Intergenerational Mobility as a Prediction Problem

Jack Blundell and Erling Risa

Stanford University (jackblun@stanford.edu) NHH (Erling.Risa@nhh.no)

Overview

- We introduce **a new descriptive measure** of intergenerational mobility
- We illustrate the use of the measure using **Machine Learning** methods in two settings:
  - Norwegian administrative data
  - British survey data

Introduction

In this project we explore a new measurement of intergenerational mobility based on the predictive power of family background over an individual's life-time income. We argue that this measure of mobility captures the core concept in a way that is both intuitive and novel. By viewing mobility as a prediction problem, we are able to draw on methods from machine learning. These methods allow for substantial flexibility of functional form and use regularization to account for 'over-fit', allowing us to extract the full predictive content of family characteristics while ensuring noise is not mistaken for signal. We compare our measure of mobility to existing measures and argue that this measure captures additional information. We explore the application of our measure to administrative data from Norway and survey data from Britain. We uncover significant heterogeneity across regions of Norway and show evidence which contradicts conventional wisdom on patterns of mobility over time in Britain.

Simple Framework

Notation:

- Y is an individual's income
- X is a vector of family background characteristics (income, education, wealth)

Our measure must capture relationship between family background characteristics X and income Y. We explore a measure based on the answer to the following:

"How much of the variation in a child's future income is explained by, or can be predicted by, their family background?"

More formally, define **Fraction of Variance Unexplained**:

$$FVU = \frac{var(Y|X)}{var(Y)} \tag{1}$$

and **Fraction of Variance Explained**:

$$FVE = 1 - FVU \tag{2}$$

FVE **is our measure**

Higher FVE values correspond to lower predictive power of family background characteristics. If only family income is used as a predictor and the true model is linear in parental / child income ranks, this measure aligns with the rank-rank correlation as popularized by Chetty, Hendren, Kline, and Saez 2014. As the number of characteristics included increases, this approaches measures used in the equality of opportunity literature popularized by John Roemer. We therefore view this measure as nesting ideas from both the intergenerational mobility and inequality of opportunity literatures.

Consider a linear model $Y = \beta X + \epsilon$ to simplify discussion. The following terms will drive the true (population) FVE:

- Parameter vector β, the relationship between each individual parental characteristic and the child outcome, driven by the covariance between each X and Y
- The variances of each component of X and Y
- The covariance between each component of X, i.e. the degree to which there exists stratification on the parental side

Our FVE measure captures all three of the above seeks to capture these three features and summarize them.

Estimation

Why not use linear regression to estimate FVE?

- Relevant factors X is **potentially high-dimensional** relative to sample size
  - The number of predictors p may be close to number of observations n
  - This gives overfit $\rightarrow$ confounding noise for signal
- **Why assume linearity**?
  - Parental factors could explain later income in complex, non-linear ways and interactions may be important

Given these two reasons, we use a variety of Machine Learning models to estimate FVE

Application 1: Norwegian Administrative Data

Our first application is using rich administrative data on all individuals born in Norway between 1970 and 1980. We obtain the various characteristics of these individuals using several administrative datasets from Statistics Norway. Parental characteristics include details on income, education, occupation and family structure. We also include several area characteristics in our model. The first figure below demonstrates the **superior performance of the ML methods**. Our outcome variable here and throughout the project is the rank of the child in the income distribution for their cohort. These "violin" plots show results for a number of cross-validation sets which approximate out-of-sample predictive performance, meaning that overfitting will lead to lower FVEs.

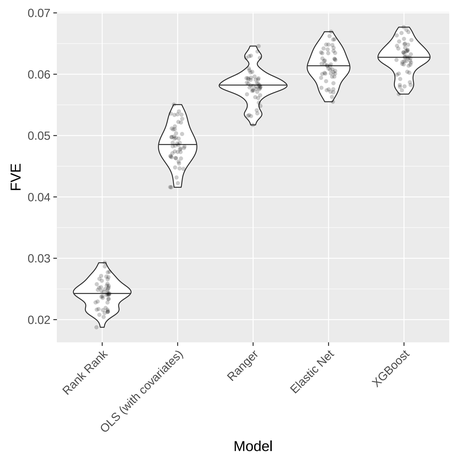

Figure 1: Model comparison for Norwegian data. Model 1 includes income alone, model 2 includes covariates linearly, remaining models use machine learning methods.

The next plot demonstrates the FVE measure estimated separately for labor market regions in Norway. We see here that there is substantial heterogeneity, with some areas seeing an FVE twice of other areas.

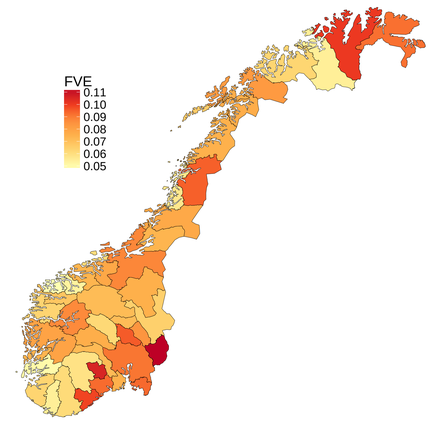

Figure 2: Estimated FVEs across labor markets of Norway

Given the geographical variation shown above, a pertinent question is whether the regional comparison here gives different results to standard measures. In the next figure we show the relationship between the FVE score and the rank-rank slope coefficient for each labor market region. The comparison measure is the coefficient from a linear regression of child income rank on parent income rank. Here we see that there is a positive relationship, but that there are areas that score quite differently on each. In current work we seek to understand better the drivers of these differences.

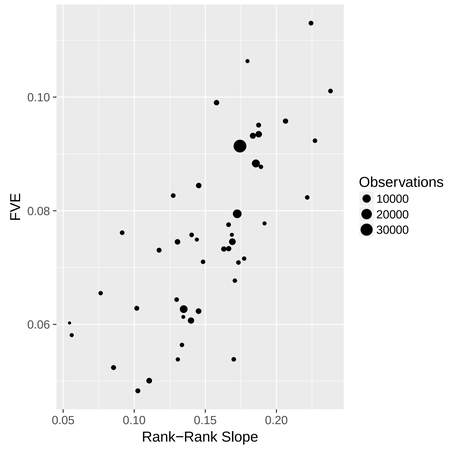

Figure 3: Comparison of FVE measures and rank-rank slopes

Application 2: British Survey Data

Our second empirical application of the new measure uses survey data from Britain. Much of what is known about intergenerational mobility in Britain stems from analyses of two surveys, the National Child Development Study (NCDS) and the British Cohort Study (BCS). We re-examine the two surveys using our new measure.

The figure below plots performance of various models in the NCDS data, one of the two surveys. We see again that ML models tend to explain more of the variation in the outcome than conventional models

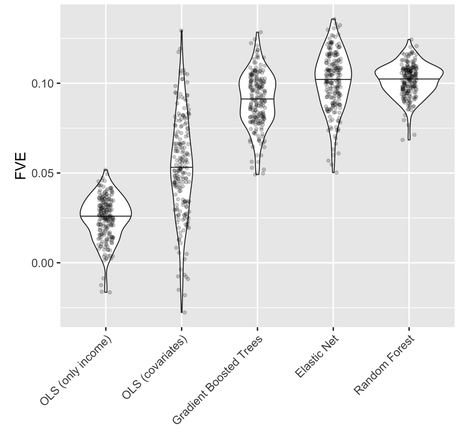

Figure 4: Model comparison for NCDS data. Model 1 includes income alone, model 2 includes covariates linearly, remaining models use machine learning methods.

Conventional wisdom among economists is that mobility fell between the NCDS cohort born in 1958 and the BCS cohort born in 1970. In the next figure we test whether our measure suggests the same pattern.

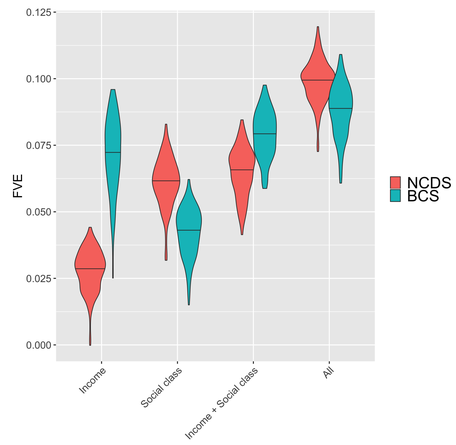

Figure 5: NCDS / BCS comparison

This plot demonstrates that if income alone is included as a predictor the FVE measure goes up, in line with conventional wisdom. However when multiple factors are included, FVE is at worst unchanged between the two surveys.

References

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez (2014). "Where is the land of opportunity? The geography of intergenerational mobility in the United States". In: *The Quarterly Journal of Economics* 129.4, pp. 1553–1623.