The Effect of Education on Wages

Jan Sklenička, Sebastian Černý*

January 26, 2025

Abstract

It is widely recognized that higher educational attainment is associated with higher wages. This study aims to empirically investigate this relationship using data from the European Social Survey. Given that education was found to be an endogenous variable, a Two-Stage Least Squares model was employed instead of the classical Ordinary Least Squares approach to obtain consistent estimates. Furthermore, the presence of heteroskedasticity was addressed by applying White's robust standard errors, ensuring valid statistical inference. The regression results indicate that advancing one level of education, such as progressing from a bachelor's to a master's degree, is associated with an approximate 67% increase in wages.

1 Introduction

The relationship between education and wage has been a topic of significant interest in labor economics, policy-making, and social science research for a long time. Education of an individual is often considered to be the key factor of an individual's earning potential, with higher levels of education associated with better employment, which implies higher wage. In our work, we aim to provide robust estimates of the effect of education on wage, while accounting for several challenges in the data such as heteroskedasticity and endogeneity. The dataset is from the European Social Survey (ESS), available at ESS Data Portal.[1] The dataset contains many socio-economic variables that help us to control for other important regressors such as gender and work experience.

Our empirical approach starts with Ordinary Least Squares (OLS) regression to estimate the relationship between education and wage. However, as noted above, we have to deal with endogeneity of education, which is mentioned in several studies. Card (1999), has discussed the endogeneity of education and the need of using appropriate econometric techniques to

[*] The authors used ChatGPT in order to expand vocabulary and writing style and for assistance with LaTeX. After using this tool, the authors reviewed and edited the content as necessary and takes full responsibility for the content of the publication.

[1] European Social Survey European Research Infrastructure (ESS ERIC) (2023) ESS9 - integrated file, edition 3.2 [Data set]. Sikt - Norwegian Agency for Shared Services in Education and Research. https://doi.org/10.21338/ess9e03_2.

obtain unbiased estimates. Additionally, study from Arabsheibani *et al.* (2010) has shown that accounting for endogeneity of education can significantly increase estimated returns, they found out that average return to education was between 6% and 8% per additional year of schooling. Thus, we extend our work by using instrumental variables: mother's education, father's education. Use of parental education as instrumental variables is justified by the study from Chevalier (2004), where he discusses their relevance in addressing endogeneity concerns. The study found that a one-year increase in parental education is associated with an increase in child's education by 0.25 to 0.40 years.

In Section 2 we describe the methodology used in our work, followed by Section 3 where we discuss the nature of the dataset and the variables that were used in both OLS and 2SLS. Lastly, in the Section 4, we present the results of our models and in the Section 5 we discuss the results, limitations and potential improvements.

2 Methodology

This Section describes the methodology used in this research paper. Subsection 2.1 discusses the Ordinary Least Squares model (OLS) and the corresponding assumptions. Subsection 2.2 addresses the issue of endogeneity of OLS using instrumental variables and the corresponding tests. The last Subsection 2.3 proposes a solution to heteroskedasticity of the data used. This section was inspired by Greene (2018).

2.1 Ordinary Least Squares

This research paper tries to explain a continuous variable. Therefore, the baseline model proposed is an OLS model defined as follows

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i, \tag{1}$$

where Y_i is the dependent variable, $X_{i1}, X_{i2}, \ldots, X_{ik}$ represent the independent variables, β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients to be estimated, and ε_i is the error term. The OLS regression model can be also written in matrix form as

$$\underbrace{y}_{(n\times 1)} = \underbrace{X}_{(n\times K)} \underbrace{\beta}_{(K\times 1)} + \underbrace{e}_{(n\times 1)}. \tag{2}$$

Using the assumption of zero correlation between the independent variables and the error term, we can derive the OLS estimator for β as

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{3}$$

The OLS assumptions are as follows:

- **Linearity:**

$$E[y \mid x] = x'\beta, \tag{4}$$

- **Exogeneity:** The expectation of the error term conditional on the independent variables is zero
$$E[\varepsilon|x_1, x_2, \ldots, x_K] = 0, \tag{5}$$
- **No Perfect Multicollinearity:** The matrix $X'X$ must be invertible
$$\text{rank}(X) = K, \tag{6}$$
- **Homoskedasticity:** The variance of the error term is constant across all observations
$$Var[\varepsilon|X] = \sigma^2, \tag{7}$$
- **Random sample:** We assume that the data come from a random sample,
- **Normality:** For small sample sizes, the error term is typically assumed to follow a normal distribution to ensure the validity of hypothesis testing and inference. However, for large sample sizes, the normality assumption is not strictly required, as the Central Limit Theorem ensures that the sampling distribution of the OLS estimators approaches normality, regardless of the distribution of the error term. This property applies to our dataset, as discussed in Subsection 3.1.

In can be shown that under those assumptions the $\hat{\beta}$ estimator is considered an unbiased, consistent, and efficient estimator of β

2.2 Instrumental Variables

However, as we suppose that we are dealing with an endogenous variable, which causes the estimator to be biased and inconsistent. We solve this problem using instrumental variables, more specifically a generalized version called two stage least square estimation (2SLS). In cases where the matrix of instrumental variables Z contains more variables than X, the product $Z'X$ results in a non-invertible matrix. Therefore, we project the columns of X onto the column space of Z as follows

$$\hat{X} = Z(Z'Z)^{-1}Z'X. \tag{8}$$

Using the instrumental variables estimator

$$\hat{\beta}_{\text{IV}} = (\hat{X}'\hat{X})^{-1}\hat{X}'y, \tag{9}$$

substituting $\hat{X}$, the Two-Stage Least Squares (2SLS) estimator is given as

$$\hat{\beta}_{\text{2SLS}} = \left[X'Z(Z'Z)^{-1}Z'X\right]^{-1} X'Z(Z'Z)^{-1}Z'y. \tag{10}$$

If we suppose that the instrumental variables satisfy the exogeneity condition $E[\varepsilon_i \mid z_i] = 0$, the finite variance condition $E[z_{il}^2] = \mathbf{Q}_{zz,ll} < \infty$, and the relevance condition $E[z_{il}x_{ik}] = \mathbf{Q}_{zx,lk} < \infty$, then we can apply the Hausman test to evaluate the validity of our endogeneity assumption The Hausman test statistic is derived from the Wald statistic and is expressed as

$$H = d' \left[\text{Var}(d) \right]^{-1} d \sim \chi^2(k), \tag{11}$$

where $d = \hat{\beta}_{2SLS} - \hat{\beta}_{OLS}$.
Hypotheses:

$$H_0 : \text{Exogeneity, OLS is consistent and efficient}$$
$$H_A : \text{Endogeneity, OLS is inconsistent, IV (2SLS) is consistent}$$

2.3 Heteroskedasticity

After rejecting the null hypothesis of the Breusch-Pagan test[2], another issue with our OLS model is the violation of the homoscedasticity assumption, resulting in heteroscedasticity. This violation compromises the efficiency of the estimator. To ensure valid statistical inference, we employ White's standard errors, as the specific form of heteroscedasticity is unknown. Although Feasible Generalized Least Squares (FGLS) could potentially improve efficiency, it requires strong assumptions about the structure of heteroscedasticity, which may not hold in our case. Therefore, White's standard errors are preferred as they provide valid inference without any assumptions on the error variance.

The variance-covariance matrix of the OLS estimator $\hat{\beta}$ is given by

$$\text{Var}[\hat{\beta} \mid X] = \sigma^2 (X'X)^{-1} (X'\Omega X)(X'X).^{-1} \tag{12}$$

White's heteroskedasticity-robust standard errors allow the estimation of the term $\frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 x_i x_i'$ using the squared residuals

$$\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i'. \tag{13}$$

Thus, the covariance matrix of $\hat{\beta}$ can be estimated as

$$\widehat{\text{Var}}[\hat{\beta}] = \frac{1}{n} \left(\frac{1}{n} X'X \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 x_i x_i' \right) \left(\frac{1}{n} X'X \right),^{-1} \tag{14}$$

where $\hat{\varepsilon}_i$ are the OLS residuals and x_i represents the independent variables for the i-th observation.

3 Data & Models

3.1 Data Preprocessing

For our work, we used the dataset from the European Social Survey (ESS) that had originally 31805 observations before pre-processing. As we were interested in the effect of education on wage we filtered out observations that had missing values for the wage represented by the

[2] See Appendix 5

Net_pay variable in the dataset. We also excluded observations where education of either the individual or his/her mother/father was "Unknown". Originally, the education variable was in form of the International Standard Classification of Education (ISCED) codes used by UNESCO to categorize education levels globally, so we converted the variable into 5 groups based on what was the highest achieved education and converted it into values from 1 to 5 as illustrated in Table 1. Treating education as numerical variable means that we have to assume that the difference between each level of education is the same.

Table 1. Education Level Classification

Education Level	**Code**
Primary Education	1
High School	2
Bachelor's Degree	3
Master's Degree	4
Doctoral Degree	5

We further used age and gender as explanatory variables, so we also filtered out observations, where age or gender was missing. The final explanatory variable that we included was working experience as it is used in many papers. Blau & Kahn (2019), discusses the importance of work experience as determinant of wage in his study on using expected work experience instead of the commonly used approach. Our dataset did not include working experience, but it included the year when the individual started working. We subtracted this year from 2024 for each observation to be at least partially able to capture the working experience of an individual. We also used the squared term of this variable to account for the non-linear relationship between working experience and education, which is a typical approach in labor economics. A classic example of such use is the Mincer earnings function, that models the natural logarithm of wage as a function of years of education, years of work experience, and the square of years of work experience(Patrinos, 2024).

Furthermore, we found out that the dataset contains outliers for wage and working experience, so we removed them based on the interquartile range (IQR) method. This approach helps to identify and exclude extreme values that could disproportionately influence the results. The IQR is a robust measure of statistical dispersion, calculated as the difference between the third quartile (Q3) and the first quartile (Q1):

$$\text{IQR} = Q3 - Q1 \tag{15}$$

Observations were considered outliers if they fell outside the following lower and upper bounds:

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} \tag{16}$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} \tag{17}$$

After removing observations with missing values and excluding outliers, the dataset still contains 19618 observations, which means that the sample is still large enough to provide us with robust estimates. For clarity, Table 2 provides an explanation of all variables as in the rest of this paper, the variable's names will be used.

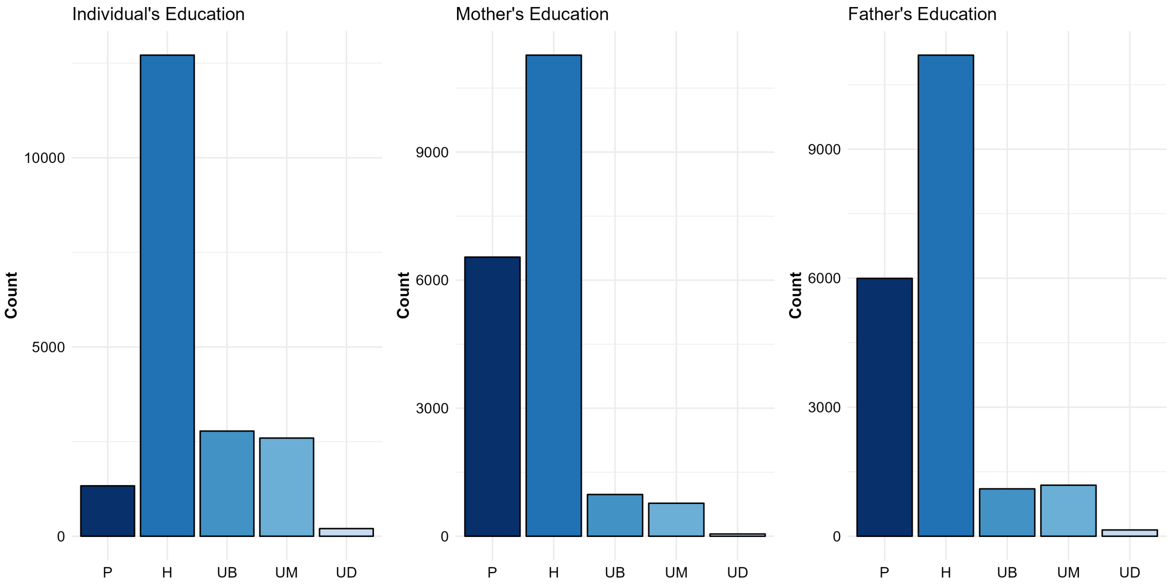
Table 2. Description of Variables Used in the Analysis

Variable	**Description**
`Net_pay`	Wage (dependent variable)
`education`	Highest achieved level of education of an individual
`educationM`	Highest achieved level of education of the mother
`educationP`	Highest achieved level of education of the father
`wexp`	Work experience of the individual
`gndr`	Gender of the individual

3.2 Summary Statistics

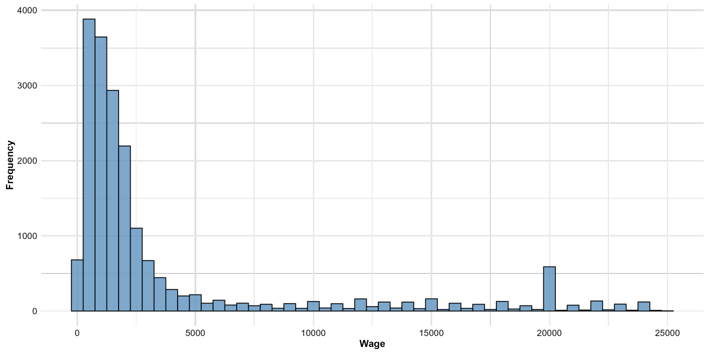
As we are using only few variables, it is possible to present plots of each variable except for gender, where the distribution is simple. The dataset contains approximately 48% of males and 52% of females, which allows for meaningful gender-based wage comparison. Figure 1 illustrates the distribution of education levels for individuals, mothers and fathers. It ranges from primary education (P) to a doctoral degree (UD). We can see that high school (H) is the most common type of education level reached. Key difference between individuals and parental distribution is that primary school is the highest level of education for roughly 6000 mothers and fathers, while it is the case for only approximately 1300 individuals. The

Figure 1. Distribution of Education Levels for Individuals and Parents



distribution is nearly identical for parents, the only noticeable difference is that there is slightly higher number of fathers with master's degree (UM) than bachelor's degree (UB), while for mothers it is the opposite case. For the distribution of wage, we look at Figure 2. The distribution is right-skewed, indicating that most individuals earn lower wages. The skewed distribution suggets the need for using logarithmic transformation in the regression to capture the relationship between wage and independent variables properly. Lastly, Figure 3 shows fairly even distribution of individual's working experience in the data. The mean of the variable is 38.96 suggesting that the observed individuals were on average relatively old.

Figure 2. Distribution of Wage



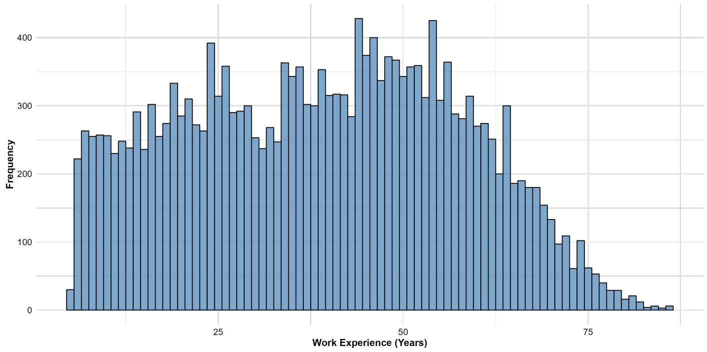


Figure 3. Distribution of Working Experience

3.3 Models

3.3.1 Baseline Model

Equation 18 is the Ordinary Least Squares (OLS) regression model, which we use as a baseline model with potential endogeneity and heteroskedasticity issues as already discussed.

$$\log(\text{Net\_pay}_i) = \beta_0 + \beta_1 \text{education}_i + \beta_2 \text{wexp}_i + \beta_3 \text{wexpSQ}_i + \beta_4 \text{gndr}_i + \varepsilon_i \tag{18}$$

3.3.2 2SLS

For the Two-Stage Least Squares we have set of two equations. First equation 19 represents the first stage, where the individual's education is regressed on instrumental variables *educationF*, *educationM* and the rest of exogenous variables. Then the estimated value of education obtained from this regression is used in equation 20, which represents the final model, to account for the endogeneity issue.

First-Stage Regression (Education Equation)

$$\text{education}_i = \gamma_0 + \gamma_1 \text{educationF}_i + \gamma_2 \text{educationM}_i + \gamma_3 \text{wexp}_i + \gamma_4 \text{wexpSQ}_i + \gamma_5 \text{gndr}_i + u_i \quad (19)$$

Second-Stage Regression (Outcome Equation)

$$\log(\text{Net\_pay}_i) = \beta_0 + \beta_1 \widehat{\text{education}}_i + \beta_2 \text{wexp}_i + \beta_3 \text{wexpSQ}_i + \beta_4 \text{gndr}_i + \varepsilon_i \quad (20)$$

4 Empirical results

First, we estimated the OLS model 18. However, as discussed in Subsection 2.2 and Section 1, there is reason to believe that the variable *education* may be endogenous. To address this concern, we proceeded with the estimation of the 2SLS model 20, where *educationM* and *educationF* were employed as instrumental variables for *education*, as specified in Equation 19.

The presence of endogeneity was formally tested using the Hausman test. The results provide strong evidence of endogeneity, as the null hypothesis is rejected with a p-value of 2×10^{-16}, confirming that *education* is indeed endogenous.

To assess the assumption of homoskedasticity, we performed the Breusch-Pagan test. The test yielded a p-value of 2×10^{-16}, indicating the presence of heteroskedasticity in the data. Consequently, we apply White's robust standard errors, as discussed in Subsection 2.3, to ensure valid statistical inference. The estimation results of the baseline OLS model (1), 2SLS model (2)and 2SLS with White's robust standard errors (3) can be seen in Table 3.

As discussed earlier, the OLS model estimates are biased and inconsistent due to the endogeneity of the variable *education*. Consequently, interpreting these results would be misleading. Instead, focusing on the estimates from Models 2 and 3, we observe that the coefficient values are the same for both models, as expected. However, the standard errors differ due to the application of White's robust standard errors, which deal with potential heteroskedasticity in the data.

For the third model (2SLS with White's robust standard errors), the coefficient on *education* is 0.515 and statistically significant at the 1% level. This suggests that moving to a higher education level (e.g., from high school to a bachelor's degree) is associated with an approximate $e^{0.515} - 1 \approx 67\%$ increase in net pay, which is in line with expectations regarding the positive impact of higher education on earnings. The variable *wexp* has a positive and significant effect (0.009), indicating that each additional year of work experience is associated with an increase in net pay of about 0.9%, which is consistent with the expected positive contribution of work experience. However, the squared term *wexpSQ* has a small negative coefficient (−0.0001), significant at the 1% level, implying diminishing returns to experience over time, which aligns with economic theory. The variable *gndr*, with a coefficient of 0.260 and statistical significance at the 1% level, suggests that males earn approximately 29.7% more than females, a result that is consistent with observed gender wage gaps in labor markets.

Table 3. Estimation results

	(1)	(2)	(3)
education	0.182^{***}	0.515^{***}	0.515^{***}
	(0.010)	(0.031)	(0.029)
wexp	0.012^{***}	0.009^{***}	0.009^{***}
	(0.002)	(0.002)	(0.002)
wexpSQ	-0.0002^{***}	-0.0001^{***}	-0.0001^{***}
	(0.00002)	(0.00003)	(0.00003)
gndr	0.243^{***}	0.260^{***}	0.260^{***}
	(0.017)	(0.017)	(0.017)
Constant	6.739^{***}	5.910^{***}	5.910^{***}
	(0.045)	(0.086)	(0.081)

Note: *p<0.1; **p<0.05; ***p<0.01

5 Conclusion

This study aims to determine the effect of education on wages using appropriate econometric techniques. Initially, we estimated an Ordinary Least Squares (OLS) model, but the presence of endogeneity in the *education* variable, confirmed by the Hausman test, necessitated the use of instrumental variables. Specifically, we applied the Two-Stage Least Squares (2SLS) method using parental education as instruments, which proved to be relevant and exogenous based on existing literature. Subsequently, the Breusch-Pagan test confirmed the presence of heteroscedasticity in our data, which was dealt with using White's robust standard errors, ensuring valid statistical inference.

The regression results indicate that education has a significant and positive impact on wages. The estimated coefficient suggests that moving to a higher level of education leads to an approximate 67% increase in wage. Work experience also has a positive effect on wages, though with diminishing returns, as indicated by the negative and significant squared term. Gender disparities were also significant, with males earning higher wages than females, reflecting gender wage gaps.

The most important limitation of this study comes from the challenge of finding a suitable dataset for our research. Consequently, we have assumed that the differences between each level of education are uniform, which may not accurately reflect reality. Treating education as a linear factor could oversimplify the relationship with wages.

References

Arabsheibani, G. R., F. G. Carneiro, A. Henley, & E. Strobl (2010): "Returns to education in four transition countries: Quantile regression approach." .

Blau, F. D. & L. M. Kahn (2019): "Expected work experience and the gender wage gap: A new human capital measure." *ILR Review* **72(1)**: pp. 1–28.

Card, D. (1999): "Chapter 30 - the causal effect of education on earnings." volume 3 of *Handbook of Labor Economics*, pp. 1801–1863. Elsevier.

Chevalier, A. (2004): "Parental education and child's education: A natural experiment." *IZA Discussion Paper* **(1153)**.

Greene, W. H. (2018): *Econometric Analysis*. Pearson, 8 edition.

Patrinos, H. A. (2024): "Estimating the return to schooling using the mincer equation." *IZA World of Labor* .

Wooldridge, J. (2012): "Introduction of econometric: a modern approach 5th ed." *South-Western College Pub* .

Appendix

Breusch-Pagan Test

The Breusch-Pagan test is conducted to detect heteroskedasticity in a regression model by detecting whether the error variance depends on the independent variables as (Wooldridge, 2012).

1. Estimate the OLS model and obtain residuals $\hat{\varepsilon}_i$.

2. Regress the squared residuals on the independent variables

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 X_{1i} + \cdots + \gamma_k X_{ki} + u_i$$

3. Compute the test statistic

$$LM = nR^2 \sim \chi^2(k)$$

4. Hypothesis testing

    - H_0: Homoskedasticity
    - H_1: Heteroskedasticity

CODE:

```
install.packages(c("AER", "lmtest", "sandwich", "MASS", "car", "systemfit", "dplyr", "stargazer"))
library(AER)
library(lmtest)
library(sandwich)
library(MASS)
library(car)
library(systemfit)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(stargazer)

data <- Data_AE3
data <- subset(Data_AE3, data$Net_pay != 0)
data$education <- data$‘highest level of educ‘
data$educationF <- data$‘fathers edu‘
data$educationM <- data$‘motherts edu‘

# Converting the education variables into 5 groups
data$education <- sapply(data$education, function(level) {
  if (level %in% c(0, 113)) {
    return("P")
  } else if (level %in% c(129, 212, 213, 221, 222, 223, 229, 311, 312, 313, 321, 322, 323, 412,
  413, 421, 422, 423)) {
    return("H")
  } else if (level %in% c(510, 520, 620)) {
    return("UB")
  } else if (level %in% c(710, 720)) {
    return("UM")
  } else if (level == 800) {
    return("UD")
  } else {
    return("Unknown")
  }
})

data$educationF <- sapply(data$educationF, function(level) {
  if (level %in% c(0, 113)) {
    return("P")
  } else if (level %in% c(129, 212, 213, 221, 222, 223, 229, 311, 312, 313, 321, 322, 323, 412,
  413, 421, 422, 423)) {
    return("H")
  } else if (level %in% c(510, 520, 620)) {
    return("UB")
  } else if (level %in% c(710, 720)) {
    return("UM")
  } else if (level == 800) {
    return("UD")
  } else {
    return("Unknown")
  }
})

data$educationM <- sapply(data$educationM, function(level) {
  if (level %in% c(0, 113)) {
    return("P")
  } else if (level %in% c(129, 212, 213, 221, 222, 223, 229, 311, 312, 313, 321, 322, 323, 412,
  413, 421, 422, 423)) {
    return("H")
  } else if (level %in% c(510, 520, 620)) {
    return("UB")
  } else if (level %in% c(710, 720)) {
    return("UM")
  } else if (level == 800) {
    return("UD")
  } else {
    return("Unknown")
  }
})

# Preprocessing
data <- subset(data, data$education != "Unknown")
data <- subset(data, data$educationF != "Unknown")
data <- subset(data, data$educationM != "Unknown")

data <- subset(data, data$gndr != 9)
data$gndr[data$gndr == 2] <- 0
summary(data$gndr)
data$age <- data$AGE
data <- subset(data, data$age < 150)
summary(data$age)
data$eduSQ <- (data$eduyrs)^2
data$ageSQ <- (data$age)^2
data <- subset(data, !(data$pdempyr %in% c(6666, 7777, 8888, 9999)))
data$wexp <- 2024 - data$pdempyr
summary(data$wexp)
data$wexpSQ <- (data$wexp)^2

#outliers
boxplot(data$Net_pay, main = "Boxplot of Net_pay")

Q1 <- quantile(data$Net_pay, 0.25)
Q3 <- quantile(data$Net_pay, 0.75)
IQR <- Q3 - Q1
```

```
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

Q11 <- quantile(data$wexp, 0.25)
Q33 <- quantile(data$wexp, 0.75)
IQR1 <- Q33 - Q11

lower_bound1 <- Q11 - 1.5 * IQR1
upper_bound1 <- Q33 + 1.5 * IQR1

# Filtering out outliers
Data_final_OUT <- subset(data, Net_pay >= lower_bound & Net_pay <= upper_bound)
Data_final_OUT2 <- subset(Data_final_OUT, wexp >= lower_bound1 & wexp <= upper_bound1)

# Converting education columns to factors with specific order
Data_final_OUT2$education <- factor(Data_final_OUT2$education, levels = c("P", "H", "UB", "UM",
"UD"))
Data_final_OUT2$educationM <- factor(Data_final_OUT2$educationM, levels = c("P", "H", "UB", "UM",
"UD"))
Data_final_OUT2$educationF <- factor(Data_final_OUT2$educationF, levels = c("P", "H", "UB", "UM",
"UD"))

# Graphical part of the code
education_df <- as.data.frame(table(Data_final_OUT2$education))
educationM_df <- as.data.frame(table(Data_final_OUT2$educationM))
educationF_df <- as.data.frame(table(Data_final_OUT2$educationF))

education_colors <- c("P" = "#08306b",
                      "H" = "#2171b5",
                      "UB" = "#4292c6",
                      "UM" = "#6baed6",
                      "UD" = "#c6dbef")

custom_theme <- theme_minimal() +
  theme(
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(size = 11, color = "black"),
    axis.text.y = element_text(size = 11, color = "black")
  )


p1 <- ggplot(education_df, aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity", color="black", show.legend = FALSE) +
  scale_fill_manual(values = education_colors) +
  labs(title="Individual’s Education", x=NULL, y="Count") +
  custom_theme
```

```
p2 <- ggplot(educationM_df, aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity", color="black", show.legend = FALSE) +
  scale_fill_manual(values = education_colors) +
  labs(title="Mother’s Education", x=NULL, y="Count") +
  custom_theme

p3 <- ggplot(educationF_df, aes(x=Var1, y=Freq, fill=Var1)) +
  geom_bar(stat="identity", color="black", show.legend = FALSE) +
  scale_fill_manual(values = education_colors) +
  labs(title="Father’s Education", x=NULL, y="Count") +
  custom_theme

final_plot <- grid.arrange(p1, p2, p3, ncol=3)

ggsave("education_plots.png", plot = final_plot, width = 12, height = 6, dpi = 300)

wage = ggplot(Data_final_OUT2, aes(x = Net_pay)) +
  geom_histogram(binwidth = 500, fill = "steelblue", color = "black", alpha = 0.7) +
  labs(
       x = "Wage",
       y = "Frequency") +
  theme_minimal() +
  theme(
    axis.title.x = element_text(size = 12, face = "bold"),
    axis.title.y = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(size = 11, color = "black"),
    axis.text.y = element_text(size = 11, color = "black"),
    plot.title = element_text(size = 18, face = "bold", hjust = 0.5)
  )

ggsave("wage.png", plot = wage, width = 12, height = 6, dpi = 300)

wexp_plot <- ggplot(Data_final_OUT2, aes(x=wexp)) +
  geom_histogram(binwidth = 1, fill="steelblue", color="black", alpha=0.7) +
  labs(
       x="Work Experience (Years)",
       y="Frequency") +
  custom_theme

ggsave("wexp.png", plot = wexp_plot, width = 12, height = 6, dpi = 300)

# Recode education levels to numeric values
Data_final_OUT2 <- Data_final_OUT2 %>%
  mutate(education_numeric = recode(education,
                                    "P"  = 1,
                                    "H"  = 2,
                                    "UB" = 3,
                                    "UM" = 4,
                                    "UD" = 5
  ))

Data_final_OUT2 <- Data_final_OUT2 %>%
  mutate(educationM_numeric = recode(educationM,
                                     "P"  = 1,
                                     "H"  = 2,
                                     "UB" = 3,
                                     "UM" = 4,
                                     "UD" = 5
  ))

Data_final_OUT2 <- Data_final_OUT2 %>%
  mutate(educationF_numeric = recode(educationF,
                                     "P"  = 1,
                                     "H"  = 2,
                                     "UB" = 3,
                                     "UM" = 4,
                                     "UD" = 5
  ))

# OLS model
ols <- lm(log(Net_pay) ~ education_numeric + wexp + wexpSQ + gndr , data = Data_final_OUT2)

par(mfrow=c(2,2))
plot(ols)
summary(ols)

# Testing for heteroskedasticity
bptest(ols)

# IV model
iv_model <- ivreg(log(Net_pay) ~ education_numeric + wexp + wexpSQ + gndr |
                    educationF_numeric + educationM_numeric + wexp + wexpSQ + gndr,
                  data = Data_final_OUT2)

summary(iv_model, diagnostics = TRUE)

# Testing for endogeneity
hausman.systemfit(iv_model, ols)

# Testing for heteroskedasticity
bptest(iv_model)

# White’s SE
robust_se <- vcovHC(iv_model, type = "HC1")
coeftest(iv_model, vcov = robust_se)
```