1. Exploratory Analysis

1.1 Data Cleaning

The dataset comprises 916 observations, which are 916 individuals in the United States, and 17 variables. Upon examining the initial 65 rows, missing values (denoted as "NaN") were identified in three columns: ‘brthord’, ‘meduc’, and ‘feduc’. To address this issue, observations containing NaN values are eliminated using the data.dropna() method.

Following data cleaning, the number of observations decreases significantly from 916 to 655. Despite this reduction, the distribution of monthly wages remains largely unchanged. Histograms and boxplots both indicate that the monthly wage distribution before and after cleaning is right-skewed, with values ranging from 115 to 3078. The mean and median of the monthly wage distribution also showed minimal changes. Prior to cleaning, the mean was 959.97, and the median was 907.5, while post-cleaning, the mean increased slightly to 988 and the median to 935. Therefore, this data cleaning process will not lead to big problem in data analysis in later sections.

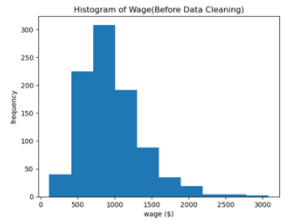


Figure 1: Histogram of Wage (Before Cleaning)

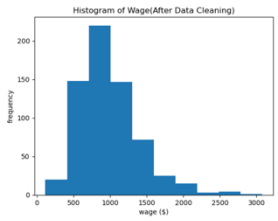


Figure 2: Histogram of Wage (After Cleaning)

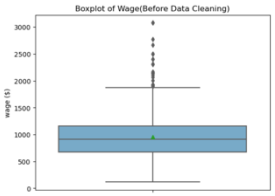


Figure 3: Boxplot of Wage (Before Cleaning)

Figure 4: Boxplot of Wage (After Cleaning)

1.2 Splitting the Data

With the data after cleaning, the data is splitted into two sets: a training set and a test set. The data will be randomly allocated, with 80% going to the training set and the remaining 20% to the test set. Consequently, the training set consists of 524 observations, while the test set comprises 131 observations.

The training set will be utilised to develop and identify the best model for predicting the monthly wage of individuals in the United States. Once the optimal model is determined, its predictive performance will be evaluated using the test set. This approach ensures that the model's accuracy is assessed on data that is not used during the training phase, providing a robust measure of its generalizability and effectiveness.

1.3 Exploratory Analysis of the Train Data

The exploratory analysis of the monthly wage in the training dataset, which comprises 524 observations, reveals several key insights. The average monthly wage is $998.08, with a minimum value of $115.00 and a maximum value of $2771.00. The lower quartile (Q1) is $700.00, the median is $948.50, and the upper quartile (Q3) is $1200.00, indicating that 50% of the observations fall between $700.00 and $1200.00. The variance of the monthly wage is 172,435.93, and the standard deviation is $415.25, reflecting significant variability in the data. The skewness of 1.05288 suggests a moderately right-skewed distribution, while the kurtosis of 1.6639 indicates slightly more pronounced tails compared to a normal distribution.

Count	Mean	Min	Lower Quartile	Median	Upper quartile	Max	Variance	Standard deviation	Skewness	Kurtosis
524	998.08	115	700	948.5	1200	2771	172435.9324	415.25	1.05288	1.6639

Figure 5: Statistic Description of Train Date

In terms of the distribution, histograms and boxplots indicate that the monthly wage in the training data set is right-skewed. This skewness suggests that there are a relatively small number of observations with exceptionally high wages, which pull the distribution's tail to the right. Consequently, most of the data points fall below the mean wage of $998.08, with fewer instances of significantly higher wages extending towards the maximum value of $2771.00. This right-skewed distribution implies that while most individuals earn wages around the central tendency measures (median of $948.50), there are outliers with substantially higher wages affecting the overall distribution shape.

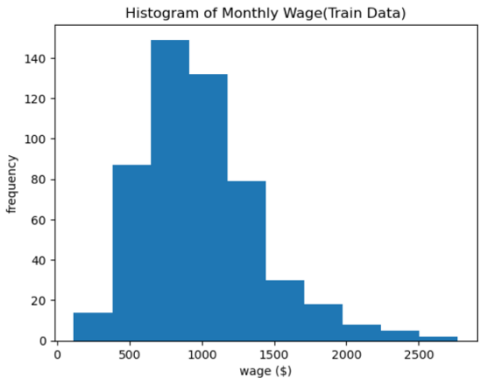


Figure 6: Histogram of Monthly Wage (Train Data)

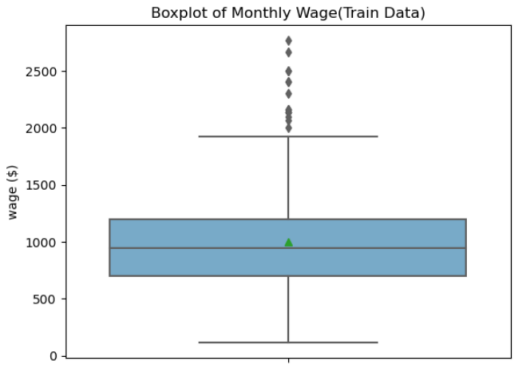


Figure 7: Boxplot of Monthly Wage (Train Data)

The correlation heatmap reveals various positive and negative correlations among the variables in the dataset. Notably, there are no strong correlations between any of the predictors and the response variable, wage. There is a moderate positive correlation between wage and education, while the relationship between wage and experience is very weak, close to zero. Other variables that show a moderate positive relationship with wage include IQ and KWW. Conversely, weak negative correlations are observed between wage and variables such as black, south, sibs, and brthord.

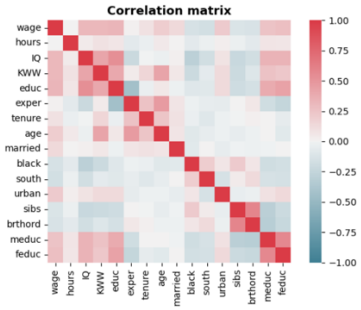


Figure 8: Heat Map of All Variables

These relationships are quantified in the summarization table of correlation (Appendix 1). As observed from the table, the strongest positive correlations with wage are only 0.34 for education, and 0.31 for IQ and KWW. Regarding the relationship between regressors, it is noticeable that there are some stronger positive correlations, such as between age and experience, education and IQ, each around 0.5.

According to the pair plot (Appendix 2), there are moderate positive relationships between wage and variables such as IQ, KWW, and education. However, these relationships are not straightforward, as the slopes exhibit increasing variability. This may imply that these relationships are non-linear. As each of these variables increases, the variance in wage

also appears to increase. This suggests that other factors may also influence the wage, contributing to the observed variability.

2. Relationship Analysis Between Education, Experience and Wage

2.1 MLR Model

The primary goal of this analysis is to understand the relationship between years of education, years of experience, and monthly wages. These relationships are investigated by fitting a Multiple Linear Regression (MLR) model to the wage data using education (educ) and experience (exper) as predictor variables.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-295.957	146.285	-2.023	0.044	-583.338	-8.576
educ	78.4469	8.433	9.302	0	61.88	95.014
exper	19.0792	4.461	4.277	0	10.315	27.843

Figure 9: Summarization of MLR Output

$$\widehat{wage} = -295.957 + 78.4469\, educ + 19.0792\, exper$$

Figure 10: Formula of MLR Output

The results indicate that, holding the variable experience constant, an additional year of education is associated with an average increase of $78.44 in monthly wage. Conversely, keeping the level of education constant, an additional year of experience is associated with an average increase of $19.08 in monthly wage.

Another important consideration is the potential multicollinearity between education and experience. Multicollinearity can complicate the interpretation of the regression coefficients by indicating that changes in one variable might be related to changes in another. In this dataset, the correlation between education and experience is moderately negative, with a value of less than -0.5 (Appendix 3). Additionally, the Variance Inflation Factor (VIF) for these variables is 1.244 (Appendix 3), which is well below the threshold of 5, suggesting that multicollinearity is not a significant issue in this case. Thus, there is no perfect collinearity between education and experience, allowing for a confident interpretation of the regression results.

2.2 LSA Assumptions Assessment

Upon conducting the residual plot of monthly wage, it is observed that there is no clear pattern in the residuals, though some outliers are present. This lack of pattern suggests that the model's assumptions of linearity is reasonably met. Further examination using a joint plot corroborates these findings, showing no discernible pattern. The average line of the residuals is close to zero, which is desirable, but the variance of the residuals is large, indicating some degree of heteroscedasticity.

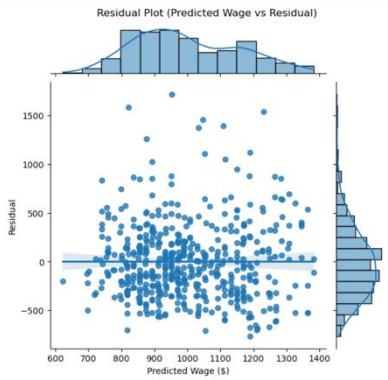


Figure 11: Residual Plot (Predicted Wage vs Residual)

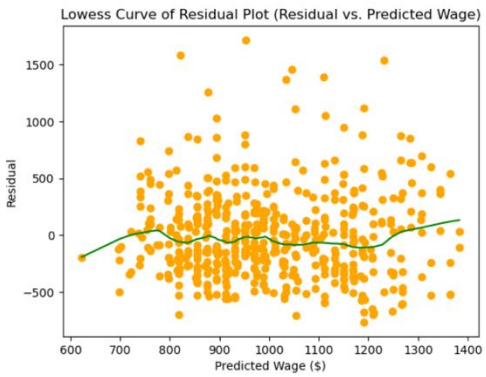


Figure 12: Lowess Curve of Residual Plot

A locally weighted scatterplot smoothing (LOWESS) curve is applied to further assess. This approach helps better understand the relationship between the predictors and the response. It is observed that the curve still averages around zero, suggesting linearity in the relationship between the predictors and the response variable for wages ranging from approximately $800 to $1300. The slope of the curve is steeper outside this range, indicating the presence of outliers. However, these outliers do not significantly affect the overall model fit. The linearity within the central range confirms that the model appropriately captures the primary trends in the data, and the impact of the outliers is minimal and does not substantially affect the model's validity.

Hence, LSA 1, which pertains to the absence of any clear pattern or trend in the residual plot against predicted wage, is not violated. The residual plot shows no discernible pattern, indicating that the residuals are randomly distributed around zero.

LSA 2 is assumed to be valid based on two residual plots of the residuals against each regressor, education and experience, respectively. The average of the residuals for each regressor is close to zero despite the presence of outliers. These outliers do not substantially affect the overall trend. However, the potential omitted variable bias (OVB) is not assessed in this section. Possible OVB will make LSA 2 violated.

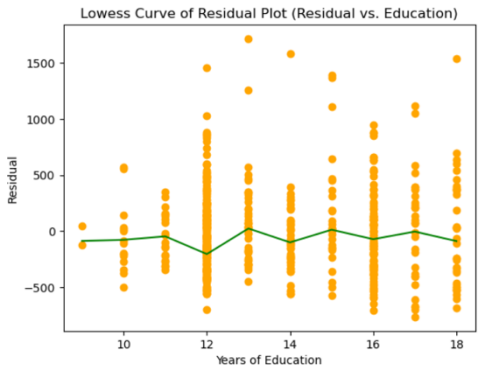


Figure 13: Residual Plot (Residual vs Education)

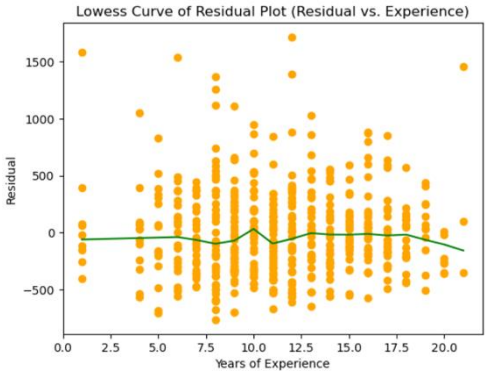


Figure 14: Residual Plot (Residual vs Experience)

LSA 3 is assumed to be true, though not explicitly assessed in this analysis. Regarding LSA 4, it is considered true. The kurtosis of wage suggests that the outliers are not significant, and the wage variable is bounded. Additionally, the regressors (education and experience) are also bounded and cannot take on infinite values.

LSA 5 is proven to be true as there is no perfect collinearity between the predictors. Based on the VIF calculated earlier, it is indicated that there is no perfect linear relationship between them. The residual plot is also utilised to determine whether LSA 6 - homoscedasticity is violated or not. Apart from some outliers, the overall variance seems constant, thus LSA6 is not violated. The use of heteroskedasticity-robust standard errors is not deemed necessary at this stage.

2.2 Goodness of Fit

Figure 15: Wage vs Education

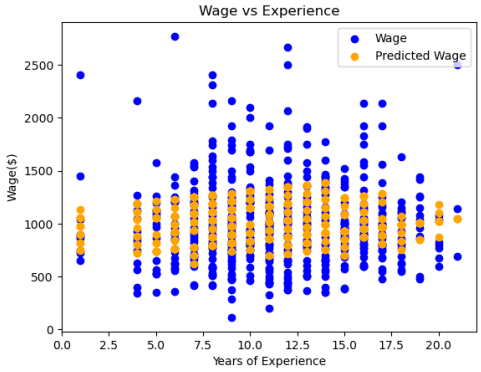


Figure 16: Wage vs Experience

The goodness of fit of the multiple linear regression (MLR) model can be assessed through the adjusted R^2 value. In this case, the adjusted R^2 value from the ordinary least squares (OLS) regression results is 0.139, or 13.9% (Appendix 4). This value indicates that only 13.9% of the variability in the response variable (monthly wage) is explained by the predictor variables (education and experience) included in the model.

Additionally, as observed in the two graphs above, the fitted values from the model do not adequately cover the range of the sample data. This discrepancy suggests that the model may not accurately capture the underlying relationship between the predictor variables and the response variable. This again emphasizes that the model may not satisfactorily explain the variability in monthly wages based solely on education and experience. Further refinement of the model is necessary to enhance its predictive accuracy and reliability.

3. OVB Identification

3.1 Finding Potential OVB

Based on the result of the MLR model including only education and experience as regressors in the last section (Appendix_), it is known that the coefficient for education and experience is 78.4469 and 19.0792 (Appendix 4), respectively. To assess the potential of OVBs, the full model, including all variables in the data set (Appendix 5), is also analysed, with findings of significant changes in the coefficient of education and experiences within the new full model compared to the previous one with only two regressors.

In the full model, the coefficient for years of education has dropped from 78.4469 to 35.2229, and the coefficient for years of experience has also changed from 19.0792 to 8.6069. The coefficients have significantly decreased compared to the previous model with only two regressors. Therefore, it is possible to suspect that other variables in the data set could have caused OVBs within the MLR in section 2.

The correlation table in section 1 (Appendix 1) is used to analyse the correlations and relationships other variables have with the two regressors' educ (years of education) and exper (years of experience). Clear and moderate positive correlations are found between IQ scores and years of father's education with years of education, with 0.53 and 0.43, respectively. These variables are also found to have some moderate negative correlation with years of experience, with -0.21 and -0.25, respectively.

The correlation of 'IQ' and 'feduc' was identified to be strongly related to at least one of the regressors 'education' and 'experience', with the highest correlations compared to other variables within the data set. Also, in real life, these are likely to be determinants of 'wage', as higher IQ is most likely to be associated with more highly skilled jobs that offer higher wages; and more years of father's experience will result in higher wage jobs in our real-life assessment. Therefore, IQ and feduc are most likely to cause OVBs within our analysis.

3.2 Forming Model with OVB

These two variables, IQ and feduc are added into a new model, together with education and experience, to make a four-variable MLR without any transformations, interactions, or non-linear effects. Regarding the output of the new model, the relationship of wage with education and experience in this new model can be found.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-295.957	146.285	-2.023	0.044	-583.338	-8.576
educ	78.4469	8.433	9.302	0	61.88	95.014
exper	19.0792	4.461	4.277	0	10.315	27.843

Figure 17: Two Variables MLR

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-587.0863	158.56	-3.703	0	-898.584	-275.589
educ	50.4082	9.852	5.116	0	31.053	69.763
exper	19.9294	4.367	4.563	0	11.35	28.509
IQ	4.4561	1.332	3.345	0.001	1.839	7.073
feduc	20.2434	5.619	3.602	0	9.204	31.283

Figure 18: Four Variables MLR

$$\widehat{wage} = -587.0863 + 50.4082\, educ + 19.9294\, exper + 4.4561 IQ + 20.2434\, feduc$$

Figure 19: Formula of Four Variables MLR Output

The estimated coefficient of education in the four-variable MLR has dropped to be lower than the 95% CI from the two-variable MLR (61.88, 95.014). Therefore, IQ and feduc are clearly causing some OVBs in this estimated effect in the two variable MLR. This conclusion also implies that the LSA 2 assumption of the two variable MLR is violated now due the presence of OVBs.

The estimated effect of experience has also slightly increased. However, the two 95% CIs have a large overlap, and the difference in estimate here can be put down to estimation error rather than OVB. In other words, though IQ and feduc could have caused some OVBs in the two variable MLR for experience, the large standard error of 375.84 makes it difficult to assess the level of OVBs caused or even if any was apparent.

3.3 LSA and Collinearity Problem Assessment

The residual vs fitted values plot suggests heteroskedasticity, with inconsistent variance throughout the values. The variance range is quite narrow for predictions of wages below $600, at around 0, then expands as the predicted wage values increase. Outliers start to appear from that value with increased frequency as the value of the predicted wage grows. This also suggests the violation of LSA 6 assumption.

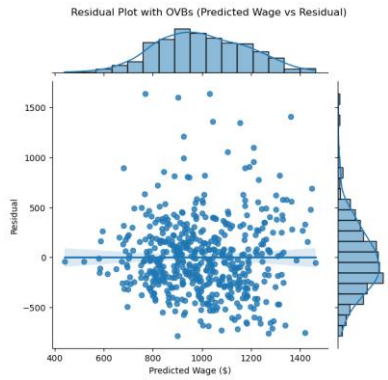


Figure 20: Residual Plot with OVBs

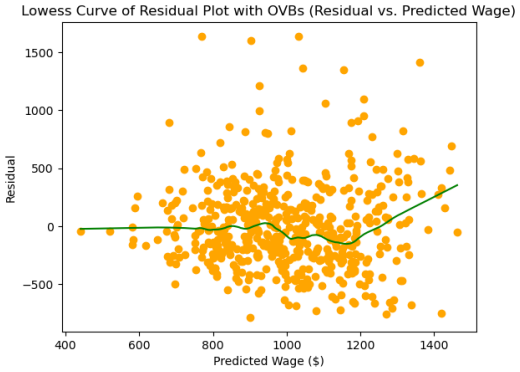


Figure 21: Lowess Curve of Residual Plot

Additionally, based on the residual plot, generally, the scatter dots are spread around the line. However, considering the high frequency of outliers appears significantly, it is likely to have a non-linear pattern in this model. Therefore, LSA 1 is likely to be violated.

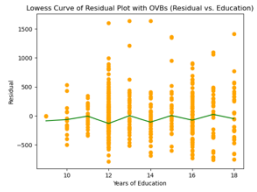


Figure 22: Residual vs Education (OVBs)

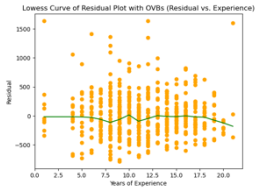


Figure 23: Residual vs Experience (OVBs)

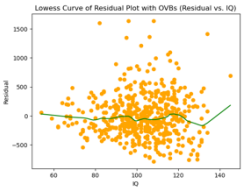


Figure 24: Residual vs IQ (OVBs)

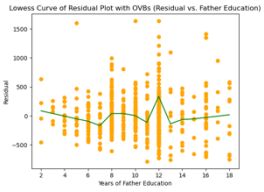


Figure 25: Residual vs Father Education (OVBs)

With four scatter plots of residuals against each regressor in the model, the average residuals for regressors—education, experience, and IQ—are close to zero. However, for the scatter plot of residuals versus father's education, the high level of outliers makes it difficult to assert that the average residual is close to zero. This indicates that LSA 2 may be violated, suggesting the possibility of omitted variables in the model. These omitted variables could prevent the true population from following a linear relationship, implying a violation of LSA 1 as well.

With regards to LSA 3, the process of data collection is unknown. However, it is possible to still acknowledge the assumption that data collection is i.i.d. As the kurtosis of wage, at 1.6639, indicates outliers are not significant, it is possible to assume that LSA 4 is not violated. Moreover, relevant variables and regressors are bounded by some natural limit. Therefore, it seems safe to assume they have finite 4th moments.

Regarding whether multicollinearity is a problem in this scenario, the correlation between all variables included in the MLR model is calculated. Based on the outcome (Appendix 6), it is noted that the correlation coefficient between the variables educ and IQ exceeds 50%, suggesting a potential presence of multicollinearity. A simple linear regression (SLR) model incorporating these two variables is employed to evaluate this issue. The adjusted R^2 stands at 0.285 (Appendix 7), indicating a relatively low explanatory power. Given this modest R-squared value, concerns regarding multicollinearity are deemed unfound in this context. Consequently, it is reasonable to assert that multicollinearity does not pose a significant concern within the examined model. This also indicates that the LSA 5 is followed since the perfect collinearity does not exist.

Finally, as mentioned before, LSA 6 is violated based on the residual plots as the variation increases as predicted wage increases, introducing the need to use hetero-robust SEs for investigation.

3.4 Goodness of Fit

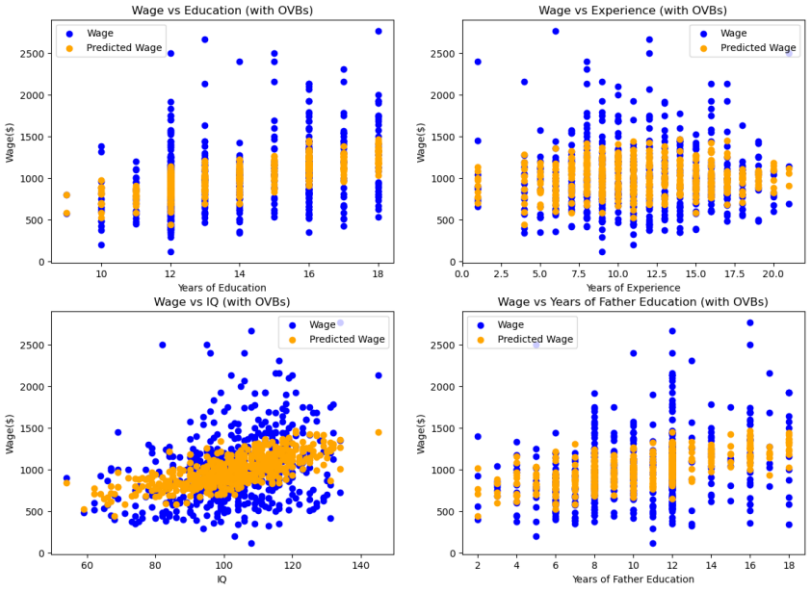


Figure 26: Wages vs Regressors in MLR

The scatters plot of wage vs each regressor suggests a very weak goodness of fit of data for the models, as the predicted values only concentrate on a part of the real values, whereas the real values have a considerably wide range compared to the predicted ones. This argument is strengthened by the fact that the adjusted R^2 in the model is very low, at 0.181, indicating only 18.1% of the variation in predicted wage can be explained by the regressors (Appendix 8).

4. Model Selection

In this section of the report, to accurately predict monthly wages, a model selection is conducted. This includes considering different effects such as interactions, transformations, and nonlinear effects. In comparing models, both the standard error of regression (SER) and adjusted R^2 are primarily considered, since adjusted R^2 accounts for the number of independent variables and provides a more accurate measure of model performance than R^2. Lower SER and higher adjusted R^2 are preferred.

4.1 Forward Selection and Reduced Model

Initially, forward selection (FS1) is carried out using the wage as predictors along with 15 basic regressors in the dataset, with education and experience being mandatory inclusions. The outcome of FS1 leads to the final full model (FS1 – full), which includes 10 regressors and yields an adjusted R^2 of approximately 23.80% (Appendix 9), with the SER calculated to be around 362.488.

However, to achieve a parsimonious model and potentially prevent issues such as overfitting, the FS1 process is terminated early during the addition of regressors. Based on the minor increase in adjusted R^2 from adding the regressor ‘south’—a mere 0.27% increase, which is significantly less compared to previous regressors—the addition process is

terminated at the regressor ‘age’. Consequently, the reduced model (FS1 – reduced model) includes regressors from 'educ' to ‘age’, with an adjusted R^2 of about 23.72% and an SER of 363.313.

Regressors Addition	adj R^2 before	adj R^2 after	Change in adj R^2
educ, exper		0.139237	0.139237
urban	0.139237	0.167961	0.028724
IQ	0.167961	0.189929	0.021968
married	0.189929	0.205582	0.015653
feduc	0.205582	0.219864	0.014282
age	0.219864	0.234519	0.014655
south	0.234519	0.237232	0.002713
KWW	0.237232	0.237862	0.00063
meduc	0.237862	0.237994	0.000132

Figure 27: The Change in adjusted R^2 in FS1

4.2 Log-transformation Effect

Transformation Effect on FS1–reduced model

Based on the pair plots between different variables analysed previously (Appendix 2), it is observed that the scatter plots for ‘wage’ vs ‘exper’ and ‘wage’ vs ‘IQ’ suggest a nonlinear relationship. Therefore, log-transformation is applied to these two regressors and the predictor ‘wage’ to potentially enhance model performance.

In the FS1-reduced model, various combinations of log-transformation are applied to these three variables. It is noted that the log-transformation on ‘exper’ did not significantly improve the model's performance (Appendix 10), so the focus is shifted to the improvements brought about by ‘log_wage’ and ‘log_IQ’.

For the models including ‘log_wage’, three different bias corrections are applied, and the one with the best adjusted R^2 and the lowest SER is selected. Specifically, two models using ‘log_wage’ as predictors utilise results from the Duan bias correction. Comparing the effectiveness of these log-transformations, the model (log_wage vs IQ – FS1 – reduced model (Duan)), which applies the transformation solely to ‘wage’ in the FS1-reduced model, demonstrates the best performance with an adjusted R^2 of approximately 24.2% and an SER of 361.493.

Combination of Log-transformation	SER	R^2	adj R^2	Number of Regressors
Wage & Log_IQ	363.4405703	0.244	0.234	7
Log_wage & IQ (Exp)	366.7262068	0.23050795	0.22006911	7
Log_wage & IQ (Duan)	361.4932863	0.25231148	0.24216841	7
Log_wage & IQ (Normal)	361.5281544	0.25216723	0.24202221	7
Log_wage & Log_IQ (Exp)	366.8173567	0.23012539	0.21968135	7
Log_wage & Log_IQ (Duan)	361.5505284	0.25207467	0.24192839	7
Log_wage & Log_IQ (Normal)	361.5830079	0.25194028	0.24179218	7

Figure 28: Log-transformation on Wage and IQ in FS1–reduced Model

Forward Selection with Log_wage

Based on the comparison above, it is reasonable to infer that much of the improvement in the model is attributed to 'log_wage'. Therefore, 'log_wage' is used as the predictor to conduct a forward selection (FS2) incorporating all other regressors in the dataset to identify a better model. The final model of FS2 included 11 regressors. After implementing bias corrections, the model with Duan bias correction (log_wage – FS2 full (Duan)) yields the best results (Appendix 11) with an adjusted R^2 of 23.9% and an SER of 362.226.

With the goal of achieving parsimony, the addition process in FS2 is terminated early at ‘feduc’ since the incremental increase in adjusted R^2 contributed by ‘tenure’ and subsequent regressors is minimal. Thus, the reduced model

from FS2 (log_wage – FS2 reduced model (Duan)), which also shows the best results with Duan bias correction (Appendix 12), contains 8 regressors, achieving an adjusted R^2 of 24.4% and an SER of 361.134.

Regressors Addition	adj R^2 before	adj R^2 after	Change in adj R^2
educ, exper		0.145899	0.145899
urban	0.145899	0.178009	0.03211
IQ	0.178009	0.201057	0.023048
married	0.201057	0.220835	0.019778
south	0.220835	0.233872	0.013037
age	0.233872	0.244374	0.010502
feduc	0.244374	0.252743	0.008369
tenure	0.252743	0.258302	0.005559
hours	0.258302	0.262807	0.004505
meduc	0.262807	0.264876	0.002069

Figure 29: The Change in adjusted R^2 in FS2

4.3 Interaction Effect

Interaction effect on FS1–reduced model

Potential interaction effects influencing the monthly wage are identified using common sense reasoning. For potential interactions with regressors included in the FS1-reduced model, each is added to the FS1-reduced model to formulate new models. Consequently, four different models based on the FS1-reduced model are established. Upon comparison, two models with interactions—specifically ‘age_married’ and ‘urban_feduc’—perform better. For ‘age_married’, married and unmarried individuals may make different career choices at different ages. For instance, married individuals may priorities economic stability at older ages, thereby influencing wages. For ‘urban_feduc’, a father with a higher education level in urban areas can provide more resources to their children, such as networking opportunities and job prospects, thereby influencing children's wages.

In the model with the interaction of ‘age_married’ (interaction – age_married – FS1 – reduced model), for married individuals, the coefficient of age is 23.1897, while for unmarried individuals, it is -7.7482 (Appendix 13). The adjusted R^2 is 23.7% and the SER is 362.825. In the model with the interaction of ‘urban_feduc’ (interaction – urban_feduc – FS1 – reduced model), for individuals living in urban areas, the coefficient of 'feduc' is 25.1741, whereas for those not in urban areas it is 1.7784 (Appendix 14). The adjusted R^2 and SER of this model are 23.9% and 362.140, respectively.

Interaction Effect	SER	R^2	adj R^2	Number of Regressors
Education_Urban	362.8487921	0.248	0.236	8
Experience_Urban	362.9117926	0.248	0.236	8
Arge_Married	362.8252175	0.248	0.237	8
Urban_Father Education	362.1401967	0.251	0.239	8

Figure 30: Different Interaction Effect in FS1–reduced model

Forward Selection with Exper_South

The interaction effect ‘exper_south’ may also influence wages in the United States, as the impact of job experience on wages could differ between the South and the North due to varying industrial focuses in these regions. For example, in the South, which focuses more on agriculture, individuals with agricultural experience may earn better wages there compared to the North.

However, since ‘south’ is not included in the regressors of the FS1-reduced model, a forward selection (FS3) is conducted, including all regressors in the dataset along with 'exper_south'. The full model in FS3, containing 11 regressors (Appendix 15), is terminated early in the addition process due to the parsimony requirement. With a low incremental increase in adjusted R^2 caused by ‘KWW’, the reduced model (interaction – exper_south – FS3 – reduced model) contains regressors from ‘exper’ to ‘exper_south’, with an adjusted R^2 of 24.1% and an SER of 361.718.

Regressors Addition	adj R^2 before	adj R^2 after	Change in adj R^2
exper, educ		0.139237	0.139237
urban	0.139237	0.167961	0.028724
IQ	0.167961	0.189929	0.021968
married	0.189929	0.205582	0.015653
feduc	0.205582	0.219864	0.014282
age	0.219864	0.234519	0.014655
south	0.234519	0.237232	0.002713
exper_south	0.237232	0.241227	0.003995
KWW	0.241227	0.242249	0.001022
meduc	0.242249	0.24236	0.000111

Figure 31: The Change in adjusted R^2 in FS3

4.4 Spline Regression

The nonlinear relationship between experience and wage is observed from the pair plot (Appendix 2), which can also be demonstrated in the Lowess curve for wage versus experience. Thus, this report considers refitting the model using regression splines such as linear spline and quadratic spline, which are the most common choices.

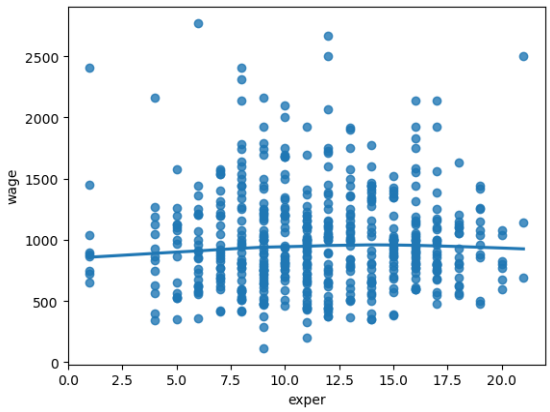


Figure 32: Wage vs Experience with Lowess Curve

Linear spline

This report adds linear spline knots at the 25%, 50% and 75% quantiles to the 'experience' data, considering these as additional potential predictors. Then a forward selection is conducted to choose an optimal model using these 3 knots and all other predictors in the training dataset. However, none of the steps are added to the model, suggesting that regression spline may not be a good choice here. but it was still worthwhile to fit the model with one, two, and three knots, respectively, to assess if the model performance improved. Moreover, other possible combinations of regressors have been tried, for example, fitting a linear spline model with 3 knots using the FS1-reduced model. From the table below, the linear spline with one knot has the highest adjusted R^2 of 0.237 and the lowest SER of 362.776, making it the best among all linear spline models.

			SER	R^2	adj R^2	P
linear spline with 1 knot - Full			362.766	0.253	0.237	11
linear spline with 2 knot -Full			363.519	0.259	0.234	12
linear spline with 3 knot -Full			363.861	0.259	0.232	13
linear spline with 3 knot - Reduced			364.231	0.245	0.231	10
cubic cubic spline with 1 knot			362.843	0.255	0.236	13
cubic linear spline with 2 knot			363.637	0.261	0.233	14
cubic linear spline with 3 knot			363.491	0.263	0.234	15

Figure 33: Performance of Different Spline Regressions

Cubic spline

Then the same process is repeated for cubic splines. As shown in the table above, the SER, and adjusted R^2 do not improve with the use of cubic spline. Overall, the linear spline with one knot is considered the best among all the regression spline models, as it keeps the model less complicated and helps avoid the overfitting problem.

4.5 Model Comparison

Optimal Model Selection

After the process of forming different models, representative or best models with different regressors or effects are summarized, also including the model formed in the previous part. Instead of R^2, adjusted R^2 is used as an indicator of model performance due to higher accuracy. The model with high adjusted R^2 and lower SER is preferred.

Therefore, based on the summary table, the log wage – FS2 – reduced model (Duan) is identified as the optimal model, followed by log wage vs IQ – FS1 – reduced model (Duan) as the second best, and interaction – exper_south – FS3 – reduced model as the third. This selection is made because among all the models, the log wage – FS2 – reduced model (Duan) has the highest adjusted R^2 of 24.4% and the lowest SER of 361.134. Concerning the number of regressors, 8 is considered a moderately low count, which also relatively satisfies the parsimony requirement. Compared to the second-ranked model, the optimal model only includes one additional regressor, which does not represent a significant difference.

Model Name	SER	R^2	adj R^2	Number of regressors
Linear - full	362.92	25.8	23.6	15
Linear - with exper and educ	385.261	14.3	13.9	2
Linear - with exper and educ and OVBs	375.845	18.7	18.1	4
FS1 - full	362.488	25.3	23.8	10
FS1 - reduced model	363.313	24.5	23.5	7
Log wage vs IQ - FS1 - reduced model (Duan)	361.493	25.2	24.2	7
Log wage - FS2 full (Duan)	362.226	25.5	23.9	11
Log wage - FS2 - reduced model (Duan)	361.134	25.5	24.4	8
Interaction - age-married - FS1 - reduced model	362.825	24.8	23.7	8
Interaction - urban_feduc - FS1 - reduced model	362.140	25.1	23.9	8
Interaction - exper_south - FS3 - reduced model	361.718	25.4	24.1	9
Linear spline - FS full - 1 knot	362.766	25.3	23.7	11
Linear spline - FS reduced - 3 knots	364.231	24.5	23.1	10

Figure 34: Summarization of Models

Assisted by two scatter plots—one comparing wage versus years of education, and the other wage versus IQ—both including the actual wage values and predicted wage values from the top three models, the performance of these models could be interpreted. In both scatter plots, the purple dots, which correspond to the fitted values from the rank three model, display comparatively poor performance. Meanwhile, the fitted values from the optimal model and the rank two model showed similar performances. Therefore, the scatter plots do not contradict the top three rankings of the models.

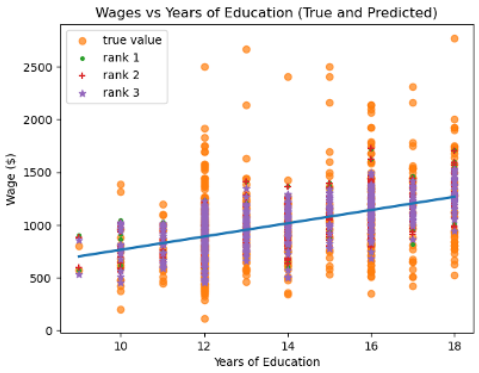


Figure 35:Wage vs Years of Education

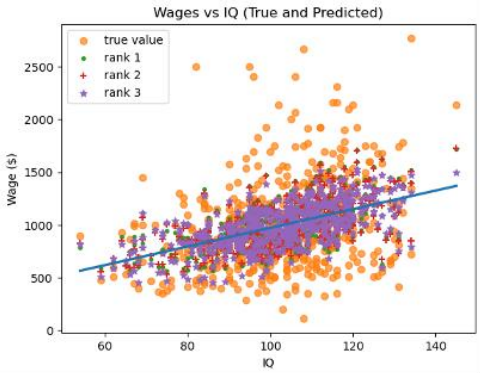


Figure 36: Wage vs IQ

Optimal Model Description

Regarding the optimal model's output, the formula could be expressed, and the relationship between 'log_wage' and each regressor could be interpreted. In this model, the only nonlinear effect arises from the log-transformation on predictor 'log_wage', and the influence of this effect on the model was revealed through the interpretation of coefficients.

$$\widehat{\log\_wage} = 4.6079 + 0.0404\,educ + 0.0121\,exper + 0.1589\,urban + 0.0045\,IQ + 0.2043\,married - 0.0999\,south + 0.0195\,age + 0.0144\,feduc$$

Figure 37: Formula of the Optimal Model

Coefficient Interpretation

(1) Intercept: When all the regressors are equal to 0, the value of log_wage is equal to 4.6079. However, this value does not make sense since zero of age, which is one of the regressors, is not reasonable.

(2) Education: Keeping all other regressors constant, an increase of 1 unit in years of education is associated with an average increase of 4.04% in monthly wage.

(3) Experience: Keeping all other regressors constant, an increase of 1 unit in years of experience is associated with an average increase of 1.21% in monthly wage.

(4) Urban: Keeping all other regressors constant, individual living in urban is associated with an average increase of 15.89% in monthly wage.

(5) IQ: Keeping all other regressors constant, an increase of 1 unit in IQ is associated with an average increase of 0.45% in monthly wage.

(6) Married: Keeping all other regressors constant, married individual is associated with an average increase of 20.43% in monthly wage.

(7) South: Keeping all other regressors constant, individual living in South is associated with an average decrease of 9.99% in monthly wage.

(8) Age: Keeping all other regressors constant, an increase of 1 unit in age is associated with an average increase of 1.95% in monthly wage.

(9) Father Education: Keeping all other regressors constant, an increase of 1 unit in years of father's education is associated with an average increase of 1.44% in monthly wage.

Optimal Model Assessment

The LSA assumptions for this optional model are assessed first to check the accuracy of the predictions. For LSA 1, assisted by the residual plot of residuals vs. fitted values of the predictor, which includes the Lowess curve, the dots

generally spread symmetrically around the curve. However, some extreme outliers exist in the scatter plot, potentially leading to a violation of LSA 1. This will be further assessed with the LSA 2 assessment.

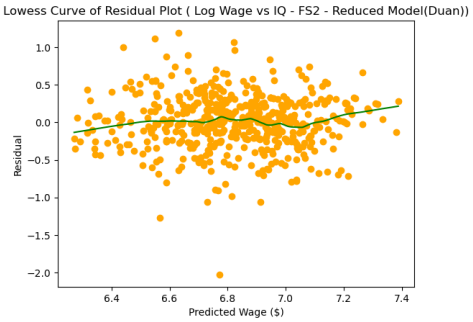


Figure 38: Lowess Curve of Residual Plot (Log_wage vs IQ – FS2 – Reduced Model (Duan))

Regarding LSA 2, using the residual plots vs. each of the regressors contained in the optimal model, it is observed that for 'urban', 'married', and 'feduc', the average of the residuals is not equal to zero (Appendix 16). This suggests that LSA 2 is violated, indicating the presence of some omitted factors in the model. Therefore, LSA 1 might also be violated since these omitted variables might imply that the true relationship in the model is not linear.

For LSA 3 regarding independence, it is challenging to prove that the data in the optimal model are independently and identically distributed (i.i.d.), so it is assumed to be true in this report. Regarding LSA 4, since the values of log_wage and all regressors are bounded, LSA 4 was satisfied in this model. For LSA 5, concerning the possible multicollinearity problem, the correlation between each regressor included in the optimal model is calculated (Appendix 17). Among these, only the correlation between ‘educ’ and ‘IQ’ exceeds 50%, at 53.4%. To assess potential collinearity between ‘educ’ and ‘IQ’, a simple linear regression between these two variables is conducted, yielding an R^2 of 28.5% (Appendix 7). Therefore, collinearity is not considered a problem in this optimal model, and LSA 5 is satisfied since there was no perfect collinearity. LSA 6 is also followed in this model since the residual plot displays constant variance overall.

Overall, LSA 1 and 2 are violated in this optimal model, implying that the prediction of wages using this model might be less accurate. However, this model would still be used as the optimal model, since its performance is better than other models evaluated.

Regarding the goodness of fit, the low value of adjusted R^2 in this optimal model indicates that the performance of this model is not quite good. Also, with the scatter plots of wage versus each regressor in the model, including the true and fitted values of wage (Appendix 18), the range of fitted values is significantly smaller than that of true values. This is consistent with the low adjusted R^2.

5. Evaluation of Goals

In alignment with the overall goals, the relationship between education versus experience and wage are identified by developing a model that provides the best predictions on the monthly wages of individuals.

Toward achieving the overall goal, sections one to four provide steps for developing the optimal model, which involves improving the original model with considerations about other possible effects such as OVBs, interaction and

transformation effects between regressors, and selecting an optimal one amongst the generated models. The optimal model that has been found so far is Log wage - FS2 - reduced model (Duan).

Regarding this process, goals a and b align, as the relationship of education versus experiences and monthly wages constitutes a significant component of the optimal model. Meanwhile, the process of developing the optimal model (goal b) allows deeper analysis of the relationship between years of education versus experience and monthly wages (goal a).

Assisted with the coefficients of education of experience in the optimal model, the trade off between these two variables can be specified. Before utilising these coefficients to analyse the trade off, the t test needs to be demonstrated on each of them to assess whether these coefficients are significantly different to zero. Specifically, the null hypothesis is that the coefficient is equal to zero, while the alternative hypothesis is that the coefficient is different to zero. Since the p values are 0 and 0.003 for coefficient of education and experience respectively, which is less than the confidence level of 0.05, the null hypothesis is rejected. Therefore, the coefficient of education and experience are significantly different from zero.

Then, these coefficients can be used to estimate the trade-off between education and experience. Keeping all other variables constant, when an individual chooses to increase the year of education by one instead of year of experience, it is expected that there will be an average increase of 4.04% in monthly wage. By contrast, keeping all other variables constant, if the experience is chosen to be increased by one year instead of education, there will be an average increase of 1.21% in monthly wage. Therefore, increasing education by one year on average will lead to more significant improvement in wage, on average.

6. Test Data in Model Selection

In the previous section, the three best model specifications were generated using a training dataset. The following section utilized the test dataset to make predictions on these models. Since the test dataset consisted of out-of-sample data, the model performance may differ due to some unseen factors in the test dataset. Therefore, this report selected the most accurate model in the test dataset as the optimal model.

Before generating predictions, it is necessary to test the assumptions and assess the goodness of fit for the three models. Since the tests for the ‘Log wage - FS2 - reduced model (Duan) model’ (optimal model based on train dataset) have already been completed, only the other two models need to be tested.

For these two models, the graphs of residual versus predictor and different regressors implied that LSA 1 and 2 have been violated (Appendix19, 20, 21 and 22), which indicated that the models formed before may have some nonlinear relationships. Also, there were some multicollinearity issues (Appendix 23) for the model including interaction effect by design, thus influencing LSA 5, but this should not be a significant problem as it is not perfect collinearity. Moreover, the significant differences between the range of true wage and predicted wage suggested poor performance in terms of goodness of fit, which aligns with the low adjusted R^2(Appendix 24 and 25). This implied that these two models did not provide adequate fit to the data. All factors mentioned above would reduce the precision of prediction. However, these remained the best models identified so far, so they would be used to generate predictions.

After generating forecast predictions using test data for the three best models, this report utilized the forecast measures such as forecast R^2, RMSFE and MAFE to assess the accuracy of these models. In order to get the revised ‘predicted wage’, bias correction (Duan’s method) needs to be conducted for the two models with log transformation on the predictor before calculating the forecast measures. Then, to generate forecast R^2, M-Z regressions for the three best models

were conducted with wage as predictor and predicted wage (revised) as regressor. Moreover, RMSFE and MAFE are calculated and summarized in the table below.

Model	RMSFE	MAFE	Forecast R^2	P
Log wage vs IQ - FS2 - reduced model (Duan)	342.486	246.772	17.3	8
Log wage vs IQ - FS1 - reduced model (Duan)	336.466	241.296	19.8	7
Interaction - exper_south - FS3	340.777	245.175	18.5	9

Figure 39: Summarisation Table for Output of M-Z regressions

Therefore, the three best models were assigned new ranks based on their performance in predictions. The 'Log wage vs IQ - FS1 - reduced model (Duan)' was ranked as the optimal model as it had highest forecast R^2(19.8), lowest RMSFE (336.466) and MAFE (241.296), and least number of regressors (7) , it was ranked as the optimal model. The 'Interaction - exper_south - FS3' model had the second-best forecast measures and was ranked second. The 'Log wage vs IQ - FS2 - reduced model (Duan)' was ranked third due to comparatively poor forecast accuracy.

After comparing the scatter plot of true and predicted wage versus education or IQ for these three models, the blue plot, representing the rank 1 model, seemed to have a slightly wider range for predicted wage than the others. However, ranking solely based on the scatter plot was challenging. Therefore, the rank based on the previous table of values was kept.

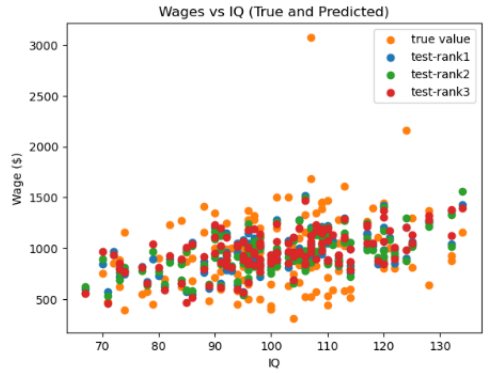


Figure 40: Wage vs IQ (Test)

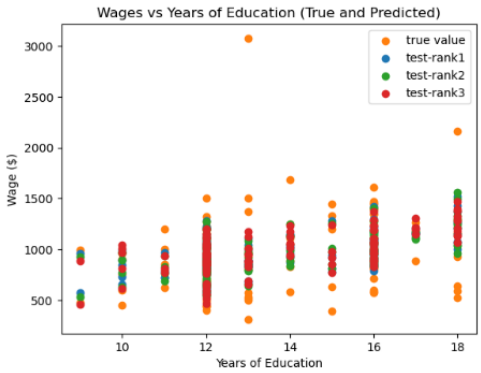


Figure 41: Wage vs Education (Test)

It is noteworthy that the rankings for the three best models have changed, which was caused by unforeseen factors in the test dataset. Furthermore, the forecast R^2 is lower than adjusted R^2 in the train dataset because the models were built based on the train dataset, but the accuracy for predictions on unseen data cannot be guaranteed.

In summary, the optimal model found in this section was 'Log wage vs IQ - FS1 - reduced model (Duan)'. This can be used to better predict wage for individuals and explain the relationship between education versus experience and wage, which aligns with our overall goal.

7. Summary and Recommendation

7.1. Summary

To summarize what this report has conducted so far, firstly the data is cleaned and split into two groups: training data for building models and testing data for generating predictions. Subsequently, the analysis on the monthly wages of 524

individuals in the US reveals that on average, people earn $998.08, ranging from $115 to $2771. A larger portion of individuals earn lower wages, while a smaller number earn significantly higher wages. Moreover, there is a moderate positive correlation between Wage and Education, while the correlation between Wage and Experience is very weak.

In section two, the analysis shows that if keeping the amount of experience constant, a person with an extra year of education can increase their monthly wage by $78.44 on average. Conversely, if keeping education level steady, each extra year of experience only leads to about a $19.08 increase in monthly wage on average. Given all the assumptions are met, this analysis helps estimate the value of getting more education instead of choosing a professional career path, to be elaborated upon later. However, as the model does not fit well with the data, relying solely on education and experience to predict wages isn't very effective.

In section three, two models are examined: one using all predictors and another using only education and experience. Differences are noted in the slope for education and experience between the two models. This could mean that some important predictors are left out of the analysis, causing mistakes. To investigate, additional factors 'IQ' and 'father's education', closely related to education or experience, are examined. Upon incorporating these factors into a new model alongside education and experience, the model's performance improves. This suggests that 'IQ' and 'father's education' may have contributed to the discrepancies observed in the initial model with only education and experience.

In section four, as the goal is to find the optimal model to predict wage, various combinations of variables are examined while ensuring that education and experience remain across all models. This involves examining effects such as converting a linear model into a log one (transformation), the joint impact of two predictors on response variable (interaction effect), and utilizing curved lines (regression spline) rather than linear ones. The optimal model found in this section is 'Log wage vs IQ - FS2 - reduced model (Duan)'. Moreover, the second-best model is 'Log wage vs IQ - FS1 - reduced model (Duan)', followed by the model 'Interaction - exper_south - FS3 - reduced model'.

Finally, this report utilises the test dataset to generate forecast predictions for the three best models and reselect the most accurate model in the test dataset as the optimal model. After assessing the performance of these predictions based on some forecast measures, the optimal model found in this section changes to the 'Log wage vs IQ - FS1 - reduced model (Duan)'. This model is believed to perform best in predicting wage and explain the relationship between education versus experience and wage, aligning with our overall goal.

The whole progress has led to the final optimal model for predicting monthly wages based on years of education and experience of individuals. Assuming that increasing the department's expenditure on education or improving early job experience results in a one-year increase in education or experience respectively. From the optimal model, it is estimated that keeping all other variables constant, if one year of education is increased, the average monthly wages of individuals will increase by 4.01%. Additionally, keeping all other variables constant, if one year of experience is increased, the average monthly wages of individuals will increase by 1.33%.

Thus, compared to work experience, spending money on increasing average years of education will result in more effective increments in monthly wage. Therefore, this report recommends the department prioritise investments in boosting the average year of education rather than emphasizing job experience.

7.2. Suggestions to Department of Education and Department of Labour:

To further incentivize individuals in the US to pursue higher levels of education, this report proposes to the Department of Education two recommendations: Student Loans Forgiveness Programs and Employer-Sponsored Education Programs.

The first suggestion targets students, involving increasing student loan forgiveness or repayment programs for degrees, diplomas, or short-term courses in high-demand fields, such as education and technology. This initiative can improve earnings prospects for graduates and adults looking to expand their education in various sectors by alleviating financial hardship associated with pursuing higher education. This practice can create up to 1.5 million new jobs and has been practised by the Biden-Harris Administration, therefore highlighting its feasibility and effectiveness (Hanson, 2023).

The second recommendation targets individuals already in the workforce. Employees within a company can enhance their educational credentials, such as pursuing Master's or PhD degrees, with the assistance of their employers. This can involve education benefits such as tuition reimbursement, access to on-site training facilities, and flexible work schedules tailored to accommodate educational commitments. To facilitate this initiative, the Department of Education should allocate a budget to companies, enabling them to sponsor their workers' educational pursuits. The benefits and practicability of this approach have been demonstrated by the practices of major companies, including IBM's internal education program. In addition, according to a research conducted by the Lumina Foundation, employees who participated in these programs experienced a 43% growth in earnings (Brower, 2022).

However, we also acknowledge the importance of job experience that can further assist academic study. Therefore, we propose to the Department of Labor to also offer apprenticeship and internship programs for students in various industries, offering structured training and on-the-job learning experiences by partnering with employers, educational institutions. A study in the US has shown that integrating education and work through apprenticeship initiatives not only diminishes youth unemployment and aligns labour supply with demand, but also improves students' academic paths: assisting students in graduating on time and gaining access to higher education than those who are not part of apprenticeship programs (de Amesti & Claro, 2021).

7.3. Recommendations on future studies:

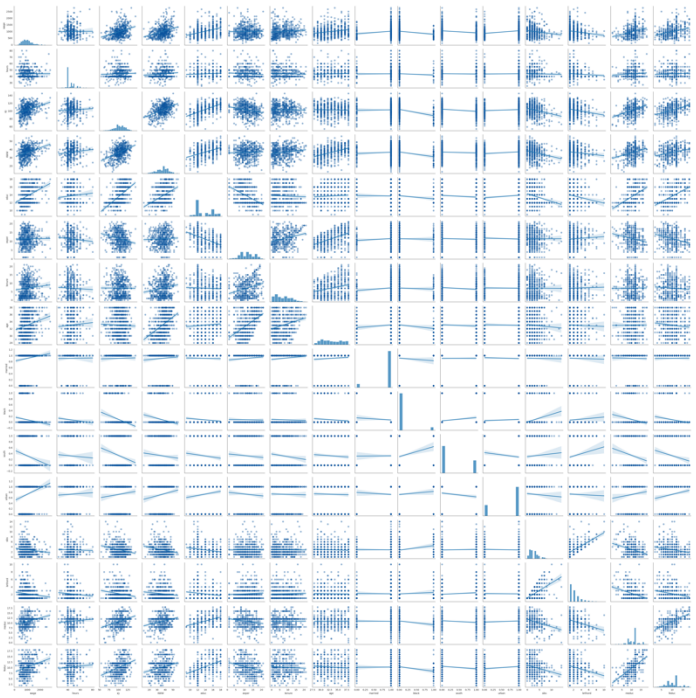
As the rationale behind this research is to identify methods to improve earning prospects, and this research has explained the relationship between education and experience on individuals’ wages, future studies can delve deeper into the relationship between the proposed actions on education and experiences and provide further empirical evidence. For example, there can be a study assessing the effectiveness of loan forgiveness programs on education through exploratory research on the education levels between individuals taking part in loan forgiveness programs and those who do not, with surveys on students of USYD as a method of data collection method.

Appendix

Appendix 1: Table of Correlation

	wage	hours	IQ	KWW	educ	exper	tenure	age	married	black	south	urban	sibs	brthord	meduc	feduc
wage	1.00	0.01	0.31	0.31	0.34	0.01	0.09	0.20	0.13	-0.15	-0.17	0.21	-0.12	-0.14	0.25	0.28
hours	0.01	1.00	0.04	0.09	0.07	-0.07	-0.03	0.05	0.01	-0.08	-0.03	0.03	-0.02	-0.08	0.08	0.06
IQ	0.31	0.04	1.00	0.42	0.53	-0.21	-0.01	-0.03	0.02	-0.29	-0.17	0.08	-0.22	-0.17	0.34	0.34
KWW	0.31	0.09	0.42	1.00	0.40	0.04	0.16	0.43	0.06	-0.21	-0.07	0.13	-0.21	-0.11	0.26	0.23
educ	0.34	0.07	0.53	0.40	1.00	-0.43	-0.03	0.06	-0.02	-0.10	-0.09	0.13	-0.19	-0.16	0.39	0.43
exper	0.01	-0.07	-0.21	0.04	-0.43	1.00	0.26	0.47	0.07	-0.02	-0.05	-0.06	0.00	0.06	-0.18	-0.25
tenure	0.09	-0.03	-0.01	0.16	-0.03	0.26	1.00	0.24	0.09	-0.03	-0.10	-0.02	-0.01	0.02	0.01	-0.06
age	0.20	0.05	-0.03	0.43	0.06	0.47	0.24	1.00	0.07	-0.09	-0.05	-0.02	-0.05	-0.02	0.00	-0.08
married	0.13	0.01	0.02	0.06	-0.02	0.07	0.09	0.07	1.00	-0.08	0.01	-0.05	-0.01	-0.02	-0.01	0.01
black	-0.15	-0.08	-0.29	-0.21	-0.10	-0.02	-0.03	-0.09	-0.08	1.00	0.20	0.07	0.22	0.07	-0.18	-0.17
south	-0.17	-0.03	-0.17	-0.07	-0.09	-0.05	-0.10	-0.05	0.01	0.20	1.00	-0.15	0.05	0.13	-0.15	-0.16
urban	0.21	0.03	0.08	0.13	0.13	-0.06	-0.02	-0.02	-0.05	0.07	-0.15	1.00	-0.04	-0.02	0.09	0.14
sibs	-0.12	-0.02	-0.22	-0.21	-0.19	0.00	-0.01	-0.05	-0.01	0.22	0.05	-0.04	1.00	0.59	-0.30	-0.21
brthord	-0.14	-0.08	-0.17	-0.11	-0.16	0.06	0.02	-0.02	-0.02	0.07	0.13	-0.02	0.59	1.00	-0.31	-0.23
meduc	0.25	0.08	0.34	0.26	0.39	-0.18	0.01	0.00	-0.01	-0.18	-0.15	0.09	-0.30	-0.31	1.00	0.57
feduc	0.28	0.06	0.34	0.23	0.43	-0.25	-0.06	-0.08	0.01	-0.17	-0.16	0.14	-0.21	-0.23	0.57	1.00

Appendix 2: Pair Plot



Appendix 3: VIF Calculation in MLR in Section 2

```
In [30]: train[['educ', 'exper']].corr()
```

Out[30]:

	educ	exper
educ	1.000000	-0.434877
exper	-0.434877	1.000000

```
In [31]: ##VIF
         1/(1-(-0.442997)**2)
```

```
Out[31]: 1.244162300299226
```

Appendix 4: OLS Regression Output for MLR in Section 2

OLS Regression Results

Dep. Variable:	wage	**R-squared:**	0.143
Model:	OLS	**Adj. R-squared:**	0.139
Method:	Least Squares	**F-statistic:**	43.30
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	4.01e-18
Time:	20:14:33	**Log-Likelihood:**	-3861.9
No. Observations:	524	**AIC:**	7730.
Df Residuals:	521	**BIC:**	7743.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-295.9570	146.285	-2.023	0.044	-583.338	-8.576
educ	78.4469	8.433	9.302	0.000	61.880	95.014
exper	19.0792	4.461	4.277	0.000	10.315	27.843

Omnibus:	96.119	**Durbin-Watson:**	1.883
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	183.965
Skew:	1.036	**Prob(JB):**	1.13e-40
Kurtosis:	5.033	**Cond. No.**	157.

Appendix 5: OLS Regression Output for Full Model in Section 3

Dep. Variable:	wage	**R-squared:**	0.258
Model:	OLS	**Adj. R-squared:**	0.236
Method:	Least Squares	**F-statistic:**	11.78
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	4.79e-25
Time:	20:14:34	**Log-Likelihood:**	-3824.0
No. Observations:	524	**AIC:**	7680.
Df Residuals:	508	**BIC:**	7748.
Df Model:	15		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-970.4756	261.812	-3.707	0.000	-1484.843	-456.109
hours	-2.2220	2.254	-0.986	0.325	-6.650	2.207
IQ	3.4937	1.419	2.462	0.014	0.706	6.282
KWW	3.2195	2.835	1.136	0.257	-2.350	8.789
educ	35.2229	10.398	3.388	0.001	14.795	55.651
exper	8.6069	5.186	1.660	0.098	-1.582	18.796
tenure	2.6195	3.376	0.776	0.438	-4.013	9.252
age	16.2755	7.034	2.314	0.021	2.456	30.095
married	169.3542	55.584	3.047	0.002	60.152	278.556
black	-65.2459	65.606	-0.995	0.320	-194.138	63.646
south	-48.0579	36.244	-1.326	0.185	-119.265	23.149
urban	152.7806	36.877	4.143	0.000	80.331	225.231
sibs	10.4937	9.345	1.123	0.262	-7.866	28.853
brthord	-17.1163	13.460	-1.272	0.204	-43.560	9.328
meduc	6.4929	7.204	0.901	0.368	-7.660	20.646
feduc	14.1718	6.274	2.259	0.024	1.846	26.498

Appendix 6: Correlation Calculation of Four Variables in MLR in Section 3

Multicollinearity assessment

```
train[['educ', 'exper', 'IQ', 'age']].corr()
```

	educ	exper	IQ	age
educ	1.000000	-0.434877	0.533993	0.058855
exper	-0.434877	1.000000	-0.213809	0.469856
IQ	0.533993	-0.213809	1.000000	-0.027972
age	0.058855	0.469856	-0.027972	1.000000

Appendix 7: SLR Between Education and IQ

Dep. Variable:	educ	**R-squared:**	0.285
Model:	OLS	**Adj. R-squared:**	0.284
Method:	Least Squares	**F-statistic:**	208.2
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	5.80e-40
Time:	20:14:34	**Log-Likelihood:**	-1072.6
No. Observations:	524	**AIC:**	2149.
Df Residuals:	522	**BIC:**	2158.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.4587	0.578	9.450	0.000	4.324	6.594
IQ	0.0802	0.006	14.430	0.000	0.069	0.091

Omnibus:	13.993	**Durbin-Watson:**	2.033
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	12.679
Skew:	0.323	**Prob(JB):**	0.00177
Kurtosis:	2.596	**Cond. No.**	732.

Appendix 8: OLS Regression Output for MLR with OVBs in Section 3

OLS Regression Results

Dep. Variable:	wage	**R-squared:**	0.187
Model:	OLS	**Adj. R-squared:**	0.181
Method:	Least Squares	**F-statistic:**	29.86
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	2.26e-22
Time:	18:00:07	**Log-Likelihood:**	-3847.9
No. Observations:	524	**AIC:**	7706.
Df Residuals:	519	**BIC:**	7727.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-587.0863	158.560	-3.703	0.000	-898.584	-275.589
educ	50.4082	9.852	5.116	0.000	31.053	69.763
exper	19.9294	4.367	4.563	0.000	11.350	28.509
IQ	4.4561	1.332	3.345	0.001	1.839	7.073
feduc	20.2434	5.619	3.602	0.000	9.204	31.283

Omnibus:	95.592	**Durbin-Watson:**	1.898
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	200.777
Skew:	0.990	**Prob(JB):**	2.52e-44
Kurtosis:	5.297	**Cond. No.**	1.02e+03

Appendix 9: Forward Selection 1 Output

```
[60]: model1_fs = forward_selected(train[['wage','hours','IQ','KWW','educ','exper','tenure','age',

you nominated variable(s) ['educ', 'exper'], the adj_r2 is: 0.139237
adding urban increases adj_r2 from 0.139237 to 0.167961
adding IQ increases adj_r2 from 0.167961 to 0.189929
adding married increases adj_r2 from 0.189929 to 0.205582
adding feduc increases adj_r2 from 0.205582 to 0.219864
adding age increases adj_r2 from 0.219864 to 0.234519
adding south increases adj_r2 from 0.234519 to 0.237232
adding KWW increases adj_r2 from 0.237232 to 0.237862
adding meduc increases adj_r2 from 0.237862 to 0.237994
final model is wage ~ educ + exper + urban + IQ + married + feduc + age + south + KWW + med
uc + 1, with adj_r2 of 0.237994
```

Appendix 10: Output for Transformation on Experience

1. Tranformation model: wage ~ log_exper

```
train['log_exper'] = np.log(train['exper'])

model_fs2_tran1 = smf.ols(formula='wage ~ educ + log_exper + urban + IQ + married + feduc +
reg_fs2_tran1 = model_fs2_tran1.fit()
resid_fs2_tran1 = reg_fs2_tran1.resid
fitted_fs2_tran1 = reg_fs2_tran1.fittedvalues
reg_fs2_tran1.summary()
```

OLS Regression Results

Dep. Variable:	wage	**R-squared:**	0.243
Model:	OLS	**Adj. R-squared:**	0.232
Method:	Least Squares	**F-statistic:**	23.61
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	7.17e-28
Time:	20:14:37	**Log-Likelihood:**	-3829.4
No. Observations:	524	**AIC:**	7675.
Df Residuals:	516	**BIC:**	7709.
Df Model:	7		
Covariance Type:	nonrobust		

Appendix 11: Log_wage FS2 Full Model with Bias Correction

```
In [108]: # SER, R2, and adj-R2 for full lwage-IQ (exp)
          res_fs_lwage_IQ = train['wage']-TS_fs_lwage_IQ
          np.sqrt(sum(res_fs_lwage_IQ**2)/(n-12)), 1 - sum(res_fs_lwage_IQ**2)/((n-1)*172435.932377796
Out[108]: (366.8322076844856, 0.23603155384348262, 0.21961816925808875)

In [109]: # SER, R2, and adj-R2 for full lwage-IQ (Duan)
          res_fs_lwage_IQ1 = train['wage']-TS_fs_lwage_IQ1
          np.sqrt(sum(res_fs_lwage_IQ1**2)/(n-12)), 1 - sum(res_fs_lwage_IQ1**2)/((n-1)*172435.9323777
Out[109]: (362.11534584021683, 0.2555520090624007, 0.2395580092571008)

In [110]: # SER, R2, and adj-R2 for full lwage-IQ (normal)
          res_fs_lwage_IQ11 = train['wage']-TS_fs_lwage_IQ11
          np.sqrt(sum(res_fs_lwage_IQ11**2)/(n-12)), 1 - sum(res_fs_lwage_IQ11**2)/((n-1)*172435.93237
Out[110]: (362.14755249482795, 0.2554195802889724, 0.2394227353342433)
```

Appendix 12: Log_wage FS2 Reduced Model with Bias Correction

```
# SER, R2, and adj-R2 for lwage-IQ-r (exp)
res_fs_lwage_IQ_r = train['wage']-TS_fs_lwage_r_IQ
np.sqrt(sum(res_fs_lwage_IQ_r**2)/(n-9)), 1 - sum(res_fs_lwage_IQ_r**2)/((n-1)*172435.932377
```

```
(366.1861980266392, 0.23425933172467406, 0.22236433105243603)
```

```
# SER, R2, and adj-R2 for lwage-IQ (Duan)
res_fs_lwage_IQ1_r = train['wage']-TS_fs_lwage_r_IQ1
np.sqrt(sum(res_fs_lwage_IQ1_r**2)/(n-9)), 1 - sum(res_fs_lwage_IQ1_r**2)/((n-1)*172435.9323
```

```
(361.13352870156433, 0.2552450584839846, 0.2436760496837359)
```

```
# SER, R2, and adj-R2 for lwage-IQ (normal)
res_fs_lwage_IQ11_r = train['wage']-TS_fs_lwage_r_IQ11
np.sqrt(sum(res_fs_lwage_IQ11_r**2)/(n-9)), 1 - sum(res_fs_lwage_IQ11_r**2)/((n-1)*172435.93
```

```
(361.1682313025967, 0.2551019193217976, 0.2435306870005829)
```

Appendix 13: Interaction Effect Age_Married

Dep. Variable:	wage	**R-squared:**	0.248
Model:	OLS	**Adj. R-squared:**	0.237
Method:	Least Squares	**F-statistic:**	22.14
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	4.47e-29
Time:	20:14:39	**Log-Likelihood:**	-3827.4
No. Observations:	524	**AIC:**	7673.
Df Residuals:	515	**BIC:**	7711.
Df Model:	8		
Covariance Type:	HC0		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-321.4548	653.012	-0.492	0.623	-1601.334	958.424
educ	37.5263	9.689	3.873	0.000	18.536	56.517
exper	9.6730	5.823	1.661	0.097	-1.739	21.085
urban	157.3959	34.372	4.579	0.000	90.028	224.763
IQ	4.4832	1.138	3.938	0.000	2.252	6.714
married	-824.7528	659.266	-1.251	0.211	-2116.891	467.385
feduc	18.9642	5.188	3.656	0.000	8.796	29.132
age	-7.7482	19.482	-0.398	0.691	-45.932	30.436
age_married	30.9379	19.676	1.572	0.116	-7.627	69.502

Appendix 14: Interaction Effect Urban_Feduc

Dep. Variable:	wage	**R-squared:**	0.251
Model:	OLS	**Adj. R-squared:**	0.239
Method:	Least Squares	**F-statistic:**	21.52
Date:	Fri, 24 May 2024	**Prob (F-statistic):**	2.64e-28
Time:	20:14:39	**Log-Likelihood:**	-3826.4
No. Observations:	524	**AIC:**	7671.
Df Residuals:	515	**BIC:**	7709.
Df Model:	8		
Covariance Type:	HC0		

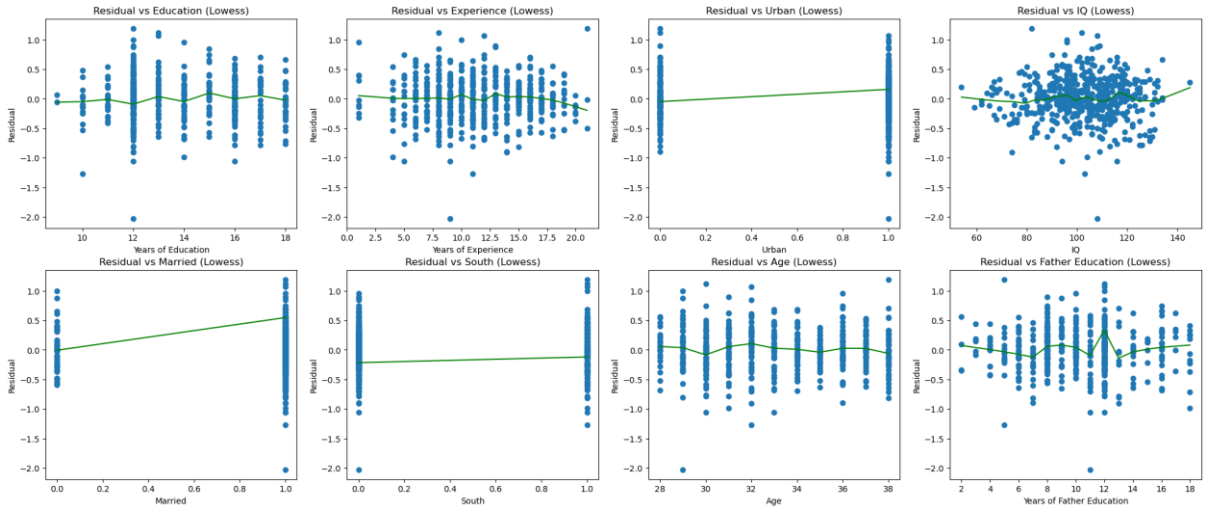
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-1086.3475	228.423	-4.756	0.000	-1534.048	-638.648
educ	35.5013	9.714	3.655	0.000	16.462	54.540
exper	8.9526	5.755	1.556	0.120	-2.327	20.233
urban	-70.6063	111.748	-0.632	0.527	-289.628	148.415
IQ	4.7472	1.115	4.257	0.000	2.561	6.933
married	170.5651	52.267	3.263	0.001	68.125	273.006
feduc	1.7784	9.800	0.181	0.856	-17.430	20.987
age	21.3998	6.774	3.159	0.002	8.123	34.676
urban_feduc	23.3957	11.367	2.058	0.040	1.118	45.674

Appendix 15: Forward Selection Output with Exper_South

```
model6_fs = forward_selected(train[['wage','hours','IQ','KWW','educ','exper','tenure','age',

you nominated variable(s) ['exper', 'educ'], the adj_r2 is: 0.139237
adding urban increases adj_r2 from 0.139237 to 0.167961
adding IQ increases adj_r2 from 0.167961 to 0.189929
adding married increases adj_r2 from 0.189929 to 0.205582
adding feduc increases adj_r2 from 0.205582 to 0.219864
adding age increases adj_r2 from 0.219864 to 0.234519
adding south increases adj_r2 from 0.234519 to 0.237232
adding exper_south increases adj_r2 from 0.237232 to 0.241227
adding KWW increases adj_r2 from 0.241227 to 0.242249
adding meduc increases adj_r2 from 0.242249 to 0.242360
final model is wage ~ exper + educ + urban + IQ + married + feduc + age + south + exper_sou
th + KWW + meduc + 1, with adj_r2 of 0.242360
```

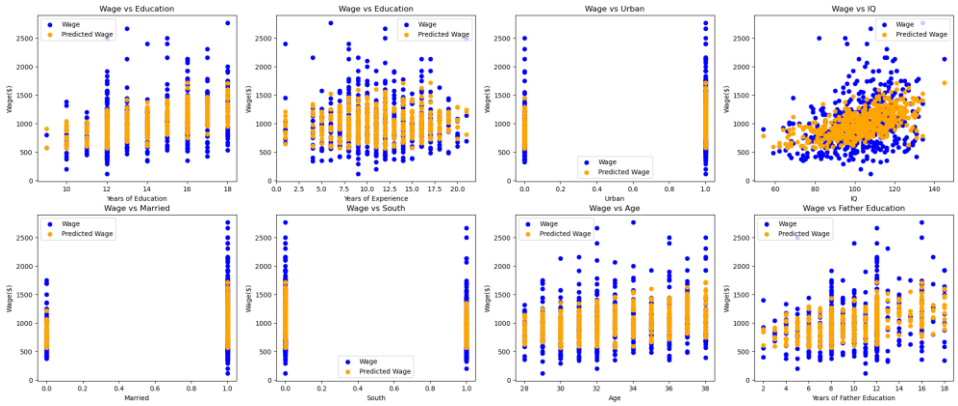
Appendix 16: Optimal Model Residual vs Regressors



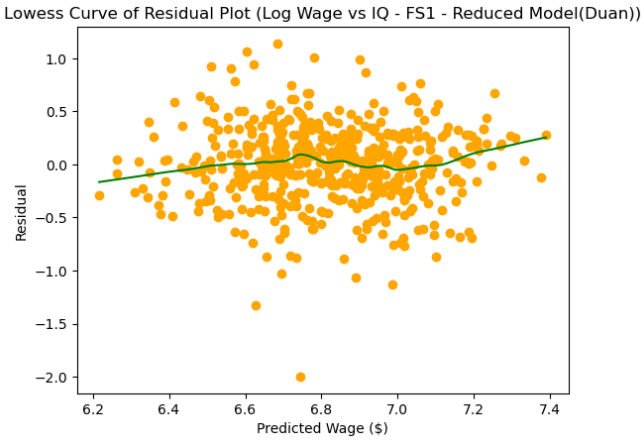
Appendix 17: Optimal Model Correlation Table

	educ	exper	urban	IQ	married	south	age	feduc
educ	1.000000	-0.434877	0.125816	0.533993	-0.017682	-0.090896	0.058855	0.430178
exper	-0.434877	1.000000	-0.059788	-0.213809	0.070744	-0.053001	0.469856	-0.248742
urban	0.125816	-0.059788	1.000000	0.079964	-0.047287	-0.145016	-0.023418	0.135768
IQ	0.533993	-0.213809	0.079964	1.000000	0.017629	-0.174662	-0.027972	0.342669
married	-0.017682	0.070744	-0.047287	0.017629	1.000000	0.005519	0.070750	0.006229
south	-0.090896	-0.053001	-0.145016	-0.174662	0.005519	1.000000	-0.053355	-0.159811
age	0.058855	0.469856	-0.023418	-0.027972	0.070750	-0.053355	1.000000	-0.080866
feduc	0.430178	-0.248742	0.135768	0.342669	0.006229	-0.159811	-0.080866	1.000000

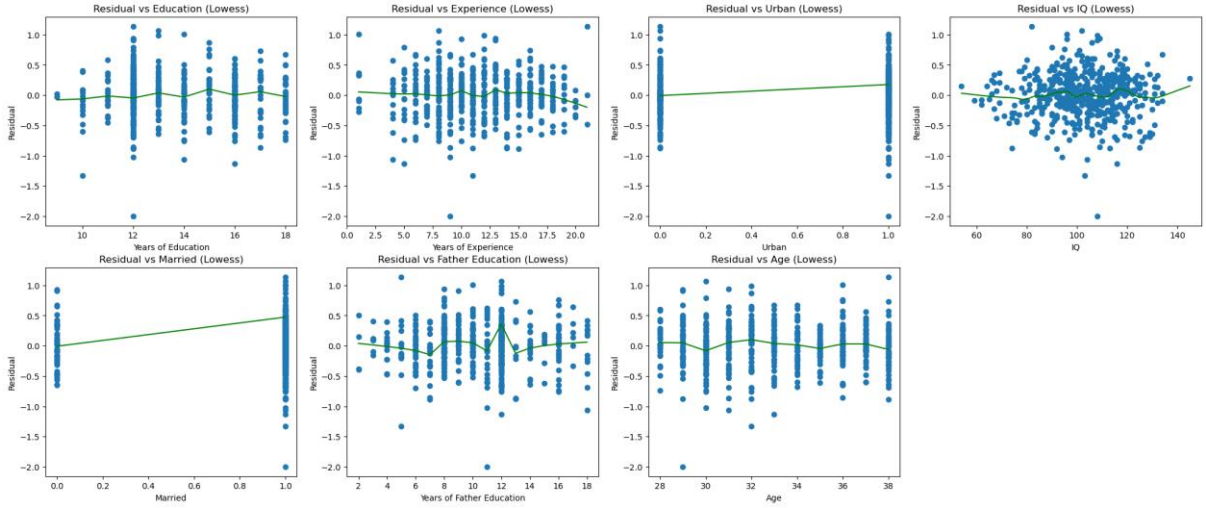
Appendix 18: Optimal Model Goodness of Fit



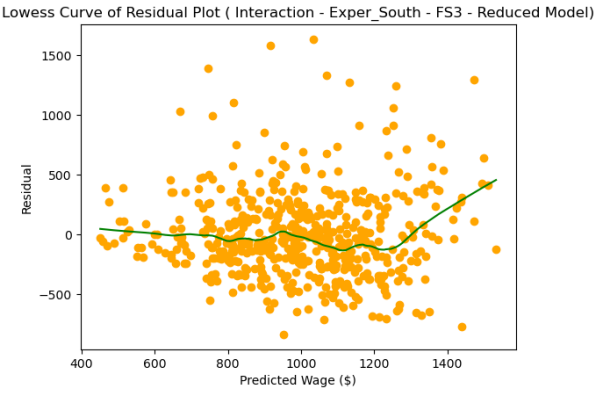
Appendix 19: Lowess Curve of Residual Plot (Log Wage vs IQ - FS1 - Reduced Model(Duan))



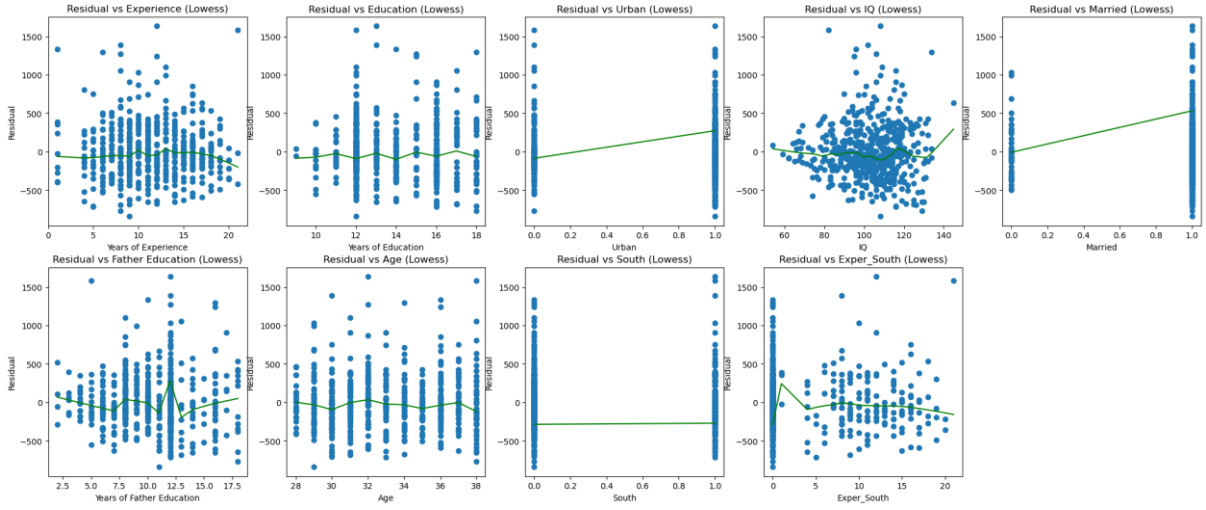
Appendix 20: Residual vs Regressors (Log Wage vs IQ - FS1 - Reduced Model(Duan))



Appendix 21: Lowess Curve of Residual Plot (Interaction - Exper_South - FS3 - Reduced Model)



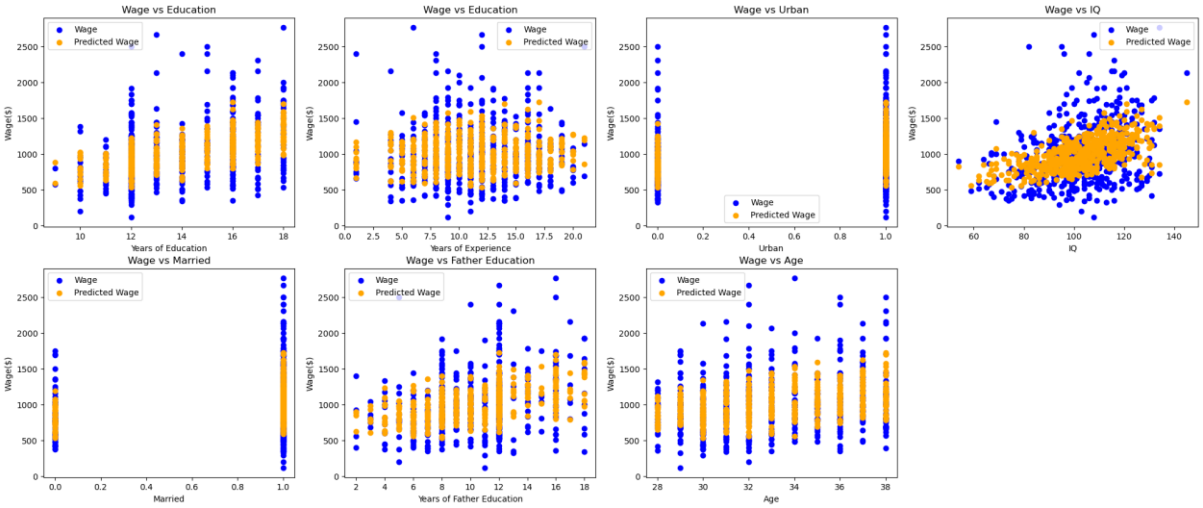
Appendix 22: Residual vs Regressors (Interaction - Exper_South - FS3 - Reduced Model)



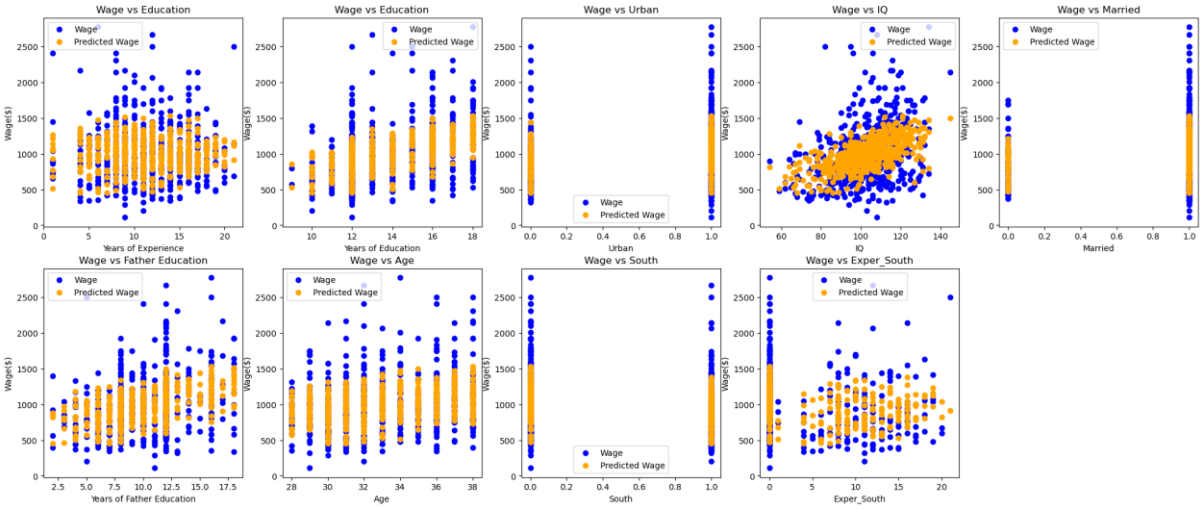
Appendix 23: Correlation Table (Interaction - Exper_South - FS3 - Reduced Model)

	exper	educ	urban	IQ	married	feduc	age	south	exper_south
exper	1.000000	-0.434877	-0.059788	-0.213809	0.070744	-0.248742	0.469856	-0.053001	0.182541
educ	-0.434877	1.000000	0.125816	0.533993	-0.017682	0.430178	0.058855	-0.090896	-0.188581
urban	-0.059788	0.125816	1.000000	0.079964	-0.047287	0.135768	-0.023418	-0.145016	-0.111809
IQ	-0.213809	0.533993	0.079964	1.000000	0.017629	0.342669	-0.027972	-0.174662	-0.202685
married	0.070744	-0.017682	-0.047287	0.017629	1.000000	0.006229	0.070750	0.005519	0.033999
feduc	-0.248742	0.430178	0.135768	0.342669	0.006229	1.000000	-0.080866	-0.159811	-0.219925
age	0.469856	0.058855	-0.023418	-0.027972	0.070750	-0.080866	1.000000	-0.053355	0.051083
south	-0.053001	-0.090896	-0.145016	-0.174662	0.005519	-0.159811	-0.053355	1.000000	0.911239
exper_south	0.182541	-0.188581	-0.111809	-0.202685	0.033999	-0.219925	0.051083	0.911239	1.000000

Appendix 24: Wage vs Regressors (Log Wage vs IQ - FS1 - Reduced Model(Duan))



Appendix 25: Wage vs Regressors (Interaction - Exper_South - FS3 - Reduced Model)



Reference List

Amesti, J., & Claro, S. (2021). Effects of Apprenticeship on the Short-Term Educational Outcomes of Vocational High-School Students. *Journal of Research on Educational Effectiveness*, *14*(3), 1–19. https://doi.org/10.1080/19345747.2021.1917026

Brower, T., PhD. (2022, January 13). Companies are offering Education Assistance—And it's the perk for happy work. *Forbes*. https://www.forbes.com/sites/tracybrower/2021/09/12/companies-are-offering-education-assistance-and-its-the-perk-for-happy-work/?sh=3f7721de39dd

Hanson, M. (2023, April 1). *Effects of cancelling student loan debt (2023): Short & Long-Term*. Education Data Initiative. https://educationdata.org/what-happens-if-student-loan-debt-is-canceled