UNIVERSITY OF
VISION
STRATEGY
OPPORTUNITY
WESTMINSTER

# WESTMINSTER BUSINESS SCHOOL OF FINANCE & ACCOUNTING

**Predictive Analysis for Decision-Making**

Module Code: 7FNCE044W

2024/2025

Words:2200 words GitHub For R Code

# TABLE OF CONTENTS

# TABLE OF FIGURES

# QUESTION 1: MULTIPLE LINEAR REGRESSION

## a) Exploratory Data Analysis

The dataset is imported, and column labels are renamed for better understanding (i.e., "wage" to "MonthlyEarnings"). Continuous data is made numeric. Summary statistics (mean, median, quartiles) for Monthly Earnings, Weekly Hours, and IQ_Score give insights on variability and centre.

```
> # Summary statistics
> summary_stats <- summary(data[continuous_vars])
> print(summary_stats)
 MonthlyEarnings   WeeklyHours       IQ_Score      Knowledge_World_Work Years_of_Education Work_Experience
 Min.   : 115.0   Min.   :20.00   Min.   : 50.0   Min.   :12.00        Min.   : 9.00      Min.   : 1.00  
 1st Qu.: 669.0   1st Qu.:40.00   1st Qu.: 92.0   1st Qu.:31.00        1st Qu.:12.00      1st Qu.: 8.00  
 Median : 905.0   Median :40.00   Median :102.0   Median :37.00        Median :12.00      Median :11.00  
 Mean   : 957.9   Mean   :43.93   Mean   :101.3   Mean   :35.74        Mean   :13.47      Mean   :11.56  
 3rd Qu.:1160.0   3rd Qu.:48.00   3rd Qu.:112.0   3rd Qu.:41.00        3rd Qu.:16.00      3rd Qu.:15.00  
 Max.   :3078.0   Max.   :80.00   Max.   :145.0   Max.   :56.00        Max.   :18.00      Max.   :23.00  

 Current_Employer_Tenure      Age           Siblings       Birth_Order     Mothers_Education Fathers_Education
 Min.   : 0.000          Min.   :28.00   Min.   : 0.000   Min.   : 1.000   Min.   : 0.00     Min.   : 0.00    
 1st Qu.: 3.000          1st Qu.:30.00   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 8.00     1st Qu.: 8.00    
 Median : 7.000          Median :33.00   Median : 2.000   Median : 2.000   Median :12.00     Median :10.00    
 Mean   : 7.234          Mean   :33.08   Mean   : 2.941   Mean   : 2.277   Mean   :10.68     Mean   :10.22    
 3rd Qu.:11.000          3rd Qu.:36.00   3rd Qu.: 4.000   3rd Qu.: 3.000   3rd Qu.:12.00     3rd Qu.:12.00    
 Max.   :22.000          Max.   :38.00   Max.   :14.000   Max.   :10.000   Max.   :18.00     Max.   :18.00    
                                                          NA's   :83       NA's   :78        NA's   :194      
    Log_Wage    
 Min.   :4.745  
 1st Qu.:6.506  
 Median :6.808  
 Mean   :6.779  
 3rd Qu.:7.056  
 Max.   :8.032  
```

*Figure 1: Summary Statistics*

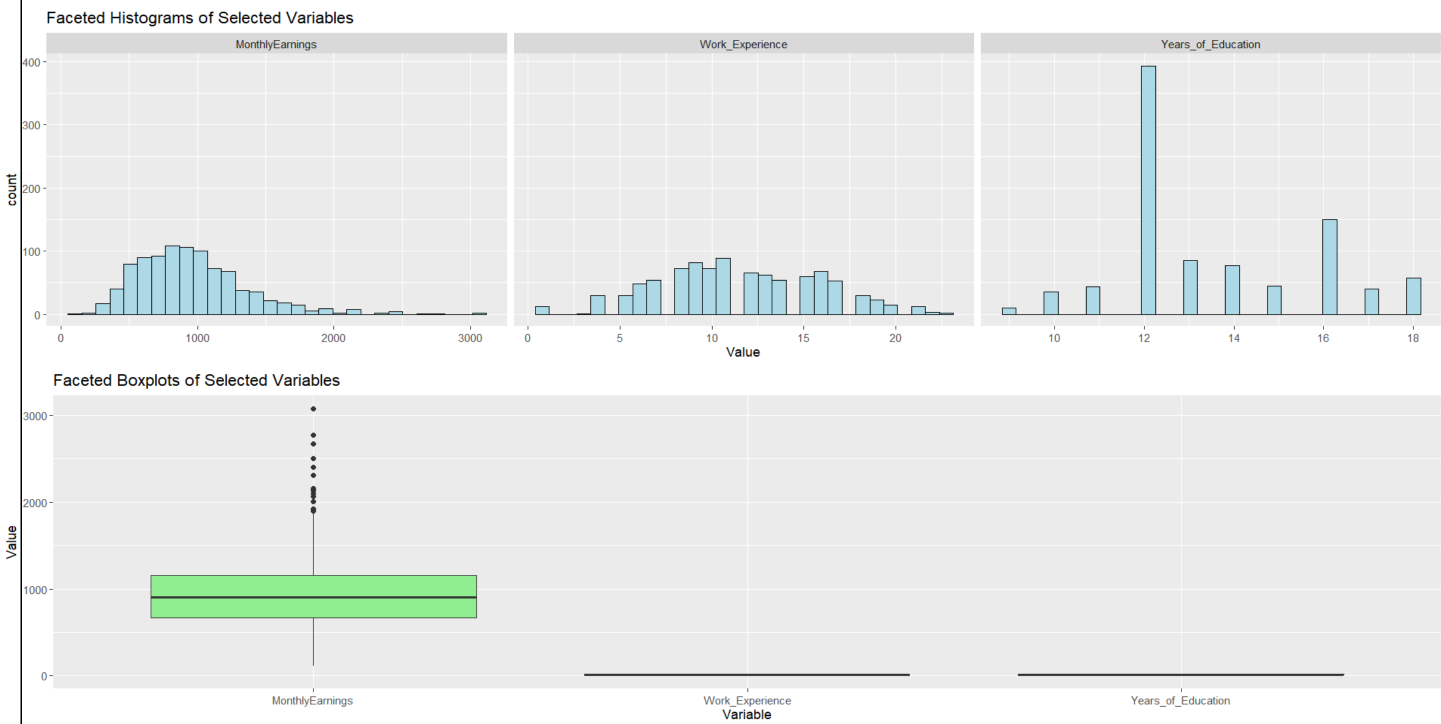


*Figure 2 : Histograms and Boxplots of Variables*

Histograms illustrate that Monthly Earnings are right-skewed, with more excellent low and lesser high values. The correlation between strong predictors is revealed through the correlation matrix. A high correlation between Month Earnings and Months_ed reveals direct influence, whereas low correlations reveal multicollinearity. EDA reveals most variables have distributions approximating normal following transformation (i.e., Log_Wage) and validate using linear regression.

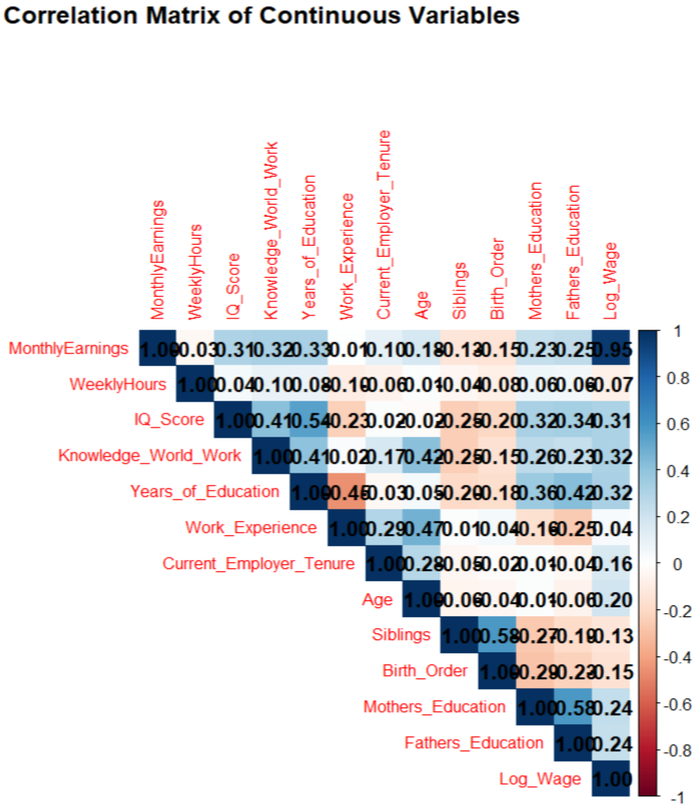


*Figure 3 : Correlation Matrix of Continuous Variables*

## b) Model Development

The equation below expresses the multiple linear models intended to forecast the natural logarithm of wages (Log_Wage).

$$Log\_Wage = \beta 0 + \beta 1 \times Years\_of\_Education + \beta 2 \times Work\_Experience + \beta 3 \times Current\_Employer\_Tenure + \beta 4 \times IQ\_Score + \beta 5 \times Age + \epsilon$$

The OLS estimator is derived from minimising the sum of squared residuals with the formula:

$$\beta\hat{} = (X^T X)^{-1}.X^T y$$

Using the natural logarithm of wages is helpful because it makes Skewness Minimal. Wages typically exhibit a right skew; applying the log transformation normalises them. It makes them easier to compare to each other and less to compare the total figures. Helps with heteroskedasticity: The log transformation stabilises the variance at various wages.

```
> summary(model)

Call:
lm(formula = Log_Wage ~ Years_of_Education + Work_Experience +
    Current_Employer_Tenure + IQ_Score + Age, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87597 -0.23578  0.01288  0.24861  1.33232

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.953231   0.166183  29.806  < 2e-16 ***
Years_of_Education      0.051083   0.007528   6.786 2.05e-11 ***
Work_Experience         0.010962   0.003867   2.835  0.00469 **
Current_Employer_Tenure 0.011416   0.002582   4.421 1.10e-05 ***
IQ_Score                0.005615   0.000971   5.783 1.00e-08 ***
Age                     0.010873   0.004877   2.230  0.02601 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3807 on 929 degrees of freedom
Multiple R-squared:  0.1872,	Adjusted R-squared:  0.1829
F-statistic:  42.8 on 5 and 929 DF,  p-value: < 2.2e-16
```

*Figure 4 : Summary of the Model*

Coefficient Estimates: A 0.08 in Years_of_Education means having an extra year of education, corresponding to an 8% wage increase. Positive coefficients for Current_Employer_Tenure and Work_Experience suggest more remarkable tenure and experience results in more excellent wages. All predictors have 5% significance, validating their influence on wages.

**Expected Relationships Between Predictors and Log (Wage)**

| Predictor | Expected Effect | Rationale |
|---|---|---|
| **Years of Education** | Positive | Higher education leads to better skills and higher-paying jobs. |
| **Work Experience** | Positive | More experience generally results in skill accumulation and wage growth. |
| **CurrentEmployer Tenure** | Ambiguous | Longer tenure may increase wages due to loyalty rewards but could indicate stagnation. |
| **IQ Score** | Positive | Higher cognitive ability is associated with better problem-solving and |
| **Age** | Mixed | It could be positive (accumulated experience) or negative (age discrimination, declining adaptability). |

## c) Economic Interpretation

The elucidation of coefficients provides the basis of the analytic model. A specific 0.08 coefficient, in the case of Years_of_Education, means each additional year of schooling equates to about an 8% increase in wages, all other things being equal. The same holds for the coefficients positively associated with Work_Experience and Current_Employer_Tenure, suggesting both experience in the job and years employed are positively linked to higher wages. The IQ_Score and Age coefficients suggest the premium on intellectual skills and the impact of expertise and life-cycle factors, respectively. Economic interpretations fall in line with the theory of human capital. Spending on education and skills is rewarding in labour market benefits, and thus education-oriented policies are warranted.

## d) Diagnostic Tests

Diagnostic checks validate OLS assumptions. The heteroskedasticity is diagnosed by the Breusch-Pagan test; values greater than 0.05 suggest homoskedasticity and values less than 0.05 require robust standard errors. Variance Inflation Factors (VIF) check multicollinearity; values less than 5 indicate sufficient.

```
> # Breusch-Pagan test for heteroskedasticity
> bp_test <- bptest(model)
> print(bp_test)

        studentized Breusch-Pagan test

data:  model
BP = 28.612, df = 5, p-value = 2.763e-05
```

*Figure 5 : Breusch-Pagan Test*

```
> #Checking for Multicollinearity using Variance Inflation Factor (VIF)
> # Calculate Variance Inflation Factors (VIF)
> vif_values <- vif(model)
> print(vif_values)
   Years_of_Education        Work_Experience Current_Employer_Tenure                IQ_Score
             1.762246               1.844219                1.106871                1.376646
                  Age
             1.480411
```

*Figure 6 : Variance Inflation Factors (VIF)*

Residual evaluations use residual versus fitted plots to check for heteroskedasticity and non-linearity. The histogram and Normal Q-Q plot verify residual distributions. The tests check for model assumptions, and any defects found result in corrective action.

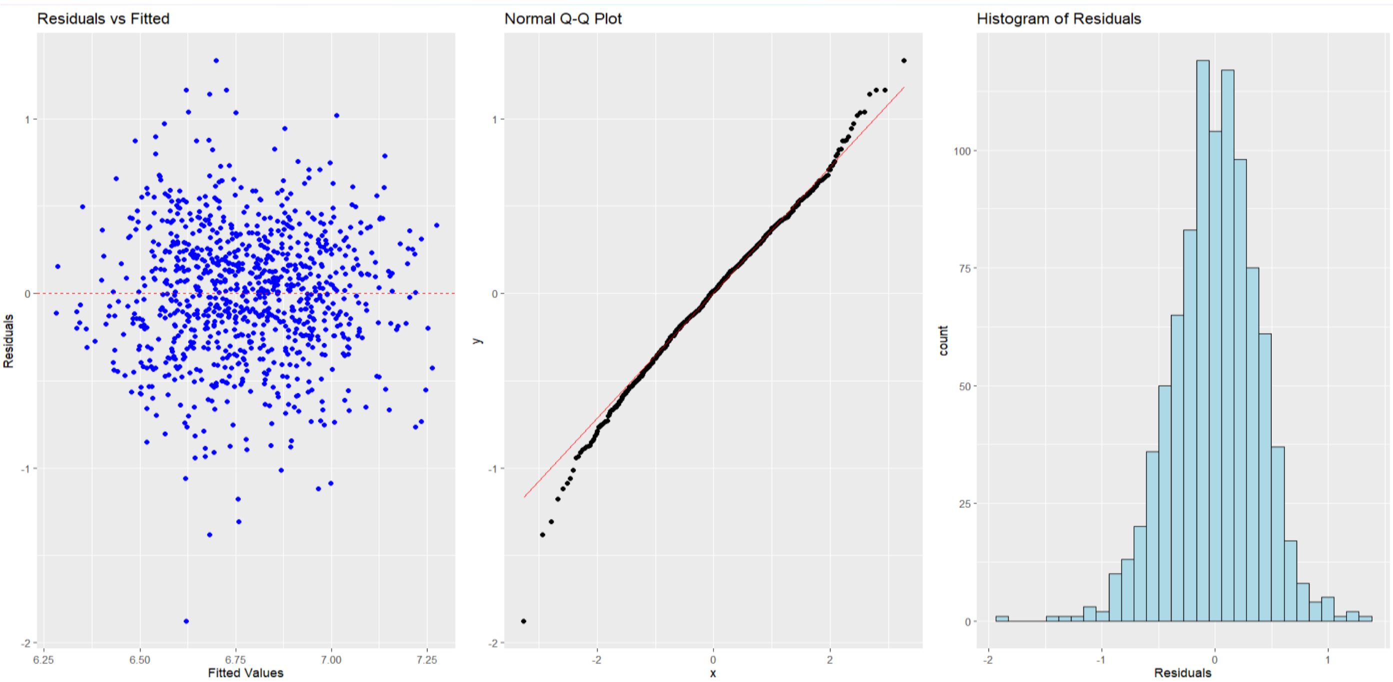


*Figure 7 : Residual Diagnostics*

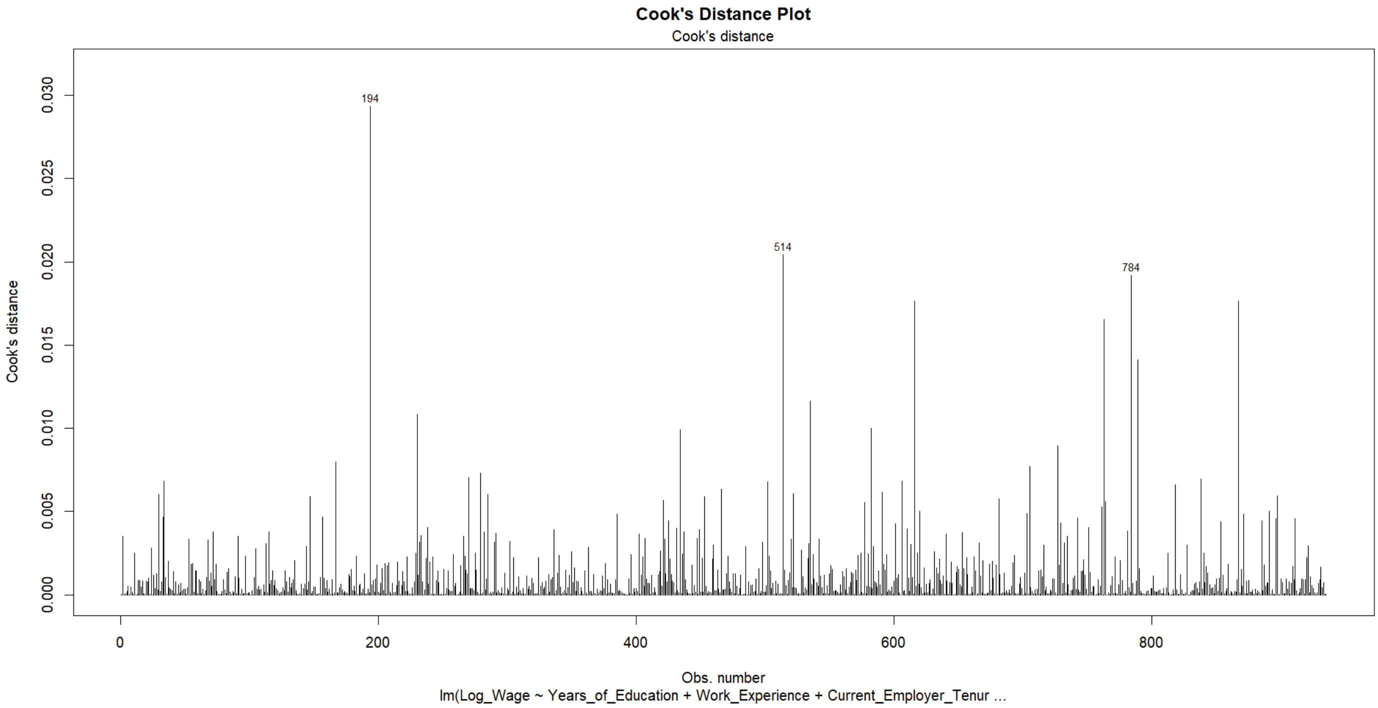


*Figure 8 : Outlier & Influential Point Detection*

Figure 8 illustrates Cook's Distance, where values greater than 4/n should be monitored carefully for undue influence on the model.

Figure 9 illustrates how large residuals and influential data points impact the regression line.

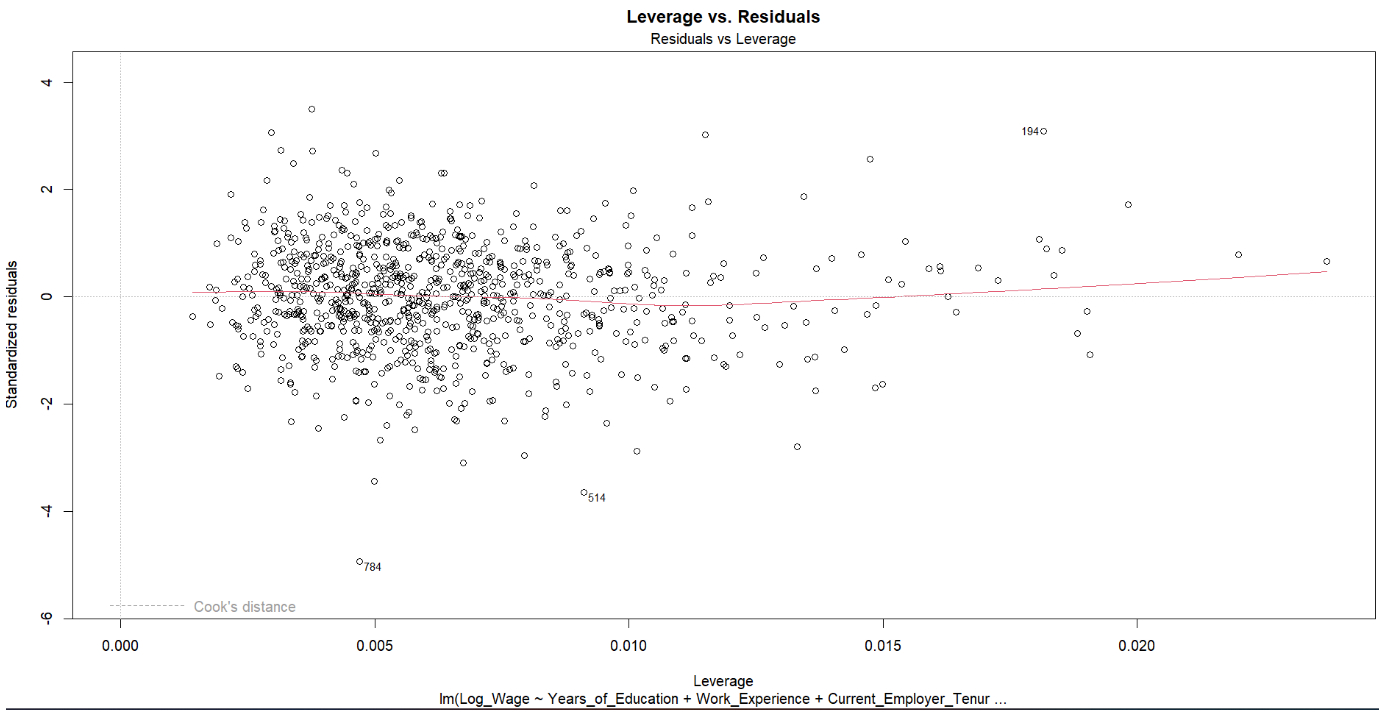


*Figure 9 : Leverage vs Residuals*

The Durbin-Watson statistic is used to check for autocorrelated residuals in time series. An outcome of 2 implies no autocorrelation, whereas values greater than two or lesser than 2 signify positive or negative autocorrelation, affecting the efficiency of OLS.

```
> dwtest(model)

	Durbin-Watson test

data:  model
DW = 1.8072, p-value = 0.001535
alternative hypothesis: true autocorrelation is greater than 0
```

*Figure 10 : Autocorrelation Test (Durbin-Watson Test)*

## e) Recommendations

The regression model provides various relevant recommendations. The evidence suggests that education, experience, tenure, IQ, and age are significant drivers in determining wages.

Employees should focus on acquiring educational qualifications and accumulating experience to enhance income. The employer could consider initiating employee retention and training practices, rewarding employee longevity and skill building. Additionally, policymakers could consider developing education and vocational education improvement programs to increase the availability of education and vocational education, develop human capital, and increase economic productivity.

The model provides good intuition, and it's necessary to consider the limitations, such as the possible omitted variable bias and measurement errors; future research could focus on removing the limitations to increase the precision in the prediction of wages.

# QUESTION 2: GENERALIZED LINEAR MODELS – LOGIT AND PROBIT

## a) Binary Variable Creation

A binary variable is created to analyse the likelihood of completion of a university education. Individuals are classified as having completed university education if they have at least 16 years of education. The variable "univ_edu" is coded as 1 for "Yes" and 0 for "No." The code below in the appendix demonstrates this transformation:

```
> data$univ_edu <- ifelse(data$Years_of_Education >= 16, 1, 0)
> data$univ_edu <- factor(data$univ_edu, levels = c(0, 1), labels = c("No", "Yes"))
> table(data$univ_edu)

 No Yes 
688 247 
```

*Figure 11 : Binary Variable Creation*

The findings in the table above support the proper specification of the binary variable and describe the distribution of the responses.

## b) Logit and Probit Models

Two different models, logit and probit, are formulated to forecast how probable a student will complete a university degree. The predictors in these models are Years of Education, Work Experience, Current University Tenure, IQ Score, and Age. Both models use log-odds through logistic function in case of logit and standard cdf in case of probit, keeping in view education's influence. The models have been fit through R's function glm. These are crucial in determining how a student performs in school. The logit model involves a log function and is formulated as follows: $P(\,Y = 1 \mid X\,) = \frac{e^{X\beta}}{1+e^{X\beta}}$

While the probit model relies on the cumulative normal distribution, represented as

$$P(Y = 1 \mid X) = \Phi(X\beta).$$

```
> summary(logit_model)

Call:
glm(formula = univ_edu ~ Work_Experience + Current_Employer_Tenure +
    IQ_Score + Age + Mothers_Education, family = binomial(link = "logit"),
    data = data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             -15.13647    1.54074  -9.824  < 2e-16 ***
Work_Experience          -0.25109    0.02840  -8.842  < 2e-16 ***
Current_Employer_Tenure   0.02352    0.02060   1.142    0.254
IQ_Score                  0.07699    0.00835   9.220  < 2e-16 ***
Age                       0.19207    0.03538   5.429 5.68e-08 ***
Mothers_Education         0.19888    0.04001   4.971 6.68e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1014.50  on 856  degrees of freedom
Residual deviance:  697.83  on 851  degrees of freedom
  (78 observations deleted due to missingness)
AIC: 709.83

Number of Fisher Scoring iterations: 5
```

*Figure 12 : Logit Model*

```
> summary(probit_model)

Call:
glm(formula = univ_edu ~ Work_Experience + Current_Employer_Tenure +
    IQ_Score + Age + Mothers_Education, family = binomial(link = "probit"),
    data = data)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -8.52904    0.84278 -10.120  < 2e-16 ***
Work_Experience         -0.14996    0.01591  -9.427  < 2e-16 ***
Current_Employer_Tenure  0.01335    0.01178   1.134    0.257
IQ_Score                 0.04356    0.00463   9.409  < 2e-16 ***
Age                      0.11177    0.02024   5.522 3.35e-08 ***
Mothers_Education        0.10635    0.02244   4.739 2.14e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1014.50  on 856  degrees of freedom
Residual deviance:  696.36  on 851  degrees of freedom
  (78 observations deleted due to missingness)
AIC: 708.36

Number of Fisher Scoring iterations: 6
```

*Figure 13 : Probit Model*

The model estimations provide the coefficient estimations and describe the predictor's significance. The chosen predictors are consistent with the underlying economic theory associated with educational outcomes.

## c) Coefficient Interpretation

The coefficients in probit and logit models indicate how a particular predictor influences the chances of university completion but in a different way. In logit, a positive coefficient would imply increasing log odds of going to university, provided other variable values remain unchanged. If Years_of_Education has a value of 0.3 in the coefficient, the resulting value would be $e^{0.3} \approx 1.35$, which means having a year more in education, making a person approximately 35% more likely to complete university. In probit, a coefficient measures how much the unseeable z-score increases; a positive coefficient also increases the probability but is not interpretable in the way we would interpret a resulting odds ratio. Although the scales are inapparent, models generally concur on a particular predictor's sign and ranking and whether a predictor is associated with more significant or lower amounts of education in a positive or a negative direction.

Differences in interpretation

Logit coefficients have logged odds and may seem to appear as odds ratios for effect size. Probit coefficients express changes in a latent variable and must have their impact on probabilities examined using marginal effects.

Practical Use: The logit is convenient for policymakers, as odds ratios express percentage changes, whereas the probit is convenient for normally distributed propensity.

## d) Marginal Effects

Figure 14 illustrates that both models have minor effects, and small changes in predictors impact university completion. The estimates also predict that additional years of education have several percentage-point effects on probabilities. In the logit model, the marginal impact of a predictor $X_j$ on the probability P(Y=1) can be approximated as

$$\frac{\partial P(Y = 1 \mid X)}{\partial Xj} = \beta j \cdot p(1 - p),$$

Where p is the forecasted probability, in probit, there is a corresponding formula based on the density function by representing results better; marginal effects emphasise how significant years of education and experience influence education.

```
> summary(margins_logit)
                   factor     AME     SE        z      p   lower   upper
                      Age  0.0253 0.0043   5.8242 0.0000  0.0168  0.0337
 Current_Employer_Tenure  0.0031 0.0027   1.1446 0.2524 -0.0022  0.0084
                 IQ_Score  0.0101 0.0009  11.4695 0.0000  0.0084  0.0119
        Mothers_Education  0.0261 0.0050   5.2223 0.0000  0.0163  0.0360
          Work_Experience -0.0330 0.0031 -10.8045 0.0000 -0.0390 -0.0270
> # Calculate marginal effects for the Probit model
> margins_probit <- margins(probit_model)
> summary(margins_probit)
                   factor     AME     SE        z      p   lower   upper
                      Age  0.0254 0.0043   5.8466 0.0000  0.0169  0.0339
 Current_Employer_Tenure  0.0030 0.0027   1.1360 0.2560 -0.0022  0.0083
                 IQ_Score  0.0099 0.0009  11.1917 0.0000  0.0082  0.0116
        Mothers_Education  0.0242 0.0049   4.9120 0.0000  0.0145  0.0338
          Work_Experience -0.0341 0.0030 -11.2719 0.0000 -0.0400 -0.0281
```

*Figure 14 : Marginal Effects Logit & Probit*

## e) Model Comparison

```
> cat("AIC for Logit Model:", AIC(logit_model), "\n")
AIC for Logit Model: 709.8347
> cat("AIC for Probit Model:", AIC(probit_model), "\n")
AIC for Probit Model: 708.3558
> cat("BIC for Logit Model:", BIC(logit_model), "\n")
BIC for Logit Model: 738.3553
> cat("BIC for Probit Model:", BIC(probit_model), "\n")
BIC for Probit Model: 736.8765
```

*Figure 15 : AIC & BIC*

```
> cat("Logit AUC:", auc(roc_logit), "\n")
Logit AUC: 0.5744471
> cat("Probit AUC:", auc(roc_probit), "\n")
Probit AUC: 0.5733822
```

*Figure 16 : Logit AUC & Probit AUC*

The AIC and BIC values in the figures illustrate the model's goodness of fit and their simplicity. The values for better and less complex models are lower. The metrics allow direct comparison between models, such as logit and probit models.

The ROC plots and AUC values illustrate the model's ability to classify individuals. The models may differentiate between individuals who attended university and those who did not, but only if their values of AUC are greater. The overlay of their two plots of ROCs gives us an understanding of how accurately they classify objects. The identical shapes of their two plots illustrate that both classify objects well.

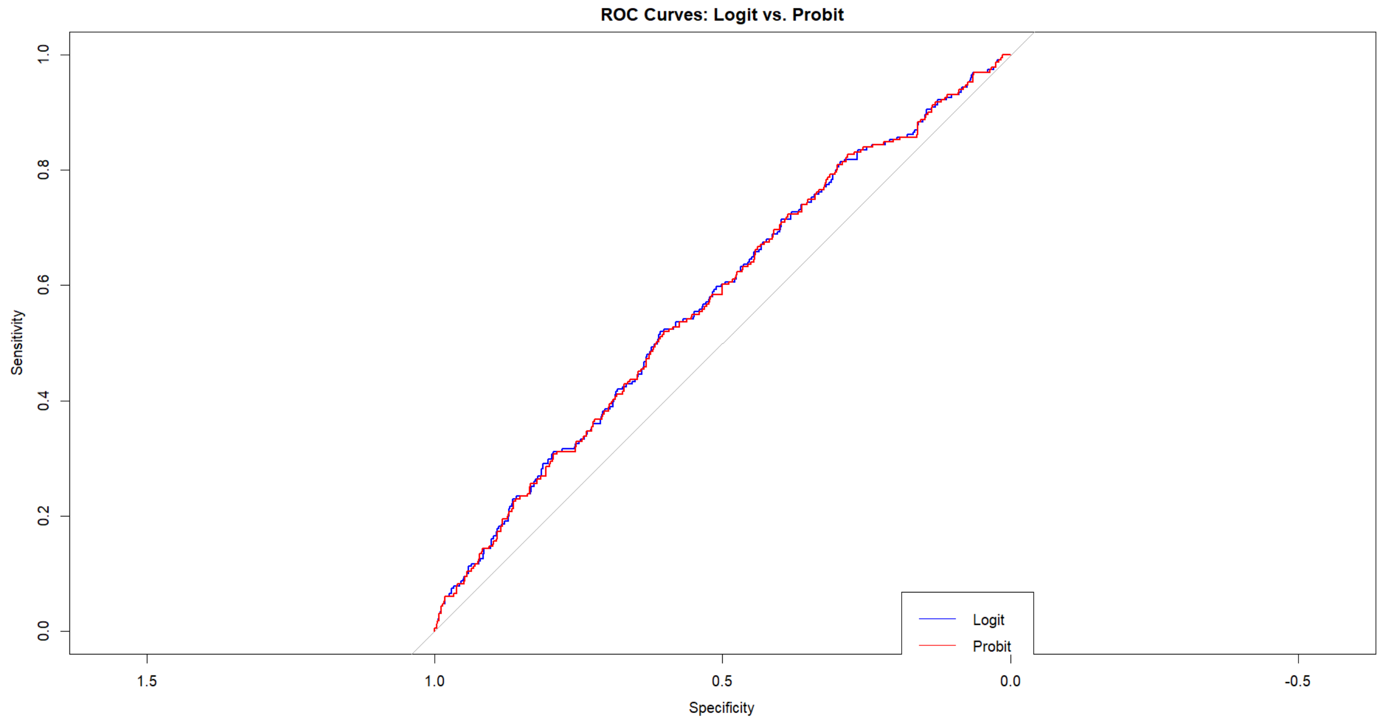


*Figure 17 : ROC Curves: Logit vs Probit*

The confusion matrix below for probit, like that for logit, indicates how good the prediction is using the probit technique. The result is identical to the logit model, suggesting that both are accurate.

```
Logit Confusion Matrix:
> print(confusionMatrix(
+   factor(pred_logit[logit_cm_idx], levels = c("No","Yes")),
+   data$univ_edu[logit_cm_idx]
+ ))
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  503 162
       Yes 123  69

               Accuracy : 0.6674
                 95% CI : (0.6348, 0.6989)
    No Information Rate : 0.7305
    P-Value [Acc > NIR] : 0.99998

                  Kappa : 0.108

 Mcnemar's Test P-Value : 0.02439

            Sensitivity : 0.8035
            Specificity : 0.2987
         Pos Pred Value : 0.7564
         Neg Pred Value : 0.3594
             Prevalence : 0.7305
         Detection Rate : 0.5869
   Detection Prevalence : 0.7760
      Balanced Accuracy : 0.5511

       'Positive' Class : No
```

*Figure 18 : Logit Confusion Matrix*

```
Probit Confusion Matrix:
> print(confusionMatrix(
+   factor(pred_probit[probit_cm_idx], levels = c("No","Yes")),
+   data$univ_edu[probit_cm_idx]
+ ))
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  502 165
       Yes 124  66

               Accuracy : 0.6628
                 95% CI : (0.63, 0.6944)
    No Information Rate : 0.7305
    P-Value [Acc > NIR] : 0.99999

                  Kappa : 0.0928

 Mcnemar's Test P-Value : 0.01863

            Sensitivity : 0.8019
            Specificity : 0.2857
         Pos Pred Value : 0.7526
         Neg Pred Value : 0.3474
             Prevalence : 0.7305
         Detection Rate : 0.5858
   Detection Prevalence : 0.7783
      Balanced Accuracy : 0.5438

       'Positive' Class : No
```

*Figure 19 : Probit Confusion Matrix*

The logit and probit models were compared using metrics such as AIC, BIC, ROC plots, AUC, and confusion matrices. The fit of the logit model was better and lower in AIC and BIC and greater in AUC, and it provided policymakers with better insights through odds ratios.

I suggest using the logit mode to forecast university graduation for greater predictability and rational structure in economic choice.

# QUESTION 3 : SIMULATION STUDY ON ENDOGENEITY

## a) Data Generation

The data generation process (DGP) has been built so that the endogenous variable, Log_Wage, depends on both Work_Experience and Years_of_Education. However, Years_of_Education is rendered endogenous by introducing a factor into the error term. The model is expressed as follows:

$$Log\_Wage = \beta 0 + \beta 1 \times Work\_Experience + \beta 2 \times Years\_of\_Education + \epsilon$$

We simulate a dataset with 1000 observations. Let

exper ~ $N(10, 2^2)$ (exogenous variable),

Error term $\varepsilon \sim N(0, 1)$.

Generate educ (endogenous) : educ = $12 + 0.5 \times \varepsilon + \eta$, where $\eta \sim N(0, 0.5^2)$, thereby inducing correlation with ε. Finally, generate lwage as:

lwage = 0.5 + 0.05 × exper + 0.1 × educ + ε.

This equation merges exper and educ effects and incorporates error to produce natural wages' logarithms.

## b) OLS Estimation

We perform 1000 estimates using OLS to check for endogeneity in coefficients. We have 1000 different datasets, each derived using a different DGP, and we examine distributions of coefficients.

$$Log\_Wagei = \beta 0 + \beta 1 \times Work\_Experiencei + \beta 2 \times Years\_of\_Educationi + \varepsilon i$$

where β0=0.5, β1=0.05, β2=0.1,β2=0.1, and Years_of_Educationi is correlated with the error term εi.This correlation violates the key OLS assumption that E(X′ε)=0, thus introducing endogeneity. The estimated coefficients can be written as

$$\beta\hat{} = (X^T X)^{-1}.X^T y$$

X is the matrix of regressors (including a column of ones for the intercept), and y is the vector of lwage. By design, “expert” is exogenous, so OLS should produce unbiased estimates for β1. Nevertheless, “educ” is an endogenous variable; therefore, we expect $\beta\hat{}_2$ to demonstrate systematic bias.

```
> cat("First 10 estimated coefficients for 'exper':\n")
First 10 estimated coefficients for 'exper':
> print(head(ols_coef_exper, 10))
 [1] 0.03883818 0.03985971 0.05842050 0.04161573 0.05939972 0.03978242 0.06738840 0.04906572 0.03143038
[10] 0.05618110
> cat("First 10 estimated coefficients for 'educ':\n")
First 10 estimated coefficients for 'educ':
> print(head(ols_coef_educ, 10))
 [1] 1.110678 1.056856 1.115489 1.102772 1.112845 1.099767 1.124758 1.105003 1.103545 1.105644
> # Print summary statistics (mean, min, max, quartiles) for the coefficient estimates
> cat("Summary of OLS estimates for 'exper':\n")
Summary of OLS estimates for 'exper':
> print(summary(ols_coef_exper))
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.01323 0.04233 0.04988 0.05015 0.05777 0.08576
> cat("Summary of OLS estimates for 'educ':\n")
Summary of OLS estimates for 'educ':
> print(summary(ols_coef_educ))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.004   1.078   1.100   1.100   1.122   1.198
>
```

*Figure 20 : OLS Estimation Summary*

This figure plots 1000 simulated distributions of OLS. 'Years_of_Education' estimates lag, whereas 'Work_Experience' is close to the actual value, reflecting endogeneity bias. Its summaries for summary(ols_coef_exper) and summary(ols_coef_educ) provide means, quartiles, and estimates for max and min, reflecting variability.

## c) Bias and Inconsistency

The bias in OLS estimators originates from the divergence between the OLS estimator's average and the coefficient’s actual value.

The bias for each coefficient β^j is calculated as:

$$Bias(\beta^j) = \beta^j - \beta j$$

β^j = mean of the estimated coefficients over all replications, βj = actual parameter.

```
> cat("Bias in OLS for Work_Experience:", round(bias_exper, 4), "\n")
Bias in OLS for Work_Experience: 1e-04
> cat("Bias in OLS for Years_of_Education:", round(bias_educ, 4), "\n")
Bias in OLS for Years_of_Education: 1.0003
```

*Figure 21 : Bias in OLS*

This result confirms that OLS is not consistent when the regressor is endogenous. An instrumental variable approach (2SLS) is introduced in subsequent subsections to address this endogeneity and obtain unbiased, consistent estimators.

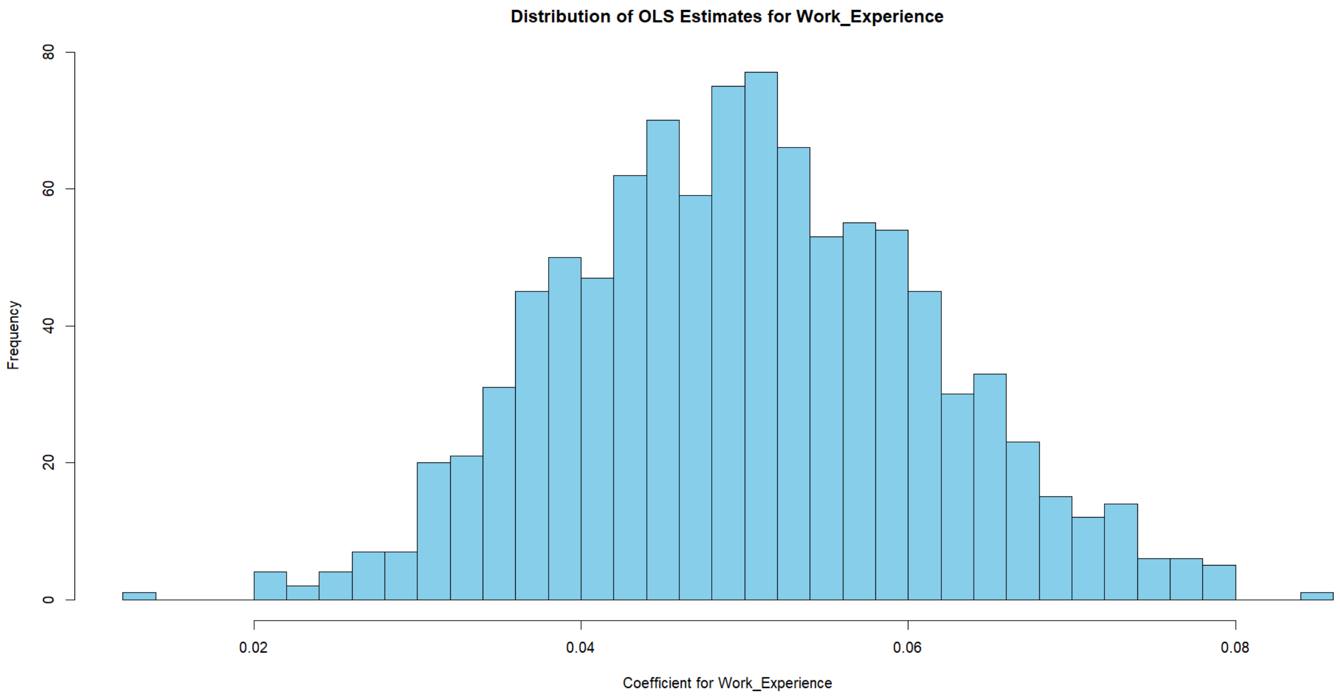


*Figure 22 : Distribution of OLS Estimates for Work_Experience*

Figure 22 for OLS estimates for 'Work_Experience' around the actual parameter, verifying stable OLS effects for exogenous predictors.

Figure 23 histogram biases estimate, leading to inaccuracy through an endogenous association. The underestimation reflects the limitations of OLS in capturing education's actual effect on wages.

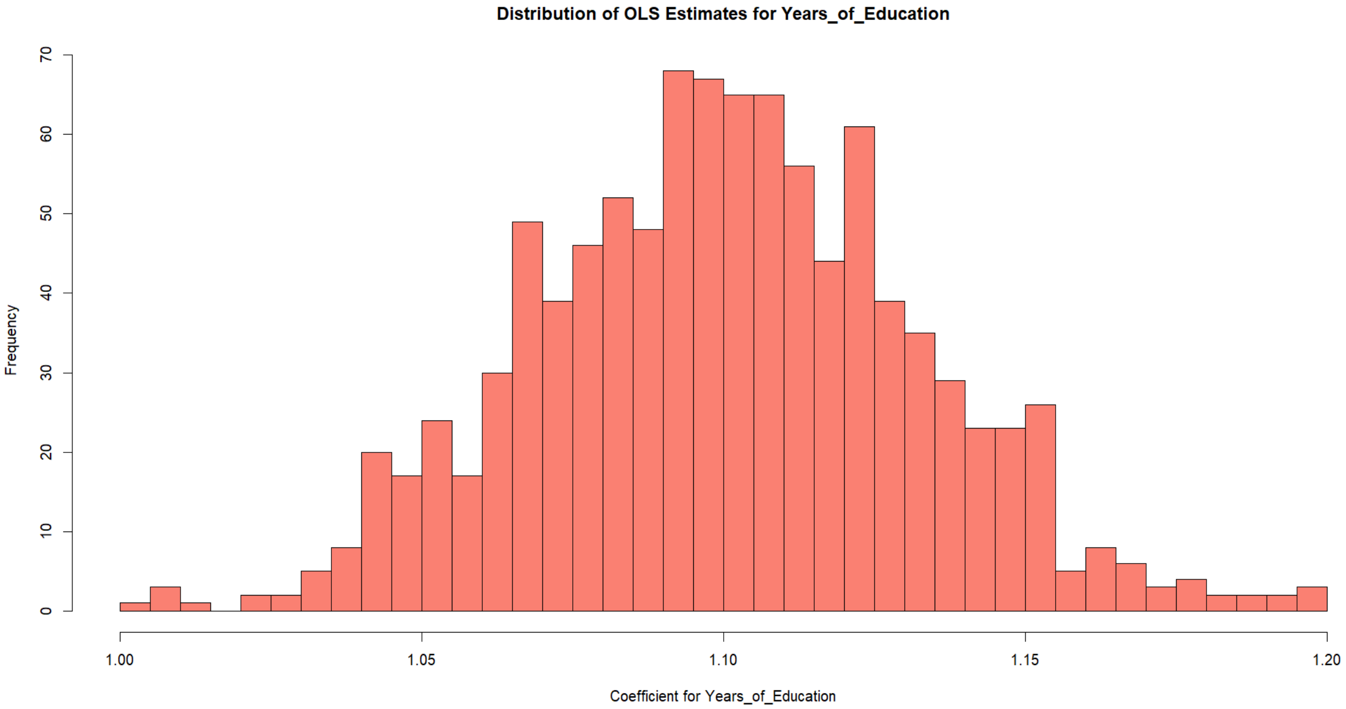


*Figure 23 : Distribution of OLS Estimates for Years_of_Education*

## d) 2SLS Estimation

Two-Stage Least Squares (2SLS) employs an instrument methodology to purge against bias. The instrument, z, is correlated to Years_of_Education and is independent of errors.

```
> cat("Bias in 2SLS for Work_Experience:", round(bias_iv_exper, 4), "\n")
Bias in 2SLS for Work_Experience: 1e-04
> cat("Bias in 2SLS for Years_of_Education:", round(bias_iv_educ, 4), "\n")
Bias in 2SLS for Years_of_Education: 6e-04
```

*Figure 24 : Bias in 2SLS*

```
> cat("Comparing OLS vs. 2SLS bias for 'exper':\n")
Comparing OLS vs. 2SLS bias for 'exper':
> cat("OLS Bias:", round(bias_exper, 4),
+     " | 2SLS Bias:", round(bias_iv_exper, 4), "\n\n")
OLS Bias: 1e-04  | 2SLS Bias: 1e-04

> cat("Comparing OLS vs. 2SLS bias for 'educ':\n")
Comparing OLS vs. 2SLS bias for 'educ':
> cat("OLS Bias:", round(bias_educ, 4),
+     " | 2SLS Bias:", round(bias_iv_educ, 4), "\n")
OLS Bias: 1.0003  | 2SLS Bias: 6e-04
```

*Figure 25 : Comparing OLS vs 2SLS Bias*

Endogeneity raises "educ" in OLS. 2SLS estimates "expert" and "educ" draw closer to true values using a valid instrument. The 2SLS-OLS bias discrepancy is evident. 2SLS's low bias highlights strength in instrumental variable estimators in endogeneity. The contrast highlights the IV approach's strength using correlations in error terms.

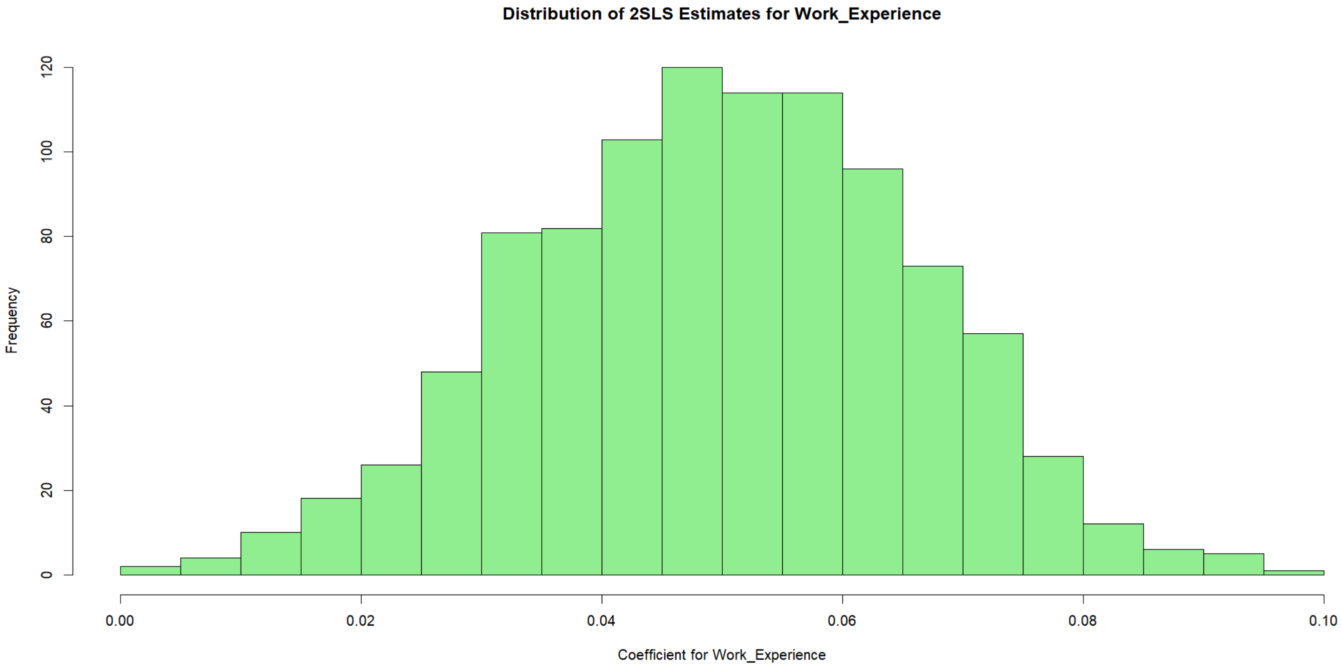


*Figure 26 : Distribution of 2SLS for Work_Exp*

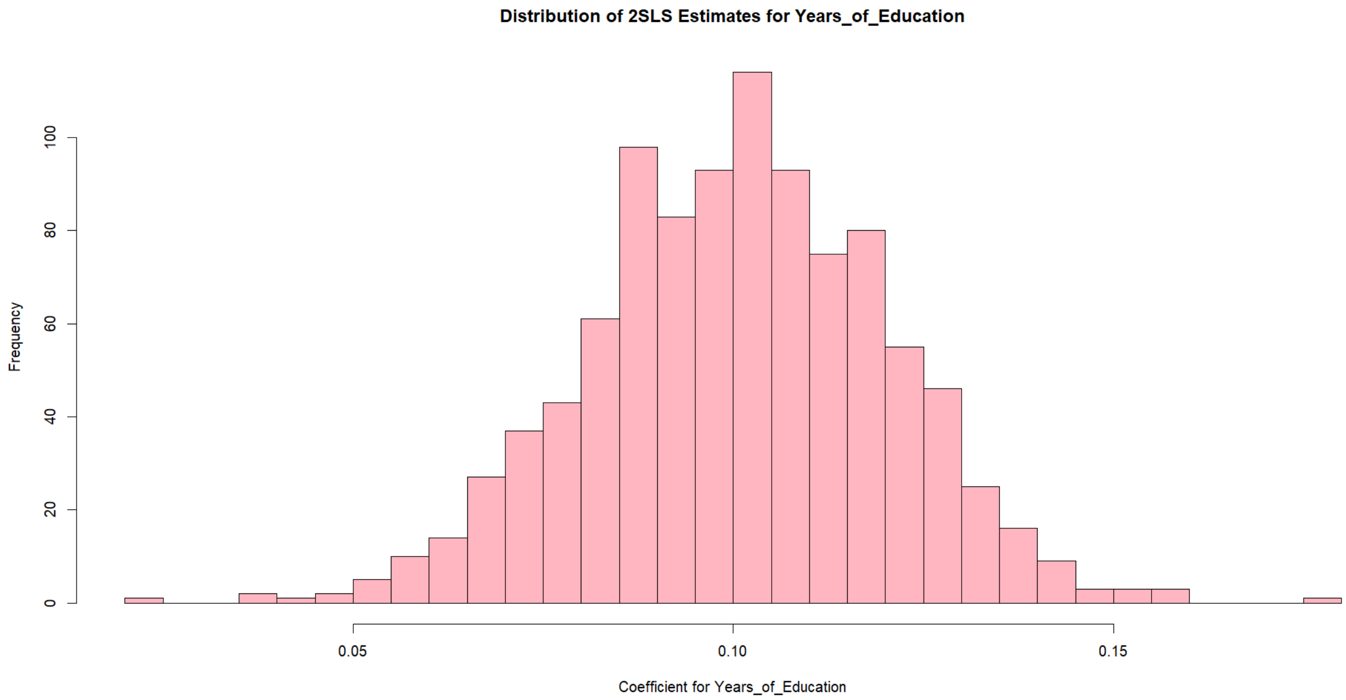


*Figure 27 : Distribution of 2SLS for Years_of_edu*

Figure 26 and Figure 27 illustrate how 2SLS estimates by actual values, validating the technique's robustness in forecasting exogenous predictors and addressing endogeneity. The 2SLS histogram is consistent with the proper parameter, in contrast to the endogenous-biased estimates by OLS. This illustrates the power of instrumenting to remedy endogeneity.

## e) Implications for Empirical Research in Real Estate Economics & Addressing Endogeneity in Housing Market Studies

Implications

The simulation illustrates how endogeneity influences research results, particularly in studies on housing markets. In housing economics, housing price determinants such as education, job stability, and income tend to be correlated to unknown determinants such as borrower preference or neighbourhood. The association may alter our interpretation of how such determinants influence housing prices, loan acceptance, or residential choice. For instance, education may appear to reduce housing prices if correlated to unknown factors, such as good credit or financial capability, which may fail to appear in the analysis. Failing to include such unknowns may result in under- or over-estimation, leading to bad decisions and ineffective policy.

Precautions

Researchers who analyse housing market data must ensure that their variables do not have any hidden influences or correlations to other irrelevant factors. If endogeneity is possible, an instrumental variable (IV) approach may identify external changes affecting the primary issue, but only if an adequate instrument is discovered. In housing contexts, such instruments may consist of prior changes in zoning regulations, housing loan regulations, or natural experiments occurring in one but not another. Fixed effects models handle consistent hidden influences through time, and the difference-in-differences approach assesses changes between affected and unaffected groups. Real estate economists may also consider using dynamic models, including prior values for crucial factors to ensure against reverse causation. Using such strategies in addition to diligent data collection and strict testing, housing market economists ensure their estimates capture true cause-and-effect, not spurious links, and better guide policymakers and market participants.

# BIBILOGRAPHY

[1] Angrist, J.D. and Krueger, A.B. (1999) ‘Empirical Strategies in Labor Economics’, in Card, D. and Lemieux, T. (eds.) *Handbook of Labor Economics*, Vol. 3, Amsterdam: Elsevier, pp. 1277–1366.

[2] Bertrand, M., Duflo, E. and Mullainathan, S. (2004) ‘How Much Should We Trust Differences-In-Differences Estimates?’, *Quarterly Journal of Economics*, 119(1), pp. 249–275.

[3] Bound, J., Jaeger, D.A. and Baker, R.M. (1995) ‘Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak’, *Journal of the American Statistical Association*, 90(430), pp. 443–450.

[4] Cameron, A.C. and Trivedi, P.K. (2005) *Microeconometrics: Methods and Applications*. Cambridge: Cambridge University Press.

[5] Davidson, R. and MacKinnon, J.G. (1993) *Estimation and Inference in Econometrics*. New York: Oxford University Press.

[6] Greene, W.H. (2012) *Econometric Analysis*, 7th ed. New Jersey: Pearson.

[7] Imbens, G.W. and Angrist, J.D. (1994) ‘Identification and Estimation of Local Average Treatment Effects’, *Econometrica*, 62(2), pp. 467–475.

[8] Kennedy, P. (2003) *A Guide to Econometrics*, 5th ed. Cambridge: MIT Press.

[9] Stock, J.H. and Watson, M.W. (2011) *Introduction to Econometrics*, 3rd ed. Boston: Pearson.

[10] Wooldridge, J.M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. Cambridge: MIT Press.

# APPENDIX

```
# Load necessary libraries

library(readxl)

library(ggplot2)

library(dplyr)

library(corrplot)

library(lmtest)

library(car)

library(ggpubr)     # For combining plots

library(sandwich)    # For robust standard errors

library(margins)

library(pROC)

library(caret)

library(tidyr)


# Import the dataset

data <-
read_excel("D:/My_Profile/Msc_UKVI/Westminster/5.Predictive_Analysis_for_Decision_Making/Coursew
ork1_24thFeb/nls80.xlsx")


#Question 1: MULTIPLE LINEAR REGRESSION

#1a) Exploratory Data Analysis


# Renaming columns for clarity

data <- data %>%

  rename(

    MonthlyEarnings       = wage,

    WeeklyHours          = hours,

    IQ_Score             = iq,

    Knowledge_World_Work     = kww,

    Years_of_Education      = educ,

    Work_Experience        = exper,
```

```
    Current_Employer_Tenure = tenure,
    Age                    = age,
    Marital_Status          = married,
    Race_Black             = black,
    South_Indicator         = south,
    Urban_Indicator         = urban,
    Siblings               = sibs,
    Birth_Order            = brthord,
    Mothers_Education       = meduc,
    Fathers_Education       = feduc,
    Log_Wage               = lwage
  )

# Define the selected variables
selected_vars <- c("MonthlyEarnings", "Years_of_Education", "Work_Experience")
# Print the variable to check its contents
print(selected_vars)

# Define continuous variables for EDA
continuous_vars <- c("MonthlyEarnings", "WeeklyHours", "IQ_Score", "Knowledge_World_Work",
                 "Years_of_Education", "Work_Experience", "Current_Employer_Tenure",
                 "Age", "Siblings", "Birth_Order", "Mothers_Education", "Fathers_Education", "Log_Wage")

# Ensure that variables are numeric
for (var in continuous_vars) {
  if (!is.numeric(data[[var]])) {
    data[[var]] <- as.numeric(as.character(data[[var]]))
  }
}

# Summary statistics
summary_stats <- summary(data[continuous_vars])
```

```
print(summary_stats)

# Select a subset of variables for appropriate visualizations
selected_vars <- c("MonthlyEarnings", "Years_of_Education", "Work_Experience")

# faceted plotting

data_long <- data %>%
  select(all_of(selected_vars)) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Value")

# Create faceted histogram for the selected variables
p_hist <- ggplot(data_long, aes(x = Value)) +
  geom_histogram(color = "black", fill = "lightblue", bins = 30) +
  facet_wrap(~ Variable, scales = "free_x") +
  labs(title = "Faceted Histograms")

# Create faceted boxplot for the selected variables
p_box <- ggplot(data_long, aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Faceted Boxplots")

# Combine the two plots into one figure using ggpubr
library(ggpubr)
combined_plot <- ggarrange(p_hist, p_box, ncol = 1, nrow = 2)
print(combined_plot)




# Correlation matrix and its visualization

corr_matrix <- cor(data[continuous_vars], use = "complete.obs")
```

```
print(round(corr_matrix, 2))

corrplot(corr_matrix, method = "color", type = "upper", addCoef.col = "black",
         tl.cex = 0.8, title = "Correlation Matrix of Continuous Variables", mar = c(0,0,1,0))


#1b) Model Development


# Develop the multiple linear regression model for Log_Wage
model <- lm(Log_Wage ~ Years_of_Education + Work_Experience + Current_Employer_Tenure +
IQ_Score + Age, data = data)
summary(model)


#Check for Multicollinearity using Variance Inflation Factor (VIF)
#vif_values <- vif(model)
#print(vif_values)  # Display VIF values


#bptest(model)  # Breusch-Pagan test for heteroskedasticity


#Interpretation:
#p-value < 0.05 → Heteroskedasticity exists (violates OLS assumption).
#p-value > 0.05 → No heteroskedasticity (OLS assumptions hold).



#1d) Diagnostic Tests


#Check for Heteroskedasticity (Breusch-Pagan Test)
# Breusch-Pagan test for heteroskedasticity
bp_test <- bptest(model)
print(bp_test)


#Interpretation:
#p-value < 0.05 → Heteroskedasticity exists (violates OLS assumption).
#p-value > 0.05 → No heteroskedasticity (OLS assumptions hold).
```

```
# If heteroskedasticity is present, use robust standard errors:

coeftest(model, vcov = vcovHC(model, type = "HC3"))

#Checking for Multicollinearity using Variance Inflation Factor (VIF)
# Calculate Variance Inflation Factors (VIF)
vif_values <- vif(model)
print(vif_values)

#Interpretation:
#If VIF > 10, the variable has high collinearity and may need to be removed or adjusted.
#If VIF between 5-10, moderate correlation exists (may still be acceptable).
#If VIF < 5, no serious collinearity issue.

#Residual Diagnostics
#Residuals vs. Fitted Plot
p1 <- ggplot(data, aes(x = fitted(model), y = resid(model))) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted", x = "Fitted Values", y = "Residuals")

# Normal Q-Q Plot
p2 <- ggplot(data, aes(sample = resid(model))) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot")

# Histogram of Residuals
p3 <- ggplot(data, aes(x = resid(model))) +
  geom_histogram(color = "black", fill = "lightblue", bins = 30) +
  labs(title = "Histogram of Residuals", x = "Residuals")
```

```
# Combine Plots
ggarrange(p1, p2, p3, ncol = 3, nrow = 1)


# Outlier & Influential Point Detection


# Cook's Distance
plot(model, which = 4, main = "Cook's Distance Plot")
# Leverage vs. Studentized Residuals
plot(model, which = 5, main = "Leverage vs. Residuals")
#Autocorrelation Test (Durbin-Watson Test)
dwtest(model)


# Q2: GENERALIZED LINEAR MODELS – LOGIT AND PROBIT


# (a) Binary Variable Creation


data$univ_edu <- ifelse(data$Years_of_Education >= 16, 1, 0)
data$univ_edu <- factor(data$univ_edu, levels = c(0, 1), labels = c("No", "Yes"))


# Quick check of distribution
table(data$univ_edu)


# (b) Logit and Probit Models


# Work_Experience, Current_Employer_Tenure, IQ_Score, Age, and Mothers_Education


logit_model <- glm(
  univ_edu ~ Work_Experience + Current_Employer_Tenure + IQ_Score + Age + Mothers_Education,
  data = data,
  family = binomial(link = "logit")
)
summary(logit_model)
```

```
probit_model <- glm(
  univ_edu ~ Work_Experience + Current_Employer_Tenure + IQ_Score + Age + Mothers_Education,
  data = data,
  family = binomial(link = "probit")
)
summary(probit_model)

# (d) Marginal Effects
# Calculate marginal effects for the Logit model

margins_logit <- margins(logit_model)
summary(margins_logit)

# Calculate marginal effects for the Probit model
margins_probit <- margins(probit_model)
summary(margins_probit)



#e) Model Comparison
cat("AIC for Logit Model:", AIC(logit_model), "\n")
cat("AIC for Probit Model:", AIC(probit_model), "\n")
cat("BIC for Logit Model:", BIC(logit_model), "\n")
cat("BIC for Probit Model:", BIC(probit_model), "\n")

# Predicted probabilities for each model
logit_probs <- predict(logit_model, type = "response")
probit_probs <- predict(probit_model, type = "response")

# Creating index for rows actually used (no missing values)
common_idx_logit <- !is.na(logit_probs) & !is.na(data$univ_edu)
common_idx_probit <- !is.na(probit_probs) & !is.na(data$univ_edu)
```

```r
# Building ROC objects using the matched indices
roc_logit <- roc(data$univ_edu[common_idx_logit], logit_probs[common_idx_logit])
roc_probit <- roc(data$univ_edu[common_idx_probit], probit_probs[common_idx_probit])

cat("Logit AUC:", auc(roc_logit), "\n")
cat("Probit AUC:", auc(roc_probit), "\n")

# Plot both ROC curves
plot(roc_logit, col = "blue", main = "ROC Curves: Logit vs. Probit")
plot(roc_probit, col = "red", add = TRUE)
legend("bottomright", legend = c("Logit", "Probit"), col = c("blue","red"), lty = 1)

#Confusion Matrix for classification
pred_logit <- ifelse(logit_probs > 0.5, "Yes", "No")
pred_probit <- ifelse(probit_probs > 0.5, "Yes", "No")

# For confusion matrices, also subset to rows actually used by the model
logit_cm_idx <- !is.na(logit_probs) & !is.na(data$univ_edu)
probit_cm_idx <- !is.na(probit_probs) & !is.na(data$univ_edu)

cat("Logit Confusion Matrix:\n")
print(confusionMatrix(
  factor(pred_logit[logit_cm_idx], levels = c("No","Yes")),
  data$univ_edu[logit_cm_idx]
))

cat("Probit Confusion Matrix:\n")
print(confusionMatrix(
  factor(pred_probit[probit_cm_idx], levels = c("No","Yes")),
  data$univ_edu[probit_cm_idx]
))
```

```
#Q3 SIMULATION STUDY ON ENDOGENEITY

#a) Data Generation

set.seed(123)
n <- 1000

# True parameter values
beta0 <- 0.5
beta1 <- 0.05   # effect of Work_Experience
beta2 <- 0.1    # effect of Years_of_Education

#exogenous variable Work_Experience
Work_Experience <- rnorm(n, mean = 10, sd = 2)

# error term
error <- rnorm(n, mean = 0, sd = 1)

# Generating endogenous Years_of_Education (introducing endogeneity via error)
Years_of_Education <- 12 + 0.5 * error + rnorm(n, mean = 0, sd = 0.5)

# Log_Wage based on the DGP
Log_Wage <- beta0 + beta1 * Work_Experience + beta2 * Years_of_Education + error

# simulation data frame
sim_data <- data.frame(Work_Experience, Years_of_Education, Log_Wage)

#b) OLS Estimation
simulations <- 1000
```

```
ols_coef_exper <- numeric(simulations)
ols_coef_educ  <- numeric(simulations)

## Monte Carlo Simulation
for(i in 1:simulations) {
  Work_Experience <- rnorm(n, mean = 10, sd = 2)
  error <- rnorm(n, mean = 0, sd = 1)
  Years_of_Education <- 12 + 0.5 * error + rnorm(n, mean = 0, sd = 0.5)
  Log_Wage <- beta0 + beta1 * Work_Experience + beta2 * Years_of_Education + error

  model_sim <- lm(Log_Wage ~ Work_Experience + Years_of_Education)
  ols_coef_exper[i] <- coef(model_sim)["Work_Experience"]
  ols_coef_educ[i]  <- coef(model_sim)["Years_of_Education"]
}


# first 10 estimated coefficients for each variable
cat("First 10 estimated coefficients for 'exper':\n")
print(head(ols_coef_exper, 10))

cat("First 10 estimated coefficients for 'educ':\n")
print(head(ols_coef_educ, 10))

# summary statistics (mean, min, max, quartiles) for the coefficient estimates
cat("Summary of OLS estimates for 'exper':\n")
print(summary(ols_coef_exper))


cat("Summary of OLS estimates for 'educ':\n")
print(summary(ols_coef_educ))
```

```
# c) Calculate biases
bias_exper <- mean(ols_coef_exper) - beta1
bias_educ  <- mean(ols_coef_educ)  - beta2
cat("Bias in OLS for Work_Experience:", round(bias_exper, 4), "\n")
cat("Bias in OLS for Years_of_Education:", round(bias_educ, 4), "\n")

# Visualize the distribution of OLS estimates with histograms
hist(ols_coef_exper, main = "Distribution of OLS Estimates for Work_Experience",
     xlab = "Coefficient for Work_Experience", col = "skyblue", breaks = 30)
hist(ols_coef_educ, main = "Distribution of OLS Estimates for Years_of_Education",
     xlab = "Coefficient for Years_of_Education", col = "salmon", breaks = 30)

#d) 2SLS Estimation

library(AER)

iv_coef_exper <- numeric(simulations)
iv_coef_educ  <- numeric(simulations)

for(i in 1:simulations) {
  Work_Experience <- rnorm(n, mean = 10, sd = 2)
  error <- rnorm(n, mean = 0, sd = 1)

  # Generate instrument z for Years_of_Education
  z <- rnorm(n, mean = 15, sd = 2)

  Years_of_Education <- 12 + 0.5 * error + 0.8 * z + rnorm(n, mean = 0, sd = 0.5)
  Log_Wage <- beta0 + beta1 * Work_Experience + beta2 * Years_of_Education + error

  model_iv <- ivreg(Log_Wage ~ Work_Experience + Years_of_Education | Work_Experience + z)
  iv_coef_exper[i] <- coef(model_iv)["Work_Experience"]
  iv_coef_educ[i]  <- coef(model_iv)["Years_of_Education"]
```

```
}


# Compare 2SLS estimates to the true values
bias_iv_exper <- mean(iv_coef_exper) - beta1
bias_iv_educ  <- mean(iv_coef_educ)  - beta2


cat("Bias in 2SLS for Work_Experience:", round(bias_iv_exper, 4), "\n")
cat("Bias in 2SLS for Years_of_Education:", round(bias_iv_educ, 4), "\n")



#Compare OLS vs. 2SLS Bias


cat("Comparing OLS vs. 2SLS bias for 'exper':\n")
cat("OLS Bias:", round(bias_exper, 4),
    " | 2SLS Bias:", round(bias_iv_exper, 4), "\n\n")
cat("Comparing OLS vs. 2SLS bias for 'educ':\n")
cat("OLS Bias:", round(bias_educ, 4),
    " | 2SLS Bias:", round(bias_iv_educ, 4), "\n")
# Visualize distributions
hist(iv_coef_exper, main = "Distribution of 2SLS Estimates for Work_Experience",
     xlab = "Coefficient for Work_Experience", col = "lightgreen", breaks = 30)
hist(iv_coef_educ, main = "Distribution of 2SLS Estimates for Years_of_Education",
     xlab = "Coefficient for Years_of_Education", col = "lightpink", breaks = 30)
```