Brian Mmari, Aphia Ishimwe, Gabriel A Samcam Vargas

UNDERSTANDING WAGE DISPARITIES IN MID-ATLANTIC REGION USING DIFFERENT TECHNIQUES.

INTRODUCTION

This paper explores different statistical techniques to understand how wages are influenced by factors such as age, education, and year. The dataset used, "Wage" dataset from the ISLP package, contains wage records along with demographic information. Traditional regression methods such as polynomial regression and spline regression, are employed to capture potential nonlinear relationships between the predictor variables and wages. Additionally, we also implement Generalized Additive Models (GAMs) to provide more flexible modeling techniques and improved interpretability. Through statistical hypothesis testing, such as ANOVA, and logistic regression, we aim to evaluate the significance of our models and predict high-earning individuals. These findings will help contribute to a broader understanding of wage distribution patterns, assisting policymakers and economists in decision-making processes.

The Wage dataset used is described below:

Wage Dataset for a group of 3000 male workers in the Mid-Atlantic region.

- year: Year that wage information was recorded
- age: Age of worker
- maritl: A factor with levels '1. Never Married', '2. Married', '3. '3. Widowed', '4. Divorced' and '5. Separated' indicating marital status
- race: A factor with levels '1. White', '2. Black', '3. Asian' and '4. Other' indicating race
- education: A factor with levels '1. < HS Grad', '2. HS Grad', '3. Some College', '4. College Grad' and '5. Advanced Degree' indicating education level
- region: Region of the country (mid-atlantic only)
- jobclass: A factor with levels '1. Industrial' and '2. Information' indicating type of job
- health: A factor with levels '1. <=Good' and '2. >=Very Good' indicating health level of worker
- health_ins: A factor with levels '1. Yes' and '2. No' indicating whether worker has health insurance
- logwage: Log of workers wage
- wage: Workers raw wage

The details above can be found at the following link: https://islp.readthedocs.io/en/latest/datasets/Wage.html

METHODS

Data Preparation and Preprocessing: The analysis begins with loading the "Wage" dataset and preparing it for modeling. The key variables include wage as the response variable, and "age", "education level, and year as predictor variables. The dataset is preprocessed by transforming categorical variables into numerical format where necessary. Because we only care about "3" categorical variables, and among them, only "education" is categorical, then data transformation is only done in this column. For instance "education" is transformed into numerical variables using one-hot encoding.

Feature Selection: Our paper chose to analyze only 3 categorical variables: "age", "Education", "year" to avoid complexity and because these variables are most determinant to someone's wage salary figures. "Age" is a continuous variable but was analyzed using polynomial and spline regression to capture its nonlinear effects on wage trends. **"**Education**"** is a categorical variable which has a strong correlation to wage levels which makes it crucial for understanding income disparities. **"**Year**"** is included to analyze change of wage trends with time, capturing economic shifts and inflation effects.

Model Training: Several machine learning models were trained using different regression techniques:

- **Polynomial Regression**: Captures nonlinear relationships by introducing polynomial terms of the predictor variable (age). The model coefficients are estimated using Ordinary Least Squares (OLS) which minimizes the sum of squared errors.
    - The poly() model used can be written mathematically as:

$$Y = \beta_0 + \beta_1 X^1 + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + e,$$

where y represents wage, x represents age, and $\beta_{0,\ 1,\ 2,\ 3,\ 4}$ are coefficients of the polynomial curve with their signs indicating the shape of the curve.

Different techniques are used to determine the best fit-degree for the polynomial above. Among the explore methods include Hypothesis testing, ANOVA tests and cross-validation. From the explored varieties, it is determined that a cubic or

quartic polynomial appears to give the best fit because their p-values are closest to 5%.

- **Spline Regression**: This is a more flexible alternative to polynomial regression that divides the data into segments and fits piecewise polynomials. Spline regression was used because it transforms the original predictor variable into multiple basis functions that define different polynomial segments. The number and placement of knots dictate how flexible the spline is. This prevents instability in high-degree polynomials while capturing local variations in wage trends. Mathematically, this can be represented by the following:

$$y = \beta_0 + \beta_1 b_1(x) + \ldots + \beta_{K+d} b_{K+d}(x) + \epsilon.$$

where k is the number of basis functions determined by the chosen knots, $b_i(x)$ represents the basis functions of the spline transformations and β_i are the coefficients to be estimated. In our example, a cubic-spline basis with 3 interior knots is fitted on the wage data.There 7 columns and 6 spline coefficients to account for the intercept. The chosen knots are at ages 33.75, 42.0, and 51.0 and these points are seen where the Spline curve changes direction sharply. The the boundaries of the natural spline graph are linear as opposed to the polynomial graph in order to produce stable estimations

- **Generalized Additive Models (GAMs)**: Generalized Additive Model (GAM) is implemented to model the relationship between wage and the predictors age, year, and education, allowing for non-linear smoothing functions while maintaining additivity. The smoothing parameter λ controls the flexibility of the splines, where low λ (e.g., 0.002) results in a highly flexible, wiggly curve, similar to interpolating splines, while high λ (e.g., 10e+6) produces a nearly straight line, similar to standard linear regression. The code fits GAMs for various λ values and visualizes how smoothness affects predictions. A grid search is performed to find the optimal λ that minimizes prediction error. To ensure consistent smoothness across variables, the degrees of freedom (df) for age and year are set to 4 by approximating the corresponding λ. The categorical variable **education** is converted into an array for proper modeling. The equation can be written as: **wage = β0 + f1(year) + f2(age) + f3(education)** where f1, f2, and f3 are non - linear smoothing functions.

- **Partial dependence plots (PDPs)** were generated for each variable to illustrate their impact on wages while holding other variables constant. The PDP for the years 2003–2009 indicates a general upward trend in wages over time. Regarding age, the PDP shows that wages tend to increase during younger years, stabilize around middle age, and begin to decline after age 60. Since education is a categorical variable, the PDP compares different education levels. Box plots reveal that individuals with higher educational degrees tend to earn higher wages.

- **Logistic regression:** A logistic regression model was used to predict high-earning individuals (wage > $250) using age and education as predictors. The logistic function maps predictions to probabilities between 0 and 1.

  $P(Y=1|X) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}}$ where β are coefficients estimated using Maximum Likelihood Estimator. This allows for the classification of individuals based on their likelihood of earning above $250.

DISCUSSION

Our lab was guided by the need to explore the relationship between wage and several key potential predictors of wage, such as age, education, and year. We aimed to apply reproducible statistical analysis.

We first used polynomial regression to model wage as a function of age. A degree-4 polynomial was selected as the optimal model based on ANOVA comparisons, balancing the risk of overfitting. This approach captured non-linearity in wage trends while maintaining interpretability through confidence intervals.

However, high-degree polynomial models have significant limitations, particularly that distant points can greatly influence the model and limitations in possible extrapolations. To avoid the pitfalls of high-degree polynomial models, we employed spines. B-spines allowed for localized flexibility while controlling for overfitting. The placement of knots at ages 25, 40, and 60 ensured that the model captured meaningful changes in wage trends. Then, Generalized Additive Models (GAMs) were used to extend regression by incorporating smooth functions for predictors, offering greater

flexibility without forcing a specific parametric form. A smoothing spline approach was applied to age and year, while education was treated as a categorical variable.

To model the probability of high earnings (when wage > 250), logistic regression was implemented. A GAM-logistic regression model was chosen for better flexibility in capturing non-linearities in age and year, improving classification performance.

Lastly, we created partial dependence plots to interpret the effects of age, education, and year on wage.

CONCLUSION

The polynomial regression models indicate that wage increases with age up to a certain point, after which it plateaus and slightly declines. This suggests that younger employees experience wage growth, but after a certain age, earnings stabilize or decrease.

The logistic regression model confirms that individuals with higher education(e.g “College Grad” and “Advanced Degree”) significantly increase chances of earning above $250.

ANOVA model selection suggests that adding more complexity to the model(e.g higher degree polynomial) improves fit up to a certain level, but beyond that, the improvement is marginal. A degree-4 polynomial model in this case is considered optimal.

The partial dependence plots show that education is the strongest predictor of wage. From the plots, we can see that people with advanced degrees yield the highest returns.

GAMs provide a more flexible way to observe wage trends over age and education. The smooth functions fitted for different predictor variables indicate that wages are affected in a nonlinear manner by both experience and qualifications.

We pledge our honor that we have not violated the honor code in this assignment!