Proposal

*Qinzhe Wang*

*3/15/2018*

Project Topic:

Predict average economic mobility in a community based on its educational effectiveness.

Introduction:

Economic mobility is the ability that an individual improves his or her economic status. It may be considered a type of social mobility, which is often measured in change in income. And economic mobility is often measured by movement between income quintiles ([Wikipedia](https://en.wikipedia.org/wiki/Economic_mobility)). The average economic mobility in a community may depend on numbers of factors, for example, location and education. Some economists suggest that in general, economic mobility is positively correlated to educational effectiveness. In this research, one of the main focuses is to look at whether the idea is correct or not by examining the link between educational quality and economic mobility. Also, if there is a relationship in between, we are able to predict economic mobility based on education level.

Questions to Address:

* What can be some representative measurements of educational effectiveness and are there any multicollinearities among those measurements? Sometimes is hard to measure education quality because the information online is incomplete. We have to choose some typical predictors by doing comparisons. Then, the multicollinearity has impacts on selecting our linear model because if a predictor in the model can be linearly predicted by another variable with a substantial degree of accuracy, we probably change our model by adding an interaction term or removing some predictors.
* How do these measurements correlate to economic mobility and how significant the relationships are? And are there any variables have impacts on the relationship between education and economic mobility? On the one hand, the degree of significance of each predictor can be determined by P-values; on the other hand, we plan to do the model selection for those predictors with large P-values to make a decision whether we need to remove these terms.
* How to select the best model? Or which model is the best one for prediction? Generally, we will compare test mean squared error to determine the best model (the lower MSE is, the better the model is). And sometimes the relationship in between cannot be linearly explained. Thus, our group wants to apply nonparametric regression, for example, kernel regression.

Data Explanation:

The dataset is taken from a comprehensive study examining factors which predict the ability of individuals to live out the American dream: move up in economic status, going from poor to rich. The dataset has 741 observations, which means there are 741 communities in total. And there are 12 columns in the given dataset:

* ID: an index to identify the community.
* Name: community names.
* Mobility **(dependent variable)**: for one who was born between 1980 and 1982, the probability that the person jumps into the top quintile from the lowest quintile at age 30.
* State: the state where the community located.
* Income: average income in the community in 2000.
* Schooling\_spending: average tuition fee in public schools in the community.
* Teacher\_student\_ratio: the ratio between the number of students and the number of teachers.
* Graduation: the residuals from a linear model “Actual Graduation rate ~ household income per capita”.
* Longitude: the longitude of the community center.
* Latitude: the latitude of the community center.
* Racial\_Segregation: the racial of segregation, separating into three levels (low, medium and high).
* Urban: logical variable, whether the community is an urban or not.

Initial Data Processing:

* We need to remove incomplete rows from the dataset in the beginning because we do not need rows that contain missing values.
* We need to scratch some specific variables that are correlated to education effectiveness from the dataset to do further analysis.
* We may also use data wrangling expressions that such as filter, group\_by, mutate, summarize, arrange, gather, separate and join to get our integrated table for analysis.

Plots:

* We may come up with a map for economic mobility comparison among cities, with aesthetics of position and color because we may see some geographic patterns.
* We may draw scatterplot matrix for all continuous variables and draw boxplots for discrete variables to visualize the predictors in exploratory data analysis part.
* We may draw heat maps of predict value (economic mobility) that against two predictors related to education (2D heat maps).

Models:

* We will do model selection to choose the best **linear model** base on cross-validation (leave-one-out or k-fold).
* We will do **nonparametric regression** (kernel regression) if linear models cannot well explain the correlation.
* We may also do **generalized additive model** (GAM) if needed.

Prediction:

After choosing the best model, if there are correlations in between, our group want to predict the economic mobility by education information and draw the heat maps (mentioned in Plots part) to see the prediction pattern.