STA 210 - Final Project

Zakk Heile & Julia Healey-Parera

INTRODUCTION

Project Motivation and Research Question

Current research supports the theory that an individual's educational attainment (number of degrees, general educational achievement) has a strong positive relationship with wages earned later in life (United States Department of Labor 2023). Operating under the assumption that this research is accurate, we can assume that proper knowledge of factors related with educational attainment would allow for better estimation of earnings later in life. Thus, this study builds upon prior research and attempts to identify longitudinal variables (variables over time) with a relationship with educational attainment.

Dataset Explanation

In order to achieve our goal of identifying variables statistically significant in relation to educational attainment, we used a dataset sourced from TidyTuesday and The UK Office for National Statistics. The UK Office for National Statistics is the United Kingdom's recognized national statistical institution that carries out the census in addition to the collection of other publicly available data. The selected dataset, titled "Educational attainment of young people in English towns," contains a variety of variables detailing the educational attainments and qualifications achieved by a given student population in each town in the UK during 2012–the key stage 4 cohort. Key stage 4, which is the American equivalent of freshman and sophomore year, is the two years in which students (typically aged 14-16) study for and take their General Certificate of Secondary Education (GCSE) exams. The students who were in key stage 4 in 2012 were followed over time for further data collection on achievement later in life.

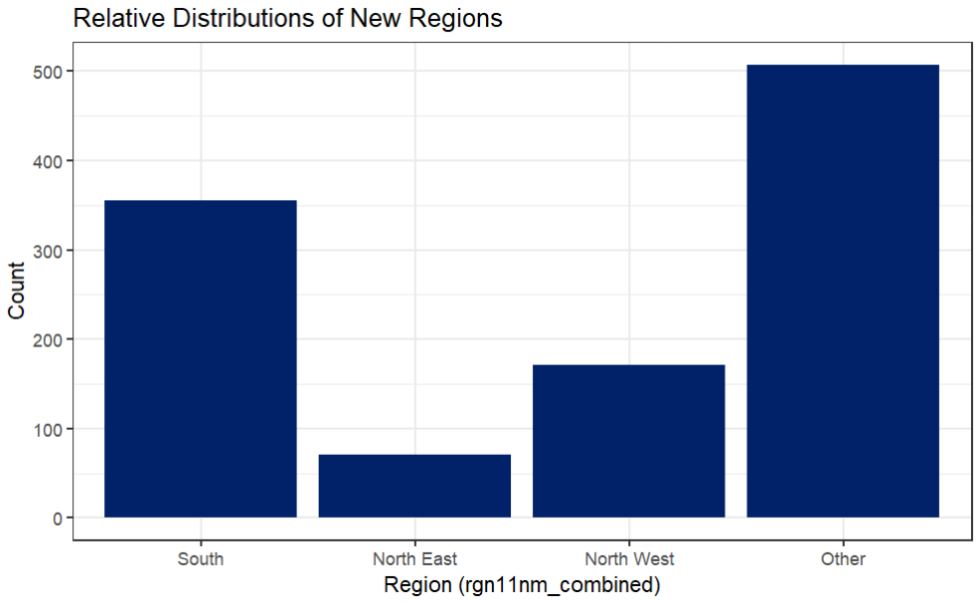
Each row in the dataset represents a given town in the UK. There are 1104 towns documented, each one listed with 30 variables relating to the educational attainment of the 2012 key stage 4 cohort as well as the respective town demographics. Our outcome variable is the town-wide average of the highest qualification (degree) achieved by each individual in the 2012 key stage cohort by the time they reached the age of 22, and our predictors are various other data about the 2012 key stage cohort before they reached the age of 22.

Relevant Variables

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

- `rgn11nm_combined` - This is the region of the UK in which the town is located. There are four factors: Northeast, Northwest, South, and Other.
- `level4qual_residents35_64_2011` - This is the proportion of the towns population aged 35 to 64 that had a qualification of level 4 or above (first year of an American bachelor's degree) in 2011. The proportions were categorized into high, medium, and low.
- `level_3_at_age_18` - This is the proportion of the town's 2012 key stage 4 cohort that achieved level 3 qualifications (equivalent of U.S. high school diploma) by the age of 18. As a proportion, this variable has a possible range of 0 to 100 but an actual range of 16.53543 to 85.71429.
- `activity_at_age_19_full_time_higher_education` - This is the proportion of the town's 2012 key stage 4 cohort that was enrolled in full-time higher education by the age 19. As a proportion, this variable has a possible range of 0 to 100 but an actual range of 7.874016 to 73.44633.
- `key_stage_4_attainment_school_year_2012_to_2013` - This is the proportion of students (in the key stage 4 cohort) in the 2012 to 2013 school year that achieved a grade of A - C on five or more General Certificate of Secondary Education (GCSE) exams. This variable is a measure of educational achievement at the age of ~16 on UK standardized tests. Because this variable is a proportion, it has a possible range of 0 to 100 but an actual range of 33.33333 to 92.85714.

The outcome, `highest_level_qualification_achieved_b_age_22_average_score`, has a range of 2.566929 to 5.142857. Actual qualification levels range from 1 (passing GCSE for grades 1, 2, and 3) to 8 (PhD or other doctorate degree). Because this variable is an average of highest qualification scores, the actual range is significantly smaller than the possible range.



Initially, the variable `rgn11nm` (region of UK the selected town is in) had many more factors; however, because not all of them were significant, we combined the categories in a statistically- and logically-sound manner. The factors South East and South West were combined into the statistically significant factor of South (used as reference level), North East (p-value of 2.03 * 10^-12) and North West ((p-value of 1.31 * 10^-9) were left as-is because their coefficient p-values were below 0.05, and all other factors were combined into factor Other (p-value of 6.11 * 10^-6). While these groupings may see arbitrary, they are not as they follow geographic groupings. This was the only variable manipulation we performed. The relative distributions of the new variable (`rgn11nm_combined`) factors are visualized above.

METHODOLOGY

In order to construct our final model, we chose variables that were statistically significant and intuitively relevant to our predictor. Some predictors were initially considered because of this intuitive relevance but were later eliminated for being statistically insignificant. These variables were:

- `population_2011` - The population of the selected town (observation) in 2011 (continuous).
- `size_flag` - The size of the town, categorically grouped into "Cities," "Large Towns," "Small Towns," etc. (categorical).
- `coastal` - Whether or not the town is coastal or not (categorical).
- `coastal_detailed` - Incorporates the town's size along with whether or not it is coastal (categorical).

- `job_density_flag` - Whether a town is working, residential, or mixed (categorical).
- `income_flag` - Whether a town is low-, mid-, or high-income (categorical).

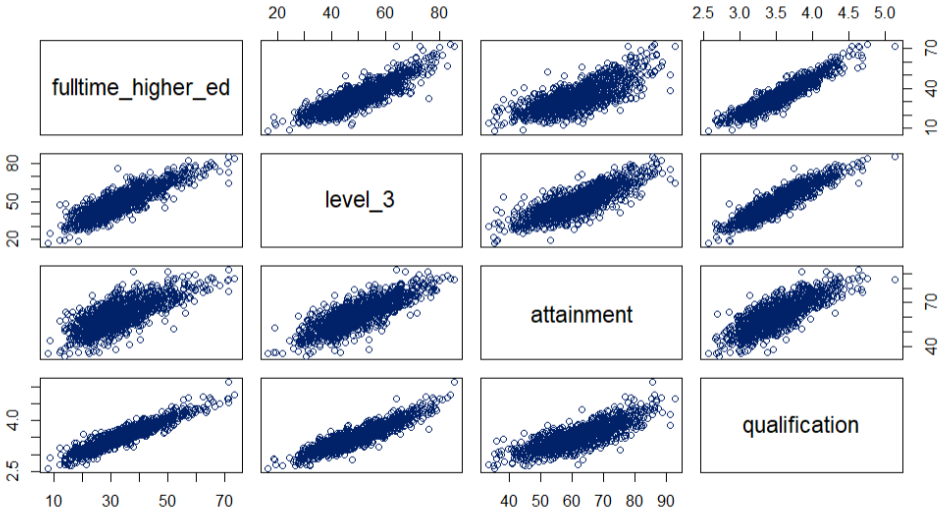
We eliminated all of the above variables for having p-values above 0.05 and thereby being statistically insignificant. We then iteratively tested for any statistically significant interaction terms between the remaining variables and found only one (`level4qual_residents35_64_2011*level_3_at_age_18`). In terms of transformations, our assumptions were sound and allowed us to conclude that no variable transformations were necessary.

Residual Standard Error: 0.09655 on 1088 DF

Multiple R-squared: 0.9282; Adjusted R-squared: 0.9275

Our linear model's R-squared metric (both multiple and adjusted) is above 92%. Over 92% of the variation in average highest qualification score is explained by our model. The residual standard error, a measure of the standard deviation of the residuals from the predicted values, is 0.09655.

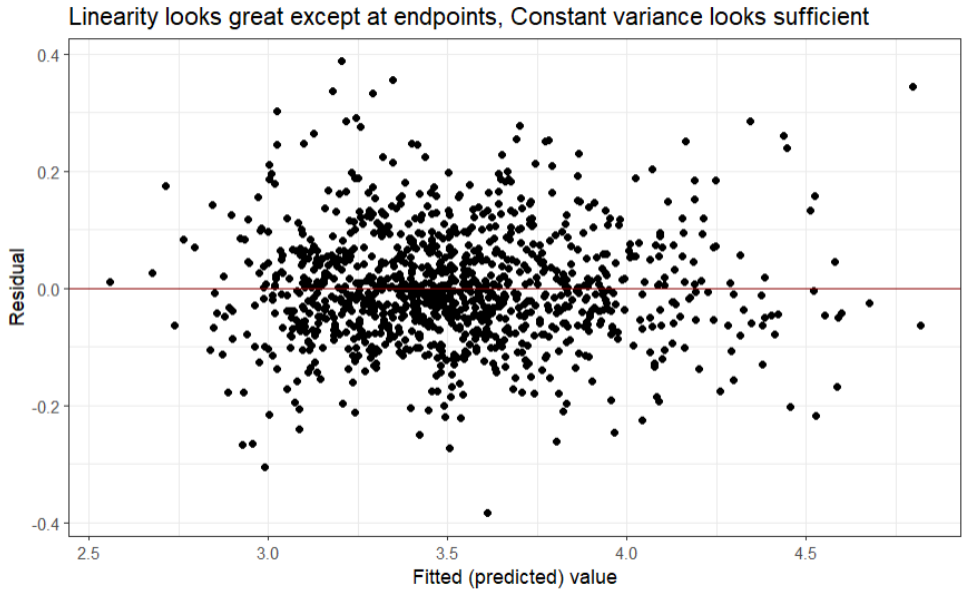
Pair Plots of Continuous Variables - No Significant Multicollinearity



Above are pair plots displaying the linear relationships between all continuous variables. Because there appeared to be strong linear relationships between all of these continuous variables, all variables were tested for multicollinearity. After statistical analysis described below, we concluded that there is no significant multicollinearity and proceeded with our study.

F-statistic: 1406 on 10 and 1088 DF, p-value: < 2.2e-16

The model F-statistic has a p-value of 2.2*10^-16, which is highly statistically significant, and all included variables are also highly statistically significant with similarly low p-values. Thus, all variables pass the test for multicollinearity. In addition, the generalized variance inflation factors were all below the statistically-accepted threshold of 10. Thus, 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 there is more variance when fitted values are 0.

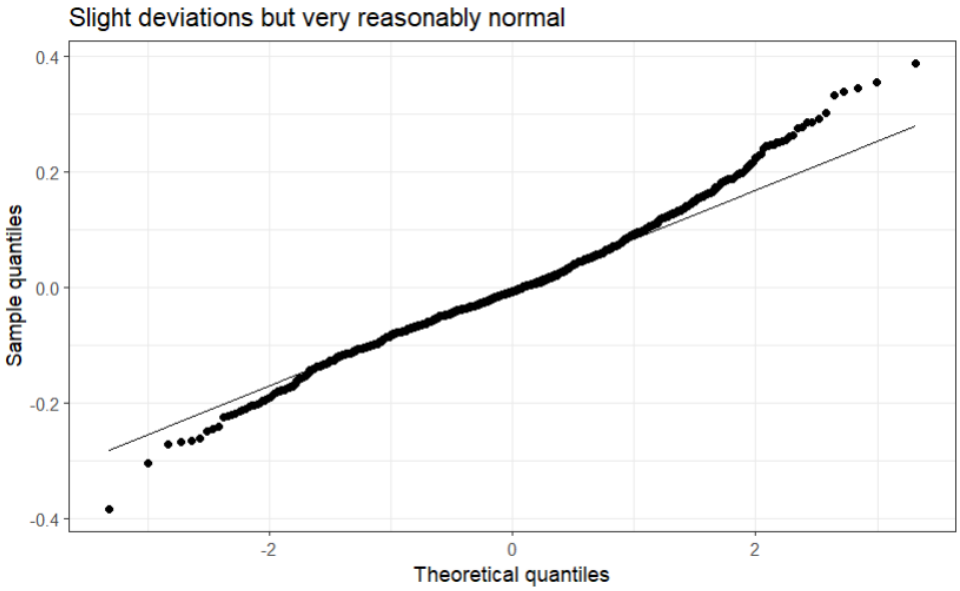
QUINTILE RANGE	~VARIANCE
2.56-3.22	0.010
3.22-3.4	0.0086
3.4-3.56	0.0077
3.56-3.78	0.0092
3.78-4.82	0.010

Code was used to find 6 quantiles, split the observations into quintiles based on those quantiles, and then calculate the variance of the residuals for each quintile. 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.

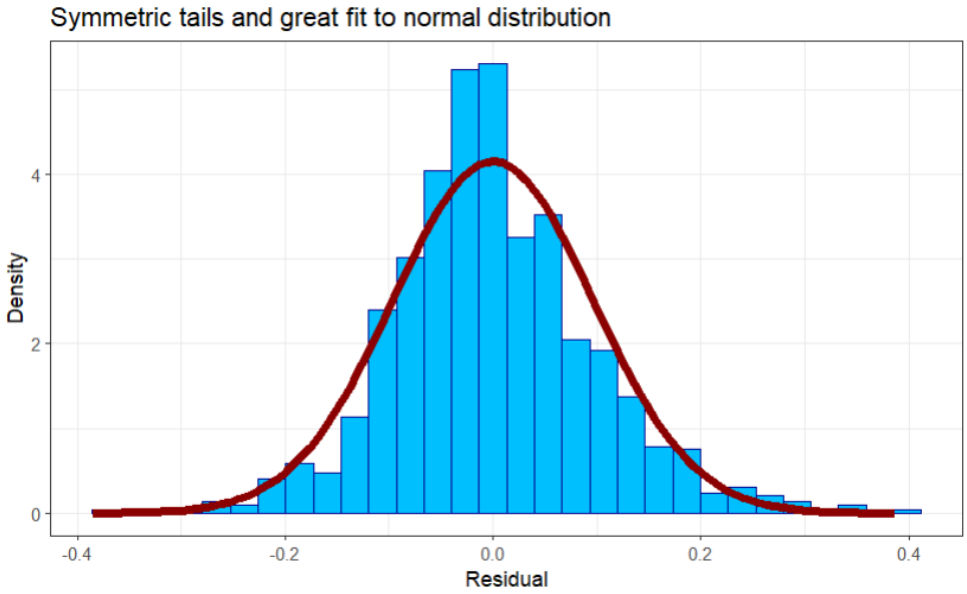
Breusch-Pagan test: BP = 27.512, df = 10, p-value = 0.00216

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. Specifically, we tried Box-Cox transformations.



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



Thus, in terms of independence of residuals, we conclude the following: The observations are unique for each town, meaning no town is double counted and each one is disjoint. Thus, the observations are fully distinct from one another. Therefore, we can state that the errors are independent, thereby implying that the predicted values are also independent.

RESULTS

We predicted the average highest education qualification for a city/town achieved by the 2012 key stage 4 cohort at the age of 22. Our predictors were all untransformed and one interaction term was included:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.8241957	0.0737405	24.738	< 2e-16	***
level4qual_residents35_64_2011Low	0.2471406	0.0764781	3.232	0.001268	**
level4qual_residents35_64_2011Medium	0.2875048	0.0783453	3.670	0.000255	***
level_3_at_age_18	0.0144596	0.0012259	11.795	< 2e-16	***
activity_at_age_19_full_time_higher_education	0.0200731	0.0006169	32.538	< 2e-16	***
rgn11nm_combinedNorth East	0.0918015	0.0129024	7.115	2.03e-12	***
rgn11nm_combinedNorth West	0.0571483	0.0093396	6.119	1.31e-09	***
rgn11nm_combinedOther	0.0323596	0.0071201	4.545	6.11e-06	***
key_stage_4_attainment_school_year_2012_to_2013	0.0031308	0.0005040	6.212	7.42e-10	***
level4qual_residents35_64_2011Low:level_3_at_age_18	-0.0033080	0.0012279	-2.694	0.007170	**
level4qual_residents35_64_2011Medium:level_3_at_age_18	-0.0040273	0.0012151	-3.314	0.000949	***

All of our coefficients are statistically significant at the 0.05 level. All predictors except for two interaction coefficients increased average highest education when each one was increased on its own, holding the other predictors constant. The two negative slopes were minimal in magnitude. Our linear model presents compelling evidence to support the idea that all of our predictors have a relationship with average highest education in a town/city, and all except the interaction terms have a positive relationship.

All variables and coefficients are explained in our Discussion. Some of our most interesting coefficients, however, are those pertaining to `rgn11nm_combined`. The interpretations for those three coefficients are as follows:

1. North East, 0.0918015 - We expect a town in the North East of the UK to have an average highest qualification level of their 2012 key stage 4 cohort at age 22 0.0918015 higher than a town in the South, all else held constant.
2. North West, 0.0571483 - We expect a town in the North West of the UK to have an average highest qualification level of their 2012 key stage 4 cohort at age 22 0.0571483 higher than a town in the South, all else held constant.
3. Other, 0.0323596 - We expect a town not in the South, North West, or North East of the UK to have an average highest qualification level of their 2012 key stage 4 cohort at age 22 0.0323596 higher than a town in the South, all else held constant.

This is further discussed below.

DISCUSSION

Variable Discussion

In discussing our model, we will first go through each variable and consider the possible reasoning behind its coefficient.

To begin, `rgn11nm_combined` indicated that all towns not in the South were expected to have higher average highest qualifications in the key stage 4 2012 cohort than towns in the South (Northeast, 0.0918015; Northwest 0.0571483; Other 0.0323596). This is in line with another article written using this dataset, "Why do children and young people in smaller towns do better academically than those in larger towns?" which found that students in more rural areas achieve higher levels of educational achievement than those in more urban areas (Office for National Statistics 2023). These findings are in line with those of this model, which illustrates that towns nearer to London (the UK's main metropolitan area, located in the South), are expected to have lower averages of highest qualifications achieved. We find it notable that the Northeast and Northwest regions have higher average highest qualification compared to the South, given that Oxford is in the South [West].

The next variable, `level_3_at_age_18`, measures the proportion of the town's students who were a part of the 2012 key stage four cohort that obtained level 3 qualifications, the U.S. equivalent of a high school diploma. The coefficient for this variable, 0.0144596, makes sense–one would expect a town with a greater proportion of students who achieved level 3 qualifications by the age of 18 to have a a higher average highest qualification level by the age of 22. `key_stage_4_attainment_school_year_2012_to_2013` has a similarly logical coefficient estimate, 0.0031308, which indicates that a town with a 2012 key stage 4 cohort that scored higher on GCSEs is expected to have achieved higher average highest qualifications by the age of 22. Thus, a town with students that performed better on standardized exams is expected to have a young adult population having obtained degrees of higher status.

The last significant base-level variable in our final model, `level4qqual_residents35_64_2011`, served as a measure of educational attainment of the adults (parents) in the community surrounding the stage 4 cohorts of the time. The coefficients for this variable, 0.2471406 (low) and 0.2875048 (medium), are not exactly what one would expect. Logically, it would make sense for these coefficients to be negative; towns with fewer adults with high-status degrees would then have correspondingly lower average highest number of degrees. However, this is not the case. Towns with both low and medium proportions of adults aged 35 to 64 with level 3 or above qualifications are expected to have higher average highest number of qualifications among 22-year olds from a standardized school year. This may be because of educational investments targeting areas without great parental education to help compensate for previous injustices and inequities.

The final variable included in our model is an interaction term between `level4qqual_residents35_64_2011` and `level_3_at_age_18`, meaning that the relationship between the proportion of students with qualifications above level 3 at age 18 and the average highest qualifications of students at age 22 depends on the number of adult residents (between the ages of 35 and 64) with qualifications above level 4. The coefficients for these terms are intuitive; the interaction terms corresponding to low and medium proportions of adults with above level 4 qualifications are negative, meaning that an increase in proportion of students with above level 3 qualifications at age 18 will correspond to the greatest increase in average highest number of qualfications at age 22 for towns in which the proportion of adults with qualifications above level 4 is high. Perhaps this is the case because students with level 3 qualifications will go on to pursue further qualifications at a greater rate if they are surrounded by adults who have obtained qualifications above level 4 given their positions as role models and influences, either actively or passively.

In general, the findings of this model are fairly tailored to the dataset used. The indicators used, though transferable in concept to other possible predictors, are unique to the data collected for this dataset and are thus difficult to generalize for other relationships. This model and dataset are also extremely unique to the structure of education in the UK. Although the concept of qualifications is transferable to the degree system of the United States, there are other patterns that are unique to the makeup of the United Kingdom. For example, the observation that educational attainment and degree obtainment is greater in smaller towns rather than cities is not in line with the observed truth in the United States, where cities are known to be centers of innovation, opportunity, and education. Because of these idiosyncrasies and others, the conclusions drawn from this study cannot be widely applied across all educational systems outside of the one being studied.

Other Limitations and Future Research

Our data potentially captures individuals' locations at the time of surveying, which could assign them to towns different from where they received their earlier education, notably year 13 and earlier, which we are interested in. Two potential causes of this would be relocating for work or college. Whether or not this is a limiting factor depends on how exactly the data was collected. Additionally, we grouped a variety of geographic regions into the Other category in

variable `rgn11nm_combined`, which increased statistical significance but reduced granularity or specific insights.

Future research could investigate how individuals perform in college (if they go) to see how well their earlier studies prepared them for it. Additional variables that would aid in predicting later educational performance are: average income in area, student:teacher ratio, school funding per pupil, and poverty percentage. We are fascinated by the reason that an increase in population is associated with a decrease in education scores, which is in contrary to our intuition for the United States education system, and believe it is worth additional exploration. Future research could deploy a multinomial model to predict regions or types of towns/cities based on education statistics as well, which would target a different question than the one we are addressing, but an interesting one nonetheless.

Works Cited (CSE)

Office for National Statistics. 2023 Jul 25. Why do children and young people in smaller towns do better academically than those in larger towns? Office for National Statistics. [accessed 2024 April 23]. https://www.ons.gov.uk/peoplepopulationandcommunity/educationandchildcare/articles/whydochild 07-25.

United States Department of Labor. 2023 Sep 6. Education pays. U.S. Bureau of Labor Statistics. [accessed 2024 Apr 24]. Available from: https://www.bls.gov/emp/chart-unemployment-earnings-education.htm

Appendix

Our code is available on Github here: https://github.com/zakk-h/STA210Final/tree/main.

The dataset used is available here: https://github.com/rfordatascience/tidytuesday/blob/master/data/2024/2024-01-23/readme.md