The Rise in Returns to Skill?

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

Senan Hogan-Hennessy*
Senior Seminar in Economics†

Pomona College, Department of Economics
425 N. College Ave.
Claremont CA 91711
May 2018

*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, with materials for replication, is available at github.com/shoganhennessy/ECON190.

The Rise in Returns to Skill? A Modern Regression Analysis of Wage Inequality in the Current Population Survey
Senan Hogan-Hennessy
Working Paper, Senior Seminar in Economics
May 2018

Abstract

A wage decomposition shows the contribution of workers' observed and unobserved characteristics, and their returns, to the distribution of income. Yet the estimated contributions and determinants of rising income inequality change with how wages are econometrically modelled. I extend the Juhn, Murphy & Pierce (1993) decomposition to multiple econometric models, including modern regression methods, on the US income distribution 1980–2016. There is a larger contribution of observed and unobserved worker characteristics in rising wage inequality than previous estimates suggest, and a smaller contribution from changing returns to worker characteristics using modern regression methods to model wages. The analysis shows the wage decomposition's vulnerability to choice of wage model, and discusses the further use of modern regression methods in econometric models of wages.

Senan Hogan-Hennessy
Pomona College
425 N. College Ave.
Claremont CA 91711

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, Murphy & Pierce (1993). The methods for wage prediction and determining observable skills are extremely important in determining this composition, and have previously only used a standard linear regression. 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 rose 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 techniques predict a 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, and modern regression models attribute even more of the rise in inequality to unobserved factors. They also exhibit a rising importance for years of education in predicting wages, with a fall in the role of gender.

The paper is structured as follows. Section 1 surveys the current literature on wage inequality and decomposition by regression approaches. Section 2 describes the March Current Population Survey (henceforth March CPS) data set, and trends in income inequality for the US. Section 3 presents the wage decomposition, and various regression prediction methods for the empirical analysis. Section 4 presents the 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. Section 6 concludes.

1 Literature Review

Wage inequality has increased dramatically since the 1950's in the US. Many studies in labour economics attribute this rise in inequality to a rise in the returns to skill. Across other subfields of economics, however, there are multiple factors that also explain rising inequality, ranging from rise in market power (Furman & Orszag 2015), to de-unionisation and supply and demand shocks (DiNardo, Fortin & Lemieux 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).

1.1 Wage Decomposition and Ordinary Least Squares

The Mincer wage equation applies 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. 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, Murphy & Pierce (1991) that focuses on wage predictions and residuals in a few standard regressions, showing the role of unobserved skills. The methods make a few notable assumptions, and crucially show the Mincer wage equation produces much worse estimates for wages (i.e. higher errors) in the late 1980's that did in earlier decades. The approach has some notable criticisms: 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 (2006*a*) 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.

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 (2006*b*) 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 today than 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. 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 variables in determining wages. Modern regression approaches provide a viable option to better address these empirical problems.

1.2 Modern Regression Techniques

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 provides a method for intuitive prediction of wages, but provides 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 relate to wages (linear or non-linear), and which have changed from those in previous decades. This problem can not be completely fixed by a simple adjustment, as noted by Lemieux (2006*a*). These issues bring into question the robustness of only using the 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. For example, a regression tree is a completely different form of regression than OLS and common econometric methods. The process involves building a decision tree by minimising an error function by splitting on available variables, relaxing any linear restrictions on the model. Bootstraps of data are used to form a random forest model, with extremely high power for prediction with less problems of over-fitting data (Breiman 2001). Labour economics has, however, been slow to use such techniques in empirical studies. Belloni & Chernozhukov (2011) and Abadie & Kasy (2017) develop novel prediction techniques, and test their properties by predicting wages in the March CPS. Chalfin, Danieli, Hillis, Jelveh, Luca, Ludwig & Mullainathan (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."

2 Wage Inequality in the US

2.1 Data

The analysis of this paper is conducted on 35 years of wage and demographic information for individuals taken from the March Annual Social and Economic Supplement of the Current Population Survey, commonly referred to as the March CPS. The 35 years span 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

Table 1: Summary Statistics, 1980-2016

Statistic	Observations	Mean	St. Dev.	Min	Max
Hourly wage, $	2,185,520	25.15	149.73	7.250	94,963.63
Annual income, $	2,185,520	50,426	50,975	4,060	1,848,079
Age	2,185,520	39.612	11.671	18	65
Years of education	2,185,520	13.7	2.7	0	22
Female	2,185,520	0.46	0.50	0	1
Race	2,185,520	1.289	0.752	0	3
Married	2,185,520	0.640	0.480	0	1
Rural	2,185,520	0.205	0.404	0	1
Suburb	2,185,520	0.381	0.486	0	1
Central city	2,185,520	0.237	0.426	0	1
Self employed	2,185,520	0.003	0.051	0	1

earnings divided by annual hours worked,[1] and at the 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.[2] The sample is restricted to full-time workers, both male and female,[3] making at least the 2015 hourly federal minimum wage ($7.25) between ages 18 and 65. The total number of observation is 2,185.520, which is extremely large (even when treated in year increments) and so is adequately large to apply modern big-data regression methods. Observations from the year 2008 are excluded due to sample size problems.[4] See Table 1 for summary statistics for real hourly and annual wage, age, years of education, and proportion female for the sample, and a collection of variables used in the prediction methods (explained in section 4.1).

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

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

[3] 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.

[4] 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 – counter-intuitively given economic conditions – so that this year is excluded in analysis.

2.2 Rising Inequality

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

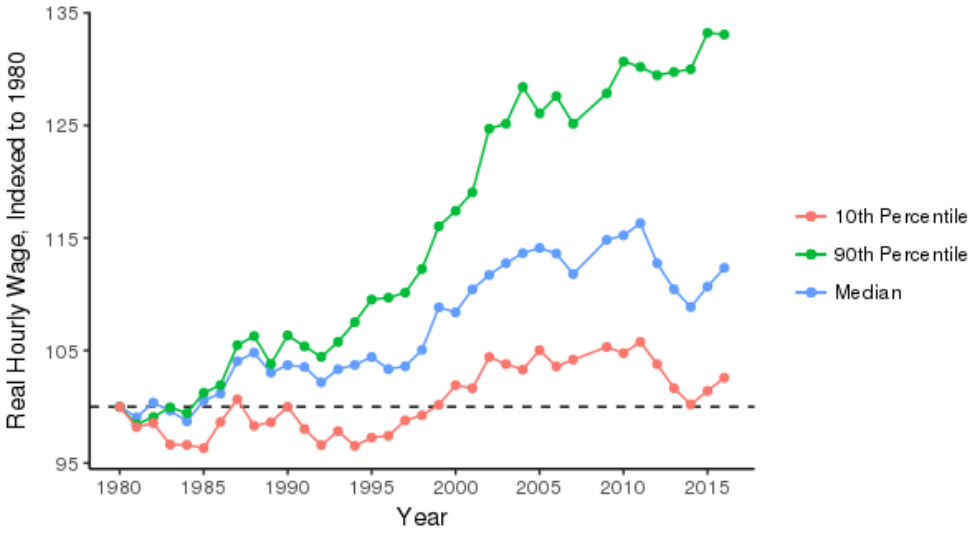
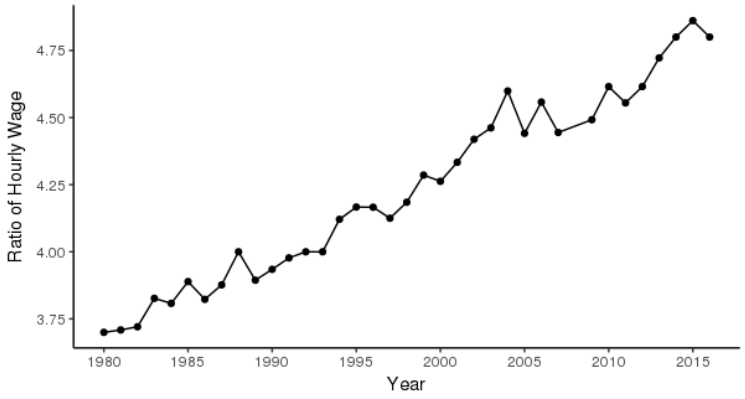


Figure 1 presents the 10th, 50th and 90th percentile of the hourly wage distribution 1980–2016. Each series is indexed to 1980 (so that each is assigned a value of 100 in 1980) for comparison between the groups and visualisation of their respective changes since 1980. The 90th percentile of he wage distribution saw a mild increase in income from 1980 until 1990, followed by a rapid rise until the mid-2000s. The median of the income distribution steadily increased until 2007 at a lower rate than the 90th percentile, while the 10th percentile saw a general stagnation across the time period, including any gains made over the 2000s vanishing in the years following 2010. 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

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



for the 36 year period, following a trend documented for preceding decades in previous research.

(a) Hourly Wage. **(b)** Annual Income.

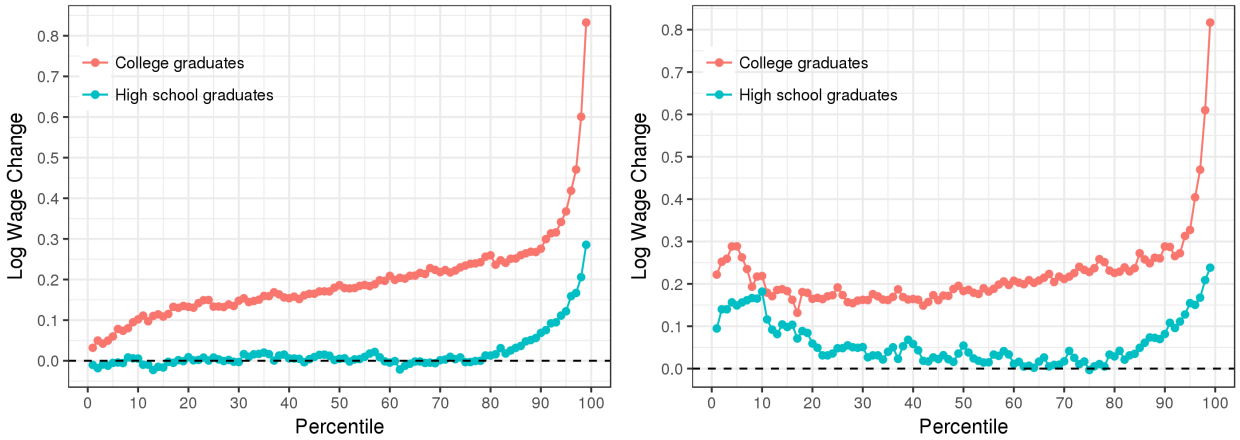


Figure 3: Wage Change by Percentile, 1980–2016.

Figure 3 shows the log wage change across the income distribution (percentiles 1 to 99) for college graduates and high school graduates.[5] High school graduates in the bottom 80% saw no change in hourly wage,[6] and only saw increase in annual income for the bottom 10%. College graduates's hourly wage increase sees a linear trend for percentile with the 10th percentile increasing by around 0.1 log points and the 90th by around 0.3 log points, with a similar yet (but not as steep) linear trend for annual income. Above the 10th percentile both measures of income rose by many multiple of any other percentile, showing a large increase in income at the top of the income distribution. Juhn et al. (1993) document the linear rise in the income distribution between the 10th and 90th percentiles between 1970 and 1985. The rise in wage inequality since then has laregly been driven by gains in income *above* the 90th percentile, however.

3 Explaining Wage Inequality

Wage Inequality has risen significantly since 1980. The role that changing returns to skills, observable or unobservable, plays in explaining this rising in inequality is not so clear cut. Equation (1) presents a standard linear regression that aims to decompose wage inequality

[5] 3 years groupings are used for 1980 and 2016, i.e. 1980-1983 vs 2013-2016, to limit variation.
[6] Likely due to sample restrictions limiting samle to an income of a minimum wage full-time worker.

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). The residual may be specified in terms of a distribution, in equation (2), where $F_{it}(.|\mathbf{X}_{it})$ is the cumulative density function for residuals of 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 and their returns (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 for residuals across a reference time period, here 1980–1985, so that neither varies year on year. Y_{it}^1, Y_{it}^2, Y_{it}^3 are then defined as follows.

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

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

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

Y_{it}^1 is the distribution of wages under fixed returns to observable characteristics, so that $\bar{F}^{-1}(.|\mathbf{X}_{it})$ and $\bar{\beta}$ are fixed and do not vary with year. Y_{it}^2 is the counter-factual distribution of wages under variable returns to observable characteristics and quantity of observable characteristics but a fixed distribution of residuals. Y_{it}^3 is 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 and thus unobserved characteristics or returns. 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.

3.1 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) takes.

3.1.1 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 takes the following form.

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

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} are standard coefficients to be estimated with residual ε_{it}. Potential experience, x_{it}, is defined as age minus years of education minus 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 after then.

3.1.2 Mincer Wage Equation, Adjusted

The Mincer wage equation may take an expanded form, as recommended by Lemieux (2006*b*). The form is presented in equation , where the quadratic form for potential experience is

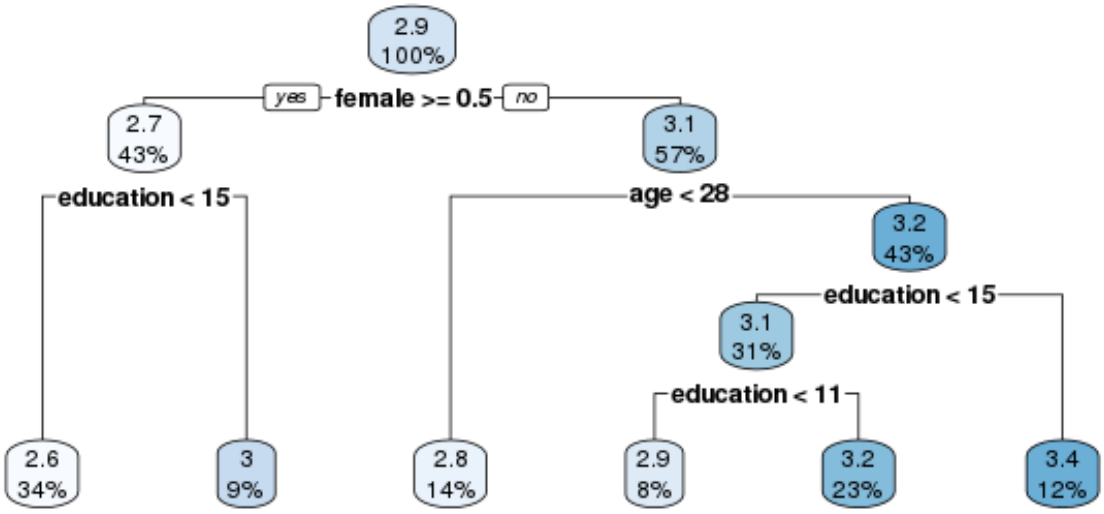
expanded to a quartic, and the years of education include a quadratic.

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

The given variables are represented as in equation (3). This functional form includes a quartic function in potential experience instead of just a quadratic and a quadratic term in years of schooling to capture the growing convexity in the relationship between schooling and wages. It follows that 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 adjusted form of the Mincer wage equation is used to predict wages while maintaining an OLS approach.

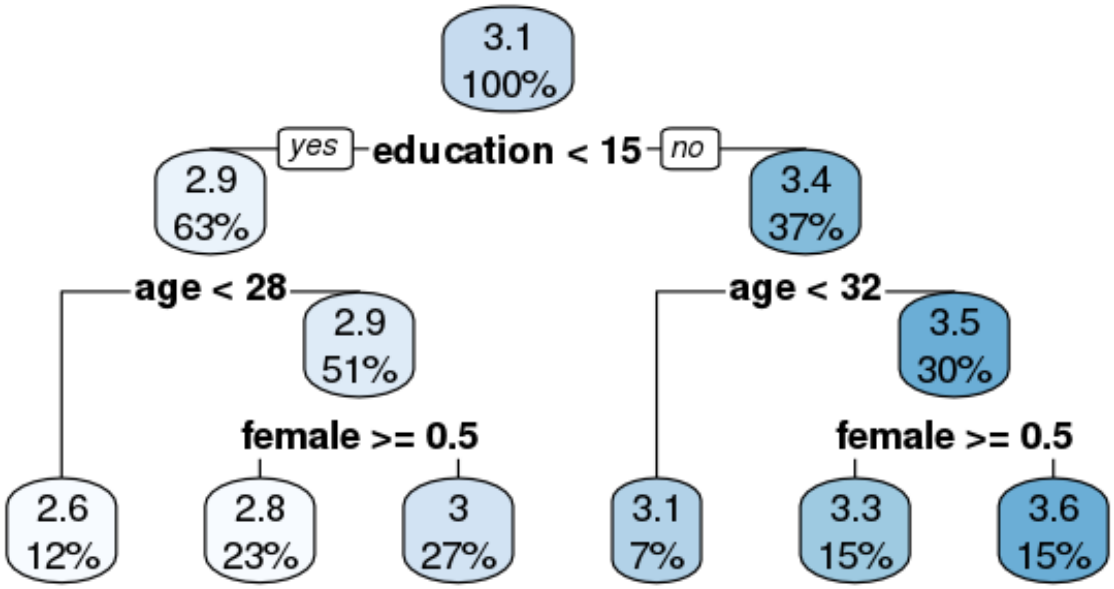
3.1.3 Tree-Based Methods

Figure 4: 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 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

Figure 5: Decision tree for real log wages, 2010-2016.



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 4 presents a decision tree to depict such a model for log wages 1980–1985, figure 5 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 generally over-fits its training data, making for a model which is not generalisable for a new data set or, indeed, the entire population. A random forest approach builds many, many trees using random samples of the entire data set[7] as training sets and cross-validating across the sample,[8] 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

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

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

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 & 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 are all the variables in the March CPS that have plausible causative relationships with income across the population. They are as follows: age, gender, race, marriage status, dummies for living in a rural area, suburb, central city, whether they are self employed, union membership and total years of education.[9,10] Work experience or a direct measure of potential experience is not available, so that selection on the age variable may indicate selection for experience in the same way that age is used as a measure of potential experience in a standard Mincer wage equation.

3.2 Residuals in Wage Inequality

Table 2 documents the distribution of residuals in five year groupings for 1980–2016,[11] 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.

The standard deviation for all models increases across the the 36 year period, being similar for the 1980–1984 grouping. Interestingly, both forms of the Mincer equation return a nearly identical distribution of residuals, increasing from 0.46 to 0.52 while the random forest prediction gives slightly lower (yet similar) values. The 90-10 percentile differential generally increases across the groupings, showing a more unequal distribution of residuals in later years. The 90-50 and 50-10 percentile differentials also increase across the years, those

[9]The March CPS uniform extracts start with 475 variables, most of which are sub-variables not useful in this project. These nine variables are those available for every year and that fit the plausible causative relationship criterion while also not causing any self-prediction problems.

[10]Missing values 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, and there are no missing values in variables used in OLS prediction.

[11]6 year grouping for 2010–2016.

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

		1980–	1985–	1990–	1995–	2000–	2005–	2010–
Mincer	S.d.	0.46	0.47	0.47	0.51	0.53	0.54	0.54
	90-10	1.17	1.18	1.18	1.22	1.26	1.28	1.30
	90-50	0.61	0.60	0.60	0.64	0.66	0.68	0.70
	50-10	0.56	0.57	0.58	0.59	0.60	0.60	0.60
Mincer,	S.d.	0.46	0.47	0.47	0.51	0.53	0.53	0.54
adjusted	90-10	1.17	1.18	1.18	1.21	1.25	1.26	1.29
	90-50	0.61	0.60	0.60	0.63	0.66	0.67	0.70
	50-10	0.56	0.57	0.58	0.58	0.59	0.59	0.59
Random	S.d.	0.41	0.43	0.44	0.48	0.50	0.51	0.52
forest	90-10	1.00	1.05	1.08	1.13	1.16	1.18	1.22
	90-50	0.52	0.53	0.54	0.58	0.61	0.62	0.65
	50-10	0.49	0.52	0.53	0.54	0.55	0.55	0.57
Observations:		275,813	266,275	271,907	247,343	341,165	308,241	474,776

this rise is not as pronounces as for the 90-10 differential.

It makes sense for the random forest method to produce lower values for inequality across the residual distribution, as this modern prediction technique is generally more accurate in predictions, thus producing smaller errors/residuals. This can be seen in a year-specific comparison of the coefficient of determination, R_t^2 for years $t = 1980, 1981, \ldots, 2007, 2009, \ldots, 2016$:

$$R_t^2 = 1 - \frac{\sum_{i=1}^{n}(Y_{it} - \hat{Y}_{it})^2}{\sum_{i=1}^{n}(Y_{it} - \bar{Y}_{it})^2} = 1 - \frac{\sum_{i=1}^{n}\varepsilon_{it}^2}{\sum_{i=1}^{n}(Y_{it} - \bar{Y}_{it})^2} \tag{9}$$

R^2 is consistently higher for random forest prediction than either form of the Mincer equation, as expected. In 1980 the random forest has an associated R_{1980}^2 of 0.35, while the Mincer equation has 0.14 and 0.15 for the adjusted form. The R_t^2 for both forms of Mincer equation rises to around 0.2 in 1985, before staying between 0.2 and 0.25 for the remaining years. The random forest's R_t^2 drops to around 0.3 in 1995, before staying around this value for the following years.

4 Results

Table 3 presents the results of a standard Mincer equation in predicting wages across the entire sample period (i.e. estimating the regression with fixed coefficients $\bar{\beta}$). The returns

Figure 6: Coefficient of Determination by Prediction Model, 1980–2016

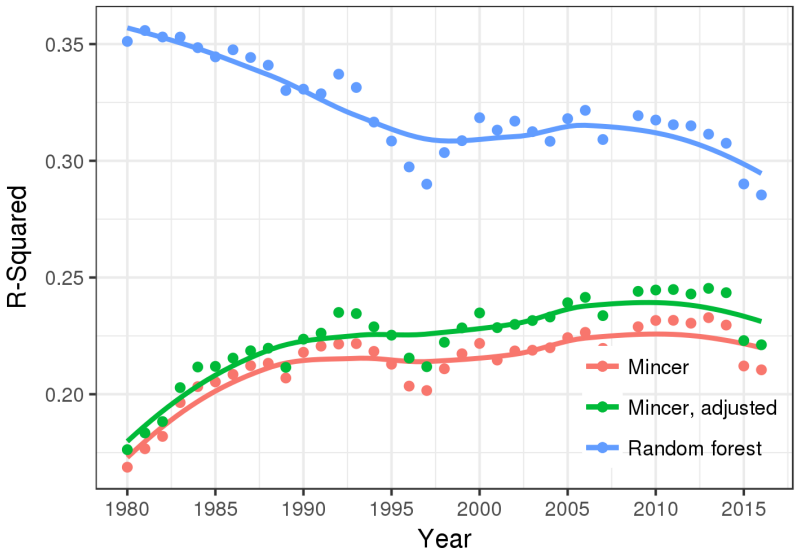
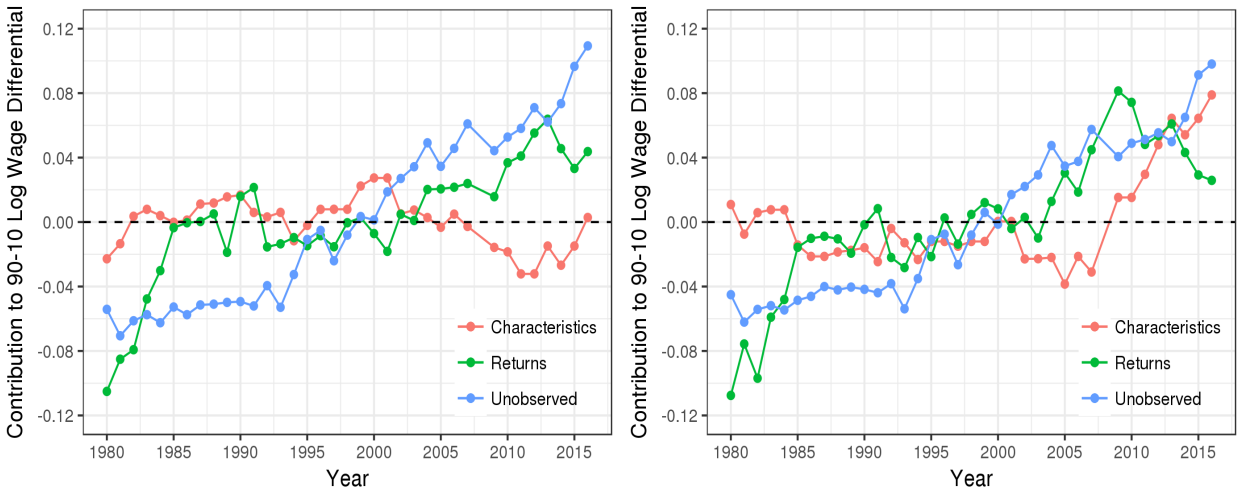


Table 3: Mincer Wage Equation Results

	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	2,185,520	2,185,520
R^2	0.222	0.230
Note:		*p<0.1; **p<0.05; ***p<0.01

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. The R^2 value in both measures of income shows how this model captures around 22% of the variation in wages for the entire sample period.

Figures 7 and 8 give the components of the 90-10th percentile log wage differential ac-



(a) Standard Mincer Equation **(b)** Adjusted Mincer Equation

Figure 7: Components of 90-10th Log Hourly Wage Percentile, 1980–2016

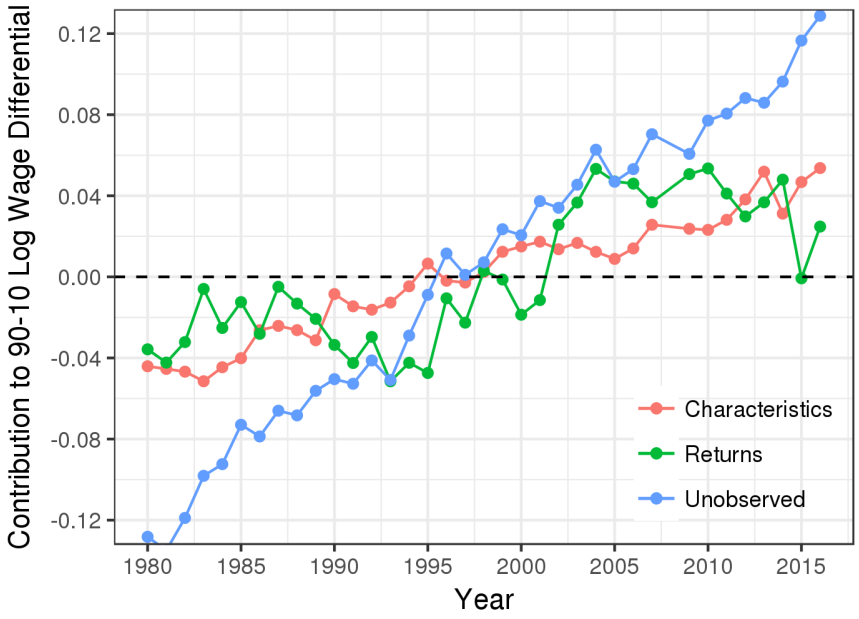


Figure 8: Components of 90-10th Log Hourly Wage Percentile for Random Forest Prediction, 1980–2016

counted for by the three components by each prediction method, minus their mean.[12,13] In every instance, unobserved characteristics rise more than any other component in the contribution to the wage differential. The Mincer regressions are similar in their decomposition of components, simulating a similar and significant rise in the component of unobserved char-

[12]Components do not sum to 1 as the graphs show change over time.

[13]The 90-10th percentile log wage differential is mathematically identical to the 90-10 wage ratio, so that the components contribute to the figures in Figure 1.

acteristics. The contribution of observed characteristics doesn't change drastically, except for the adjusted form for after 2005 where the contribution rises significantly. Returns to observed characteristics rise rapidly 1980–1990 in the standard Mincer regression,[14] with a very similar profile for the adjusted equation. After 1990 returns to observed characteristics do not continue on the same rapid increase in either regression, though do follow a broad increase in contribution.

The decomposition using random forest prediction differs markedly from the Mincer decompositions in the component change for observed characteristics. There is a steady rise over the entire time period of the role of observed characteristics in contributing to the wage differential, disagreeing completely with the standard regression's components. Returns to observable skill follow a similarly stagnant path before rising in contribution in the 2000s. And lastly, the contribution of unobserved characteristic and returns rises consistently and more significantly than in any other prediction method.

Table 4: Observed and Unobserved Components of Changes in Inequality 1980–2016

	Differential	Total Change (1)	Observed Quantities (2)	Observed Prices (3)	Unobserved Differences (4)
Mincer					
	90-10	0.264	0.006	0.049	0.147
	90-50	0.193	0.027	0.020	0.103
	50-10	0.070	-0.021	0.029	0.043
Mincer, adjusted	90-10	0.264	0.065	0.067	0.132
	90-50	0.193	0.013	0.024	0.100
	50-10	0.070	-0.052	0.042	0.032
Random forest	90-10	0.264	0.091	-0.043	0.229
	90-50	0.193	0.044	-0.018	0.148
	50-10	0.070	-0.047	0.024	0.081

Table 4 presents the total change in the log wage differential for 1980–2016,[15] with contribution of each component. The components in columns 2–3 theoretically sum to the total change in column 1 yet there is some error in practice, yielding non-perfect sums as noted by Autor (2009). Unobserved characteristics is the largest component for all prediction tech-

[14] An almost exact replication of the trend by Juhn et al. (1993), as expected.
[15] Again using 3 year groups to minimise small-sample variation.

niques, but is estimated to be a much larger component in random forest prediction than in the others. The Mincer equations estimate change in observed prices as the second biggest component to the wage differential, while observed quantities barely contributing. The role is reversed in random forest prediction, where observed returns even contributing negatively to the wage differential i.e. the conterfactual distribution of wages under varying returns is estimated to have a *lower* wage differential in 2016 than in 1980 – which is clearly not observed.

Figure 9 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 a trend line, while the role of age stays constant and gender falls. This result supports findings 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.

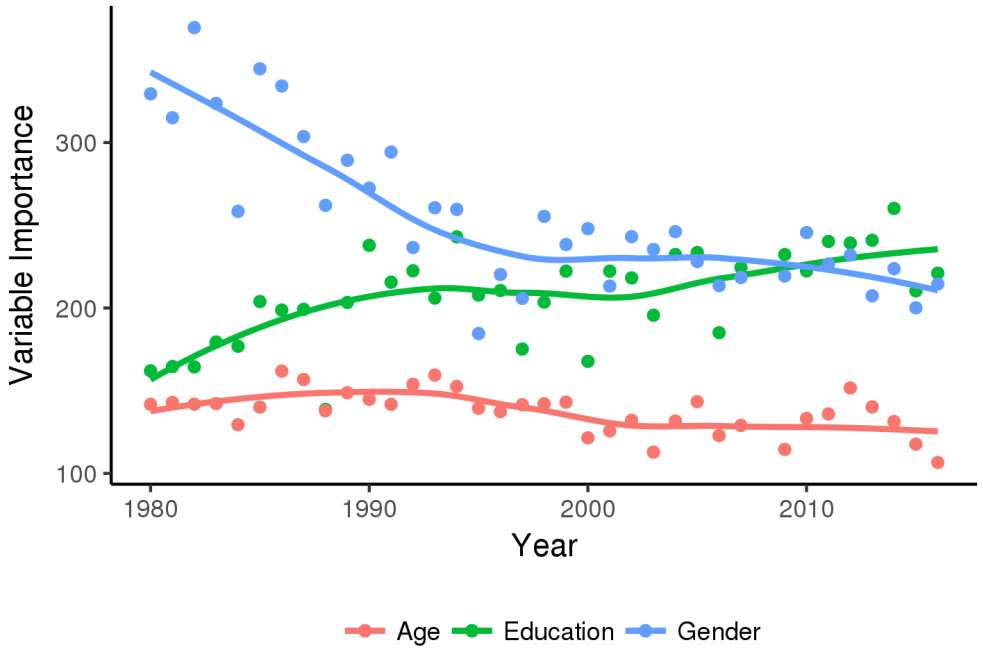


Figure 9: Variable Importance in Random Forest Prediction, 1980–2016

5 Discussion and Limitations

This paper provides estimates for a greater role of unobserved skill in explaining rising inequality for 1980 and onwards than has been estimated in previous years. This reinforces the interpretation of Juhn et al. (1993) of the rise in returns to unobserved skill before 1985, extending the same findings and interpretation to 2016. The role of unobserved skill in explaining rising inequality seems to be even higher today than it was before 1985, when estimated by standard linear regressions and even more so with modern regression techniques. There are many possible interpretations of these findings. Modern prediction methods better use given data and characteristics to predict wages, so that if they cannot attribute much of the rise in inequality to any of these observed returns or characteristics, then they attribute an even greater component of inequality to unobserved characteristics. It follows that even with more covariates, and more efficient use of them, the rise in wage inequality is even less explainable than was previously estimated in labour economics.

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?

Standard econometric methods in the ordinary least framework have decades and decades of research and analysis depicting their mathematical properties. These properties range from simple consistency, unbiasedness in the Guass-Markov theorem to numerous important and more complicated properties. Modern prediction methods are extremely good at predicting variables, even only using variables determined to have economic significance. They do not, however, have the same research on their exact statistical properties. For example, consider the property of consistency. This means that asymptotically a parameter estimator converges in probability to its true value, and has a short proof for a standard OLS regression under some standard assumptions. It's not so clear what form this property even takes in modern regression techniques,[16] let alone what conditions must be satisfied for it to apply.

Modern regression techniques, including random forests, are designed to (and successfully do) produce better estimates for response variables with computational algorithms, but

[16]For example in a random, is that forming the optimal number of splits in variables, optimal number of bootstrapped trees, or whatever leads to the best predictions?

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 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.

6 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 to a rise in returns to skill, observed (such as education and experience) and unobserved. Modern prediction techniques attribute the rise in inequality even more so 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 estimates that observed characteristics have contributed steadily more to inequality since 1980, and unobserved characteristics and returns have increased at a significantly higher rate. Furthermore, years of education is increasingly important in predicting wages, supporting the interpretation that rise in returns to observed skill have also steadily risen.

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.

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. However, there is no consensus for what is driving an increase in the role of unobserved characteristics and returns in explaining wage inequality. The modern prediction methods provide better estimates for wages than standard econometric methods, but attribute the rise in inequality more to what is unobserved than in previous research. 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

Abadie, A. & 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. & Chernozhukov, V. (2011), High dimensional sparse econometric models: An introduction, *in* 'Inverse Problems and High-Dimensional Estimation', Springer, pp. 121–156.

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

Center for Economic and Policy Research. March CPS Uniform Extracts, Version 1.0. Washington, DC. (2016).

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

DiNardo, J., Fortin, N. M. & Lemieux, T. (1996), 'Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach', *Econometrica* **64**(5), 1001–1044. **URL:** *http://www.jstor.org/stable/2171954*

Furman, J. & 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. & 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.* pp. 107–143.

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

Lemieux, T. (2006*a*), 'Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?', *American Economic Review* **96**(3), 461–498.

Lemieux, T. (2006*b*), The "mincer equation" thirty years after schooling, experience, and earnings, *in* 'Jacob Mincer A Pioneer of Modern Labor Economics', Springer, pp. 127–145.

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', NBER, pp. 41–63.

Mullainathan, S. & 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.