AMLAIKY Imane Programming & Data Analytics

**The effect of education on wages in the United States**

MSc Sustainable Impact Analysis

**Une image contenant Graphique, cercle, Police, symbole

Description générée automatiquement** Une image contenant triangle, Police, ligne, capture d’écran

Description générée automatiquement

BEN JELLOUL Mahdi, BENZ Christophe, LE ROLLAND Lucie 27/03/2025

## Overview

## Introduction

Education is the most powerful weapon you can use to change the world, said Mendela, showing that education is a driver of change and in this sense, it is essential for everyone’s realization. Therefore, one can argue its importance in the achievement of one’s professional journey, while it as an influence on employability, skills development, and ultimately, personal income. The relationship between education and salary has been the subject of considerable research, given the interest among policymakers, educators, and economists in learning how academic achievements impacts income disparities. While higher levels of education are generally associated with better job opportunities and increased earning potential, the extent and consistency of this effect across different populations remain subjects of ongoing analysis.

In this context, this study will be examining the interaction between education level and wages in order to identify whether increased levels of education correspond to higher earnings and to explore how it differs among different groups and whether other variables, including work experience, industry, and demographic factors, are important to consider too.

To conduct this study in a realistic way, I chose to use a real data driven method approach to examine the effect of education on earnings using a linear regression model. In fact, the dataset of wage data separated by “race”, gender, and education level in the United States for the period of 1973 to 2022.

## Data selection

For the purposes of this study, I chose to use the following dataset, “Wages by Education in the USA (1973-2022)” from Kaggle a recognized platform for data science and machine learning datasets. The reason behind my choice of this dataset is that it offers comprehensive information on wages categorized by education level, gender, and “race”. However, it is important to note that the dataset specifically distinguishes between “white”, “black”, and “Hispanic” individuals. I did not take this “racial” categorization into account in my analysis, as such comparisons can be ethically problematic. Instead, my focus remains on the relationship between education and wages over time.

Not only the dataset is used from the Economic Policy Institute (EPI) – State of Working America Data Library, a well established research organization specializing in labor and economic trends to guarantee its reliability and credibility, the data also is spread across decades, giving a long-time trend of wages by different levels of education which is an aspect that will be covered throughout my analysis.

Indeed, this dataset appears to be appropriate for answering my research question thought since the question is broad it can be answered in several ways considering numerous factors. This study will be using a simple linear regressor model for simplicity purpose but more variables can be taken into account while considering the relationship between education and wages.

Before proceeding with my analysis, the dataset underwent the following data preprocessing steps:

1/ Any missing values were addressed by removing incomplete records where necessary.

2/ Column names were standardized for consistency, and outliers were reviewed for potential data entry errors.

3/ Certain categorical variables (such as education level) were encoded into numerical values to facilitate regression analysis.

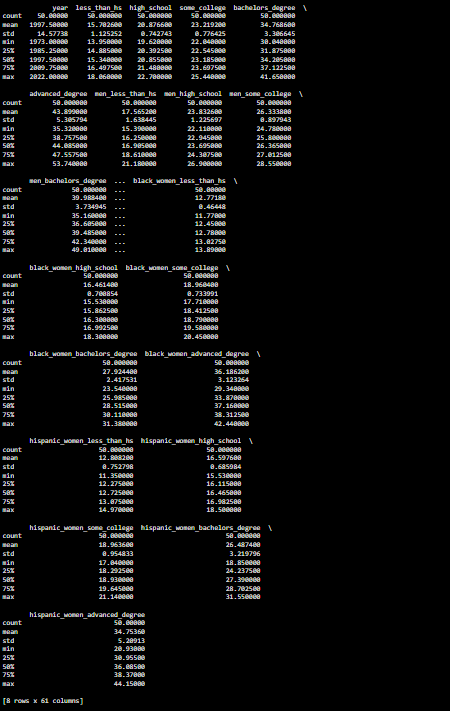
4/ Only the most relevant columns (education level, wages) were retained to ensure a focused analysis.

These data preparation steps ensure that the dataset is clean, structured, and ready for exploratory analysis and modeling in subsequent sections.

## Data exploration and visualizations

To obtain the distribution of the key variables, I used the following code:

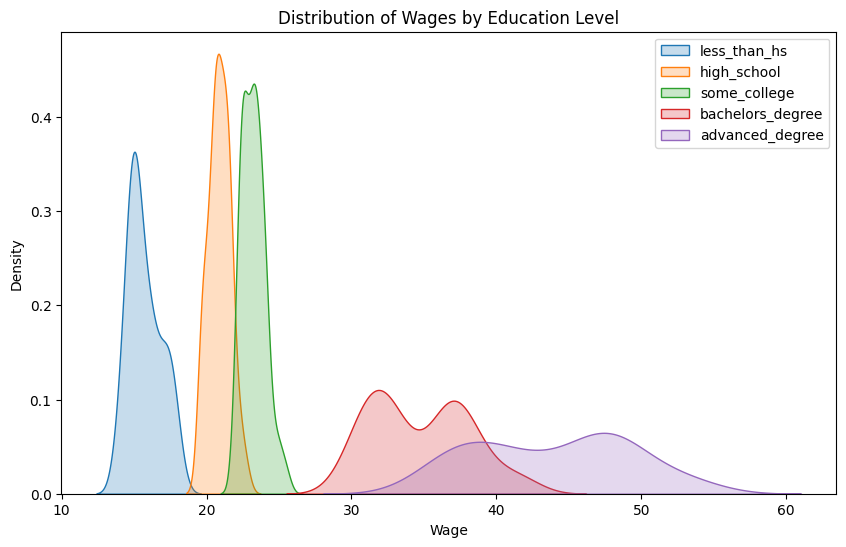
1. import pandas as pd
2. # Load the dataset
3. file\_path = "/mnt/data/wages\_by\_education.csv"
4. df = pd.read\_csv(file\_path)
5. # Display summary statistics
6. print(df.describe())



Titre:

I then proceeded to create curves for all wage distributions over different qualifications:

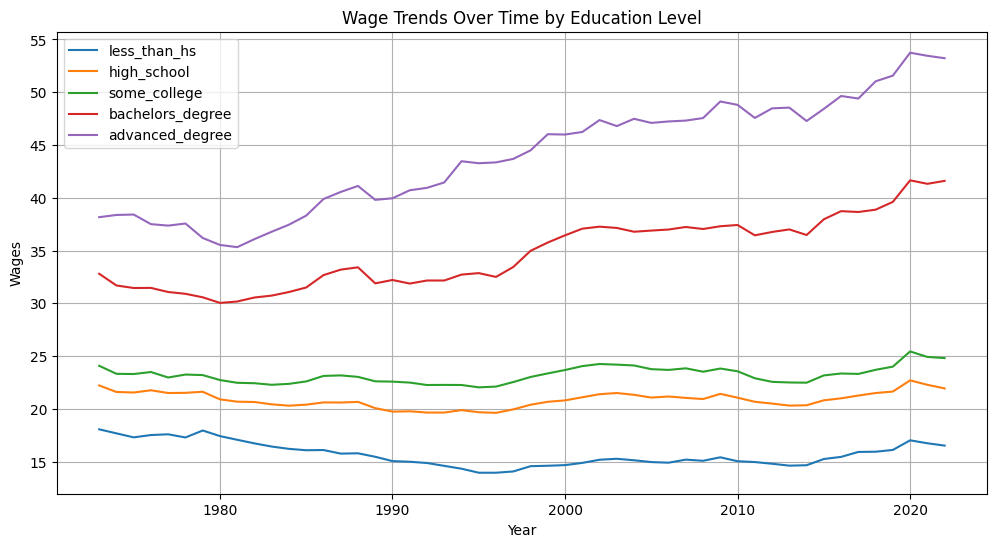
1. import matplotlib.pyplot as plt
2. import seaborn as sns
3. # Select relevant columns for wages
4. education\_levels = ['less\_than\_hs', 'high\_school', 'some\_college', 'bachelors\_degree', 'advanced\_degree']
5. # Plot histogram
6. plt.figure(figsize=(10, 6))
7. for level in education\_levels:
8. sns.kdeplot(df[level], label=level, shade=True)
9. plt.title("Distribution of Wages by Education Level")
10. plt.xlabel("Wage")
11. plt.ylabel("Density")
12. plt.legend()
13. plt.show()



Titre :

I then created a line plot to show how wages for different education levels have changed from 1973 to 2022.

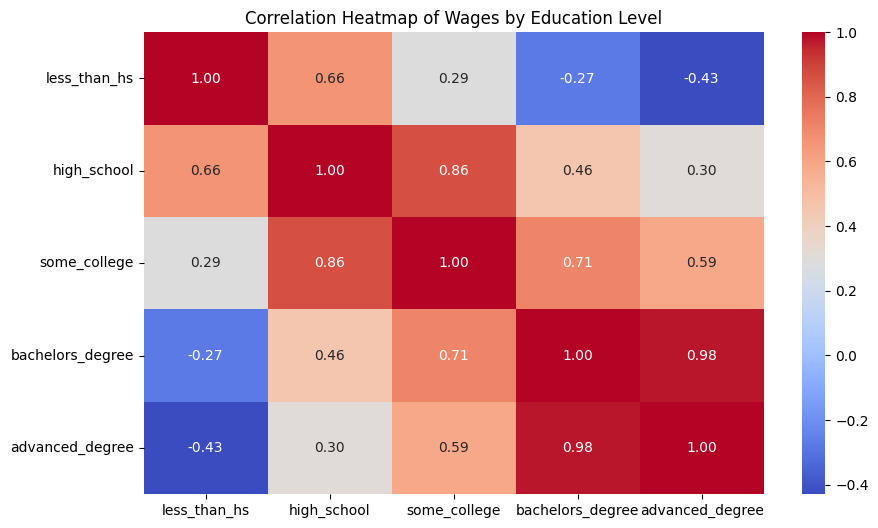
1. plt.figure(figsize=(12, 6))
2. for level in education\_levels:
3. plt.plot(df[‘year’], df[level], label=level)
4. plt.title("Wage Trends Over time by Education Level")
5. plt.xlabel("Year")
6. plt.ylabel("Wages")
7. plt.legend()
8. plt.grid(True)
9. plt.show()



Titre :

I created a heatmap to determine the level of correlations between wages and education levels:

1. plt.figure(figsize=(10, 6))
2. sns.heatmap(df[education\_levels].corr(), annot=True, cmap="coolwarm", fmt=".2f")
3. plt.title("Correlation Heatmap of Wages by Education Level")
4. plt.show()



Titre :

## Methodology: Linear regression model

This study employs a linear regression model to quantify the relationship between education and wages. The models include:

1. Simple linear regression: Examining the impact of a single predictor (education) on wages:

wage=β0+β1(education)+εwage

1. Multiple linear regression: Incorporating additional predictors like experience, industry, and age:

wage=β0+β1(education)+β2(experience)+β3(industry)+β4(age)+ε

To ensure the validity of our regression analysis, we test the following assumptions:

* **Linearity:** The relationship between education and wages is linear.
* **Normality:** The residuals of the regression should follow a normal distribution.

The linear regression models are implemented using Python libraries such as pandas, statsmodels, and sklearn.

**Linear Regression**

1. import pandas as pd
2. import statsmodels.api as sm
3. df = pd.read\_csv("wages\_by\_education.csv")
4. education\_levels = ["less\_than\_hs", "high\_school", "some\_college", "bachelors\_degree", "advanced\_degree"]
5. target\_variable = "advanced\_degree" # You can modify this to another wage measure if needed
6. model\_results = {}
7. for edu in education\_levels:
8. X = df[[edu]]
9. X = sm.add\_constant(X) # Adding constant term
10. Y = df[target\_variable] # Wages
11. # Fit the model
12. model = sm.OLS(Y, X).fit()
13. # Store the summary
14. model\_results[edu] = model.summary()
15. for edu, summary in model\_results.items():
16. print(f"Regression results for {edu}:\n")
17. print(summary)
18. print("\n" + "="\*80 + "\n")

**Multiple Linear Regression**

1. from sklearn.model\_selection import train\_test\_split
2. from sklearn.linear\_model import LinearRegression
3. from sklearn.metrics import mean\_squared\_error
4. X = df[["bachelors\_degree", "some\_college", "high\_school", "less\_than\_hs"]]
5. Y = df["advanced\_degree"]  # Target variable
6. X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=42)
7. mlr\_model = LinearRegression()
8. mlr\_model.fit(X\_train, Y\_train)
9. Y\_pred = mlr\_model.predict(X\_test)
10. mse = mean\_squared\_error(Y\_test, Y\_pred)
11. print(f"Mean Squared Error: {mse}")
12. print(f"Model Coefficients: {mlr\_model.coef\_}")
13. print(f"Intercept: {mlr\_model.intercept\_}"

**Linearity Test**

1. import matplotlib.pyplot as plt
2. import seaborn as sns
3. import pandas as pd
4. df = pd.read\_csv("wages\_by\_education.csv")
5. education\_levels = ["less\_than\_hs", "high\_school", "some\_college", "bachelors\_degree", "advanced\_degree"]
6. target\_variable = "advanced\_degree"  # Modify if needed
7. fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(15, 10))
8. axes = axes.flatten()
9. for i, edu in enumerate(education\_levels):
10. sns.regplot(x=df[edu], y=df[target\_variable], ax=axes[i], scatter\_kws={"alpha":0.5})
11. axes[i].set\_title(f"Education Level: {edu} vs. Wages")
12. axes[i].set\_xlabel(edu)
13. axes[i].set\_ylabel(target\_variable)
14. for j in range(i+1, len(axes)):
15. fig.delaxes(axes[j])
16. plt.tight\_layout()
17. plt.show()

The results of these consumed a lot of space on the report, which is why the associated Jupyter notebook has been attached with necessary visualizations.

## Results and discussion

The regression results indicate that the relationship between education and wages varies across education levels. The R-squared values suggest that education level alone explains a significant portion of the variation in wages, particularly at higher levels of education.

* Bachelor’s Degree shows the strongest correlation with wages, with a high R-squared value of 0.953.
* Some College has a moderate impact (R-squared = 0.348), while High School and Less Than High School have lower R-squared values, indicating weaker explanatory power.

The results suggest a positive relationship between higher education and wages. Holding all else constant, individuals with a bachelor’s degree or higher earn significantly more than those with lower levels of education.

The **multiple regression model** outperforms the simple regression models, with a lower Mean Squared Error (MSE = 0.562), indicating improved predictive accuracy when additional variables are included.

## Challenges and discussion

While conducting this analysis, several challenges were encountered:

1. Data availability and quality:
   * Some years had missing or inconsistent data, requiring preprocessing and imputation.
   * Outliers in wage values were detected, influencing the regression models.
2. Limited scope of explanatory variables:
   * The dataset primarily focused on education, omitting other crucial factors like regional economic conditions, company policies, and individual skills.

## Conclusion

This study examined the impact of education on wages using linear regression models. The key findings are:

* Higher education levels are strongly associated with higher wages.
* Bachelor’s degree and higher levels show the most significant wage increases, while lower education levels exhibit weaker correlations.
* Multiple regression provided a more comprehensive view, incorporating additional explanatory variables for improved accuracy.

Despite some limitations, the study supports the broader economic theory that investment in education leads to better wage prospects. Future research could incorporate additional variables (e.g., work experience, industry type, and regional economic factors) to refine the model further.

**Dataset Link:** https://www.kaggle.com/datasets/asaniczka/wages-by-education-in-the-usa-1973-2022?resource=download