common guideline suggests that a VIF exceeding 5 or 10 implies high chance of multicollinearity. Meaning that the explanatory variables significantly overlap, making it challenging to determine their individual impact. When any parameter has a high VIF, we consider dropping them, combining them with highly correlated variables, or transforming them to reduce this overlap (Tay, 2017).

4.4.4 Homoskedasticity

One of the assumptions is also homoskedasticity, meaning that the residuals must have equal variance, otherwise, we have heteroskedasticity and the assumption is violated. To check this assumption, we examine the Breusch-Pagan test. This test runs an auxiliary regression where the squared residuals from our main model are regressed on a set of predictor variables. Under the null hypotheses, these predictors do not explain any change in the squared residuals, implying homoskedasticity. Hence, a small p-value indicates that heteroskedasticity is present (Ilori and Tanimowo, 2022).

In addition, we plot the residuals against each explanatory variable to examine if these plots show any signs of heteroskedasticity. Hence, if the residuals are not spread out uniformly, we suspect non-constant variance, and thus the assumption is violated.

4.4.5 Normal residuals

We inspect the normality of the residuals by plotting a QQ-plot with a histogram and employing the Jarque-Bera test. If the residuals align along the 45-degree line of the QQ-plot, then the residuals follow a normal distribution. However, any systematic deviations from this line suggest a violation of this assumption.

Moreover, the Jarque-Bera test examines whether the skewness and the kurtosis of the residuals are significantly different from the normal distribution (skewness = 0 and kurtosis = 3). It yields a statistics that follows χ^2 distribution under the null hypotheses of normality. A p-value under the 5% significance level implies that the residuals deviate from normality, violating this assumption (Jarque and Bera, 1980).

4.5 Prediction

After specifying our overall model, we apply three methods to predict wages: linear regression, K-nearest neighbors (KNN) regression, and regression trees.

KNN is a non-parametric method, hence it does not rely on any assumption, making it more flexible than linear regression. To predict an observation $X = x$, KNN identifies the K training observations closest to x and takes the average outcome among those neighbours, as shown in formula 3 (Imandoust and Bolandraftar, 2013).

$$\hat{Y}(x) = \frac{1}{K} \sum_{x_i \in N_k(x)} y_i \tag{3}$$

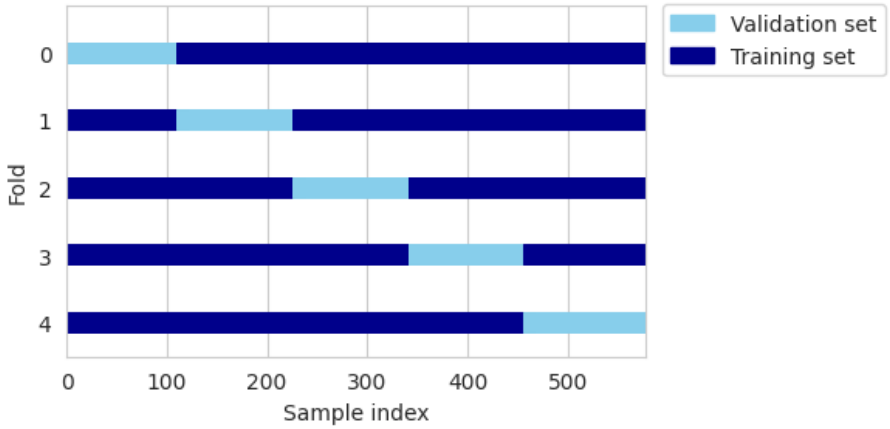
where y_i is the dependent variable, K is the number of nearest neighbors, x_i is the regressor and $N_k(x)$ are the K training observations that are closest to x .

We determine the optimal value for K using 5-fold cross-validation on the training set. Each of the five folds is split into a validation and training subset, for illustration see Figure 3. While the model is trained on the training observations, the validation subset is used to compute the mean squared error (MSE). After repeating this process for each candidate K, we identify the "elbow" point in the average MSE curve across folds and select that K. Furthermore, we keep the test set separated until the final performance check.

Regression trees recursively split the predictor space into distinct regions, using the mean wage in each region (leaf node) as the prediction. This method is also non-parametric and does not require a fixed model form (Breiman et al., 1986). Rather than applying 5-fold cross-validation to find the right tree complexity, we fix the maximal depth directly. This allows us to simply demonstrate how the regression trees work in practice.

Once we choose the settings for each model, we evaluate the final performance. By comparing MSE across linear regression, KNN, and regression trees in the test set, we determine which approach generalizes best to new and unseen observations (James et al., 2013).

Figure 3: Splitting each fold into validation and training subset



Notes: For simplicity, the validation subset is displayed in a single block in each fold. In practice, it is usually randomly distributed through data set.

5 Results

5.1 Parents' education variable

Firstly, we test whether mother's and father's education affect the wage equally by estimating two regressions. To build the two models we use model 1 with $p = 1$ and $p = 2$, respectively:

1. Single regressor model (parents' education):

$$lwage = \beta_0 + \beta_1(meduc + feduc) + \varepsilon = \beta_0 + \beta_1 peduc + \varepsilon$$

2. Model with 2 separate regressors (mother's and father's education):

$$lwage = \beta_0 + \beta_1 meduc + \beta_2 feduc + \varepsilon$$

We then perform a Wald test to assess whether the two variables could be combined (i.e, the null hypothesis is $H_0 : \beta_1 = \beta_2$). The resulting p-value of 0.94 exceeds the significance level, so we fail to reject the null hypothesis of equal coefficients. Consequently, our assumption holds, and we can combine mother's and father's education into a single regressor.

5.2 Modeling and checking assumptions

We start with a single linear regression, model (1), on the training set. Then, we gradually add each category to this model until we end up with our final model, model (4). For each specification, we analyze the p-values and the $\bar{R}^2$, and in addition, we perform the t / F-test, as already stated. Table 7 maps these results.

Table 3: **OLS Regression Results:** Dependent variable = *lwage*

	(1)	**(2)**	**(3)**	**(4)**
constant	6.409	5.774	5.204	4.967
	(0.069)	(0.127)	(0.158)	(0.159)
peduc	0.018	0.008	0.009	0.009
	(0.003)	(0.004)	(0.003)	(0.003)
educ	–	0.025	0.046	0.044
		(0.009)	(0.010)	(0.010)
IQ	–	0.005	0.005	0.005
		(0.001)	(0.001)	(0.001)
exper	–	–	0.025	0.023
			(0.004)	(0.004)
urban	–	–	–	0.185
				(0.036)
married	–	–	–	0.195
				(0.051)
t-stat	5.81	–	5.74	–
F-stat	–	20.25	–	20.57
p-value	0.00	< 0.001	0.00	< 0.001
$\overline{\boldsymbol{R}}^{\mathbf{2}}$	0.054	0.105	0.152	0.205
Notes: Standard errors in parentheses.				

Since model (1) and model (3) introduce only one new variable, we perform a t-test: $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$, where β_i is the coefficient of the newly added variable. As indicated in Table 7, both t-tests return the corresponding p-values smaller than our significance level, so we reject the null hypothesis for each specification. In other words, *peduc* in model (1) and *exper* in model (3) explain a significant part of the variation in wage. Thereby, we add them to our final linear regression model.

Consequently, models (2) and (4) add multiple variables at once, and thus we need to perform the partial F-test: $H_0 : \beta_i = \beta_j = 0$ vs $H_1 : \beta_i \neq \beta_j \neq 0$. β_i and β_j again represent the coefficients of the variables that we test. Both F-tests yield small p-values, indicating that these variables enhance the fit of the model. Furthermore, compared to other models, model (4) has the highest value for $\bar{R}^2$ and also the lowest values for BIC and AIC and for the sum of squared residuals (*BIC = 566.8, AIC = 536.3, SSR = 83.53*). Hence, we continue with model (4) as our final linear regression model, see formula (4).

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} + \hat{\beta}_5 x_{i5} + \hat{\beta}_6 x_{i6}, \qquad j = 1, 2, 3, 4, 5, 6 \qquad (4)$$

where $\hat{y}_i$ is *lwage*, $\hat{\beta}_0$ is the estimated intercept, $\hat{\beta}_j$ is the estimated slope coefficient of the j-th

regressor, x_1 is *peduc*, x_2 is *educ*, x_3 is *IQ*, x_4 is *exper*, x_5 is *urban*, x_6 is *married*.

5.2.1 Linearity

To check linearity, we plot the residuals against each explanatory variable and employ the Ramsey Reset test. The residual plots do not reveal any non-linear patterns. Additionally, the p-value ($p = 0.51$) that is above our significance level does not indicate a strong evidence of model misspecification. Hence, linearity is not violated.

5.2.2 Multicollinearity

To examine the multicollinearity we compute VIF. In table 4, the intercept shows a high VIF, which is very common, and it does not indicate any problematic collinearity. All other variables exhibit low VIF values (with $VIF < 2$), suggesting that multicollinearity is minimal and the model's regression coefficients are likely stable.

Table 4: VIF for Selected Variables

Variable	**const**	**peduc**	**educ**	**IQ**	**exper**	**urban**	**married**
VIF	99.052321	1.352368	1.811255	1.485151	1.276981	1.010946	1.015594

5.2.3 Homoskedasticity

To assess homoskedasticity, we use the Breusch-Pagan test and plot the residuals against each explanatory variable. This test returns relatively large p-value ($p=0.14$) which is larger than significance level. Hence, we fail to reject the null hypothesis of constant variance, indicating no strong evidence of heteroskedasticity. Furthermore, the plots do not reveal any patterns such as cone and funnel shapes that would suggest heteroskedasticity, thus this assumption is satisfied.

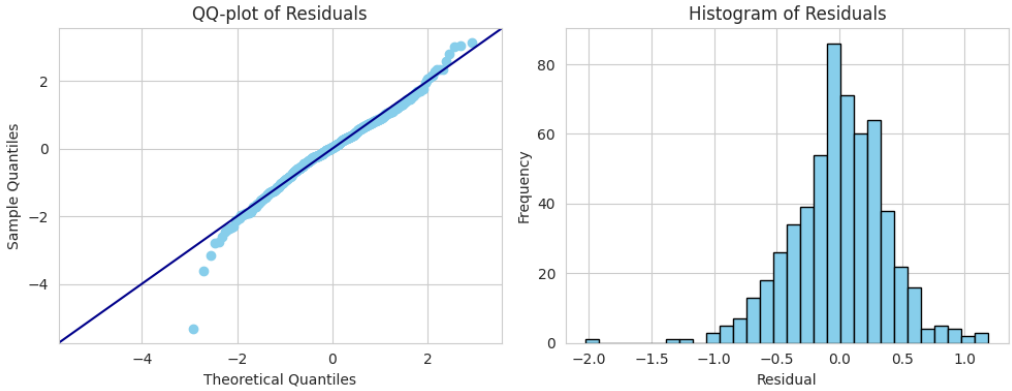
5.2.4 Normal residuals

We test the normality of residuals by conducting the Jarque-Bera test, and by examining both a QQ-plot and a histogram, see Figure 4. Although the QQ-plot shows that a majority of the residuals follow a near-normal distribution, the histogram and the Jarque-Bera test suggest a violation of this assumption. As we can observe in Table 5, the p-value is very low, the kurtosis exceeds 3, and the skewness is slightly negative. The histogram in Figure 4 also highlights this heavy left tails, which strengthens the evidence of non-normal residuals.

Table 5: Jarque–Bera test results

Statistic	**Value**
Jarque–Bera statistic	84.45
p-value	< 0.001
Skewness	-0.40
Kurtosis	4.70

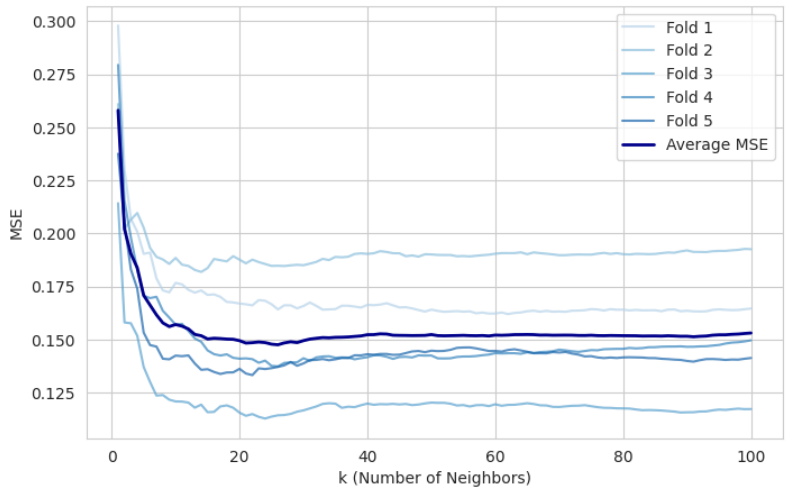
Figure 4: QQ-plot and Histogram of residuals from our final model (4)



5.3 Prediction

Figure 5 shows the average MSE for a range of K values in the KNN model. The curve's "elbow" occurs at $K = 9$, giving a test set MSE of 0.123. Consequently, we chose $K = 9$ as the optimal number of neighbors. This means that to predict a new observation for *lwage*, KNN identifies the 9 closest points in the training set and takes their average as the prediction.

Figure 5: Choosing K for KNN based on 5-fold cross-validation



Moreover, Figure 6 displays the structure of our regression tree. Each node presents the average *lwage* value in that region (leaf node) and the associated squared error.

In conclusion, we compare the test set MSE of linear regression, KNN, and regression tree. As shown in Table 6, the linear regression model returns an MSE of 0.09, whereas the KNN model with $K = 9$ yields 0.123 for MSE. The regression tree, with a maximum depth of three, gives a MSE of 0.117. Overall, linear regression slightly outperforms KNN and the regression tree for our dataset, suggesting a linear relationship between predictors and wages.

Figure 6: Regression tree

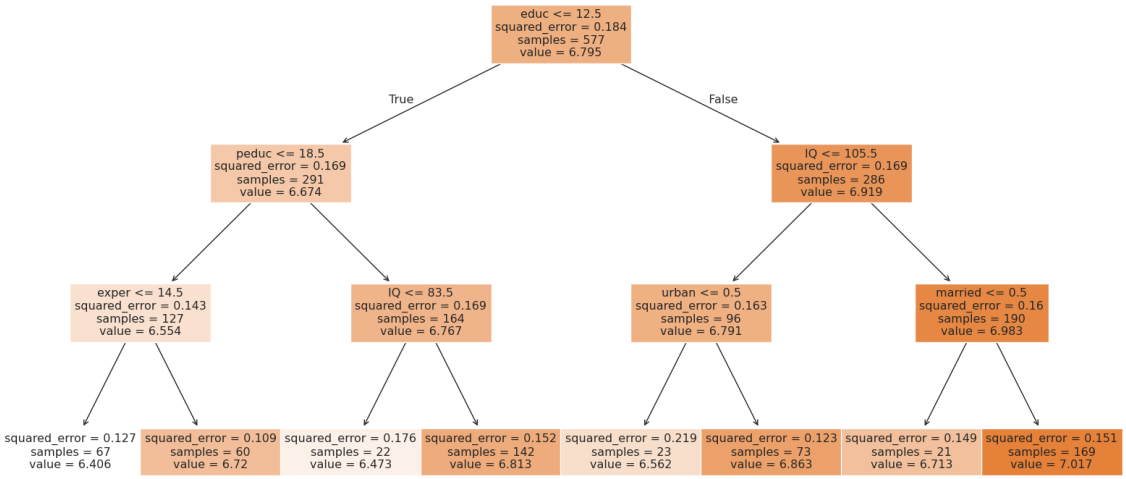


Table 6: MSE for different prediction methods

Model	**Test MSE**
Linear Regression	0.0908
KNN (K = 9)	0.1230
Regression Tree	0.1167

6 Simulation study

In our analysis, we assume that the residuals follow a normal distribution when estimating the linear regression model. However, as previously noted, our final model violates this assumption, as the residuals have heavier tails than the normal distribution. Therefore, we conduct a simulation study to investigate whether the efficiency of the OLS estimates is reduced and whether the predictive outcomes differ under these conditions.

To start with, we design a Monte Carlo simulation (MCS) that generates new data. MCS employs random sampling to estimate potential outcomes of uncertain events by simulating the process multiple times (Ratick and Schwarz, 2009).

We construct a linear regression model, shown in Equation (5), and run 10,000 simulations for each sample size with $n \in \{50, 100, 200, 500\}$. In our setup, we treat the empirically estimated coefficients as the "true" parameters and generate new predictors under a normal distribution with $\mu = 21.06$ and $\sigma = 5.45$. We chose these parameters based on the previous results from our analysis. To incorporate heavy tails of residuals, we use the Student's t-distribution with low degrees of freedom, specifically $\nu = 8$, allowing for heavier tails than a normal distribution.

$$y_i = 6.4091 + 0.0182x_{i1} + \varepsilon_i, \tag{5}$$

where y_i is a dependent variable, x_{i1} is a regressor, ε_i is an error term.

In each simulation, we fit an OLS model to obtain coefficient estimates. Hence, we evaluate whether OLS remains unbiased and consistent. We then take average of these outcomes over

10,000 runs. Furthermore, we fit a KNN model on the same simulated data to compare its predictive performance with that of linear regression under local averaging. This procedure assesses how robust OLS is to heavy-tailed errors and whether, in practice, a nonparametric method might outperform a linear model in finite samples with heavy tails and outliers.

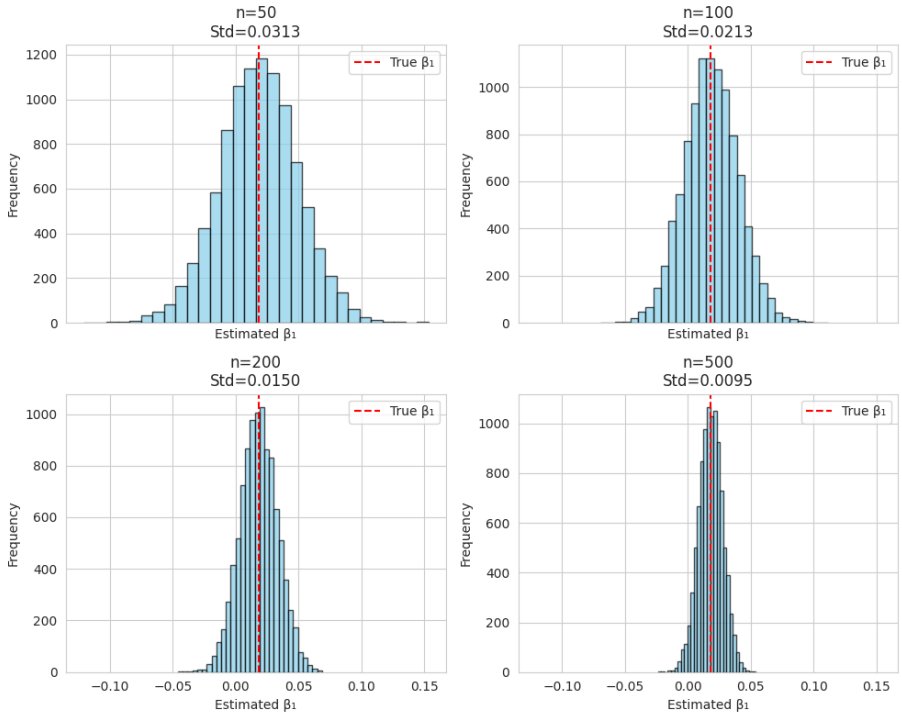
Additionally, we repeat the same process but replace the linear specification with a nonlinear model, see Equation (6), to determine whether the KNN approach yields superior performance compared with a nonlinear regression model under non-normal residuals.

$$y_i = 6.4091 + 0.0182x_{i1} - 0.0005x_{i2}^2 + \varepsilon_i, \qquad j = 1, 2 \tag{6}$$

where y_i is a dependent variable, x_{ij} is a j-th regressor, ε_i is an error term.

As we can observe in Figure 7, our simulation study demonstrates that two fundamental properties of OLS still hold even under heavy-tailed errors. Firstly, the estimated β_1 remains close to its true value. This confirms that OLS is unbiased in expectation despite heavier-than-normal tails of residuals. Secondly, the variance of these estimates shrinks with increasing sample size, confirming the consistency of OLS estimates. To conclude, even under heavy-tailed Student-t distributed residuals, OLS continues to produce unbiased and consistent estimates.

Figure 7: Sampling distribution of β_1 estimates across sample sizes



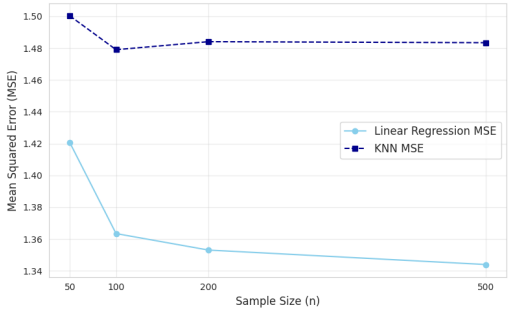
Furthermore, Figure 8 indicates that when the true data-generating process is approximately linear, KNN does not match linear regression's performance at larger sample sizes. Although we test higher values of K to improve KNN, this does not significantly change its overall accuracy. While the two approaches perform closely for smaller sample sizes, linear regression ultimately produces lower MSE as n grows, reflecting the advantage of a correctly specified parametric form.

A similar pattern appears in the non-linear setting, see Figure 9. Although raising K improves KNN's performance for small sample sizes, the non-linear approach still outperforms KNN for large sample sizes. This suggests that simple local averaging struggles to capture deeper patterns in the data. In contrast, the non-linear model, which might match the data-generating process

more closely, performs better at larger sample sizes. To conclude, these results suggest that a well-specified parametric model typically outperforms a non-parametric approach under both linear and nonlinear conditions.

Figure 8: Prediction performance of Linear Regression Model vs KNN Regession

Prediction methods with K = 9



Prediction methods with K = 26

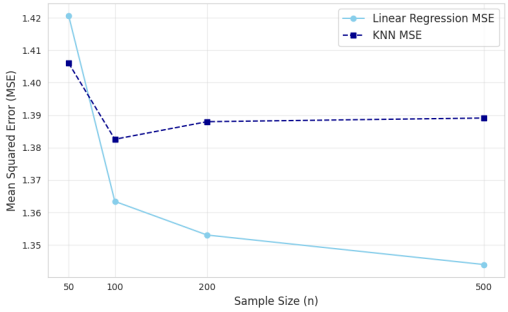
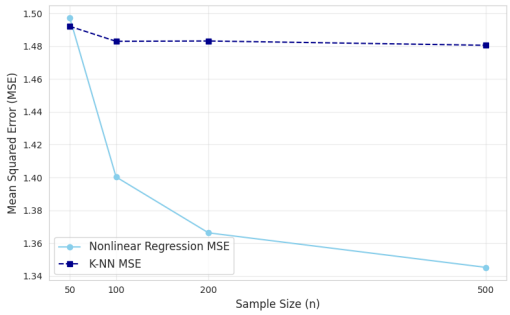
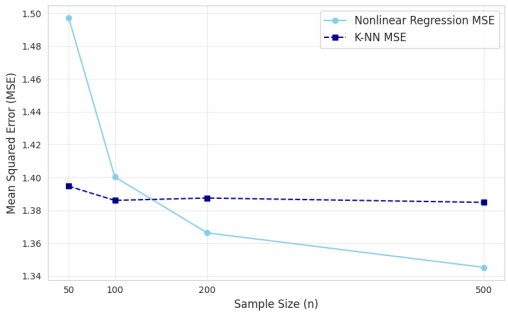


Figure 9: Prediction performance of Nonlinear Regression Model vs KNN Regression

Prediction methods with K = 9



Prediction methods with K = 26



7 Conclusion

In our effort to uncover the relationship between wages and parental education, we find a strong linear relationship. The simple linear regression model that only includes parental education has a low $\bar{R}^2$, and the effect of parental education diminishes after including the individual's education and IQ. This suggests that parental education might have an indirect effect on wages. Despite this, parental education remains significant. Specifically, a one-year increase in parental education is associated with approximately a 1.8% increase in an individual's wage.

The effects of IQ, work experience, marital status, and urban residency further improve our model. The final linear regression model, which includes all the mentioned variables, outperforms all other specifications. As for the prediction methods, we also consider KNN and regression trees. After comparing the different methods, linear regression proves to be superior on the basis of MSE as a measure of accuracy. This finding indicates that a parametric approach is more effective for this study.

Moreover, our simulation study results confirm that, despite heavy-tailed errors from the Student-t distribution, key properties including unbiasedness and consistency of OLS estimates still hold. Furthermore, when comparing KNN to linear or nonlinear regression, KNN does not achieve lower MSE, even if we adjust the number of neighbors, and especially at larger sample sizes. This outcome highlights the advantage of using a well-specified parametric model to more effectively capture the data-generating process.

8 Discussion

This study investigates whether parents' education substantially influences wages. First, we examine whether wages are mainly dependent on parents' education or if additional variables better explain earnings. Our results suggest that, while *peduc* plays a significant role, its effect shrinks once we add additional factors. Thus, wages are not solely determined by parental education. This confirms our initial hypothesis and supports previous research (Card, 1999).

Throughout our study, we face several issues that are not addressed and require further analysis. Firstly, we delete the missing observations, which might bias the results if the missingness is associated with specific subgroups. Future research should investigate alternative strategies to mitigate potential bias in handling missing data.

Secondly, we observe non-normal residuals in the linear regression, potentially affecting parameter estimates and hypothesis testing.

Finally, the chosen predictors explain only about 20.5% of the variation in *lwage*, indicating that other variables contribute to wage determination. We also do not distinguish genetic factors from parental resources that might shape children's earnings. This could affect our overall analysis. Hence, future work could incorporate such distinctions to achieve a more complete understanding of intergenerational wage determinants.

Given these results, researchers should continue to analyze wage models, as a better understanding of these relationships can help improve policies. For instance, knowing that family background strongly influences a child's future earnings could motivate targeted educational investments to help reduce wage gaps and improve economic growth.

A Appendix

In addition to the four linear regression models that we already have, we also consider a fifth model, see Table 7. In this model the variables *IQ* and *KWW* are combined into one variable. To test if we can can combine them, we perform a Wald test (i.e, the null hypothesis is $H_0 : \beta_1 = \beta_2$). Since the p-value exceeds 0.05, we fail to reject the null hypothesis, and thus we can combine *IQ* and *KWW* into a single variable named *cognitive*. Before merging, we first standardize these two variables as they do not have the same measures. We then add *tenure* and *hours* to the model. Because these variables are introduced jointly, we perform the partial F-test, which yields a p-value below 0.05, so we reject the null hypothesis. This means that we should include these variables in the model.

When comparing the fourth and fifth model, we see that the fifth model has a greater $\bar{R}^2$, and lower AIC and BIC. However, this model also violates the assumption of homoskedasticity, while the fourth model does not. Furthermore, the fifth model includes eight variables, compared to only six in the fourth model. Since the difference between the performance of the two models is relatively small, we decide not to include the fifth model in our analysis, instead we select the fourth model as our final model.

Table 7: **OLS Regression Results:** Dependent variable = *lwage*

	(4)	**(5)**
constant	4.967	5.833
	(0.159)	(0.198)
peduc	0.009	0.009
	(0.003)	(0.003)
educ	0.044	0.034
	(0.010)	(0.010)
IQ	0.005	–
	(0.001)	–
exper	0.023	0.016
	(0.004)	(0.004)
urban	0.185	0.171
	(0.036)	(0.035)
married	0.195	0.197
	(0.051)	(0.051)
cognitive	–	0.080
	–	(0.003)
tenure	–	0.010
	–	(0.003)
hours	–	-0.007
	–	(0.002)
F-stat	20.57	9.910
p-value	< 0.001	< 0.001
$\overline{R}^2$	0.205	0.229
AIC	536.3	520.5
BIC	566.8	559.8
Notes: Standard errors in parentheses.		

References

Becker, Gary S. (1975). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. 2nd. University of Chicago Press.

Bennett, Derrick A. (2001). "How can I deal with missing data in my study?" In: *Australian and New Zealand Journal of Public Health* 25.5, pp. 464–469.

Borjas, George J. (2020). *Labor Economics*. 8th. McGraw-Hill Education.

Breiman, Leo, Jerome Friedman, Richard A. Olshen, and Charles J. Stone (1986). *Classification and Regression Trees*. Wadsworth Brooks/Cole.

Card, David (1999). "The Causal Effect of Education on Earnings". In: *Handbook of Labor Economics, Volume 3A*. Elsevier.

Heckman, James J., Lance Lochner, and Petra E. Todd (2006). "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond". In: *Handbook of the Economics of Education*. Elsevier.

Ilori, Omotayo Oluwatosin and Fatai Olalekan Tanimowo (2022). "Heteroscedasticity Detection in Cross-Sectional Diabetes Pedigree Function: A Comparison of Breusch-Pagan-Godfrey, Harvey and Glejser Tests". In: *International Journal of Scientific and Management Research* 5.12, pp. 150–163.

Imandoust, Sadegh Bafandeh and Mohammad Bolandraftar (2013). "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background". In: *International Journal of Engineering Research and Applications* 3.5, pp. 605–610.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer.

Jarque, Carlos M. and Anil K. Bera (1980). "Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals". In: *Economics Letters* 6.3, pp. 255–259.

Mincer, Jacob (1974). *Schooling, Experience, and Earnings*. National Bureau of Economic Research; distributed by Columbia University Press.

Ratick, S. and G. Schwarz (2009). "Monte Carlo Simulation". In: *International Encyclopedia of Human Geography*. Elsevier, pp. 175–184.

Stock, James H. and Mark W. Watson (2015). *Introduction to Econometrics*. 3rd.

Tay, Richard (2017). "Correlation, Variance Inflation and Multicollinearity in Regression Model". In: *Journal of the Eastern Asia Society for Transportation Studies* 12, pp. 1752–1761.

Volkova, V. and V. Pankina (2013). "The research of distribution of the ramsey reset-test statistic". In: *Applied Econometrics* 32, pp. 44–55.

Wooldridge, Jeffrey M. (2019). *Introductory Econometrics: A Modern Approach*. 7th. Cengage Learning.