STA 210 - Final Project

Zakk Heile & Julia Healey-Parera

Introduction

Project motivation + research question

It is no surprise that early childhood education affects eventual employment. Current research supports the theory that the quality of education received during primary and secondary school has a strong positive relationship with wages earned later in life (Lee & Lee, 2024). Operating under the assumption that this research is accurate, we can thus assume that proper prediction of educational attainment in primary and secondary schooling would allow for estimation of earnings later in life. Thus, this study builds upon prior research and attempts to create a model to predict educational attainment given indicators grouped by regional location (i.e. town).

Dataset explanation

In order to achieve our goal of creating a model for predicting educational attainment, we used a dataset sourced from TidyTuesday and The UK Office for National Statistics. The UK Office for National Statistics is the recognized statistical institution of the nation that carries out the census for England and Wales in addition to the collection of a multitiude of other data made publicly available. The selected dataset, titled "Educational attainment of young people in English towns" details the educational score of each town in the UK using attainment scores from the 2012 key stage 4 cohort of that town. Key stage 4, which is the American equivalent of freshman and sophomore year, are the two years in which students (typically aged 14-16) study for and take General Certificate of Secondary Education (GCSE) exams. Thus, our outcome variable is a level of educational attainment at a time of standard evaluation for students in the UK with an equal metric.

Relevant variables

The variables included in our final model were tested for statistical significance as related to the outcome variable of educational attainment. This is outlined in our methodology section. The five final relevant variables are as follows:

- population_2011 -
- Highest_level_qualification_achieved_b_age_22_average_score -

- level_3_at_age_18 -
- activity_at_age_19_employment_with_earnings_above_10_000 -
- key_stage_4_attainment_school_year_2012_to_2013 -

Lee, Hanol & Lee, Jong-Wha. (2024). Educational quality and disparities in income and growth across countries. Journal of Economic Growth. 1-29. 10.1007/s10887-023-09239-3.

In our dataset, education score ranges from -10.03 to 11.87, population ranged from 5003 to 1085810, highest qualification ranges from 2.57 to 5.14, percent obtained level 3 ranged from 16.54 to 85.71, percent earning >10,000 pounds at 19 ranged from 7.06 to 48.98, and percent of students that achieved 5 or more C or higher certificates ranged from 33.33 to 92.86.

Complete case analysis – get rid of all observations with NAs for one or more variables that we are using in our model.

Link to dataset: https://github.com/rfordatascience/tidytuesday/blob/master/data/2024/2024-01-23/readme.md

https://github.com/zakk-h/STA210Final

Predict key_stage_4_attainment_school_year_2012_to_2013 or highest_level_qualification_achieved_by_a (proportion) - need Beta regression or transform response into log odds and transform back after

Predict activity_at_age_19_full_time_higher_education - logistic

Predict logistic - activity_at_age_19_employment_with_earnings_above_10_000

Population_2011 vs. size_flag

Income_flag vs. job_density_flag

rgn11nm vs coastal

Highest_level_qualification_achieved_b_age_22_average_score

Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions

Methodology/Final Model Results

Our linear model's R-squared metric (both multiple and adjusted) is above 96%. Over 96% of the variation in education scores are explained by our model.

The residual standard error, a measure of the standard deviation of the residuals from the predicted values, is 0.7.

The F-statistic has a p-value of 2.2*10^-16, which is highly statistically significant. We can reject the null hypothesis of all coefficients being 0 with extremely high confidence.

The residual plot shows great symmetry across the residual = 0 line, indicating that the linearity assumption in the parameters is indeed met. Visually, homoscedasticity looks reasonable, but when fitted values are 0, there is more variance. Code was used to find 11 quantiles, split the observations into deciles based on those quantiles, and then calculate the variance of the residuals for each decile. From there, we quantitatively assessed homoscedasticity, and each decile's variance was reasonably close to all others, with the exception being where the fitted values are 0. The Breusch-Pagan test finds statistically significant evidence for heteroscedasticity. However, we moved forward with this model because the heteroscedasticity does not bias the estimates for the coefficients. We also explored transformations in the predictors but did not find them worth pursuing.

The quantile-quantile plot shows that the residuals roughly follow a normal distribution, with the left tail deviating slightly and seeming heavier, but when combined with the histogram, appear sufficiently normal for our purposes.

Independence of residuals: The observations are unique per region, meaning no regions are double counted and each is disjoint, so the observations are fully distinct from one another. Additionally, the data is not collected over time, which could be a problem if the observations overlapped. Thus, the errors are independent, which implies the predicted values are independent.

We predicted an aggregate education score that ranged between -10.028 and 11.872 in our dataset. Our predictors were all untransformed: population of a town/city + average of the highest level of qualification at age 22, the proportion of employed 19-year-olds in that town/city earning more than 10,000 pounds, the proportion of students in that town/city earning more than 5 certificates with grades no less than C. All of our coefficients are statistically significant at the 0.05 level. All predictors except for population increased education scores when each one was increased on its own, holding the other predictors constant. The population slope is negative, though it is very small, is for a one-person increase in population. The slope is much more meaningful when you consider a larger increase like one thousand people.

Discussion

VARIABLE DISCUSSION JULIA

- Summary – what is the impact of each coefficient upon educational attainment and why?
  - Highest_level_qualification_achieved_b_age_22_average_score
    - Higher later educational attainment -> more likely
  - Activity_at_age_19_employment_with_earnings_above_10_000