FACTORS OF HIGH VERSUS LOW INTERGENERATIONAL MOBILITY RATES WITHIN THE UNITED STATES

MARIA MOYA

ABSTRACT

This study will expand on estimates by county and birth cohort gathered by, Chetty, Hendren, Kline, Saez and Turner (2014): Intergenerational Mobility Estimates by County and Birth Cohort (click here). Their paper uses federal income tax records to identify the incomes of more than 40 million children and their parents between 1996 and 2012. They measure the childs income up until they are approximately. The data-set categorizes these children by their corresponding birth cohort from 1980-1988 and the county they reside in.

We will consider a linear model through neural networks which holds no assumptions about the relationship between the independent and dependent variables, rather their relationship will be learned through an iterated process(neural networks optimize cost minimization through gradient descent). This would allow us to also make nonlinear comparisons, if they exist such as the relationship between log childs income and the log of their parents income and it may help alleviate the biased between observed and unobserved factors that determine mobility. The type of neural networks that will be chosen for this study is backward propagation. This study will predominately investigate the predicted classification of high versus low mobility rates by county. Previous studies have used regression analysis to predict intergenerational mobility by county. This paper considers alternative approaches such as a neural network and random forest. Thus we will also consider these alternative algorithms and compare the predictive power to that of the standard linear regression.

In regards to the neural network, we will make use of the BFGS algorithm as a cost optimization method.To control for overfitting, we will impose regularization in which

we normalize our cost function and add an additional term to that penalizes overly complex models.

1. Introduction

The intergenerational mobility rate tends to be heterogeneous among different counties in the United States (Chetty 2014). Thus parental outcomes that directly impact a childs future expected earnings, are influenced by the county in which they reside in. The greater the mobility rate, the higher the probability that a child reaches the top national income quartile. Existing work regresses a childs earning potential relative to their parents income and other factors that determine mobility. However these studies fail to capture the casual effects among these determinants that relate to upward mobility. These limitations are due to endogeneity among the factors that determine mobility and the correlation between unobserved and observed factors.

This study seeks to show alternative methods of drawing predictions rather the linear regression model which is the default tool of choice within Economics. We will compare our predictive power of a neural network and random forest to that of a linear regression. Afterwards, we will run a sensitivity analysis for each given feature by considering a correlation matrix.

The median absolute mobility rate is 43, therefore our output variable denotes "High" for a county if its absolute mobility rate is greater than or equal to 43 and low if it is less than 43.

This dataset contained a total of features 35 however 27 of which had 3,321 missing values per feature. This paper will use the kNN imputation method to fill in the missing values within our dataset since it is computationally inexpensive relative to other machine learning algorithms and tends to outperform most traditional statistic approaches such as interpolation.

2. Related Work

This paper provides an extension to Raj Chetty, Nathaniel Hendren,Patrick Kline and Emmanuel Saez work on the heterogenity of integenerational mobility on a county by county basis across the US. They find that intergenerational mobility rate tends to be heterogeneous among different counties in the United States (Chetty 2014). However these studies fail to capture the casual effects among these determinants that relate to upward mobility. These limitations are due to endogeneity among the factors that determine mobility and the correlation between unobserved and observed factors.

This essentially means that given MLRM, $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in} + u_i$ or $y = X\beta + u$ that there exist a correlation between the regressor and error, thus $E[u|x] = 0$

Subsequent papers consider other approaches regarding intergenerational mobility, however much like all of economics, they all use the standard MLRM. Therefore this paper not only seeks to introduce different tools that evoke causality, moreover I want to compare the predictive power of a neural network and random forest.

3. Neural Network

Neural networks hold no assumptions about the relationship between the independent and dependent variables, rather their relationship will be learned through an iterated process(neural networks optimize cost minimization through gradient descent). This would allow us to also make nonlinear comparisons, if they exist such as the relationship between log childs income and the log of their parents income and it may help alleviate the biased between observed and unobserved factors that determine mobility. The type of neural networks that will be chosen for this study is backward propagation.

The dataset has 28,243 observations and 35 features(so X is a 28,243 by 34 matrix. I will use the last column of my dataset as my y variable, absoluteupwardmobility(which is the high and low variable. After applying my linear threshold unit, this should be a binary variable in which 1= high and 0=low).Thus y is a 28,243 by 1 vector.

Since X has 34 features the model will have n=34 inputs or neurons. We will first consider the case in which we have 1 hidden layer and k=35 hidden units. We will only consider one hidden layer for this study since the data is not overly complicated. Deeper neural networks are necessary when the data is much more sophisticated (such as the spiral dataset from assignment 5). Thus this will yield a total of $k \cdot n + K = 35 \cdot 34 + 35 = 1225$ synapses. Moreover, this model will have 1225 weights. These hyperparameters must be specified since the algorithm will only learn the weights. The weights are randomly assigned

The structure will multiply each input by a weight(in the first layer). This will yield a corresponding z (in the second layer). Thus $X \cdot W^{(1)} = z^{(2)}$ where X is J by k (where J=28,243), $W^{(1)}$ is n by k and $Z^{(2)}$ is J by n. We then apply the sigmoid activation function, $f(\cdot) = a = \frac{1}{1+e^{-1}}$ on each component in $z^{(2)}$. This will yield a new equation $a^{(2)} = f(z^{(2)}$. Afterwards, we take the dot product of $a^{(2)}$ with $w^{(2)}$ which reprsents the weights in the 2nd layer(which is k by 1). This yields another equation$z^{(3)} = a^{(2)}W^{(2)}$. Applying the sigmoid function on $z^{(3)}$ yields our equation for the estimated y value, $\hat{y} = f(z^{(3)})$. In order to train our data, we must first consider our cost function $C = \sum \frac{1}{2}(y-\hat{y})^2$. The objective is to minimize cost(or our error). We will use gradient descent through Back Propagation to minimize cost(which tells us the direction in which we want to increase or decrease w values in order to minimize cost), thus given the other four equations we can rewrite the cost function as $C = \sum \frac{1}{2}(y - f(f(X \cdot W^{(1)} \cdot W^{(2)}))^2$. We minimize cost when $\frac{\partial C}{\partial w} = 0$. Since we have two layers of weights, $W^{(1)}$(n by k) and $W^{(2)}$ (k by 1) we have to consider the partial of cost wrt both of them separately.

$$\frac{\partial C}{\partial W^{(2)}} = \frac{\partial \sum \frac{1}{2}(y-\hat{y})^2}{\partial W^{(2)}}$$

Consider,

$$\frac{\partial \frac{1}{2}(y-\hat{y})^2}{\partial W^{(2)}} = -(y-\hat{y})\frac{\partial \hat{y}}{\partial W^{(2)}}$$

Recall $\hat{y} = f(z^{(3)})$, therefore

$$= -(y-\hat{y})\frac{\partial \hat{y}}{\partial Z^{(2)}}\frac{\partial \hat{Z^{(3)}}}{\partial W^{(2)}}$$

$$= -(y - \hat{y}) f'(z^{(3)}) \frac{\partial \hat{Z^{(3)}}}{\partial W^{(2)}}$$

Let $\delta^{(3)} = -(y - \hat{y}) f'(z^{(3)})$ and note that $\frac{\partial \hat{Z^{(3)}}}{\partial W^{(2)}} = a^{(2)}$, hence

$$= \sum (a^{(2)})^t \delta^{(3)}$$

To get $\frac{\partial C}{\partial W^{(1)}}$ we do more math and get $\frac{\partial C}{\partial W^{(2)}} = X^t \delta^{(3)} (W^{(2)})^t f'(z^{(3)})$ Next I have to consider regularization to control for overfitting in which we normalize our cost function and add an additional term to that penalizes overly complex models. This model will use the standard L2 regularization type to account for overfitting and a rate of 0.01.

4. Random Forest

Another alternative algorithm that we will consider for this study is the Random Forest. First, decision trees in general tend to automatically impute missing values with either a surrogate attribute or from randomly drawn samples. Random forest algorithms are particularly useful for datasets with many features rather than performing a ID3. For instance, if we had a dataset with 10,000 features, we would have to evaluate 10,000 information gains which is of course computationally expensive.

Another key feature that Random Forest algorithms provide is its ability to reduce bias by considering multiple trees and combining them.Note we normally sacrifice variance for the reduction in bias. However, this algorithm controls for this increase in variance by taking the average of the uncorrelated trees. Boostraping is the notion of randomly drawing different subsets of features from our training data (which is an alternative approach to cross validation). We will draw B boostrap samples and formulate B trees. It imposes randomness by assuming each tree gets a different subset of the data set(therefore the trees should be uncorrelated with one another by construction). This allows us to evaluate the confidence of whether or not you are estimating a good parameter. Randomness also occurs during the feature splits. Each node is limited to the all possible thresholds within a given subset(note that the best split is still determined by its corresponding information gain). The notion of combining these trees is known as

"bagging", in which we are bagging tree instead of other estimators.Since each tree contains a different data set, each tree will have a different probability for the classification. For instance tree 1 may generated a $P(c|T_1(x)) = 0.75$ probability of classifying it as High while tree 2 may have a $P(c|T_2(x)) = 0.82$ probability of classifying it as High. We then combine all trees by the average of their corresponding probability distribution, or:

$$g_c(x) = \frac{1}{t}\sum_{j=1}^{t} P(c|T_j(x))$$

This is how we control for the increase in variance. Since each tree is uncorrelated with one another by construction, once we average over all probability distribution this will reduce the variance and hence allow for a better estimation of the predicted value.

This is particularly interesting considering previous studies faced issues with bias and inconsistency for their OLS estimators in which $E(u_i|x_i) \neq 0$. Therefore, though difficult to interpret, random forests is an ideal approach to measure causality.

5. Results

Unfortunately I ran into issues with trying to implement the kNN inputation method whether using R(apparently I did not have enough memory) or Python (my editor would not recognize the fancyimputer package). Therefore, I impute the median of each given feature to their corresponding missing values (I used median over mean due to potential outliers in the data).

The neural network generated a predicted accuracy of 0.56. However this low accuracy may be due to the tuning parameters. The linear regression algorithm outperformed the neural network generating a predicted accuracy of 0.67. However, unsurprisingly, the random forest algorithm outperformed all three algorithms with a predicted accuracy of 0.92. Thus it would be interesting to consider testing regressions derived from random forest rather than the standard OLS or time series model within economics.

In regards to the sensitivity analysis among features, I decided to create a coefficient matrix(click here to view the correlation matrix).

Since counties are identified by their fips code, I was able to include the county feature. Moreover, I converted the HighLowMobility feature to a binary variable where 1 equals High and 0 equals low. Unfortunately the correlation matrix provides little information besides which feature strongly influence another. For instance, there exist a strong correlation between HighLowMobility and the child's median income. Relative to other variables, county and upward mobility is slightly strong.

6. Conclusion

Despite the neural network's shortcoming of predictive power relative to the MLRM, I will continue to investigate the optimal tuning of its parameters. Overall, given that neural networks hold no assumptions between the independent and dependent variable rather they are learned through an iterated process may be an alternative approach to analyzing a causal model. On the other, random forest provide complete causality due to the randomness from boostraping which allows for unbiased estimates and averaging uncorrelated trees which further eliminates the issue of high variance.However it is difficult to generate inferences from a random forest model. Nonetheless, the fact that it holds a stronger predictive power over MLRM should be a good indicator that economics needs to consider alternative tools for drawing predicted estimates. I would like to further expand this study in the future and figure out how I can use random forests to explain *why* integenerational mobility rates are higher in some counties over others.

References

[1] James J. Feigenbaumy, *"Intergenerational Mobility during the Great Depression"*, Cambridge, MA: Harvard University Press, 2015

[2] Raj Chetty, *"Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States"*, The Quarterly Journal of Economics Vol. 129, 2014.

[3] Philip Oreopoulos, *"The Long-Run Consequences of Living in a Poor Neighborhood"*, Quarterly Journal of Economics, 2003

[4] Sean F. Reardon, *"Measures of Income Segregation"*, CEPA Working Papers, 2011

[5] Gary Solon, *"Intergenerational Income Mobility in the United States"*, Economic Review, 82, no. 3 1992

[6] William J. Wilson, *"The Truly Disadvantaged: The Inner City, the Underclass,and Public Policy"*, Chicago: University of Chicago Press, 1987

[7] Debraj, Ray *"Uneven Growth: A Framework for Research in Development Economics"*, Journal of Economic Perspectives, 24, no. 3, 2010

[8] Susan, Athey *"Estimation and Inference of Heterogeneous TreatmentEffects using Random Forests"*, Standford University 2015

[9] Caroline, Rodriguez *"The treatment of missing values and its effect in the classifier accuracy"*, Department of Mathematics, University of Puerto Rico at Mayaguez

[10] Robert J, Sampson *"Assessing Neighborhood Effects: Social Processes and New Directions in Research"*, Annual Review of Sociology, 28 20-2 2002