Upward Intergenerational Social Mobility

Leon Luo, Mark Fu, Tony Pappas

July 2021

Project Overview

This project uses statistics and model building to predict the Intergenerational Social Mobility Rate, as defined in Chetty et al. 2020. This project, an intersection of data science and social science, consists of data collection, statistical analysis, data revision and merging, and model building. The project first generates correlation matrices to observe correlations between mobility rate and independent variables, such as parent median income, kid median income, college tiers, median income differences, and percent of parents earning in 1st quintile, and the dependent variable, social mobility rate. Through the investigation in correlation with additional exploratory analysis with graphs and plots, unnecessary variables are filtered out and more accurate multivariable regression models are built.

Related Work

Our data and analysis drew from an Opportunity Insights paper examining intergenerational mobility by Chetty, Friedman, Saez, Turner, and Yagan.

During the planning of this project, we acquired a more precise understanding of social mobility on Lumen Learning, a teaching website where key social-related terms are defined. This website provides a clear explanation of the important factors that influence social mobility, including financial and material resources, job markets, and educational opportunity.

In addition to getting an idea of how social mobility is characterized and analyzed, we studied a research paper on social mobility: "How do we characteristically measure and analyze intergenerational mobility?", written by Florencia Torche. From there, we further developed our knowledge of different types of mobility and the approaches to measure mobility rate.

Lastly, we used Opportunity Insights, a database established by Harvard professors, which utilizes data from publicly available federal government records. We compared different datasets related to social mobility, scrutinized all the presented variables, and descriptions of different studies.

Initial Questions

Our research question is: **What socioeconomic factors influence social mobility rate in the United States?** As we moved into exploratory analysis after obtaining necessary variables, we wanted to know the extent of each feature's influence on the dependent variable: social mobility rate. After generating correlation matrices and multiple linear regression analysis, we sought to test the influence of variables such as parent median income, kid median income, and college tier. For future study, we would like to consider more independent variables, such as location, race, gender, and measure the indication of variance in social mobility rate on wealth gap, and the impact of policy on mobility rate within society.

Data

Once we decided to research social mobility, we looked for datasets that yielded enough features to make a multiple linear regression model, where the dependent variable would be social mobility. Social mobility would be either portrayed as parent income subtracted from kid income or as a rate, such as the probability of a kid having an income in the top quintile of income distributions while having parents earning in the bottom quintile. Some key features we hoped to use as independent variables were education level, race, and housing. Many datasets only had one of the factors we were looking for but we eventually found several datasets that matched our needs. Opportunity Insights, a research institute focused on finding solutions for economic disparities, had several datasets that had features such as mean incomes, median incomes, and social mobility rates to make our dependent variable. They also collect government data on location, education level, race, birth year, and number of people in each family. All financial data was normalized to the 2014 US dollar value.

Unfortunately, We were able to merge datasets with only education level and parent income, and child income as dependent variables. The merge took place on both features **state** and **par_q1** (fraction of parents in the bottom quintile for income distribution). Before the merge we cleaned up the datasets by eliminating extraneous variables, which were almost all string data types. The only string features we kept were **state** and **tier** (rank for level of college/university). The **tier** column was converted to integers from 1 (most prestigious schools) to 14 (not attending college between the ages 19-22) by creating a library pointing each level description to a number. This was a key step because now we could use **tier** as a variable in correlation matrices and our multi-linear regression model. Another pre-merge change was multiplying the second dataset **par_q1** column by 100 and then rounding both **par_q1** columns to the fourth decimal point. This was because the first dataset measured **par_q1** as a percent and the second dataset presented it as a decimal. To finalize the merged dataset we divided the parent and kid median incomes by 1000 so that our visualizations would be easier to interpret.

Exploratory Analysis

Following the merge of our two main datasets we graphed correlation matrices to determine which variables we would want to use for our multi-linear regression model. Collinearity was discovered after running a correlation matrix for the merged dataframe (fig 1). Both parent median income and kid median income had high correlation and indicated collinearity. As a result, parent median income and kid median income could not be used as independent variables in the same multi-linear regression model. The matrix revealed positive correlations between **normed_mr_kq5_pq1** (mobility rate) and **k_median**, as well as a negative correlation between parent median income and college tier or kid median income. This means that as the college tier decreases to 1, the highest school level, median income increases.

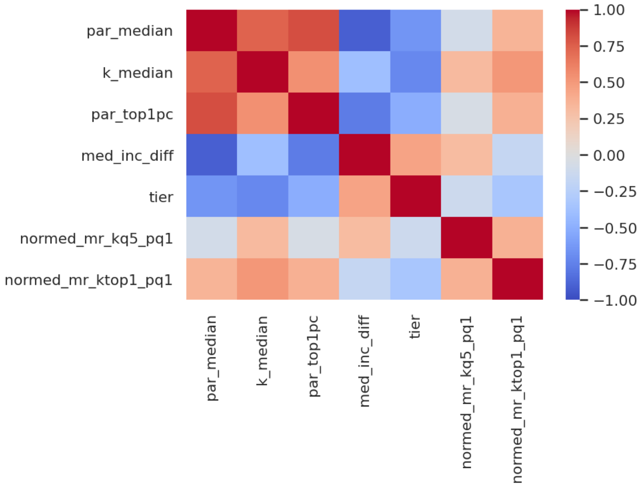


Figure 1: Correlation matrix. Note strong correlation between k_median and par_median, indicating collinearity.

After creating the correlation matrices, we looked at other descriptive statistics for our variables. This process involved compiling averages, standard deviations, and counts for our independent and dependent variables. We were able to verify our distribution graphs (histogram plots) based on this information. The most important distribution graphs were for the difference between kid and parent median incomes and the difference between kid and parent mean incomes. The median income difference histogram plot revealed that most of our data was negative or that most kids had lower income than their parents (fig 2). The mean income difference revealed outlier points such as a mean parents income of 1.6 million dollars with a mean kid income of 16 thousand dollars. We determined that the parent income features were in fact representations of household income and could consist of more than one earner. Due to this discovery we made the parent and kid median income features independent variables and made the social mobility rate our dependent variable.

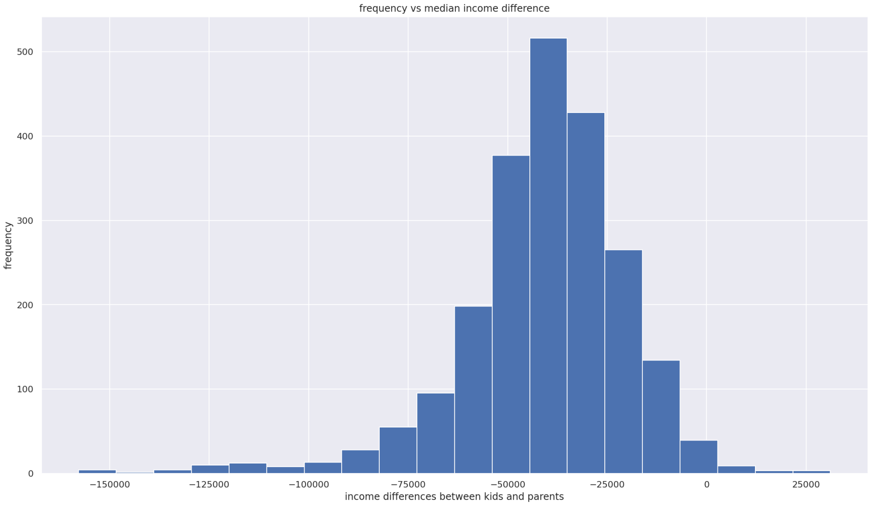


Figure 2: Histogram shows majority of children make less than parents. This is accounted for because parent income is defined by household, while child income is defined by individual.

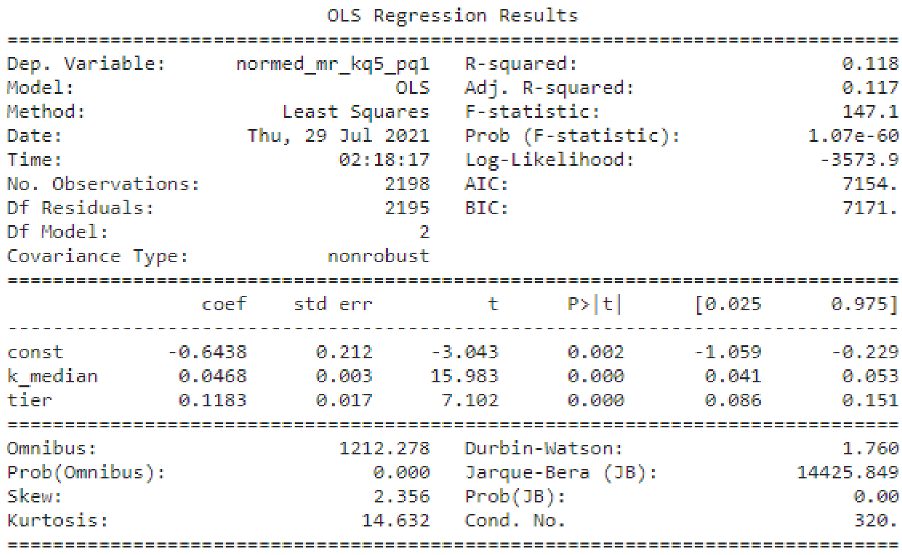
Predictive Analysis

We chose to use multiple linear regression to build our predictive model. This choice was made due to the numerous advantages multiple regression provided over other modeling techniques; namely, we were interested in examining the strength and importance of the relationships between each predictor and mobility. By using multiple regression, we were able to gain these insights from the generated coefficients.

We chose **k_median**, **par_median**, and **tier** as predictors and **normed_mr_kq5_pq1** as the response variable. **k_median** represents average child income, **par_median** average child income, **tier** college tier, and **normed_mr_kq5_pq1** is the mobility rate, defined as the joint probability of an individual earning in the top quintile and their parents earning in the bottom quintile.

As noted in the Exploratory Analysis section, our **k_median** and **par_median** predictors clearly exhibited collinearity. We chose to resolve this issue by producing two separate regression models, each incorporating one of the collinear predictors.

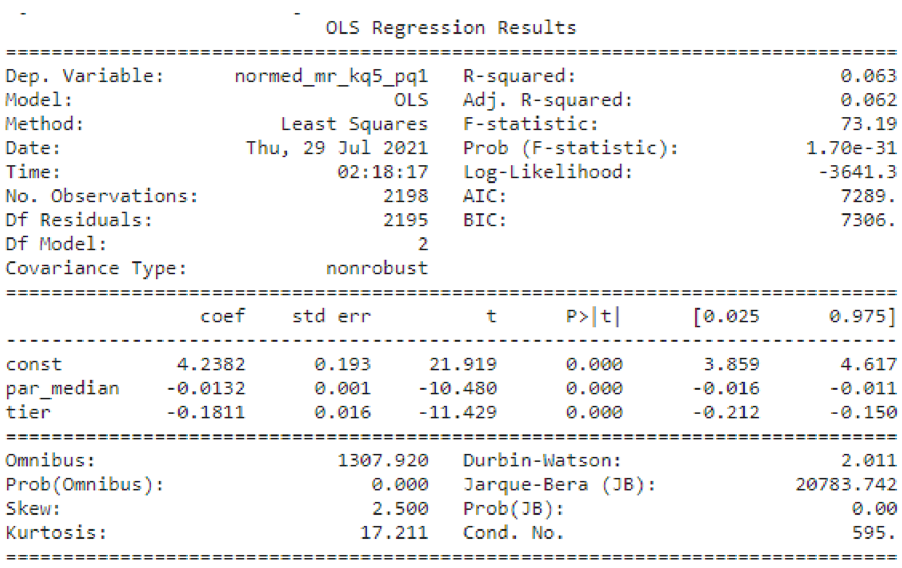
Our chosen method of regression was OLS (Ordinary Least Squares) regression. The results are shown below.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      normed_mr_kq5_pq1   R-squared:                       0.118
Model:                            OLS   Adj. R-squared:                  0.117
Method:                 Least Squares   F-statistic:                     147.1
Date:                Thu, 29 Jul 2021   Prob (F-statistic):           1.07e-60
Time:                        02:18:17   Log-Likelihood:                -3573.9
No. Observations:                2198   AIC:                             7154.
Df Residuals:                    2195   BIC:                             7171.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.6438      0.212     -3.043      0.002      -1.059      -0.229
k_median       0.0468      0.003     15.983      0.000       0.041       0.053
tier           0.1183      0.017      7.102      0.000       0.086       0.151
==============================================================================
Omnibus:                     1212.278   Durbin-Watson:                   1.760
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14425.849
Skew:                           2.356   Prob(JB):                         0.00
Kurtosis:                      14.632   Cond. No.                         320.
==============================================================================
```

Figure 3: Regression including predictors **k_median** and **tier**. Note that despite a lower $R^2 = 0.118$, the low pvalues all indicate significant results.

The first regression included the predictors **k_median** and **tier**. The R^2 value was 0.118, indicating a fairly low accuracy score, but we suspect this can be attributed to high variance in the data. Indeed, all P values were all less than the standard 0.05 threshold, indicating significant correlation for all predictors.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:      normed_mr_kq5_pq1   R-squared:                       0.063
Model:                            OLS   Adj. R-squared:                  0.062
Method:                 Least Squares   F-statistic:                     73.19
Date:                Thu, 29 Jul 2021   Prob (F-statistic):           1.70e-31
Time:                        02:18:17   Log-Likelihood:                -3641.3
No. Observations:                2198   AIC:                             7289.
Df Residuals:                    2195   BIC:                             7306.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.2382      0.193     21.919      0.000       3.859       4.617
par_median    -0.0132      0.001    -10.480      0.000      -0.016      -0.011
tier          -0.1811      0.016    -11.429      0.000      -0.212      -0.150
==============================================================================
Omnibus:                     1307.920   Durbin-Watson:                   2.011
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            20783.742
Skew:                           2.500   Prob(JB):                         0.00
Kurtosis:                      17.211   Cond. No.                         595.
==============================================================================
```

Figure 4: Regression including predictors **par_median** and **tier**. Note that despite a lower $R^2 = 0.063$, the low pvalues all indicate significant results.

The second regression included the predictors **par_median** and **tier**. The R^2 value was 0.063, indicating an even lower accuracy score than the first model, but we again attribute this to high variance in the data. Indeed, all P values were all less than the standard 0.05 threshold, indicating significant correlation for all predictors.

Conclusion

We conclude that child income is positively correlated with mobility. This makes intuitive sense; if the child has a higher income, then there will be a higher chance for them to earn more than their parents. Furthermore, parent income is negatively correlated with mobility. This also agrees with intuition; if parent income is high, then it will be more difficult for the child to earn more than their parents. Finally, school tier is both positively and negatively correlated with mobility, depending on the regression model. This interesting result is not actually contradictory and can be explained by examining **const** in the models. In the first model, with tier showing positive correlation, const = -0.64, while in the second model, with tier showing negative correlation, const = 4.2. Thus, the two regressions fit different intercepts but still move in the same direction.

Our research is not perfect; it is likely that we failed to account for several predictors. In fact, there are several additional questions yet to be answered.

- How would social mobility be affected by location, race, gender?
- What should parents do to set the next generation up for positive social mobility?
- What does a wealth gap look like in terms of social mobility?
- How can social mobility solve a wealth gap?
- How does public policy affect social mobility?

We hope future research may answer these questions.

Works Cited

Boundless. (2021). Boundless sociology. Lumen. https://courses.lumenlearning.com/boundless-sociology/chapter/social-mobility/.

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2021, January 15). Income segregation and intergenerational mobility Across colleges in the United States. Opportunity Insights. https://opportunityinsights.org/paper/undermatching/.

Policy solutions to the American Dream. Opportunity Insights. (2021, July). https://opportunityinsights.org/.

Torche, F. (2013). How do we characteristically measure and analyze intergenerational mobility? Stanford Center on Poverty and Inequality. http://cpi.stanford.edu/_media/working_papers/torche_how-do-we-measure.pdf.