The Rise in Returns to Skill?

A Modern Regression Analysis of Wage Inequality in the Current Population Survey (CPS)

First Draft

Senan Hogan-H.[*]

Senior Seminar in Economics, Pomona College[†]

May 2018

Abstract

The rise Wage inequality has been well-documented as, yet evidence that attributes this to a rise in return to skill use a standard Ordinary Least regression approach to predict wages. Today, we have access to newer regression techniques which are much better at prediction than a standard econometric approach. This paper decomposes wage inequality for 1980–2016, expanding the Juhn et al., 1993 decomposition method to modern regression techniques. The analysis showing a diminished role of unobserved characteristics or skill in contributing to rising inequality, and how the decomposition depends heavily on the specific prediction methods used.

[*] This paper was completed in accordance with requirements for the Pomona College Department of Economics Senior Seminar, Spring 2018. I am grateful for the comments and guidance of Michael Kuehlwein, Pomona College Department of Economics.

[†] This project's Github repository, which hosts all contributing materials, is available here.

1 Introduction

Wage inequality has been rising drastically since the 1960s, possibly due to many factors. Multiple significant studies decompose wage inequality attributing a large portion of the rise in inequality to a rise in returns to skill and unobservable characteristics, first shown by Juhn et al. (1993). The methods for wage prediction and determining observable skills are extremely important in determining this composition, and have previously only used standard linear regressions. This paper uses modern regression techniques to expand the Juhn et al., 1993 decomposition method to modern regression techniques. The random forest prediction method produces more accurate estimates of wages than the standard linear regression approach, leading to a different composition of observed characteristics in explaining wage inequality. This analysis thus shows how dependent the wage decomposition is to the exact prediction method used, an issue that has not before been examined rigorously.

Wage inequality rises consistently from 1980, following an accelerated rise for those with at least a college degree compared to the rest of the population. The distribution of residuals in wages rises regardless of prediction method, also becoming more unequal across the time period. However, the modern prediction techniques predict a much less dramatic rise in inequality across the residual distribution. Standard linear models predict that unobserved skill contributes the most of any component to rising inequality, but modern regression models attribute less of the rise in inequality to unobserved factors and more to observed skill and characteristics. They also exhibit a rising importance for years of education in predicting wages, with a fall for gender and age.

The paper is structured as follows. Section 2 surveys current literature on wage inequality and decomposition by regression approaches. Section 3 describes a framework for expanding the regression approaches to this topic with specifications of the regression approaches used, both common-place and novel, as well as a description of the March CPS data set they are applied to. Section 4 presents the empirical results of each approach. Section 5 discusses the findings of the paper, with lessons to learn for studies that use predictive models and regression approaches in the study of wage inequality.

2 Literature Review

Wage inequality has increased dramatically since the 1950's in the US. Many studies in labour economics attribute this this rise in inequality a rise in return to skill. Across the other subfields of economics, however, there are multiple factors that also explain rising inequality, ranging from rise in market power (Furman and Orszag, 2015), to de-unionisation and supply and demand shocks (DiNardo et al., 1996), to “skill-biased technological change” (Acemoglu, 1998, 2002). The literature that attributes a large percentage of the rise in wage inequality to an increase in returns to skill mainly relies on regression approaches pioneered by Mincer (1958, 1974). Today, we have a greater repository of econometric and statistical methods to draw from in predicting wages and wage inequality by regression approaches. These newer methods provide an avenue to better estimate the role of returns to skill in explaining rising wage inequality.

Juhn et al. (1993) showed a drastic rise in wage inequality between 1963 and 1989, using the March CPS – a representative data set for the US population. The analysis extends the decomposition method of Juhn et al. (1991) that focuses on the residuals in a standard regression to predict wages, showing the role of unobserved skills. The methods make a few notable assumptions, primarily primary choice of model to predict wages. Yun (2009) questions the validity of using the change in distribution of residuals to explain discrimination in the decomposition methods of Juhn et al. (1991, 1993), while Lemieux (2006a) attributes the rise in wage inequality to a secular increase in experience and education, attributing the results of Juhn et al. (1993) to composition factors and noisy data.

The Mincer wage equation was popularised in labour economics by Mincer (1958, 1974), where the standard Ordinary-Least-Squares (OLS) regression method is used to predict wages by simple function of potential experience and years of education. This method was noted as being successful in explaining wages and wages inequality while maintaining an intuitive basis, and has become a benchmark model in labour economics as a result. Juhn et al. (1993) use this approach to predict wages, crucially using the fact that the method produces much worse estimates (i.e. higher errors) in the late 1980's that did in earlier decades. However, today the Mincer wage equation is noted as in need of adjustments to capture changes in the US economy and acknowledge advances in empirical labour economics. Lemieux (2006b) specifically proposes functional adjustments to accommodate changes in the economic relationship between years of education and wages seen in wage data until the 1990's. However, the US economy is very different to how it was in the 1950's, especially in terms of size, complexity and technology. Today, supply chains are increasingly complicated and dependent on global markets, entire industries have risen and fallen with ramifications for structural employment, and the Internet has revolutionised many sectors of the economy. Labour economic studies that use the original Mincer wage function to predict wages in the modern economy do not fully acknowledge the rise in complexity or resulting changes in the determinants of wages,[1] and would do better to take advantage of advances in regression methods that can better accommodate.[2]

Regression practices have advanced tremendously in the last few decades, and since Dr Jacob Mincer began modern labour economics as a field. The original Mincer wage equation does provide a method for intuitive prediction of wages, though an overly simplified version of the story. There are many variables – observed, unobserved, or even unobservable – with possible explanatory power for wages in the US economy. There are also many ways in which these variables explain wages (linear or non-linear), and certainly in different ways to how they did in previous decades. This problem is not be completely fixed by a simple adjustment, as noted by Lemieux (2006a). These issues bring in to question the robustness

[1] That is the simple linear equation may only capture changes through an increase in estimated returns to education, experience and potential experience, yet may not acknowledge any change in the composition or functional form of them and other important variables in determining wages.

[2] More advanced regression practices will also have the advantage of predicting wages without onerous assumptions in the underlying data and economy, as in Juhn et al. (1993).

of only using simple OLS approach in analyses that require prediction of wages with given data.

Regression by machine learning practices provide a viable avenue to expand the literature, and more robustly estimate returns to skill in the US economy. A regression tree is a completely different form of regression than OLS and other common econometric methods. The process involves building a decision tree by minimising an error function, allowing any level of non-linearities in model estimation. Bootstraps of data are used to form the model forms a random forest model, with extremely high power for prediction with less problems of over-fitting data (Breiman, 2001). Labour economics has so far been slow to use such techniques in empirical studies. Belloni and Chernozhukov (2011) and Abadie and Kasy (2017) develop novel prediction techniques, and example their properties by predicting wages in the March CPS. Importantly, Chalfin et al. (2016) demonstrate the benefit of predictive power of tree-based machine learning methods in productivity of public sector workers. Gains in predictive power from these methods are noted as "large both absolutely and relative to those from interventions studied by standard causal analyses in microeconomics." Lastly, the March CPS is an extremely large data set, especially when treated in 5 year increments,[3] and so provides an adequate place to use newer regression practices as outlined by Varian (2014).

This paper applies multiple regression-based predictive models – including the novel random forest method – to a representative data set of the United States, suitably large for big-data practices. The results are presented in the wider context of previous research on the rising returns to skill, and the role of using newer regression and big-data practices in the field of labour economics.

[3] See Section 3.1 for an expanded discussion of this issue.

3 Wage Inequality in the US

3.1 Data

The analysis of this paper is conducted on 35 years of wage and demographic information for individuals taken from the March the Annual Social and Economic Supplement of the Current Population Survey, commonly referred to as the March CPS. The 35 years refer to 1980–2016 (without 2008), with data referring to the 12 months preceding the March survey. Uniform extracts of the March CPS are taken, in full, from publicly available hosting by the Centre for Economic Policy Research (Version 1.0, 2016).

Analysis of inequality refers to wage information at the hourly level, defined as annual earnings divided by annual hours worked,[4] and annual level, defined as total income in the 12 months preceding, as specified. Wages are adjusted according to the CPI Research Series Using Current Methods (CPI-U-RS), set to 2015 dollars.[5] The sample is restricted to a sample meant to be representative of full-time workers, both male and female.[6] Only full-time employed workers are included, making at least the 2015 hourly federal minimum wage ($7.25), positive total annual income, and between the ages 18 and 65. Observations from the year 2008 are excluded due to sample size problems.[7] See Table 1 for summary statistics for real hourly and annual wage, age, years of education, and proportion female for the sample.

[4] Weekly hours worked times by amount of weeks worked in a year.

[5] These specifications for wages are provided in full by the CEPR Extracts.

[6] Whereas Juhn et al. (1993) analyse only men's wages in order to remove effects of rising women's labour force participation. This study however considers a later time period when women participation is relatively similar and so includes women.

[7] 2008 has remarkably few observations that fit the above criteria. Specifically the years 2007 and 2009 have 77,907 and 75,175 observations respectively, whereas 2008 has 13,720. This led to unusually high averages and quantiles for wages in the year 2008 compared to years either side – counterintuitively given economic conditions – so that this year is excluded in analysis.

Table 1: Summary Statistics, 1980-2016

Statistic	Observations	Mean	St. Dev.	Min	Max
Hourly wage, $	2,185,521	25.35	335.72	7.25	444,241.80
Annual income, $	2,185,521	50,426	50,976	4,060	1,848,079
Age	2,185,521	39.61	11.67	18	65
Years of education	2,185,521	13.70	2.69	0	22
Female	2,185,521	0.46	0.50	0	1

3.2 Rising Inequality

Figure 1 presents the 10th, 50th and 90th percentile of the hourly wage distribution 1980–2016. Each series is indexed to 1980 (so that each is assigned a value of 100 in 1980) for comparison between the groups and visualisation of their respective changes since 1980.

Figure 1: Indexed Real Hourly Wage by Percentile, 1980-2016.

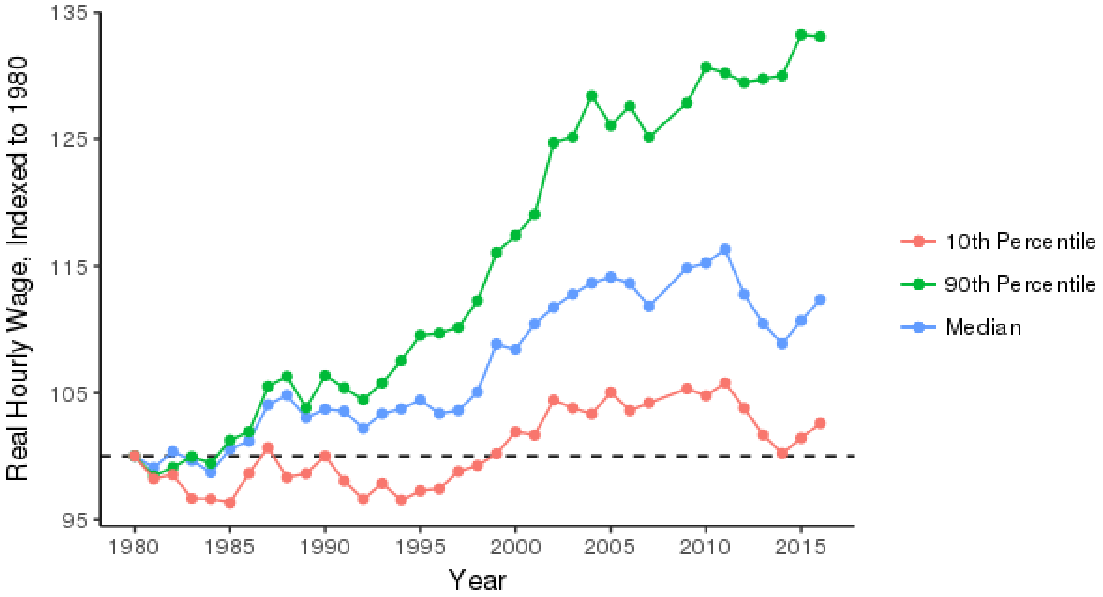
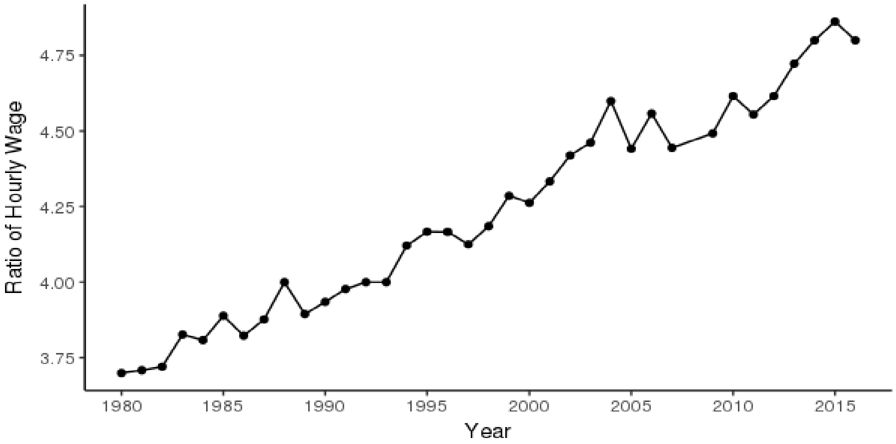


Figure 2 presents the ratio between the 90th and 10th percentile of the hourly wage distribution 1980–2016. The series shows a clear and persistent increase for the 36 year period, following a trend documented for preceding decades in previous research.

Figure 2: Ratio of Wage Between 90th and 10th Percentiles, 1980-2016.



4 Explaining Wage Inequality

Wage Inequality has risen significantly since 1980, and for years preceding. The role of changing returns to skills, observable or unobservable, plays in explaining this rising in inequality is not quite so clearly cut. Equation 1 presents a standard linear regression that aims to decompose wage inequality in the form of a standard wage equation.

$$Y_{it} = \mathbf{X}_{it}\beta_t + \varepsilon_{it} \tag{1}$$

Y_{it} represents the log weekly wage for individual i in year t, $\mathbf{X}_{it}$ a vector of observable characteristics and β_t the vector of coefficients representing returns to observable skills. ε_{it} is the standard residual, the component for wages that are not otherwise explained by specified variables in the given regression method (unobserved characteristics). This residual may be specified in terms of position of the residual distribution, in equation 2, where $F_{it}(.|\mathbf{X}_{it})$ is the cumulative density function for individuals with observed characteristics $\mathbf{X}_{it}$ in year t.

$$\varepsilon_{it} = F_t^{-1}(\theta_{it}|\mathbf{X}_{it}) \tag{2}$$

Decomposition of wage inequality involves attributing changes in inequality to three factors across time: changes in observable characteristics (i.e. changes in $\mathbf{X}_{it}$), changes in returns to observable characteristics (i.e. changes in β_t), and changes in unobserved characteristics (i.e. changes in ε_{it}). To demonstrate this decomposition define β to be the returns to observable characteristics and $F(.|\mathbf{X}_{it})$ the cumulative distribution distribution for residuals across a reference time period, here 1980–1985, so that neither varies year on year. So that Y_{it}^1, Y_{it}^2, Y_{it}^3 may be defined as follows.

$$Y_{it}^1 = \mathbf{X}_{it}\beta + F^{-1}(\theta_{it}|\mathbf{X}_{it}) \tag{3}$$

$$Y_{it}^2 = \mathbf{X}_{it}\beta_t + F^{-1}(\theta_{it}|\mathbf{X}_{it}) \tag{4}$$

$$Y_{it}^3 = \mathbf{X}_{it}\beta_t + F_t^{-1}(\theta_{it}|\mathbf{X}_{it}) = \mathbf{X}_{it}\beta_t + \varepsilon_{it} = Y_{it} \tag{5}$$

Y_{it}^1 is the distribution of wages under fixed returns to observable characteristics, Y_{it}^2 the counterfactual distribution of wages under variable returns to observable characteristics and quantity of observable characteristics but a fixed distribution of residuals, Y_{it}^3 the distribution of wages where all components may vary leading to equality of the observed distribution.

The last step of the decomposition defines $(Y_{it}^1 - \bar{Y}_i)$ as the component of difference in inequality between year t and across the sample time period due to change in quantity of observable characteristics, $[Y_{it}^2 - (Y_{it}^1 - \bar{Y}_i)]$ the marginal contribution of change in returns to observable skill, and $(Y_{it}^3 - Y_{it}^2)$ the marginal contribution of change in residuals. Note the following identity that recovers the observed distribution of wages.

$$(Y_{it}^1 - \bar{Y}_i) + [Y_{it}^2 - (Y_{it}^1 - \bar{Y}_i)] + (Y_{it}^3 - Y_{it}^2) = Y_{it}^3 = Y_{it} \tag{6}$$

It follows that the observed distribution of wages can be decomposed to three discrete components. Juhn et al. (1993) present this framework for predicting wages with a very specific method of linear regression for predicting wages.

4.1 Prediction Methods

The residual, ε_{it}, and its distribution, $F_t(.|\mathbf{X}_{it})$, is crucial in these methods to document the rise in returns to unobservable skill. It follows that the methods used to generate the residuals and the way that "observed" skills are defined dramatically influence their estimated returns. An important choice to be made is the choice of regression method and form that equation 1, section 4 takes.

4.1.1 Mincer Wage Equation

The vast majority of labour economics studies that decompose wages use the Mincer wage equation to predict wages. The approach applies Ordinary Least Squares regression algorithm to the Mincer earnings function, which dates back to some of the first labour economics studies that focus on wage inequality in Mincer (1958, 1974). The function to be estimated by this approach takes the following form.

$$Y_{it} = Y_0 + \rho_t s_{it} + \beta_{1t} x_{it} + \beta_{2t} x_{it}^2 + \varepsilon_{it} \tag{7}$$

Here, Y_{it} represents the log wage for individual i in year t, s_i years of education, and Y_0 the standard intercept, ρ_t, β_{1t}, β_{2t} standard coefficients to be estimated with error term ε_{it}. Potential experience, x_{it}, is defined as age minus years of education – 6, i.e. $x_{it} = Age_{it} - s_{it} - 6$. This model is extremely influential in labour economics to describe and predict inequality in wages in the US population. Its influence comes in part from its theoretical foundations and simplicity in interpretation, yet is documented as being only accurate in predicting wages for the 1950s, and much less so for beyond.

4.1.2 Mincer Wage Equation, Adjusted

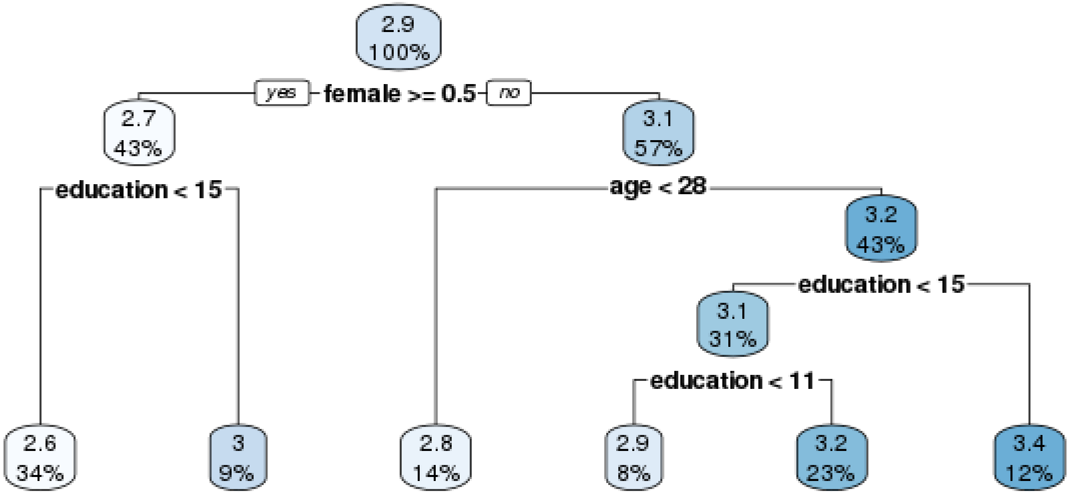
The Mincer wage equation may take an expanded form. The form is presented in equation 4, where the quadratic form for potential experience is expanded to a quartic, and the years of education include a quadratic.

$$Y_{it} = Y_0 + \rho_{1t} s_{it} + \rho_{2t} s_{it}^2 + \beta_{1t} x_{it} + \beta_{2t} x_{it}^2 + \beta_{3t} x_{it}^3 + \beta_{4t} x_{it}^4 + \varepsilon_{it} \tag{8}$$

The given variables are represented as in equation 3. This functional form better captures advances in the economy, and so is noted as better predicting ages for later decades following the 1950's than the original form. However, the returns to observable characteristics (i.e. estimated coefficients) are harder to interpret from an economic or intuitive point of view. This form is also noted as accounting allowing estimates for coefficients to vary by time period.[8] This adjusted form of the Mincer wage equation is used to predict wages while maintaining an OLS approach.

4.1.3 Tree-Based Methods

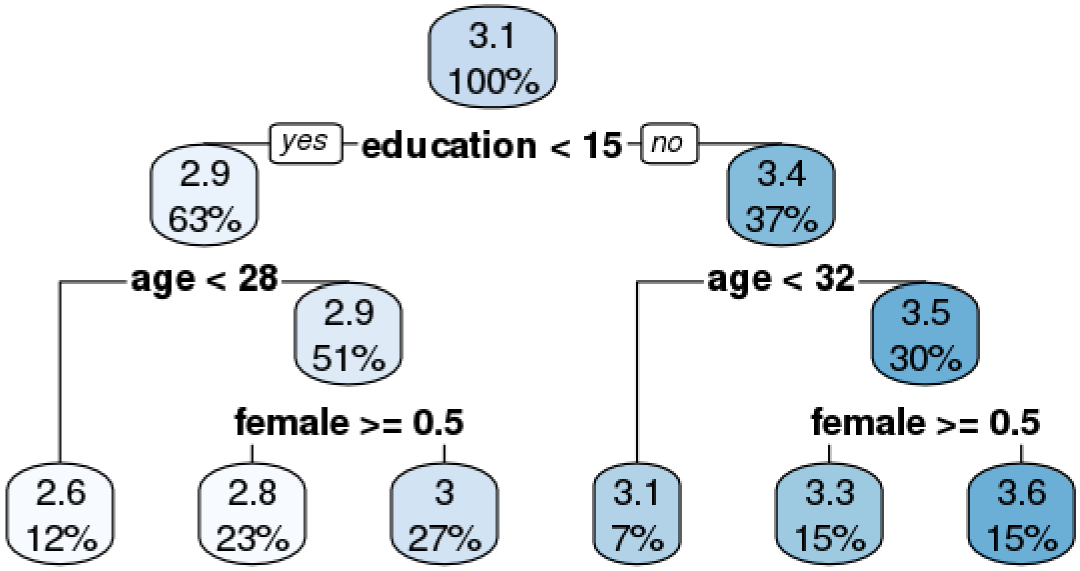
Figure 3: Decision tree for real log wages, 1980-1985.



Tree-based methods are a more modern approach to regression and prediction. The process involves building a tree in the following process. A data set of interest is taken, with a variable

[8] Lemieux (2006b) propose a similar functional form, where cohort effects are included in place of year-specific coefficients. Equation 4 is more appropriate for the decomposition, and so is used ahead of the form proposed by Lemieux (2006b).

Figure 4: Decision tree for real log wages, 2010-2016.



to be predicted (called the *response* variable) and a number of variables used to predict it (called the *predictor* variables). A set of algorithms is applied to the data set, where an error function is minimised (similar to the OLS framework) for a defined amount of decisions that predict the response variable. The decisions take the form of a true-false condition on given variables, leading to a final prediction conditional on all the given conditions. Many methods involve using one data set or a subsample of a dataset as a *training* sample, where a model is built (also called *trained*) on the training set before being applied to a *test* sample or new dataset to form predictions.

Figure 3 presents a decision tree to depict such a model for log wages 1980–1985, Figure 4 for 2010–2016. The regression tree approach allows for non-linearities and interactions between variables. For the first tree, a woman with less than 15 years of education is predicted a log hourly wage of 2.6, a man over the age of 28 with more than 15 years of education is predicted 3.4. The second tree predicts a log wage of 2.6 for someone who has less than 15 years of education and is younger than 28, 3.1 for someone with more than 15 years of education who is younger than 32.

The decision tree method often over-fits for the data set which it is built on, making for a model which is not generalisable for the entire population. A random forest approach builds many, many trees using random samples of the entire data set[9] as training sets and cross-validating across the sample,[10] making for a model which is less likely to over-fit the training data set. This model specifically uses observations left out by sampling methods at each formed tree to form an estimate of the accuracy and so select technical parameters to maximise prediction accuracy. The approach makes a model extremely good at capturing non-linearities and interactions in given data, to form a method extremely good at predicting a response variable. Such non-linearities are displayed in Figures 3 and 4, where splitting according to specific variables – and possibly multiple times in the same variable – produces predictions, with no specific conditions or linearities needed.

It is important to be careful about predictor variable selection. For example, hourly wage rate can be extremely well predicted by annual income. However, this prediction process has no economic significance above demonstrating the well-known fact that annual income and hourly wage are very, very highly correlated. Mullainathan and Spiess (2017) note this issue of self-prediction and others in the use of big-data practices in economics. The variables selected for random forest prediction each have plausible causative relationships with income measures. They are as follows: age, gender, race (where available), citizenship status (where available), marriage status, dummies for living in a rural area, suburb, central city, whether they are self employed, union membership and total years of education.[11] Work experience or a direct measure of potential experience is not available, so it should be noted that selection on the age variable may indicate selection for experience (similar to the use of age in a standard Mincer wage equation).

[9] This process is called bootstrapping, where at each step of the process a sample (with replacement) of the data set is taken.

[10] Cross validation is a method to chose best value of parameter in cost function minimisation, a technical detail in building the prediction model.

[11] See Appendix 1 for a summary table of these variables. It should be noted that missing values (which are common in citizenship status or race) are coded as 0 to be included in tree prediction. This transformation has no numerical effect in the tree-based methods other than ensuring they may be included in the analysis.

4.2 Residuals in Wage Inequality

Table 2 documents the distribution of residuals in five year groupings for 1980–2016,[12] for the prediction methods: standard Mincer equation, adjusted Mincer equation, random forest prediction. The standard deviation of the generated residuals is shown, as well as differences between the 90th and 10th, 90th and 50th, 50th and 10th percentiles.

Table 2: Inequality Measures Based on Regression Model Residuals for Hourly Wage

		1980–	1985–	1990–	1995–	2000–	2005–	2010–
Model 1.	S.d.	0.24	0.24	0.26	0.27	0.28	0.27	0.28
	90-10	1.15	1.20	1.19	1.22	1.30	1.27	1.26
	90-50	0.60	0.65	0.61	0.63	0.70	0.68	0.66
	50-10	0.55	0.55	0.57	0.58	0.61	0.59	0.60
Model 2.	S.d.	0.23	0.21	0.26	0.26	0.31	0.25	0.30
	90-10	1.13	1.20	1.18	1.19	1.27	1.24	1.23
	90-50	0.60	0.65	0.61	0.62	0.68	0.66	0.65
	50-10	0.53	0.55	0.57	0.58	0.60	0.58	0.59
Model 3.	S.d.	0.23	0.21	0.26	0.26	0.31	0.25	0.30
	90-10	1.05	1.12	1.14	1.15	1.24	1.21	1.17
	90-50	0.54	0.60	0.56	0.62	0.66	0.66	0.60
	50-10	0.51	0.52	0.58	0.53	0.58	0.54	0.57
Observations:		1302	1222	1241	1113	1548	1382	2192

Note: This analysis is performed on a 10,000 subsample, as computation was too slow before the deadline in the entire dataset.

The standard deviation for all models increases across the the 36 year period, being practically identical at 0.23 for the 1980–1984 grouping. Interestingly the adjusted Mincer equation returns identical standard deviation values to the random forest prediction. The 90-10 percentile differential increases to a maximum in the 2000–2004 grouping in all three models, noting that the residual difference is lowest in the random forest estimation at every step. The 90-50 and 50-10 percentile differential increase over the sample period at a much lower rate than the 90-10 differential. This shows a distribution of residuals in all models increasing at the upper extreme across the sample period.

[12] 6 year grouping for 2010–2016.

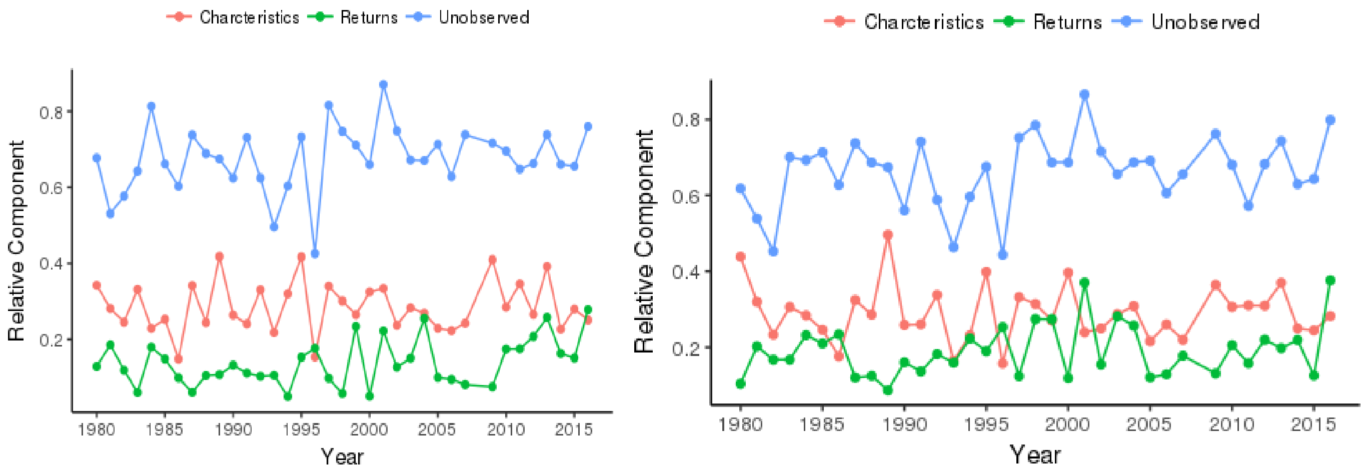
5 Results

Table 3 presents the results of a standard Mincer equation in predicting wages. The returns to education are fairly standard, at around 9 to 10% increase for an additional year of education, with a significant and negative coefficient for the quadratic term of potential experience. Noticeably, the R^2 value for predicting both measures of income, showing how this model captures only around 22% of the variation in wages for the sample time period.

Table 3: Mincer Wage Equation Results

	Dependent variable:	
	Log hourly wage	Log annual income
	(1)	(2)
Years education	0.090^{***}	0.104^{***}
	(0.0001)	(0.0002)
Potential experience	0.032^{***}	0.057^{***}
	(0.0001)	(0.0001)
(Potential experience)^2	-0.001^{***}	-0.001^{***}
	(0.00000)	(0.00000)
Constant	1.392^{***}	8.511^{***}
	(0.002)	(0.003)
Observations	2,185,521	2,185,521
R^2	0.222	0.230
Note:		*p<0.1; **p<0.05; ***p<0.01

Figures 5 and 6 give part of the 90-10th percentile log wage differential accounted for by the three components by each prediction method. In every instance, unobserved characteristics are the largest component contributing to the wage differential. The standard OLS Mincer regressions are almost identical in their decomposition of components, simulating a small yet significant rise in the component of unobserved characteristics, and a similar fall in the component of observed characteristics. Returns to observed characteristics fall from around 0.4 in 1980 to around 0.3, seeing a large degree of variation over the time period, an on average a fall in importance.



(a) Standard Mincer Equation

(b) Adjusted Mincer Equation

Figure 5: Components of 90-10th Log Hourly Wage Percentile, 1980–2016

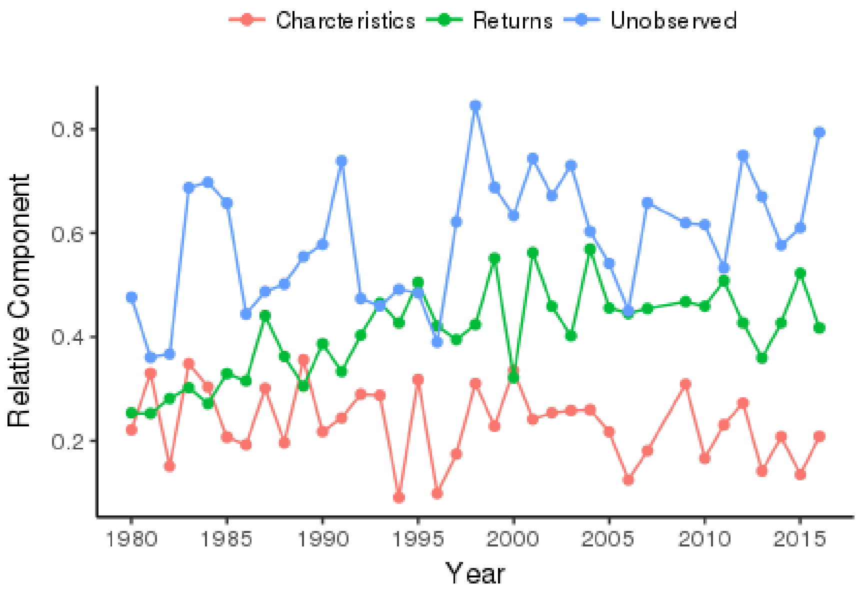


Figure 6: Components of 90-10th Log Hourly Wage Percentile for Random Forest Prediction, 1980–2016

Random forest predictions are starkly different to the OLS Mincer wage predictions. Returns to observable characteristics increase consistently across the sample period, while the contribution of observable characteristics is stagnant over the sample period. This rise in return to observable characteristics continues the rise documented for 1980–1985 by Juhn et al.

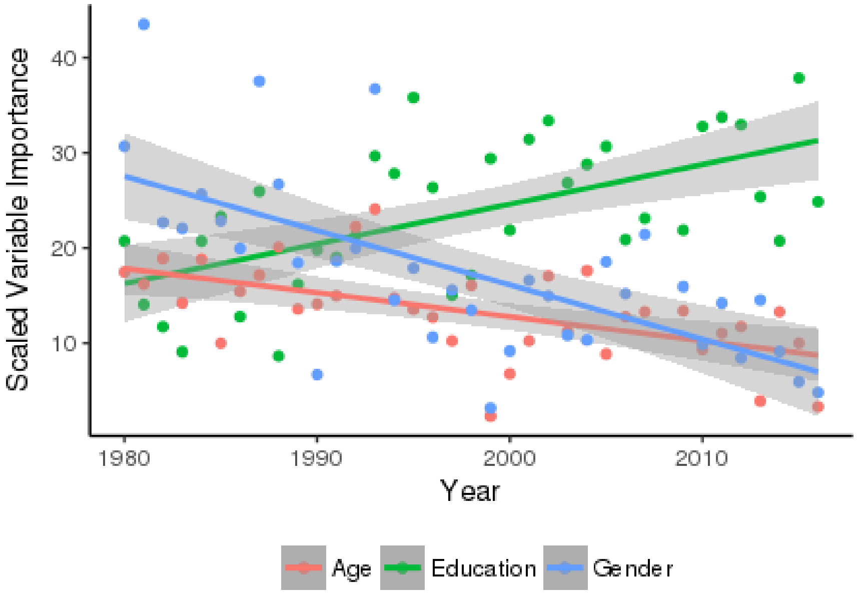


Figure 7: Variable Importance in Random Forest Prediction, 1980–2016

(1993), but is only shown so clearly in a decomposition that uses random forest prediction. The role of unobserved characteristics is, while still the largest component, comparatively lower than the other models. This result comes from the fact that the random forest uses a larger amount of the given information in a more efficient manner, so that less variation is naturally attributed to unobserved characteristics.

Figure 7 documents variable importance in random forest prediction of log hourly wages for the three variables which were most important: age, years of education and gender. Age is likely selected on so frequently because of its high correlation with experience in the work force so that older workers have higher wages. Over the sample period the role of education in predicting wages increases, shown by the line of best fit, while the role of age and gender fall. This result supports findings in figure 6 that estimate a rising role of returns to observable characteristics over the sample period. This graph also gives a representation of the diminishing role of gender in determining wages, or in other words, the decrease (though not disappearance) of large wage differences between male and female wages. [13]

[13] Note: these analyses were completed on a 10,000 subsample of available data due to computational times. They will be run on the entire sample for the next draft.

6 Discussion and Limitations

This paper provides estimates for a lesser role of unobserved skill in explaining rising inequality than has been previously estimated. This reinforces the interpretation of Juhn et al. (1993) of the rise in returns to skill before 1985, extending the same findings and interpretation to 2016. The role of unobserved skill in explaining rising inequality seems to be lower today than it was before 1985, even when estimated by standard linear regressions. There are multiple interpretations of this finding. Today, job-searching and wage decisions are aided by more advanced technology (not least the internet) so that it's possible that job-searching and wage decisions are more efficient and so more aligned with observable characteristics (and less so with unobserved characteristics). The role of years of formal education, especially at the university level, has become more important in explaining higher wages at the top of the income distribution. This is, of course, an observed characteristic, so can also explain the relatively smaller role of unobserved characteristics in rising inequality. Finally, it is also possible that the unobserved characteristic of discrimination (outside of gender which is observed) in the work force has become less common in the labour market, explaining the relative fall in unobservable characteristics in wage inequality.

This analysis has focused on the use of different prediction models in an extremely famous decomposition of wages and wage inequality. The decomposition method was, of course, built to consider a linear regression in prediction methods. The exact equation for the procedure in section reflects this, exampling the novel process of applying other prediction methods to this decomposition. The comparison across models leads to the question: which prediction models provide the correct estimate for the role of the components in explaining rising wage inequality. Modern regression techniques, including random forests, are designed (and successfully) produce better estimates for response variables with computational algorithms, but they can use any number of variables to predict other variables. This is commonly called data-mining, where any data can be found to have some form of statistical pattern to predict a response variable though the process may have no economic intuitive or interpretable significance. This issue has been addressed in these methods by using a well-defined set of intuitive and economically significant variables to predict wages (and in the process avoiding self-prediction). Future research on using computational statistics in economic research will need to address the question of what should be considered predictor variables – the same as choosing what should be considered observable characteristics in this paper. If we can one day collect so many more variables on people that have reasonable causative relationship for wages, should they all be considered observable characteristics? If not, then is the line drawn at only observing gender, years of education and potential experience as in the Mincer wage equation?

The random forest method was chosen as one of the best prediction methods that is easily reproducible in standard statistical software that also has a fairly intuitive basis for understanding. The field of computational statistics and machine learning is progressing extremely fast, so that newer or adjusted prediction methods with better accuracy are already available. This means that the statistical methods of elaborating wage decompositions to tree-based methods is only scratching the surface of using new modern prediction methods beyond a linear regression framework in these methods.

Lastly, it in unclear in this wage decomposition where the role of other factors than the three components identified fit in. For example, recent research has concentrated on the role of constrained competition and rising barriers to entry in rising inequality in the US. Perhaps this factor fits in to unobserved skill, or to observed skill for those with higher education and positions in concentrated industries or large companies. Future research on the components of rising inequality at the individual level may need to provide a greater framework to consider such larger scale issues.

7 Conclusion

This analysis documented a rise in wage inequality for 1980–2016, continuing the well-documented rise for 1960–1980. The trend to greater inequality is attributed primarily to a rise in returns to skill, observed (such as education and experience) and unobserved. Modern prediction techniques attribute the rise in inequality much less to unobserved characteristics than standard linear models, exhibiting the dependence of the Juhn et al. (1993) decomposition method to specific prediction methods. Specifically, random forest prediction estimates that returns to observed characteristics have steadily risen since 1980, while returns to to unobserved characteristics have increased at a lower rate. Furthermore, years of education is increasingly important in predicting wages, supporting the interpretation that rise in returns to observed skill has steadily risen.

The basic rationale for the rise in returns to observable skill is the growth for demand for highly skilled workers and skill-biased technological change. The modern prediction methods capture the rise in returns to education and observable skill than standard econometric methods, attributing the rise in inequality to more so than in previous research. It follows than returns to unobservable characteristics, play less of a role in rising inequality than previously estimated. Future research on the role of unobserved and observed skill in determining the level of inequality in the US – as well as all economic research that uses or has used standard linear models for prediction purposes – requires more attention to the choice of prediction methods used.

References

(2016). Center for economic and policy research. march cps uniform extracts, version 1.0. washington, dc.

Abadie, A. and Kasy, M. (2017). The risk of machine learning. *arXiv preprint arXiv:1703.10935.*

Acemoglu, D. (1998). Why do new technologies complement skills? directed technical change and wage inequality. *The Quarterly Journal of Economics*, 113(4):1055–1089.

Acemoglu, D. (2002). Technical change, inequality, and the labor market. *Journal of economic literature*, 40(1):7–72.

Autor, D. (2009). Mit 14.662 graduate labor economics ii spring 2009 lecture note 2: Educational production and wage structure.

Belloni, A. and Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*, pages 121–156. Springer.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., and Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27.

DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.

Furman, J. and Orszag, P. (2015). A firm-level perspective on the role of rents in the rise in inequality. *Presentation at A Just Society Centennial Event in Honor of Joseph Stiglitz Columbia University.*

Juhn, C., Murphy, K. M., and Pierce, B. (1991). Accounting for the slowdown in black-white wage convergence. *In M. H. Kosters (ed.), Workers and Their Wages: Changing Patterns in the United States, AEI Press:Washington D.C.*, pages 107–143.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of political Economy*, 101(3):410–442.

Lemieux, T. (2006a). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, 96(3):461–498.

Lemieux, T. (2006b). The mincer equation thirty years after schooling, experience, and earnings. In *Jacob Mincer A Pioneer of Modern Labor Economics*, pages 127–145. Springer.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of political economy*, 66(4):281–302.

Mincer, J. A. (1974). Schooling and earnings. In *Schooling, experience, and earnings*, pages 41–63. NBER.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Yun, M.-S. (2009). Wage differentials, discrimination and inequality: A cautionary note on the juhn, murphy and pierce decomposition method. *Scottish Journal of Political Economy*, 56(1):114–122.

8 Appendix I.

Table 4: Extended Summary Statistics, 1980-2016

Statistic	Observations	Mean	St. Dev.	Min	Max
Hourly wage, $	2,185,521	25.352	335.722	7.250	444,241.800
Annual income, $	2,185,521	50,426.500	50,976.060	4,060.050	1,848,079.000
Age	2,185,521	39.612	11.671	18	65
Female	2,185,521	0.460	0.498	0	1
Race	2,185,521	0.571	0.851	0	3
Citizenship	2,185,521	0.597	0.490	0	1
Married	2,185,521	0.640	0.480	0	1
Rural	2,185,521	0.205	0.404	0	1
Suburb	2,185,521	0.381	0.486	0	1
Central city	2,185,521	0.237	0.426	0	1
Self-employed	2,185,521	0.003	0.051	0	1
Union member	2,185,521	1.042	0.647	0	3
Years of education	2,185,521	13.696	2.694	0.000	22.000