ISLP Non-linear Statistical Modeling

Gabriel A Samcam Vargas, Aphia Ishimwe, Brian Mmari
SML301, The Center for Statistics and Machine Learning

Background

Limitations:: Traditional linear models often fail to capture complex, non-linear relationships in real-world data.

Problem Statement: To address this challenge, we employ advanced Regression Techniques, including:

- Polynomial Regression (PR)
- Splines (including B-Splines (BS) and Natural Splines (NS))
- Generalized Additive Models (GAMs)
- Logistic Regression (LR)

This project explores the relationship between age, education, and wage using advanced regression techniques.

Linear modeling: Mathematical representation of a situation with a constant rate of change.

Polynomial modeling: type of regression analysis where a polynomial in x represents the relationship between the independent variable x and the dependent variable y.

Aims

In this lab, the goal is to develop a comprehensive non-linear model to explore predictors of wage.

Specifically, we will explore polynomial regression models, step functions, splines, & GAMs, among others.

Methodology

Dataset & Preprocessing: The dataset - Wage Dataset - comprises of a group of 3000 male workers in the Mid-Atlantic region. The response variable was wage, while the explanatory variable was either education or age. Missing values were dropped. Age was transformed using polynomial regression to capture non-linear relationships.

Model Selection & Implementation:

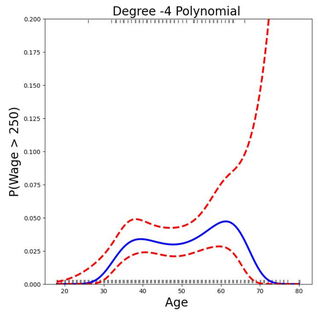
- Polynomial regression was used to fit the relationship between age and wage.
- B-splines and Natural splines were used to improve model flexibility.
- GAMs were fitted using pygam to model wage as a function of age, year, and education.
- Logistic regression was used to predict high-earner status(wage > $250).

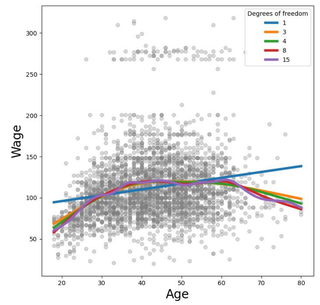
Model Evaluation & Performance Metrics:

- OLS Regression Summary was used to check coefficient significance(p-values).
- **ANOVA Analysis** compared different polynomial degrees and spline transformations.
- **Confidence Intervals** were computed to assess the robustness of predictions.
- **Generalized Cross Validation (GCV) Score** was used to optimize GAMs.

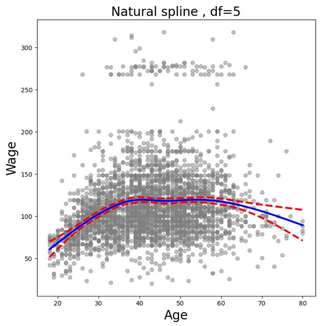
Results

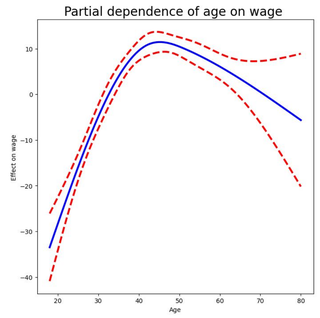
- **Polynomial model:** A degree r polynomial model was used to model wage as a function of age, selected as the most optimal model based on ANOVA comparisons with alternatives. Here, we captured non-lineriarity in wage trends.



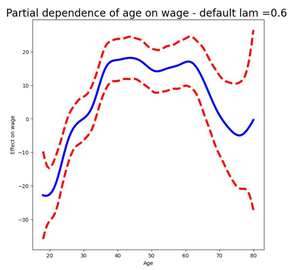


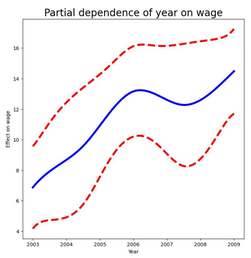
- **Splines:** Splines were then employed to address the risk of overfitting with high-degree polynomial. Placement of knots at ages 33.75, 42, and 51 were done to ensure capturing of meaningful changes in wage trends.

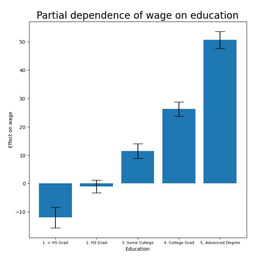




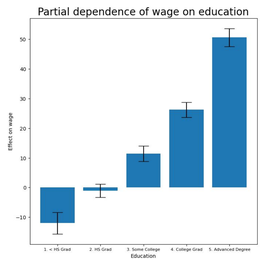
- **Generalized Additive Models (GAMs)**: GAM's were used to extend regression by incorporating smooth functions for predictors. This offered flexibility without forcing a specific parametric form. Smoothing spline was applied to age and year, education was treated as a categorical variable hence PDP compared different degree levels of education.

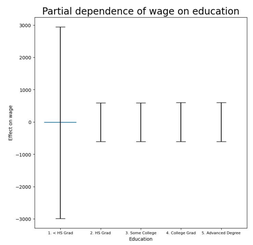


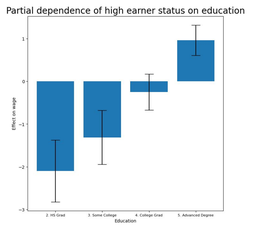




- **Logistic regression:** Lastly, GAM-logistic regression was implemented to model the probability of high earnings, which improved the classification performance.







Conclusions

Logistic Regression Model: The model confirms that individuals with higher education(e.g "College Grad" and "Advanced Degree") have significantly higher chances of earning above $250.

Partial Dependence Plots: The plots suggest that education is the strongest predictor of wages, with HS Grad the least determinant while Advanced Degree being the most determinant.

Polynomial Regression Model: The model confirms that wage increases with age up to a certain point, and then plateaus, and after slightly declines.

GAM Model: The model allows for more flexible, non-linear relationships between predictors and the target variable.

Future Work

- Incorporating job type, region, and work experience could improve wage predictions.
- Exploring Random Forests or Gradient Boosting Models could enhance predictive accuracy.
- Combining SHAP (SHapley Additive Explanations) with GAMs could provide detailed feature importance insights while maintaining interpretability.

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor. An Introduction to Statistical Learning : with Applications in Python. July 5, 2023.

Acknowledgements

- Thank you to our TA- James for providing assistance on both our project report and poster.
- We would also like to say thank you to our course instructor for providing feedback in organizaing our poster.
-

.