Predicting Educational Outcomes

Milestone II Project, SIADS 694 & 695
Fall 2021 / Winter 2022
Adrian Wallace (bluewall@umich.edu), Tanner Olson (tanolson@umich.edu), & Susannah Trevino (sustrev@umich.edu)
https://github.com/bluewallumich/Predicting-Educational-Outcomes

Overview

Through learning comes the ability for an individual to improve the quality of their life. The great Nelson Mandela went as far to say, “Education is the most powerful weapon which you can use to change the world.” However, educational opportunities globally are scattered. To see if we can highlight the disparity of education across the world, we will dive into education attainment, nation expenditures on education, and world university ranking data to compare countries' education opportunities. Our hypothesis that we will be testing is that countries that spend more on education for their citizens have both better educational outcomes and a higher quality of life. Also, as expenditure data is incomplete for many countries, we will also substitute country GDP for education expenditure to test if countries with higher GDP also have higher educational outcomes.

Data Sources

Multiple data were utilized for this project, including:

#	Data Set	Description	Provider	Years	Links
1	World University Rankings	Ranks top universities in world	Center for World University Rankings	2011-2016	https://cwur.org/
2	Country Expenditures	List % of GDP spent on various initiatives by country	World Bank	1995-2011	https://data.worldbank.org/indicator/SE.XPD.TOTL.GD.ZS
3	Educational Attainment	Level of education by different demographics within a country	World Bank	1985-2015	https://databank.worldbank.org/source/education-statistics:-Education-Attainment
4	GDP	Normalized GDP by country	World Bank	1960-2022	https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Data Cleaning

Although the data was downloadable in CSV files from the providers, the data was raw and unfiltered. The first step in the data cleaning process was to ensure country names were consistent across the files to allow for merging data-sets once data files were cleaned up. After

looking through the files and finding inconsistencies, a script was created to loop through the file's country column and update varying names to consistent names.

Now with a consistent country naming convention in place, the next step was to dive deeper into each individual data-set to better explore the data to begin to understand how to best approach cleaning and merging.

- **Country Expenditures**:This is the first file that we dove into as it consists of the % of GDP that a given country spends on education. The expenditure file only had 37 unique countries represented with a row signaling all school spending, k-12, and university expenditures for the given country. For this project, we filtered on all school spending. Once filtered, there were gaps from year to year making it hard to decide which years to focus on. In order to control for this, we decided to take the mean of expenditures from 2000 to 2011 to account for any potential missing years. After these operations, we were only left with 30 out of the original 37 countries in our data set to further explore total expenditures.
- **GDP:** The GDP file was next to explore. GDP data was missing considerably less data points and had a representation of 266 countries.To merge the expenditure data, we took the mean of the years 2000 to 2011 as this data set would ultimately be joined with the Country Expenditure data detailed previously. The product of a country's % of GDP spending and their actual GDP would give us the actual spending of a nation on school investments.
- **Educational Attainment:** This data set consisted of 79k total rows of education attainment questions for 188 different countries. Many countries were missing answers from multiple questions and years. As each country had around 500 different survey questions and responses, we first had to determine which questions would help us answer our question of how a country's wealth and expenditure on education impacts educational outcomes for a nation. After sifting through all of the questions, it was determined that we would use the number of years of school for the population age 25+, number of years of school for the population age 25+ female, and percentage of population 25+ with no education to go deeper on. After filtering on these questions, we followed a similar approach and filtered the data from 2005-2015. A reason for this is that data before 2005 was largely missing. We are alright with the gap between data sets as well as we are making the assumption that actual school output is delayed post GDP and expenditure investments as well. Thus a 5 year delay in schooling outcomes vs investments, smoothed out by taking the mean of multiple years, should not greatly impact outcomes.
- **World University Rankings:** For this dataset, we grouped the list of all universities by country to merge the number of countries and mean college rankings of those countries.

Merging Data Sets

Once these data sets were cleaned by smoothing out the data with averaging over multiple years, dropping NAs, and making the country column consistent for all data sets, we merged the data sets into two different data frames. One data frame consisted of educational outcomes and expenditures. The other education outcomes and GDP. These files act as the base files for allowing us to better visually explore the data and begin running models on the data to see if we can extract helpful patterns to assist us in answering our research questions.

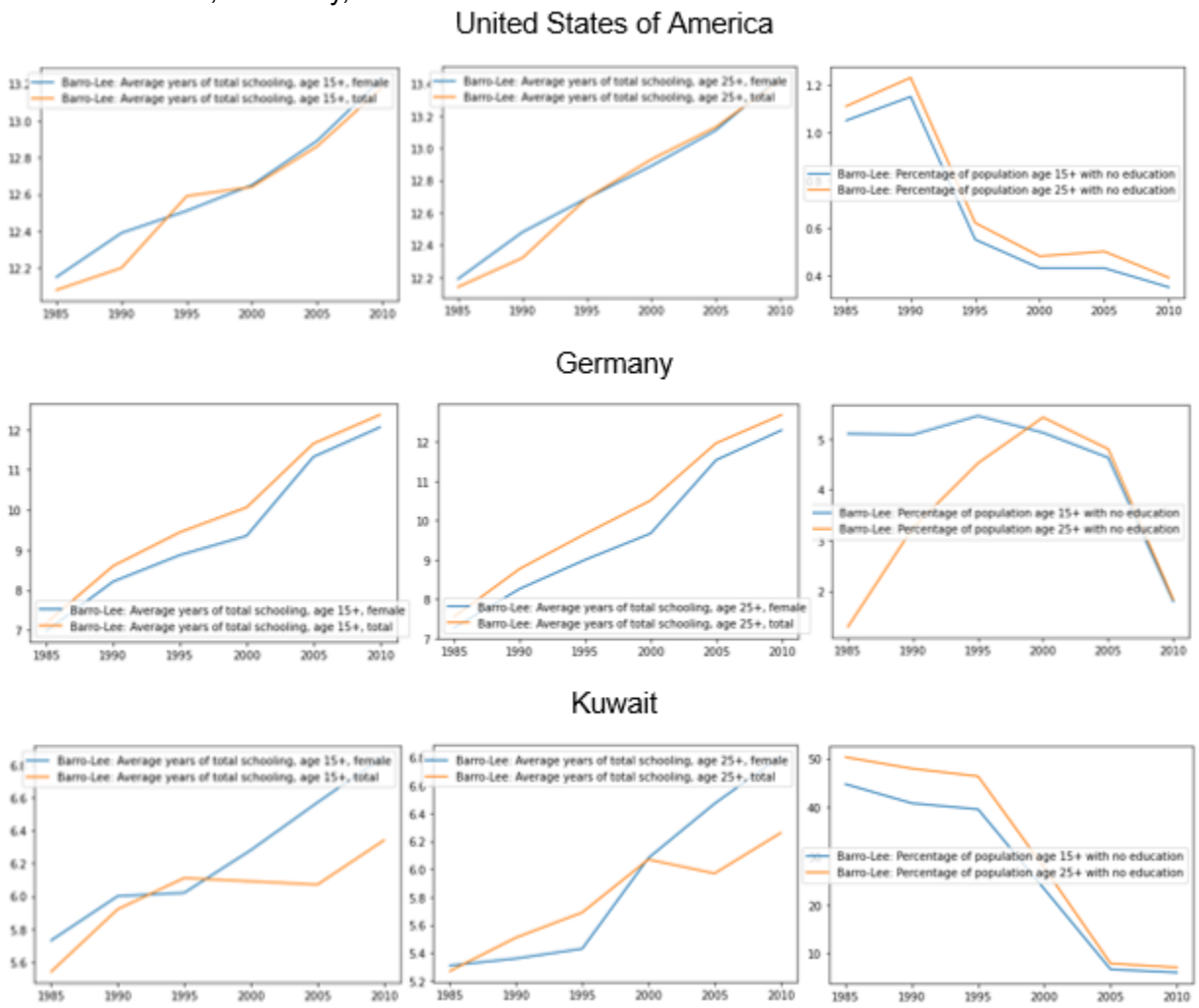
Shortcomings of Data Set

Although we have good compliance for country GDP data, there is not great compliance around school expenditures by each country (~30 countries). Having a small sample size is not optimal here and will have to be factored in when we run our models. Also, we are defining educational

outcomes by the number of years of schooling a person had the opportunity to have at age 25+. Number of years of schooling does not necessarily mean quality of education.

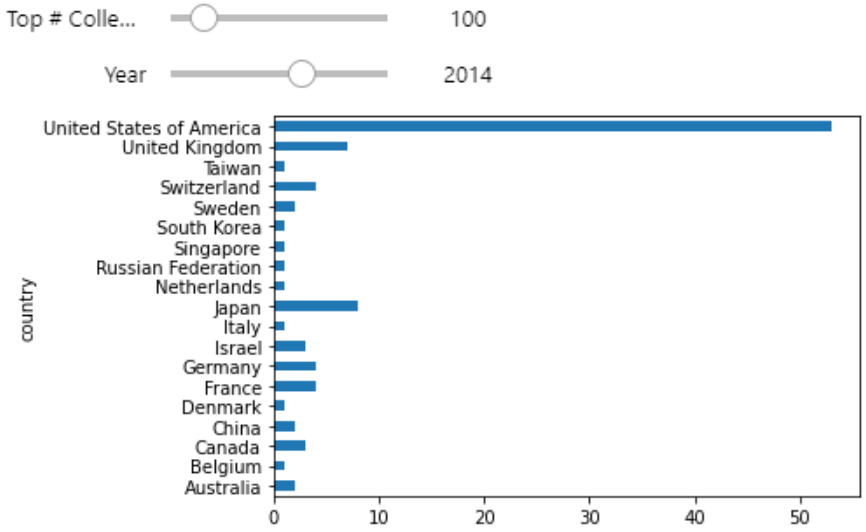
Data Exploration

To better understand the data we're working with, we created interactive visuals to explore different facets of our dataset through matplotlib and ipywidgets. For the Educational Attainment dataset, we separated the data by country and isolated two questions per graph to compare trends. To demonstrate this, we've highlighted three countries in the below visual: the United States of America, Germany, and Kuwait.



Many of the available countries demonstrate trends of increasing years of total schooling and decreasing percentages of no education, with varying starting values and slopes. Germany is an interesting outlier for "no education." This may be due to hidden variables or data collection variances related to East and West Germany's reunification in 1990. Kuwait is an interesting outlier for the gender differences demonstrated in average years of schooling. We can see at both the 15+ and 25+ age levels, except for a dip around 1995, females have a higher average than overall. This may also be due to world events, as the early 1990s saw significant warfare and political upheaval in Kuwait and the surrounding region. Although we may not know all of the contributing factors for each country's data, we can intuit that there is the potential for many hidden variables.

Additionally, our data exploration revealed an ethical consideration regarding potential biases. We created an interactive visual to model how many countries would be included if we included all ranked colleges regardless of their country. No matter our year or cutoff point for how many top colleges to consider, the United States has far and away the majority of ranked institutions. Therefore, it is important to keep in mind that our available data is heavily skewed towards the USA.



Unsupervised Learning

Motivation
The goal for the unsupervised learning portion was to see if there were any preliminary patterns that would give insight into whether or not GDP and expenditures have an impact on educational outcomes.

Data Source
For the unsupervised learning portion, we leveraged the data sets that were cleaned and merged as previously described in the section before. However, manipulations were done on the base datasets in order for them to be prepared to run through an unsupervised learning model. The rows of the data consisted of three rows per country. Each country had a row for the following questions and survey response answers:

1) Barro-Lee: Average years of total schooling, age 25+, total
2) Barro-Lee: Average years of total schooling, age 25+, female
3) Percentage of population age 25+ with no education

In order to further clean the dataset to prepare it for the unsupervised learning portion, these questions were pivoted so that they were represented as columns instead of rows. With this transformation, each country now only represented one row of data but had all the needed data points as columns.

The columns we were most interested in for trying to find any preliminary patterns before we ran supervised models were 1) Average years of total schooling, age 25+, total, 2) avg_2000_2015 which is a column containing the average GDP per country, and 3) school_expenditures which

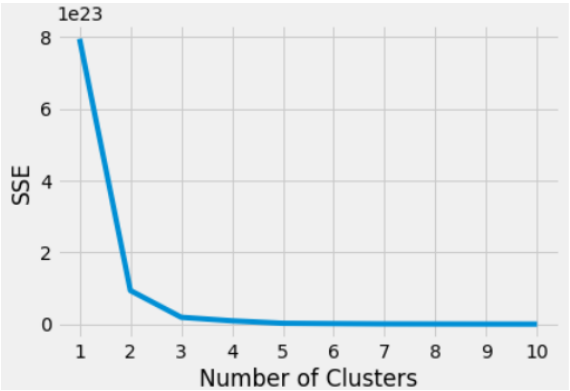
is a column containing how much money a country on average spent specifically towards educational expenditures. The original dataset containing country responses to survey questions was around 79k rows of data with 30 columns with more than 50% of the cells being NaNs. The modified data set we have for running our models in the form explained is 143 rows with 8 columns. Each row represents a unique country. Lastly as previously explored, the expenditure and GDP data was smoothed out by taking the average between 2000-2015 while the educational assessment data is the average between 2005-2015. The five year gap is driven by the assumption that GDP growth and school expenditures investments are delayed in their actual impact to educational outcomes.

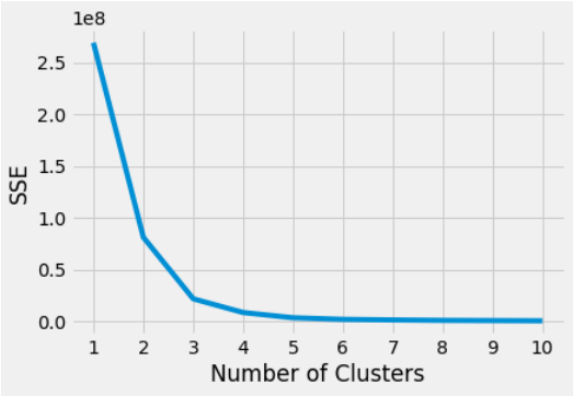
Unsupervised Learning Methods

As we were interested in seeing if there were any patterns between a country's GDP, School Expenditures, and educational outcomes, we leveraged k-means clustering. K-means clustering was leveraged due to its ability to generally form tight clusters, flexibility, and easy visual capability to see centroids of clusters. By being able to see the centroids of the cluster, we will be able to see directionally if countries with higher GDP and School expenditures have better educational outcomes.

To prepare the data for the k-means model, we first need to scale the data to better normalize. This helps normalize the data between the features we are going to run through the model. In this case we are going to run two models: 1) School Expenditures and Educational and Outcomes and 2) GDP and Educational Outcomes. Post scaling, we have the data prepped to run through the models. In order to tune for the optimal number of clusters, we leveraged the “elbow method”. The elbow method calculates the WCSS (within-cluster sum of squares). By plotting the WCSS against the number of clusters, we can visually see which cluster has the most rapid improvement (this is the “elbow”). This is essentially the ideal cluster for the specific dataset as it helps reduce and tune the model to reduce noise.

For 1) School Expenditure and Educational Outcomes, the optimal number of clusters is 2 as seen below (left). For 2) GDP and Educational Outcomes, the optimal number of clusters is 3 as seen below (right).

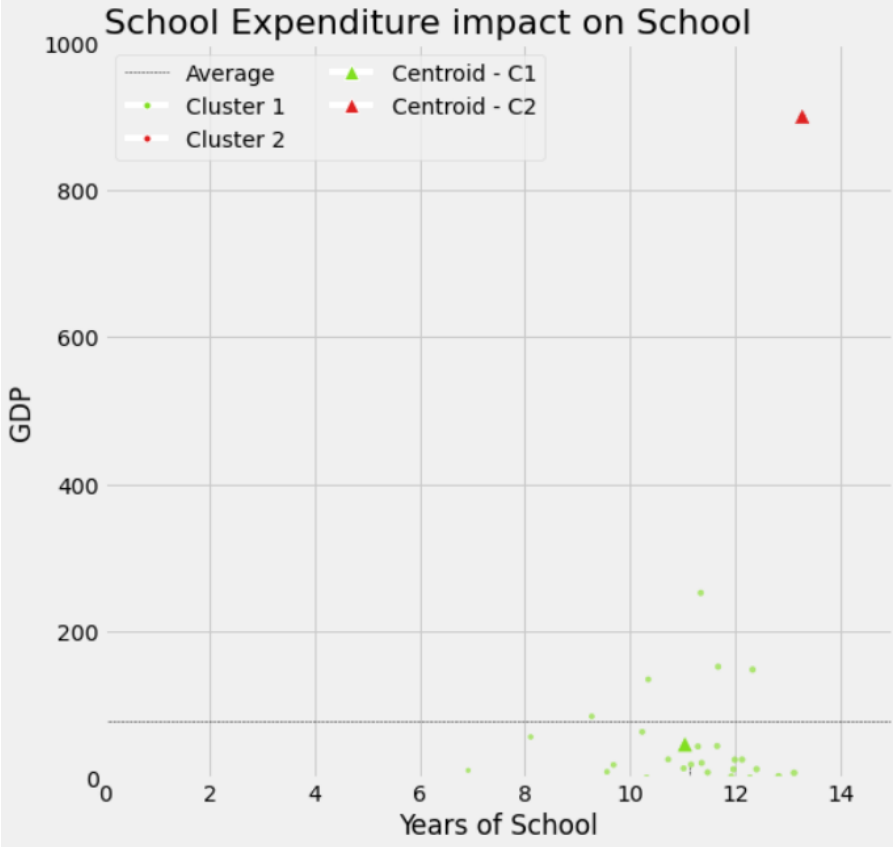




Now with the number of clusters to run each k-means model with to look at the data, we ran the models which produced the following clusters:

Note: The top right centroid is the US. They spend considerably more on education than any other country and have the highest average years of education for persons age 25+. The

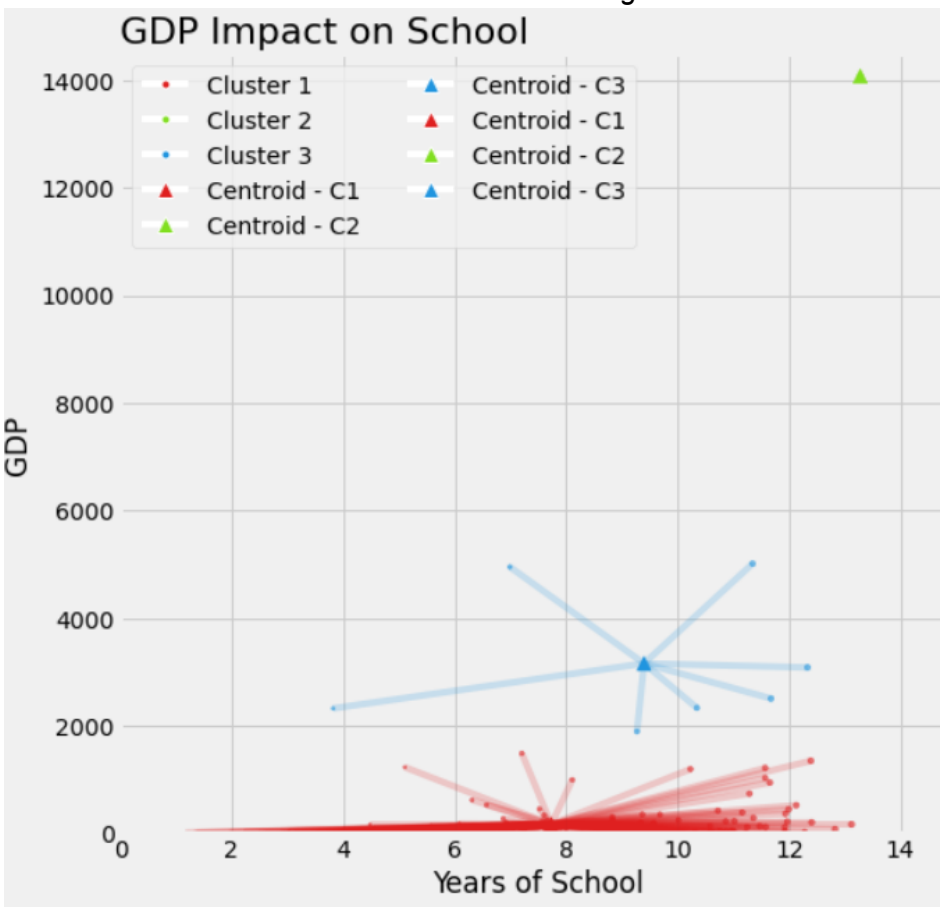
centroid for the other clusters is around 11 years of schooling for the US 13, however, there are some countries that appear to be close to 13 years of schooling and spend considerably less.



cluster	female_25_plus	All_25_plus	All_25_plus_no_edu	cen_x	cen_y
0	10.879615	11.042692	2.655577	11.042692	45.882424
1	13.275000	13.275000	0.445000	13.275000	899.541961

Note: We see a similar pattern for GDP as well. The US is the top right green centroid indicating that the US has the highest GDP and appears to be the highest average educational outcomes for the country as a whole. It visually appears that the other two clusters follow the belief the higher the GDP the more educational opportunities on average for schooling, however, there

are obvious countries with low GDP that have high educational outcomes.



cluster	female_25_plus	All_25_plus	All_25_plus_no_edu	avg_2000_2015	country_id	cen_x	cen_y
0	7.348037	7.733741	19.471926	163.317689	70.792593	7.733741	163.317689
1	13.275000	13.275000	0.445000	14074.337562	136.000000	13.275000	14074.337562
2	9.125000	9.396429	6.435000	3154.151372	65.714286	9.396429	3154.151372

Unsupervised Learning Conclusion

From the above visuals, it appears we have directional supporting data to infer that School Expenditure and GDP impact educational outcomes. Because there is such a disparity between some of the countries vs other countries for School Expenditures and GDP, it appears that there is more work to do to better understand how and if we can better normalize the data and if we should remove outliers. There are also countries that spend low amounts of money comparatively on School Expenditures and have low GDP who have very good educational outcomes. More should be done to dive into some of these edge-cases to understand what is driving these countries to have quality education with lower amounts of resources. This could be helpful in understanding how other countries who fall into a similar category with School Expenditures and GDP could model their education system after to have better outcomes with limited resources.

Supervised Learning

Motivation
The goal for the supervised learning portion was to try and find the correlation between dependent and independent variables. No features in the dataset could be considered confounding as each was impacted by GDP. Patent is an exception as some countries with low GDP had more patents.

Data Source
For the supervised learning portion, we leveraged the data sets that were cleaned and merged as previously described in the Data Sources section. However, manipulations were done on the base datasets in order for them to be prepared to run through supervised learning models. The following three columns were obtained from the unsupervised learning section:

1) Barro-Lee: Average years of total schooling, age 25+, total
2) Barro-Lee: Average years of total schooling, age 25+, female
3) Percentage of population age 25+ with no education

The Center for World University Rankings dataset was used to provide ranking data across eleven features. A subset was created, taking the schools whose national rank was equal to 1. The subset DataFrame was grouped and averaged by country. After the initial pivoting was completed in the unsupervised learning section, the resulting DataFrame was merged with the Center for World University Rankings subset DataFrame. The resulting DataFrame left only 59 columns while still allowing for 11 feature columns.

The following columns were also used in the supervised learning process:
1) Score
2) Institution
3) Country
4) Quality of Education
5) Broad Impact
6) Patents (instrument)

Supervised Learning Methods & Evaluation
As we were interested in seeing the correlation between country GDP and country features, several supervised models were used to determine the impact of GDP on female education. IV Two-Stage Least Squares was chosen due to the modeling error that occurred when creating the DataFrame. Uncertainty was introduced when the dataset was simplified to 59 rows. Therefore, the full unsupervised dataset was used as a comparison.

As we were interested in seeing if there was any correlation or impact GDP has on Average years schooling for females, age 25+, these two variables were chosen from both datasets to be the exog and dependent variables respectively. Score was used as the endogenous variable and patents was used as the instrument variable (This is in reference to the shortened dataset. The full dataset only uses the dependent and exogenous variables.)

The subset data tells us that p is significant at 0.1, 0.01, and 0.05 levels. This is due to the lack of data (59 rows) used and will be discussed in the failure section. The full dataset is also

significant but with 143 observations, it is hard to tell if there is still model error. This could be determined in the future using an unfiltered dataset.

Sub Dataset

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
avg_2000_2015	-0.0007	0.0002	-4.3083	0.0000	-0.0010	-0.0004
score	0.2073	0.0088	23.480	0.0000	0.1900	0.2246

Endogenous: score
Instruments: patents
Robust Covariance (Heteroskedastic)
Debiased: False

Full Dataset

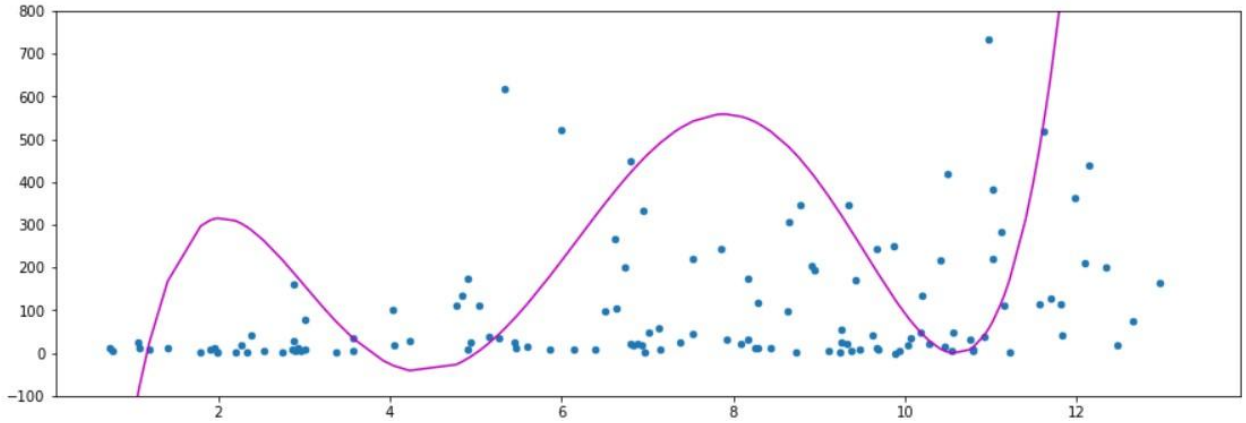
Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
avg_2000_2015	0.0020	0.0007	2.7304	0.0063	0.0006	0.0034

Other models were used to determine how sensitive the data was. Information gained from the Polynomial model tells us that the data has very low sensitivity. In other words the data is subject to the noise of other features. At degree a of 5 we get a sign wave pattern that hits about 9 points. At higher degrees we start to hit more data points but not without the cost of adding noise.

Combined Failure Analysis

R2_Score: 0.3228102486210609
Mean_squared_error: 1274982.5354354023
Root_mean_squared_error: 1129.15124559795
Mean_absolute_error: 538.688875646782



The Unsupervised and Supervised analysis fails on several points. Averaging features such as patents left mismatch data. For example a country with low GDP might have more patents than

a high GDP country. The countries with high GDP also had colleges which were ranked very poorly. This skewed the data and made it look like low GDP countries were well off.

Discussion

Part A Learnings

In the supervised learning portion we went beyond linear regression models to determine if GDP affects average female attendance over 25+. We learned that there were some disparities between countries with high and low GDP. With a low p, all models (IV2SLS, Polynomial, Lasso, GaussianNB, 5-Fold Cross Validation) seemed to be significant. The exception to this was the low sensitivity produced by the small amount of data produced by merging. This was a surprise as we thought the high GDP producing countries would remain at the upper levels. Also the difference between the Mean Squared Error and Mean Absolute Error became too high as more data was taken away. With more time we would have been able to find a way to maintain most features and data points.

Part A Ethical Issues

There are always internal and external factors that affect a countries or schools' status. Apart from that, the collection and parsing of data can become misleading when multiple datasets are merged. If we lied with data, such as removing features like country or even misort the data, it could be used to lead potential students to attend college in the poorest countries.

Part B Learnings

From the k-means unsupervised model, it is clear that there is a directional relationship between school expenditures and GDP on a country's education attainment. This became more obvious when we added centroids in the visual to see the average of each cluster. With more time and resources, we would have extended our solution to better account for outliers. For example, the United States GDP made the US an outlier as it was not close in value to any other country. Also, there were several countries that had low GDP and investments in schooling, however, they still had excellent educational attainment results. With more time, it would have been insightful to run a model specifically on this cluster to see if we could pull out specific patterns for lower GDP nations.

Part B Ethical Issues

An ethical issue that could arise for part B is how we are dealing with NaN values. As countries around the world have various internal and external factors that impact data collection and quality (e.g. conflict), a more robust understanding of these potential factors at the country level may help speak more to the data. For example it may be better to exclude countries altogether that were experiencing hardships during the years we averaged out results.

Statement of Work

This project was the result of a collaborative effort. Team members utilized both synchronous and asynchronous communication channels to discuss progress, roadblocks, and next steps. Github was utilized to maintain version control and allow all team members to view and contribute to the codebase. Adrian led the supervised learning section. Tanner led the unsupervised learning section. Susannah led the data exploration and visualization sections. All team members contributed to data cleaning and writing the final report. All team members reviewed each other's work for accuracy and clarity throughout the process.

References

Agarwal, A. (2018, October 8). *Polynomial Regression*. Medium. https://towardsdatascience.com/polynomial-regression-bbe8b9d97491

Kirenz, J. (2021, December 27). *Lasso Regression with Python*. Jan Kirenz. https://kirenz.com/post/2019-08-12-python-lasso-regression-auto/

linearmodels—Linearmodels v4.25 documentation. (n.d.). Retrieved January 30, 2022, from https://bashtage.github.io/linearmodels/index.html

scikit-learn developers. (n.d.). *1.1. Linear Models*. Scikit-Learn. Retrieved January 30, 2022, from https://scikit-learn/stable/modules/linear_model.html

Scikit-learn: Machine learning in Python—Scikit-learn 1.0.2 documentation. (n.d.). Retrieved January 30, 2022, from https://scikit-learn.org/stable/