ESSAYS ON MACHINE LEARNING APPLICATIONS IN ECONOMICS: CAUSAL INFERENCE AND PREDICTION

A Dissertation presented to

the Faculty of the Graduate School

at the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

YONG BIAN

Dr. David Kaplan, Dissertation Supervisor

MAY 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

ESSAYS ON MACHINE LEARNING APPLICATIONS IN ECONOMICS: CAUSAL INFERENCE AND PREDICTION

presented by Yong Bian,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. David Kaplan

Dr. Saku Aura

Dr. Peter Mueser

Dr. Christopher Wikle

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Dr. David Kaplan, for his greatest help in my research life. Without his advise I would not accomplish this dissertation; without his support I would not achieve the goal of getting a PhD degree. I am so grateful that I have such a nice advisor who has endless knowledge and is always being supportive.

Secondly, I would like to thank my other committee members, Dr. Saku Aura, Dr. Peter Mueser and Dr. Christopher Wikle for the valuable revising suggestions they provide for my dissertation. I would also like to thank my coauthors Dr. Wei Kong and Dr. Dawei Li for their collaborating and hard work. I'm so lucky that I have wonderful working experience with them.

With special mentions to Cheng Qian and Jing Song for studying together; Dr. Xiqian Wang and Fangda Wang for great suggestions on my research and all my friends. It was great to have them as peer fellows during my PhD journey.

Last but not least, I would express my deep thanks to my parents, Hongwei Bian and Yanli Huai for their endless support and love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

This study includes three chapters related to machine learning applications with focus on different empirical topics. The first chapter talks about a new method and its application. The second chapter focuses on young economics professors salary issues. While the third chapter discusses scientific paper publication values based on text analysis and gender bias.

In the first Chapter, I give a discussion of Double/Debiased Machine Learning (DML) which is a causal estimation method recently created by Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) and apply it to an education empirical analysis. I explain why DML is practically useful and what it does; I also take a bootstrap procedure to improve the built-in DML standard errors in the curriculum adoption application. As an extension to the existing studies on how curriculum materials affect student achievement, my work compares the results of DML, kernel matching, and ordinary least squares (OLS). In my study, the DML estimators avoid the possible misspecification bias of linear models and obtain statistically significant results that improve upon the kernel matching results.

In the second chapter, we analyze the effects of gender, PhD graduation school rank, and undergraduate major on young economics professors' salaries. The dataset used is novel, containing detailed and time-varying research productivity measures and other demographic information of young economics professors from 28 of the top 50 public research universities in the United States. We apply double/debiased machine learning (DML) to obtain consistent estimators under the high-dimensional control variable set. By tracking the first 10 years of their professional work experi-

ence, we find that there barely exist effects on young faculties' salaries from the above three factors in most of the experience years. However, the gender effect on salary in experience year 7 is both statistically significant and economically significant (large enough in magnitude to have a practical meaning). In experience years 5 to 7, which are also near most faculties' promotion years, the gender effects are obvious. For both PhD graduation school rank and undergraduate major, the estimates for experience years 7 to 9 are large in magnitude; however they do not possess statistical significance. Overall, the effects tend to expand with years of experience. We also discuss possible economic mechanisms and reasons.

In the third chapter, we build machine learning and simple linear models to predict academic paper publication outcomes as measured by journal H-indices, and we discuss the gender bias associated with these outcomes. We use a novel dataset with paper text content and each paper's associated H-index, authors' genders, and other information, collected from recently published economics journals. We apply term frequency–inverse document frequency vectorization and other Natural Language Processing (NLP) tools to transfer text content into numerical values as model inputs. We find that when using paper text content to predict an H-index, the prediction power is around 60% in our classification model (4 tiers) and the root mean squared error is around 44 in our regression model. Moreover, when controlling for paper text, the gender causal effect hardly exists. As long as the paper contains similar text, gender does not influence the change in H-index. Additionally, we give real-world meanings associated with the models.

Chapter 1

Double Machine Learning Discussion and Application to the Causal Effect of the *California Math* Curriculum

1.1 Introduction

In empirical studies, economists often use linear models to achieve their analyzing goals and typically ordinary least squares (OLS) to estimate their linear models, however, there might be problems with it. Traditional causal effect analysis methods such as fixed effects, difference-in-differences, and two-stage least squares (2SLS) are common in empirical studies, and most of the time, they are conducted by linear models. However, linear is too strong of an assumption. When causal meanings are the purpose of modeling, linear partial effect estimations could have the limitations in reflecting the true causality which is not constant or with complicated forms. Even

though one may argue that with polynomials and interaction terms a linear model can achieve the purpose of capturing non-constant and heterogeneous effects as well, the number of regressors is restricted by the sample size. The number of parameters must be considerably less than the number of observations in order to get precise estimates. OLS performs poorly when the sample size is limited and the number of parameters is large or even increasing with sample size.

Linear models can help achieve different goals when it is used to different situations. OLS regression captures the global feature of the data, but is less flexible revealing details. Sometimes the population relations could be complicated and interact in a nonlinear way; then, establishing one linear model is very difficult. In statistical modeling, the goal is to fit the data and make good predictions, one could build local nonlinear functions and sum them up just like the idea from Kernel Smoothing Methods, but it would be very tedious and with low efficiency in economics empirical studies. Because in economics, people care more about explanation than simply a well fitted model. So that economists will just go for other non-parametric such as kernel matching techniques or would just use a linear model instead even if the linear model has very low prediction power or fit the data badly but it has good explanation properties.

Even if a linear model is a final "best" choice among many empirical techniques to an economics researcher, it may still be hard for him to build a solid linear based model to choose what interaction terms, year dummies, and fixed effect terms to add into a model. It depends on subjective judgement or convention. For example, most of the time researchers will include as many regressors (the so-called characteristic variables) as possible in their regressions for the purpose of reducing bias and control-

ling variance. However, increasing the number of independent variables in an OLS regression is at a risk of having multicollinearity and overfitting problems. Moreover, when the sample size is smaller than the number of parameters, even if interest is only in a small part of the parameters in the model, the existence of nuisance parameters still causes poor performance of traditional OLS regression.

Recently, machine learning technology has come into economists' sights, and they want to connect these techniques with causal inferences. Among them, a newly developed method named Double/Debiased Machine Learning (DML)[1] is a powerful one as a causal estimation technique. It nests the prediction power of machine learning to obtain consistent causal estimators under high dimensional covariates.

In this paper, I make several contributions. First, I explain how DML develops and connects to machine learning, discussing how it remedies the limitations of traditional models. Second, I conduct an application of DML. My application serves as an extension to a recent elementary school math curriculum effect analysis conducted by Koedel et al. (2017). In their paper, they use the traditional estimation methods of kernel matching and restricted[2] OLS as causal analysis tools and conclude that positive effects exist in elementary math textbook adoptions on students' achievement in California. Among the four textbooks they study, *California Math* outperforms the other three and will increase the students' performance by 0.05 to 0.08 student level standard deviations compared to its alternatives. In their paper, the authors express concerns about the models they use, one of which is about the linear setting. My work extends their study by estimating a partially linear model with DML. I obtain more efficient causal estimates compared to theirs. Finally, my work extends

[1]Originated by Chernozhukov et al. (2018).

[2]Restricted to the part of the sample with common support.

the built-in DML standard errors to provide clustered standard errors.

1.2 Methodology

I will use two models in my application, both of them follows the methodology in (Chernozhukov et al., 2018). And I will in this section briefly explain their main modeling idea. One is a partially linear model, with the binary treatment variable D linearly adding to a nonparametric function $g_0(.)$ of control variables $\mathbf{X}$. The other is a more general one with the binary treatment variable D being included in a totally nonparametric function. The partially linear model is as defined in (Chernozhukov et al., 2018) equation (1.1) and (1.2)

$$Y = D\theta_0 + g_0(\mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0, \tag{1.2.1}$$

$$D = m_0(\mathbf{X}) + V, \quad \mathbb{E}[V \mid \mathbf{X}] = 0. \tag{1.2.2}$$

This model measures a constant causal effect to everyone; whereas the more general nonparametric model allows the effect to be heterogeneous across individuals. Letting binary variable D be involved in the g_0 function, as defined in (Chernozhukov et al., 2018) euqation (5.1) and (5.2)

$$Y = g_0(D, \mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0, \tag{1.2.3}$$

$$D = m_0(\mathbf{X}) + V, \quad \mathbb{E}[V \mid \mathbf{X}] = 0. \tag{1.2.4}$$

The parameter of interest is the average treatment effect (ATE) and when CIA is satisfied, ATE is equal to the model parameter θ_0,

$$\theta_0 = \mathbb{E}[g_0(1, \mathbf{X}) - g_0(0, \mathbf{X})]. \tag{1.2.5}$$

$\mathbf{X}$ affects the treatment through $m_0(\mathbf{X})$ and affects the outcome variable through $g_0(D, \mathbf{X})$. Both g_0 and m_0 functions are nonparametric, complicated and unknown. Unconfoundedness or the conditional independence assumption (CIA)[3] must be satisfied if the goal is to identify the causal effect.

The main idea of DML is to build a Neyman-orthogonal score function and split the sample doing cross fitting (see Appendix A.1 for more details). The score has to satisfy both a moment condition and an orthogonality condition to overcome the regularization bias. The stricter requirement on the score function makes DML different from traditional methods. Meanwhile, sample splitting is also important to remove the bias from overfitting. Because in the model it requires estimations of nuisance parameters like $g_0(\cdot)$ and $m_0(\cdot)$ as well as causal parameters and they are not estimated simultaneously. Such that using different part of data for estimating different part is needed.

In the estimation of both models above, the sample of $(D, \mathbf{X})$ needs to be independent, identically distributed (iid). Sample splitting plays an important role. First, let's divide the sample into K folds randomly, such that each subsample I_k contains

[3]Unconfoundedness means conditioning on $\mathbf{X}$, the counterfactuals $Y(0)$ and $Y(1)$ are uncorrelated with treatment D. CIA means conditioning on $\mathbf{X}$ the choice of D is statistically independent with U, the model error term. They are similar concepts, I treat them as same in my paper.

N/K number of observations, where $k \in 1, \ldots, K$. The final DML estimator $\tilde{\theta}_0$ solves

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{nk}[\psi(D, \mathbf{X}; \tilde{\theta}_0, \hat{\eta}_{0k}] = 0, \tag{1.2.6}$$

(equation (3.4) in (Chernozhukov et al., 2018)), where $\psi(\cdot)$ is the Neyman-orthogonal score function; $\hat{\eta}_{0k}$ represents the estimator of nuisance parameters associated with $g_0(.)$ and $m_0(.)$; $\mathbb{E}_{nk}$ is the empirical expectation over kth fold of the data. For each subsample I_k, its corresponding auxiliary sample is used to construct an ML estimator of $\hat{\eta}_{0k}$.

1.3 Empirical Application: Background and Data

It has been showed that textbook materials have important educational effects on student achievement (Hadar, 2017; van den Ham and Heinze, 2018). But which textbook materials to choose in each subject are barely studied.

Recently, Koedel et al. (2017) published a study about effects of curriculum materials on student achievement based on California data. The basic backgrounds are as followed. In California, the curriculum adoption process is partially centralized. The state initiates a list of textbooks for a particular subject at a certain year. Each district has the right to adopt any textbook from the list, or they can choose not to adopt at all (do not use textbooks on the list). Moreover, school textbook adoptions in California are required to be reported as a result of the 2004 *Eliezer Williams et al. vs. State of California et al.* court ruling and resulting legislation. In math, the adoption process roughly moved together by the state's initiation. The majority of

districts make their textbook adoption decision when state adoption happens.

Koedel et al. (2017) focus on elementary math textbooks adopted in California in fall 2008 and fall 2009. The data is originated from schools' 2013 School Accountability Report Cards (SARC).[4]

Table 1.1: Descriptive Statistics for *California Math* and Composite Alternative

Variable	*California Math*	Composite Alternative
School Outcomes		
Preadoption Grade 3 math score	0.06	−0.06
Preadoption Grade 3 ELA score	0.07	−0.07
School Characteristics		
%Female	48.9	48.7
%Economically disadvantaged	56.0	57.8
%English learner	28.0	30.1
%White	29.9	28.6
%Black	6.3	7.1
%Asian	7.2	7.8
%Other races	56.6	56.5
Enrollment	429.5	415.2
2008 adopter	53.7	46.2
School-area characteristics (census)		
Median household income (log)	10.9	10.9
Share low education	19.3	21.4
Share missing census data	1.2	1.2
District outcomes		
Preadoption Grade 3 math score	0.00	−0.02
Preadoption Grade 3 ELA score	−0.12	−0.08
District characteristics		
Enrollment	6075.5	5341.0
n(Schools)	602	1276
n(Districts)	92	224

Note: The information showed in this table is summarized from Koedel et al. (2017) Table 1.

Table 1.1 shows basic descriptive statistics of the 4 math textbooks studied by Koedel et al. (2017). The test score values in this table are the average (school level

[4]For more detailed information please refer to Koedel et al. (2017).

and district level) standardized data. Column "*California Math*" shows those schools adopted *California Math* with characteristics from either 2007 or 2008 (preadoption), and at least one test score[5] of Grade 3 from 2009 to 2013. Column "Composite Alternative" shows the average of those schools who adopted the other three textbooks[6]. Outcome variable Y here is Grade 3 math score, and except Grade 3 ELA score, all the other characteristic variables are denoted by vector $\mathbf{X}$. School-level data is used throughout this study for the reasons discussed in (Koedel et al., 2017).

Koedel et al. (2017) use traditional techniques for acquiring the effects and point out what is the limitation in regards of their modelings. They first calculate the propensity score for each observation using a probit model. Based on the propensity score, they select the part of sample with common support. That means in both treated and control groups, only individuals from the same range of propensity scores are being selected as modeling sample. They apply kernel matching, restricted OLS and residualized matching separately to estimate treatment effect of the four textbooks on student achievement.[7] Among all four commonly used elementary math textbooks they study, *California Math* published by Houghton Mifflin outperforms others and has the highest treatment effect. They also point out that in their restricted OLS model by imposing a linear form they can obtain more statistically precise results but introduce bias into their estimation.

Upon knowing *California Math* works best among all four textbooks, they con-

[5]All test scores mentioned here are standardized student test scores. The standardized test score is obtained from the universe sample data collected from California Department of Education (CDE).

[6]Namely, *enVision Math California*; *California Mathematics: Concepts, Skills, and Problem Solving*; *California HSP Math.*

[7]The results are shown in Table 1.2. For more detailed information of these methods please refer to Koedel et al. (2017).

duct a quasi-experimental study on *California Math* with respect to the composite alternative. They make adopters of *California Math* to be in the treatment group and all other three textbook adopters as alternative composite in the control group.

Table 1.2 summarizes results of 4 years treatment effects after adoption. The result of Year 1 gives the comparison between students who used the newly adopted textbooks only in their Grade 3. Year 2 gives the comparison between students who used the newly adopted textbooks in their Grade 2 and Grade 3. Year 3 and Year 4 gives give results for students who used adopted textbooks for all three grades.

In order to make sure the estimates have causal meanings, Koedel et al. (2017) justify the CIA condition by doing falsification tests. They estimate two types of models. In the first model, they estimate the pre-adoption curriculum effects on students. The only difference is the time of test scores, instead of using scores after adoption they choose scores of 3 to 6 previous adoption years. As in the main model, schools adopting *California Math* is set as the treatment group and the composite alternative as the control group. Results show that there is no effect on student scores for all pre-adoption years. In the second model, They use Grade 3 English test scores as outcome variable and estimate effects for not only all pre-adoption years same as in the first model but also all 4 years effects after adoption like in the main model. They also report no effect. Therefore they argue the conditional independence is satisfied.

Koedel et al. (2017) express their surprises on seeing the treatment effects are not increasing over time. Even though the model results are out of expectation, they gave some possibilities behind the scene: moderate dosage effect, increased exposure time in earlier grades dose not have enough effect on grade three test scores; unstable quality of different grade's curriculum materials.

Table 1.2: Koedel et al. (2017) Estimated Effects of *California Math* on Grade 3 Mathematics Achievement Relative to the Composite Alternative

Variable	Year 1	Year 2	Year 3	Year 4
Treatment: *California Math*				
Control: Composite Alternative				
Treatment Effect: Kernel matching	0.063	0.083	0.061	0.070
	(0.054)	(0.051)	(0.059)	(0.059)
Treatment Effect: Restricted OLS	0.050**	0.064**	0.049**	0.058**
	(0.019)	(0.023)	(0.023)	(0.023)
Treatment Effect: Residualized matching	0.050**	0.065**	0.052**	0.060**
	(0.020)	(0.024)	(0.024)	(0.026)
No. of districts/schools (*California Math*)	92/597	89/588	91/595	90/590
No. of districts/schools (composite alternative)	213/1,143	214/1,145	216/1,146	213/1,144

Note: A result from Koedel et al. (2017). Standard errors are obtained by estimation on a 250 bootstrap sample being district level clustered. Year 1 means after 1 years adoption, grade 3 students have used the newly adopted book for one year; Year 3 means after 3 years adoption, grade 3 students have used the book for 3 years. The estimates are converted from school level to student level standard deviation by multiplying 0.45, which is a transformation ratio as explained in Koedel et al. (2017). $^{*}p \leq .10$, $^{**}p \leq .05$.

Moreover, they also mentioned that the potential problems could come from the limitation of linear model. For example Koedel et al. (2017) use linear Probit model to calculate propensity scores where prediction is the purpose. However, it is hard to believe linear models can work as well as the true conditional expected function (CEF) or its non-parametric approximation. Moreover, in this California textbook adoption data set, it's impossible to write a saturated model because the variables are not all discrete and the number of variables is relatively large.

DML estimates CEF nonparametrically; under the CIA condition, it can give consistent treatment effect estimators. It allows discrete and continuous variables and can deal with high-dimensional nuisance parameters. With DML, there is no need to separately calculate propensity score and then do a OLS regression or Kernel Matching. Therefore, in order to make up the concerns of Koedel et al. (2017), I apply this newly developed DML to estimate textbook causal effects and compare my results with the original ones. The improvement on the accuracy of estimation will be showed in next section.

1.4 DML Application and Results

Tables 1.3 to 1.6 list DML estimation results of the student achievement effects from using *California Math* for 4 years after adoption.

Columns in each table displays six different ML methods for getting g_0 and m_0 estimators. In "Lasso", I use all characteristic variables listed in Table 1.1 with six order polynomials of school level enrollment, six order polynomials of district level enrollment and eight order polynomials of income, and all their second order

interaction terms. For all the other methods I use all variables in their original level (no interacting terms or powered terms). "Reg.Trees" fits a single decision tree, the hyper parameter is chosen by 2-fold cross validation. "Forest" runs Random Forest, it takes an average over 1000 trees. "Boosting" uses boosted regression trees with 2-fold cross validation. "Neural net" uses two neurons and set logistic loss function for classification and linear for regression. "Ensemble" combines "Lasso", "Boosting", "Random Forests" and "Neural Net" and take their average. The last column "best" runs differently, at each time of splitting, method(s) will be selected respectively for giving best estimates of g_0 and m_0, and use each selected method to estimate them separately. Therefore, in "best", DML could end up using different methods in g_0 and m_0 estimations.

Each table reports two sets of results. Panel A displays the general interactive DML model, as in Equation (1.2.3) and Equation (1.2.4); panel B displays partially linear DML model, as in Equation (1.2.1) and Equation (1.2.2). Within each model, an estimate of ATE and its standard error are reported. "se(median)" reports standard errors using median method adjusting splits variations.[8] "se" reports median standard error across the 10 splits.

DML standard errors, "se(median)" and "se" are calculated under the i.i.d. sampling assumption, however for our data, it is more proper to use a clustered standard error would. So, I take a bootstrap procedure so to get clustered standard errors for the DML estimates. First, bootstrap the whole sample in district level (randomly select district repeatedly and include all schools within each selected district in the bootstrap sample), obtaining 50 bootstrapped samples; then, for each of the 50 boot-

[8]For median method details please check Chernozhukov et al. (2018) definition (3.3).

strapped sample get a DML estimate, use the 50 estimates to calculate a standard deviation. The clustered standard errors are reported under "Clustered.se" in each result table.

In Table 1.7, I list ensemble DML results for comparison. A main purpose of applying DML is to relax the linear setting, so an interactive model would be more general and representative compared to partially linear model. Interactive DML gives smaller clustered standard errors compared to Kernel Matching and OLS. DML fulfill its promise of being a more efficient nonparametric estimator due to its two core strategy (orthogonalization and sample splitting). For the point estimates, interactive DML is quite similar to Kernel Matching in year 1 and Year 3; however, the effects seem to be fairly stable over the years in interactive DML instead of an increase and decrease trend in Kernel Matching. DML results meet what to expect of how an effect develop in reality. A stable effect is more common than a fluctuating effect. Therefore, no matter from realistic meaning aspect or model setting aspect, DML results give us a new window of finding and discussing how the true effect looks like.

In Year 1, interactive DML and Kernel Matching are similar but OLS estimate is much smaller. The reason could be the model misspecification bias from linear models. Or because that OLS only use original level of the variables with limited second order terms, without any interaction terms and further higher order terms. Moreover, the difference between OLS and interactive DML reminds me to think about the concern from Koedel et al. (2017). And that linear models capture the general data features, it is stable however lack of flexibility makes it worth to run a new set of nonparametric practice as DML do.

Table 1.3: Effect of *California Math* After 1 Years' Adoption

	Lasso	Reg.Trees	Forest	Boosting	Nnet	Ensemble	best
A. Interactive Model							
ATE	0.051	0.040	0.039	0.047	0.094	0.065	0.084
se(median)	(0.015)	(0.055)	(0.019)	(0.022)	(0.04)	(0.025)	(0.034)
se	(0.01)	(0.026)	(0.008)	(0.015)	(0.017)	(0.008)	(0.02)
Clustered.se	(0.031)	(0.034)	(0.016)	(0.019)	(0.045)	(0.015)	(0.016)
B. Partially Linear Model							
ATE	0.046	0.024	0.047	0.047	0.042	0.029	0.053
se(median)	(0.011)	(0.026)	(0.019)	(0.014)	(0.014)	(0.018)	(0.020)
se	(0.010)	(0.017)	(0.018)	(0.014)	(0.012)	(0.018)	(0.018)

Table 1.4: Effect of *California Math* After 2 Years' Adoption

	Lasso	Reg.Trees	Forest	Boosting	Nnet	Ensemble	best
A. Interactive Model							
ATE	0.068	0.083	0.057	0.065	0.098	0.068	0.065
se(median)	(0.016)	(0.032)	(0.015)	(0.017)	(0.028)	(0.013)	(0.036)
se	(0.014)	(0.026)	(0.009)	(0.015)	(0.021)	(0.009)	(0.023)
Clustered.se	(0.026)	(0.027)	(0.018)	(0.019)	(0.039)	(0.019)	(0.018)
B. Partially Linear Model							
ATE	0.059	0.062	0.074	0.076	0.057	0.081	0.080
se(median)	(0.015)	(0.05)	(0.023)	(0.016)	(0.020)	(0.023)	(0.024)
se	(0.012)	(0.018)	(0.019)	(0.014)	(0.014)	(0.019)	(0.019)

1.5 Conclusion

I find that the effect of using *California Math* is stable over years. There is no increase in effects as expected by Koedel et al. (2017), and they don't observe an increasing effect in their study either. One reason could be the moderate dosage effect which is hard to detect; or the dosage effects was dominated by the most recent textbook usage, the increased expose on earlier grades is not important compared to grade 3 contemporary usage results. Another possible reason is that the quality of textbook is not stable from year to year. Considering the DML's advantage in dealing with high dimensional data and gives consistent estimator, the first reason is more plausible.

Table 1.5: Effect of *California Math* After 3 Years' Adoption

	Lasso	Reg.Trees	Forest	Boosting	Nnet	Ensemble	best
A. Interactive Model							
ATE	0.011	0.042	0.041	0.061	0.077	0.057	0.049
se(median)	(0.033)	(0.047)	(0.017)	(0.025)	(0.043)	(0.016)	(0.037)
se	(0.014)	(0.03)	(0.01)	(0.019)	(0.024)	(0.01)	(0.028)
Clustered.se	(0.047)	(0.032)	(0.018)	(0.021)	(0.029)	(0.018)	(0.019)
B. Partially Linear Model							
ATE	0.041	0.036	0.054	0.056	0.086	0.039	0.061
se(median)	(0.018)	(0.025)	(0.023)	(0.018)	(0.018)	(0.023)	(0.023)
se	(0.013)	(0.019)	(0.020)	(0.016)	(0.014)	(0.021)	(0.021)

Table 1.6: Effect of *California Math* After 4 Years' Adoption

	Lasso	Reg.Trees	Forest	Boosting	Nnet	Ensemble	best
A. Interactive Model							
ATE	0.037	0.015	0.039	0.055	0.087	0.057	0.038
se(median)	(0.024)	(0.049)	(0.018)	(0.028)	(0.059)	(0.017)	(0.046)
se	(0.014)	(0.033)	(0.009)	(0.017)	(0.024)	(0.01)	(0.034)
Clustered.se	(0.044)	(0.035)	(0.019)	(0.020)	(0.031)	(0.018)	(0.019)
B. Partially Linear Model							
ATE	0.051	0.028	0.047	0.057	0.050	0.058	0.058
se(median)	(0.014)	(0.027)	(0.022)	(0.015)	(0.015)	(0.023)	(0.021)
se	(0.013)	(0.019)	(0.020)	(0.015)	(0.014)	(0.020)	(0.020)

Table 1.7: Effects of *California Math* on Grade 3 Mathematics Achievement Relative to the Composite Alternative: Compare with DML results

Variable	Year 1	Year 2	Year 3	Year 4
Treatment: *California Math*				
Control: Composite Alternative				
Treatment Effect: Kernel matching	0.063	0.083	0.061	0.070
	(0.054)	(0.051)	(0.059)	(0.059)
Treatment Effect: Restricted OLS	0.050^{**}	0.064^{**}	0.049^{**}	0.058^{**}
	(0.019)	(0.023)	(0.023)	(0.023)
Treatment Effect: Interactive DML (Ensemble)	0.065^{**}	0.068^{**}	0.057^{**}	0.057^{**}
	(0.015)	(0.019)	(0.018)	(0.018)

Note: $^{*}p \leq .10$, $^{**}p \leq .05$.

The effect is more severe on the contemporary use of a textbook when the test score is considered as a achievement measure.

OLS model may not be a good choice in making inferences in this situation. The OLS modeling effects after first year adoption is smaller than those from Kernel Matching or DML. That could be a consequence of model misspecification. Even though Kernel Matching is a nonparametric method and free from model specification problems, when using it in practice as Koedel et al. (2017) did, the estimates are not statistically significant, the standard errors are high.

Therefore, DML outperforms Kernal Matching in providing a statistical significant estimator and beats OLS in linear model restriction.

Chapter 2

Are Young Economics Professors' Salaries Affected by their Background?

2.1 Introduction and Motivation

The literature in the past 40 years contains many studies of salaries in academia. However, not many focus specifically on economics, and most lack depth or breadth. There are some studies that concentrate on particular influencing factors and their relationships with salary, such as gender, race, seniority, citation and administrative service (Moore, Newman, and Turnbull, 1998; Bratsberg, Ragan, and Warren, 2003; Li and Koedel, 2017; Hilmer, Ransom, and Hilmer, 2015). For example, Moore, Newman, and Turnbull (1998) point out that with research productivity held constant, the negative relationship between seniority and earnings disappears. Carlin, Kidd, Rooney, and Denton (2013) discuss gender differences in the role of productivity in

determining salary; however, they cover all disciplines and only include one university. Hilmer, Ransom, and Hilmer (2015) focus on economics faculty but not specifically young professors and only discuss how citation history affects an individual's salary. Others focus on mechanisms. For example, Kahn and Lange (2014) consider how employers learn employees' true productivity over time as well as how true productivity itself changes over time.

Good measures of research productivity as control variables are important. Because we are curious about the causal effect on salary of gender, PhD school rank and undergraduate major that might be correlated with many other things (for example, family and education background, research ability, etc.) that have an influence on salary as well, good research productivity measures as controls are helpful to identify the effects. As Katz (1973) stated, "It became clear that research ability, publication record, and national reputation were the most important factors influencing salary and promotion decisions." However, many researchers treat each person's productivity as constant over time (Ginther and Hayes, 2003; Claypool, Janssen, Kim, and Mitchell, 2017) or only use a few rough or self-reported productivity measures (Yang and Webber, 2015; Sax, Hagedorn, Arredondo, and Dicrisi, 2002). In such cases, there is reason to doubt the causality of estimated coefficients of treatment variables, due to the lack of variability of productivity measures.

Unlike full professors who already have many publications and established reputations, it is hard to precisely value young economics professors' research abilities, which is important in determining salary. In economics, it takes longer to get a paper published than in many other scientific disciplines. Even after acceptance by a journal, it could take one year (or longer) to get published. Therefore, young economics

professors may have fewer publications compared to professors in other disciplines at the same career stage.

Because publication quantity is so limited for most young economics professors, employers might look into more detailed information on their existing research and publication information. For example, employers may examine numbers of citations, coauthors, and pages. Hilmer, Ransom, and Hilmer (2015) give reasons why citations are an important research ability measure in social science fields. Moreover, the above factors may interact with other detailed journal quality information to influence the professors' salaries, for example, number of pages published in "tier one" journals, pages in "tier two" journals, etc. This intuition is further supported by research; e.g., Liebowitz and Palmer (1984) write, "Where articles are published can affect one's promotion, tenure, and salary at one's present job." As Gibson, Anderson, and Tressler (2017) mention, separating journal rank into different tiers and collecting information by journal tier would be practical. In this study, we include journal tier-separated publication information to better control for research productivity and to improve previous studies' estimates. Also, Gibson, Anderson, and Tressler (2017) argue that the number of citations has a different impact on salary at different universities depending on their ranking. Our focus on public research universities could help reduce the estimation bias from heterogeneity in the university level.

The above facts led us to construct a novel dataset that contains detailed productivity measures along with other background and demographic characteristics of young economics professors. We contribute to the literature by using this dataset to estimate the effects on salary associated with gender, PhD graduation school rank, and undergraduate major (economics and STEM). We have collected data from most

of the top 50 public research universities in the United States, from calendar year 2008 to 2014, focusing on younger economics faculty. We collected personal education and demographic information as well as detailed measurements of research productivity: publications, citations, coauthors, pages written, and journal rank. [1]

Another concern is model specification. Specifically, most studies use linear regression (Altonji and Pierret, 2001; Ginther and Hayes, 2003; Yang and Webber, 2015; Hilmer, Ransom, and Hilmer, 2015; Carlin et al., 2013). There can be model misspecification bias especially when we have a high-dimensional data, even if conditional independence holds.

Our analysis of how the three variables of interest affect young economics professors' salaries is based on professional work experience years (Johnson and Stafford, 1974; Altonji and Pierret, 2001; Perna, 2001; Toutkoushian, Bellas, and Moore, 2007) instead of calendar years. In each sub-dataset of experience year 1 to 10, we collected data on individuals who obtained the same academic work experience during calendar year 2008 to 2014. Approximately 80 variables are captured. In order to estimate causal effects under this high-dimensional control variable set, we apply the double/debiased machine learning (DML) method (Chernozhukov et al., 2018; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, and Newey, 2017). By not setting specific functional forms, we are able to take the advantage of machine learning techniques nested in the DML method in dealing with high-dimensional data, letting the method choose which and how to use variables from the high-dimensional control set to get consistent causal effect estimators.

Next, we explore possible mechanisms behind the causal effects. For example,

[1] Full list of variables are shown in Table 2.3.

Altonji and Pierret (2001) and Altonji (2005) model how employers may statistically discriminate based on group types. Negotiation power (Oberfichtner, Schnabel, and Töpfer, 2020; Claypool et al., 2017; Gerhart and Rynes, 1991) might be another source of causality as well as mobility differences (Blackaby, Booth, and Frank, 2005; Kidd, O'Leary, and Sloane, 2017).

In the rest of this chapter, we will give an in-depth analysis of whether gender, PhD graduation school and undergraduate major make a difference on young economics professors' salaries. By tracking over the first ten years of young economics faculty's professional experience, we are able to discover how the effects change with experience and research productivity. By applying DML, we provide consistent estimators of the effects and discuss possible mechanisms associated with them.

2.2 Literature Review

We found that many papers discuss salary disparity problems, but we focused on those related to young economics faculty members' salary payments. In particular, we collected knowledge on how young economics professors' salaries are determined and the mechanisms at play as documented in these studies.

People might be discriminated against, as reflected in their salary. Altonji and Pierret (2001) show that employers perhaps start with statistical discrimination against people who newly enter the labor force based on easy-to-observe characteristics that are highly correlated with productivity; these employers then gradually learn the workers' true productivity and adjust their salaries accordingly. In their paper, they use experience time variable t to interact with variables, such as education, race and

Armed Forces Qualification Test (AFQT) to track how those factors affect salary payments over time. They found that coefficients on easy-to-observe variables decrease whereas those on hard-to-observe variables increase with additional years of working experience; employers have limited information on early year labor force workers, and there is statistical discrimination on the basis of education. This study relates to our research with respect to how we look at the problem. In addition, we examine the determinants of salary change and track the effects over time. Moreover, we discuss causality instead of just using one unchanged variable to represent productivity as in the case of Altonji and Pierret (2001). Finally, we rely on time-varying productivity measurements as controls.

In a later paper, Altonji (2005) extends the previous paper to examine different job types. One advantage of studying the academic labor market is that it is well defined; each person under study has a PhD degree, which excludes many kinds of heterogeneity (Johnson and Stafford, 1974). Altonji (2005) found that employers learn more about workers in high-skill jobs as time progresses. Research-related academic jobs belong to high-skilled job types, so according to Altonji (2005), a research university faculty member's true productivity will be gradually learned by employers, and their changing research productivity affects their salary.

Besides the employer learning process, we focus on productivity change and how that affects salary. In a more recent paper, Kahn and Lange (2014) show how a worker's productivity changes over time and emphasize that two mechanisms are important: they use a dynamic structure in their model that nests both employer learning and productivity heterogeneity. They also point out: "The pure Employer Learning model predicts that salary payment correlates more with past than with

future performance measures because firms rely on past but not future performance measures to set current pay. In contrast, the pure dynamic productivity heterogeneity model implies that pay correlates similarly with past and future performance." This means employers will not only rely on what they observe regarding an individual's past productivity but also what they expect of their future performance to decide salary. That is the reason we collected current and cumulative productivity measures. Employers can actually observe forthcoming information from a paper being published at a future date. Therefore doing a robustness check is also needed; by including next years' productivity information into each experience year model, we were able to verify our model specification (details in Section 2.7).

However, most studies do not take productivity change into consideration and lack good measures of it. Our dataset can make up this defect by including multi-dimensional and time-varying research productivity measures. Because a good measure of productivity is lacking, people have to treat each person's productivity as constant over a lifetime as in (Ginther and Hayes, 2003) and (Claypool et al., 2017). An important drawback of the most commonly used approach in the literature, as noted by Kahn and Lange (2014), is to treat the Armed Forces Qualifying Test (AFQT) as a proxy for true productivity. The score does not change over time, whereas true productivity may. Kahn and Lange (2014) also point out that traditional data sources make it difficult to distinguish between fixed productivity and changing productivity because they lack independent productivity measures. That is why our new dataset is valuable. Unlike Kahn and Lange (2014), who focus on a labor market with a single employer and a productivity measure that is subjective and discrete (taking integer values of one through four), we include detailed measures in our dataset of

young professors' productivity. These measures include number of pages in papers, number of coauthors, number of citations; current and cumulative information; and information by journal ranking tiers.

There are many other studies discussing how salary is influenced by particular factors, such as seniority, citation, and department rank. Those gave us a guide on what information we should include in our dataset and how we should construct it. Moore, Newman, and Turnbull (1998) estimated the relation between seniority (years at the same university) and salary in research universities after controlling for individuals' publications and education, and they found that seniority is negatively related to salary. Bratsberg, Ragan, and Warren (2003) presented a study on a panel of professors from five universities that further confirmed the negative relation between seniority and faculty pay by controlling for research productivity. However, Barbezat and Donihue (1998) hold an opposite opinion on seniority with a detailed study on different subgroups of people. Given that information, we decided to include seniority information as a control in our dataset. Hamermesh, Johnson, and Weisbrod (1982) concentrated on citations and provided evidence that citations provide an effective productivity measure. However, one limitation of their study is that they only use data from seven universities. Ehrenberg, Pieper, and Willis (1998) and Formby and Hoover (2002) proved that department ranking is correlated with salary payments, such that a higher ranked PhD program usually means higher pay. As a contrast and through extension, we collected data from the top 50 research universities in the United States thought to correlate with salary payments. This data includes detailed information on citations, seniority, department rank, as well as other demographic information (Claypool et al., 2017), such as visiting status, geographic location, age,

and school (details in Section 2.3).

Survey data are widely used. Ehrenberg, Pieper, and Willis (1998) used survey data collected by the American Economics Association (AEA) questionnaire, from 1974/75 to 1980/81, studying whether lower tenure probability universities would pay more to new assistant professors. However, they pointed out that the response rate was not very high; less than half of the departments reported salary information of newly accepted assistant professors. The authors admitted that due to the limitation of available data their results might encounter biases. Because of the use of survey data, there could be self-selection problems as well. Sax et al. (2002) use national survey data of college and university faculties. The productivity measure they used only consisted of self-reported numbers of published article papers over the past two years. The possibility that highly productive faculty happen to have published less than usual or no papers at all during the two survey years makes their result less convincing. Ehrenberg (2002) applied institutional data to analyze the salary difference between public universities and their private counterparts. However, they used aggregated data, which may not be able to detect individual-level effects. Li and Koedel (2017) used self-collected data to analyze the wage differences based on racial/ethnic and gender dimensions among public university faculty in six different disciplines. Similar to Claypool et al. (2017), they used one-year cross sectional data with one-time wage decomposition.

Next, we examined specifically how people study gender, graduate school rank, and undergraduate major in the literature.

Gender. The disparity of salary between male and female faculty in academic fields has been studied extensively. According to Zhang (2018), in the same academic

job level, males are paid less then females. Mitchell and Hesli (2013) and Sax et al. (2002) noted some people have the impression that women professors have less incentive to take part in research, that they take on more service, and have higher teaching loads. We want to find whether the difference is causal or just simply statistical correlation. Many studies show that gender may be correlated with productivity measures. Smart (1991) proved that gender can influence salary through different channels such as academic rank, working age, and whether it is a male-dominated discipline. As noted by Ghosh and Liu (2020), Yang and Webber (2015), and Sax et al. (2002), being female was negatively related to publishing refereed papers and the total number of publications. There are still unexplained positive gender salary gaps between men and women faculties even with controls for education, productivity and other background information (Perna, 2001; Toutkoushian and Hoffman, 2002; Toutkoushian and Conley, 2005; Blackaby, Booth, and Frank, 2005; Claypool et al., 2017). Johnson and Stafford (1974) focused on gender difference in academic salaries, which they attributed to a combination of statistical discrimination and human capital differences over work lifetime. Sax et al. (2002) discussed the gender gap on productivity. Recently, Claypool et al. (2017) showed that among political science faculty, males had superior negotiation power that helped to improve their salary. Li and Koedel (2017) pointed out that academic field, experience, and research productivity are the three major factors that drive wage differences. However, those factors cannot fully explain the gender difference. Ginther and Hayes (2003) note that the gender gap in salaries was not significant and that if there was discrimination on salaries, it came through promotions. Later studies like Weisshaar (2017) and Chen, Kim, and Liu (2008) also confirm the gender gap in promotion is prevalent in many fields.

PhD Graduation School Rank is a comprehensive index that can reflect quality of education a student receives. Thus, students who graduate from higher ranked universities may also obtain higher potential human capital such that there is positive effect on salary from graduation school rank (Claypool et al., 2017). Graves, Marchand, and Thompson (1982) proved that there is an association among department ranking, number of publications, and salary. However, they only show that full professors' salaries are strongly related to publication and ranking. Stock and Alston (2000) found that there is a positive correlation between program rank and initial salaries, even after controlling for information on qualifications, such as research and teaching. Those findings help us to construct one of our research goals: finding more precise causal estimates of graduation rank; we expand the analysis significantly by tracking 10 years of experience.

Undergraduate Major is another interesting factor we investigate to determine its effect on salary. It usually is the case that an economics professor holds a PhD degree in economics. Admissions to PhD programs in economics considers economics talent, where quantitative ability is an important item to be considered (Jones, Schuhmann, Soques, and Witman, 2020). Moreover, quantitative ability has been perceived to be very important in determining research productivity (Grove and Wu, 2007). Zhang (2005) showed that social science, business, and art undergraduate majors receive higher salaries, whereas math, biological, and other science majors receive relatively low pay. In this study, we investigate whether the choice of undergraduate major also affects research productivity in a professor's prolonged research career after graduation.

Since the numbers of control variables is quite large in our new dataset, applying

the double/debiased machine learning (DML) methodology is needed to consistently estimate causal/treatment effects. Chernozhukov et al. (2018) and Chernozhukov et al. (2017) proved that by using Neyman-orthogonal scores and sample splitting, DML estimates are consistent for the true causal parameters under high-dimensional control space.

2.3 Background and Data

In this study, we use a novel dataset that aggregates salary data from top public research universities in the United States. The focus is on young economics faculty members that are matched to personal educational and demographic characteristics and a detailed account of research productivity. The data includes young faculty members in the 2008–09 academic year that were assistant professors or still graduate students to ensure that our target sample is a full list of the young economics faculty members.[2] Thus, anyone who was an associate or full professor in the initial year was excluded. Then, we tracked the full list of young faculty members forward to the 2014–2015 academic year as long as they were still employed in the system of public universities.[3]

For each faculty member, information on salary, educational background, demographic characteristics, work experience, and research productivity was collected. The available salary data for the full list of young faculty members each year is from the

[2]We only focus on the tenure-track faculty members. All the non-tenure-track, including "teaching faculty" or "instructors," emeritus, adjunct, and visiting faculty are excluded.

[3]For the job-hopper, if she transferred to a public university, we still track her performance until 2014–2015 academic year. Otherwise, if she entered to a private school or the industry, she was out of the sample.

state's web page. The time unit of analysis is one year.[4] Also, all salaries are in 2008 dollars adjusted by the Consumer Price Index (CPI).

Individual educational and demographic characteristics were obtained by the most recent CV of faculty members. We extensively describe these characteristics based on their undergraduate and graduate education. The main measures for undergraduate educational background are school rank and major. We consider the contribution of undergraduate education to the foundation of economic sense and math. Thus, we create two dummy variables as measures of major: one is whether the major is economics and the other is whether the major is STEM. For the measure of graduate educational background, not only is school rank and major/concentration included but also total years of graduate education; the total years calculation includes time in masters degree programs, time pursuing degrees in ECON/non-ECON fields, and the total years spent pursuing a PhD.

In addition, we used information on work experience, including the rank of the department that the young economics professor works at, the number of years the position of assistant professor has been held, the number of years until they were promoted to associate/full professor, the number of years faculty members worked at the current school, and the year they're on a visit to other schools.

The number and quality of journal publications is the main measure of research productivity in the economics profession. How to value journal publications is a question of interest within academia in economics. Faculty publication record is an important criterion in determining promotions. To evaluate research productivity, we collect

[4]For some schools, the salary data was reported per academic year. For others, the salary data was reported per calendar year. No matter what cases are, the length of time for each observation was adjusted to one year.

journal publication information from faculty CVs and the Google Scholar database.[5] To measure the worth of a journal publication, we included journal rank, number of pages, number of co-authors, and number of citations. In calculating the publication value of each faculty member per year, the trade-off between quality and quantity is another difficulty we encountered. To measure the quality of publications, we adopted the journal rankings from IDEAS 2015[6] and divide the rankings into five tiers. Publishing in a top-ranked journal is quite different from a lower ranked journal. Thus, we set corresponding variables to value publications for different tiers separately. To consider the effects of both quality and quantity of publications, a synthetic variable called "publication value" is created by number of publication pages times $log((1 + r(max))/Journal\ rank) * log(2 + number\ of\ citation)/(1 + number\ of\ coauthors)$, where $r(max)$ is the max rank of all journals. We also consider cumulative publications, average productivity of publications, and best-valued publication to capture the research productivity of a faculty member. "Average productivity of publications" is calculated by $Cumulative\ publication\ value/(Year - PhD\ graduation\ year + 1)$. Best-valued publication measures the best performing research, defined as the best ranking among cumulative publications for each faculty member.

Table 2.1 lists the 28 public universities that faculty members worked in, from 17 different states. These universities are diverse in terms of school quality. We planed to collect faculty information from top 50 public universities in the U.S., because of the limitation of data availability such as missing values in variable, we end up having information from 28 of them. The criteria used to measure school quality is based

[5] We only take journal articles into account, not "working papers" (or "discussion papers," etc.), not book chapters (unless it's NBER something).

[6] https://ideas.repec.org/top/top.journals.all.html

Table 2.1: Public Universities Faculty Members Worked in

University of Arizona	University of Illinois at Chicago
University of California, Berkeley	University of Illinois at Urbana–Champaign
University of California, Davis	Purdue University
University of California, Irvine	University of Maryland
University of California, Los Angeles	University of Michigan
University of California, Riverside	Michigan State University
University of California, Santa Barbara	University of Minnesota
University of California, Santa Cruz	University of Missouri
University of California, San Diego	Rutgers University–New Brunswick
University of Florida	Ohio State University
Florida State University	University of Texas at Austin
Georgia State University	University of Virginia
George Washington University	University of Washington
Iowa State University	University of Wisconsin–Madison

on the ranking from IDEAS during 2011–2014. Around 93% of universities ranked in the top 200.[7] The sample includes 363 faculty members from the departments of economics.[8] All of the faculty members in the sample are employed by top public universities in US.[9] The dataset includes information on all of the young faculty from the year 2008[10] or the year of starting to be an assistant professor.[11] Also, we tracked them forward to 2014, unless they left the system of public universities. We focus on professors with experience amounting to 10 years or less, and we consider all observations for the same year of experience as one "cohort." The statistics for each cohort are summarized in Table 2.2.

Table 2.3 shows information collected in our dataset. We collected the data by individual, and keep track of them within the year range of 2008 to 2014 as long as

[7]Among these universities, 21.4% universities ranked at top 30, 25% universities ranked at top 30–50, 25% universities ranked at top 50–100, 21.4% universities ranked at top 100–200.

[8]Considering the gap among different fields, this study examines the faculty members from department of economics, excluding the departments that are in business schools.

[9]Note that our data is collected from a public website and different states may have difference criterion in reporting salaries.

[10]For the faculty members if in academic year 2008–2009, they were assistant professors

[11]For the faculty members if in academic year 2008–2009, they were still graduate students

Table 2.2: Summary of Statistics, by Experience

Experience	Number of Faculty	Male	Graduate School Rank	Undergraduate Econ
0	99	0.74	36.40	0.71
1	118	0.72	39.94	0.69
2	120	0.72	45.83	0.70
3	125	0.72	42.10	0.70
4	135	0.72	45.10	0.70
5	135	0.76	50.84	0.70
6	134	0.73	51.53	0.69
7	111	0.74	51.86	0.64
8	80	0.71	59.65	0.64
9	70	0.76	54.95	0.67
10	51	0.76	50.41	0.76

they are working in the population pool shown in Table 2.1.

2.4 Methodology

We will follow Chernozhukov et al. (2018), use their DML model as our main methodology. As shown above, we have high-dimensional control variables included in our datasets and a relatively small sample size for each working experience year. In order to avoid overfitting, multicollinearity, and other model problems, before the data is put into the model, principal component analysis (PCA) is conducted on the control variables. PCA can help reduce the dimension and mitigate multicollinearity in the explanatory variables. Next, DML is applied to the model

$$Y = D\theta_0 + g_0(\mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0, \tag{2.4.1}$$

$$D = m_0(\mathbf{X}) + V, \quad \mathbb{E}[V \mid \mathbf{X}] = 0. \tag{2.4.2}$$

Table 2.3: Variables in the dataset

Dummy Variable (=1 if Yes)	***Continuous variable***	***Continuous variable***
Visiting	PhD Graduation School Score(USnews_2013)	Current number of publication in tier 3
South school	School rank in current year	Current number of publication in tier 4
Graduate major in economics	Speed measure of promoting to associate professor	Current number of publication in tier 5
Undergraduate major in economics	Average productivity of publication	Cumulative publication value
Undergraduate in STEM	Best rank of publication in current year	Cumulative publication value in tier 1
Male	Best rank of publication till current year	Cumulative publication value in tier 2
Associate Professor	Current publication value	Cumulative publication value in tier 3
Full Professor	Current publication value in tier 1	Cumulative publication value in tier 4
Associate and Full Professor	Current publication value in tier 2	Cumulative publication value in tier 5
Salary total	Current publication value in tier 3	Cumulative number of total publication paper pages
Salary base	Current publication value in tier 4	Cumulative number of publication paper pages in tier 1
	Current publication value in tier 5	Cumulative number of publication paper pages in tier 2
	Current number of total publication papers pages	Cumulative number of publication paper pages in tier 3
Factor Variable	Current number of publication papers pages in tier 1	Cumulative number of publication paper pages in tier 4
Year	Current number of publication papers pages in tier 2	Cumulative number of publication paper pages in tier 5
PhD graduation year	Current number of publication papers pages in tier 3	Cumulative number of total citation
Assistant Professor starting year	Current number of publication papers pages in tier 4	Cumulative number of citation in tier 1
Associate Professor starting year	Current number of publication papers pages in tier 5	Cumulative number of citation in tier 2
Full Professor starting year	Current number of total citation	Cumulative number of citation in tier 3
School code	Current number of citation in tier 1	Cumulative number of citation in tier 4
	Current number of citation in tier 2	Cumulative number of citation in tier 5
	Current number of citation in tier 3	Cumulative number of total coauthors
Continuous variable	Current number of citation in tier 4	Cumulative number of coauthors in tier 1
Experience year	Current number of citation in tier 5	Cumulative number of coauthors in tier 2
Age	Current number of total coauthors	Cumulative number of coauthors in tier 3
Salary	Current number of coauthors in tier 1	Cumulative number of coauthors in tier 4
CPI adjusted salary (base=2008)	Current number of coauthors in tier 2	Cumulative number of coauthors in tier 5
Seniority (total)	Current number of coauthors in tier 3	Cumulative number of total publication
Seniority (continuous)	Current number of coauthors in tier 4	Cumulative number of publication in tier 1
Total years of graduate education	Current number of coauthors in tier 5	Cumulative number of publication in tier 2
Total Years of PhD education	Current number of total publication	Cumulative number of publication in tier 3
PhD Graduation School Rank(IDEAS_2013)	Current number of publication in tier 1	Cumulative number of publication in tier 4
PhD Graduation School Rank(USnews_2013)	Current number of publication in tier 2	Cumulative number of publication in tier 5

Note: "Associate Professor starting year" and "Full Professor starting year" has been set at 3000 if they were not an associate\full professor as of CV date. "Seniority (total)" is calculated by current year minus the year they first started at the current school. "Seniority (continuous)" is considered as years of the most recent continuous period. "Total years of graduate education" includes masters degrees, degrees in non-ECON fields, etc. "Speed measure of promoting to associate professor" is calculated by $1/(\textit{Associate Professor starting year} - \textit{Assistant Professor starting year} + 1)$ if $\textit{Associate professor} == 1$. "Average productivity of publication" is calculated by $\textit{Cumulative publication value}/(\textit{Year} - \textit{PhD graduation year} + 1)$. "Current publication value" is publication value for each person in current year: base on the value of each publication. Publication value is calculated by number of publication pages times $log((1+r(max))/\textit{Journal rank}) * log(2+\textit{number of citation})/(1+\textit{number of coauthors})$, where $r(max)$ is max of journal rank. Tiered variables are continuous and partitioned by journal rank. Tier 1 include journals ranked from 0 to 10; tier 2 include journals ranked from 11 to 50; tier 3 include journals ranked from 51 to 150; tier 4 include journals ranked from 151 to 300; tier 5 include journals ranked 301 and after. We take IDEAS 2015 journal rank here as criterion.

Equation (2.4.1) and (2.4.2)[12] are partial linear conditional expectation functions (CEF). The variable of interest (gender, school rank, etc.) is linearly involved, denoted by D, and all of the other characteristic variables lumped in a nonparametric function $g_0(\mathbf{X})$. Characteristic variables are allowed to be correlated with treatment variables, as is summarized by (2.4.2), another CEF. We call parameters in $g_0(\mathbf{X})$ and $m_0(\mathbf{X})$ nuisance parameters, since we do not care much about them. The nuisance parameter space is allowed to be increasing with sample size and could be infinite in dimension. We neither know nor specify function forms for g_0 or m_0. They could be nonlinear and\or complicated. $D\theta_0$ in (2.4.1) is a linear restriction, assuming that the effect is homogeneous to every case in the dataset. It can be used for both binary and continuous treatment variables.

More generally, let treatment D in (2.4.1) go to the nonparametric function[13] together with $\mathbf{X}$,

$$Y = g_0(D, \mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0 \tag{2.4.3}$$

allowing full interactions between treatment and controls so that heterogeneous effects can be obtained from this model. According to Chernozhukov et al. (2018) a partial linear model can be applied to both dummy treatment variables and continuous variables; a general nonparametric model can only be applied to the dummy treatment variable. So both models are used in the analysis of gender and undergraduate economic major effects; only the partial linear model is used in the analysis of graduate school rank effects because the rank variable is continuous.

Because it is hard to know what the true CEFs are, the flexibility of $g_0(\mathbf{X})$ and

[12]Equation (4.1) and (4.2) are showed as (1.1) and (1.2) in Chernozhukov et al., 2018.
[13]Shown as Chernozhukov et al. (2018) Equation (5.1)

$m_0(\mathbf{X})$ may help us get closer to the true CEFs as much as possible. However, since machine learning techniques are used here in the estimation of $g_0(\mathbf{X})$ and $m_0(\mathbf{X})$, another important point worth noting here is that we care about the unconfoundedness or conditional independence assumption (CIA)[14]. We are interested in causal meanings so we need CIA to hold. Only when CIA holds, the DML estimate of θ_0 has causal meaning.

The essential idea of DML is (1) to build a Neyman Orthogonal score[15]; the score is a function that needs to satisfy not only a moment condition but also an orthogonality condition to overcome the regularization bias; and (2) to remove the bias caused by overfitting by doing sample splitting.

The story starts with a general moment condition[16],

$$\mathbb{E}(\psi(D, \mathbf{X}; \theta_0, \eta_0)) = 0 \tag{2.4.4}$$

where ψ is a vector of score functions; it could be in any form: a maximum likelihood score function, a GMM moment function, and so on; η_0 denote the true value of nuisance parameters included in g_0 and m_0, $\eta_0 \in \tau$ where τ is the nuisance parameter space. The score function must satisfy an additional condition that its Gateaux derivative $D_r[\eta - \eta_0]$ exists, and it is non-sensitive to the change of nuisance parameters

[14]Unconfoundedness means conditioning on $\mathbf{X}$, the counterfactuals $Y(0)$ and $Y(1)$ are uncorrelated with treatment D. This definition is first articulated by Rosenbaum and Rubin (1983). CIA means conditioning on $\mathbf{X}$ the choice of D is statistically independent with U, the model error term, as defined by Angrist and Pischke (2009). According to Wooldridge (2010), they are similar concepts, so we treat them as the same in this paper.

[15]Introduced by Neyman (1959)

[16]Chernozhukov et al. (2018) equation (2.9)

η towards any direction. The Gateaux derivative[17] is

$$D_r[\eta - \eta_0] := \partial_r\{\mathbb{E}[\psi(D, \mathbf{X}; \theta_0, \eta_0 + r(\eta - \eta_0)]\}, \quad \eta \in \tau$$

where $r \in [0, 1)$. $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \tau$ and at $r = 0$, then the orthogonality condition[18] is

$$\partial_\eta \, \mathbb{E} \, \psi(D, \mathbf{X}; \theta_0, \eta_0)[\eta - \eta_0] = 0 \tag{2.4.5}$$

The two conditions (2.4.4) and (2.4.5) together make the DML method different from others[19] and be able to consistently estimate causal effects.

It is necessary to make sure observations from $(D, \mathbf{X})$ are i.i.d.(independent, identically distributed). Also, sample splitting plays an important role. Before any estimation, we should divide the sample into K folds randomly, such that each subsample I_k contains N/K number of observations, where $k \in 1, ..., K$. For each subsample I_k, we construct an ML estimator of $\hat{\eta}_{0k}$ and the final DML estimator of interest is solved by

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}_{nk}[\psi(D, \mathbf{X}; \tilde{\theta}_0, \hat{\eta}_{0k}] = 0$$

Chernozhukov et al. (2018) equation (3.4). When the kth fold as main subsample is used to obtain a DML estimator $\check{\theta}_{0k}$, all of the rest of the subsamples (except the kth fold) becomes the auxiliary sample corresponding to the main one. So for a K fold splitting, we do have K different main samples and K different auxiliary samples. For each time of estimation, we use the kth subsample I_k as the main sample to estimate

[17] This defined as in Chernozhukov et al. (2018) the equation below (2.1).
[18] Chernozhukov et al. (2018) equation (2.3)
[19] Traditional methods like OLS only need the one moment condition (2.4.4).

θ_0 and its corresponding auxiliary sample to estimate m_0 and g_0. Then, we take the average over K estimates. In this partial linear model, a rough estimation procedure would be: (1) split the sample into several folds, treat one fold as a main sample and the rest as auxiliary sample; (2) use Machine Learning (ML) to predict D given $\mathbf{X}$ and estimate $g_0(\mathbf{X})$ using auxiliary sample; (3) estimate the parameter of interest θ_0 using the main sample. (4) do this procedure K times and take the average of K estimates, which will be the DML estimator.

2.5 Empirical Strategy

The empirical study focuses on the effects on salary associated with three interesting factors: gender, a binary variable as "male"[20]; young economics professors' graduation school rank, a continuous variable as "PhD graduation school rank"[21]; and their undergraduate Major, binary variables as "undergraduate major in economics" and "undergraduate major in STEM"[22]. We investigated how those factors causally affect salary. We denote salary as Y, its potential outcomes as (Y_0, Y_1) when the treatment factor is a binary variable. Also, we denote the treatment\causal variable as D and the control variables as $\mathbf{X}$. Once a factor was analyzed (as treatment variable), we treated all the other variables in Table 2.3 as controls. For both the binary and continuous treatment\causal variables, a partial linear structural model

$$Y = D\theta_0 + g_0(\mathbf{X}) + U \tag{2.5.1}$$

[20]A variable name in Table 2.3.
[21]A variable name in Table 2.3.
[22]Variable names in Table 2.3.

can be applied. Apart from (2.5.1), a more general structural model

$$Y = g_0(D, \mathbf{X}) + U \tag{2.5.2}$$

can be used when D is a binary variable, where g_0 denotes an unknown\nonparametric function. Model (2.5.2) is more general than (2.5.1) since it allows treatment variables to interact with all other variables. When CIA[23] and Overlap Assumption[24] are satisfied, the average causal effect of a continuous variable is $ACE \equiv \theta_0$; the average treatment effect of a binary variable is $ATE \equiv \mathbb{E}[Y_1 - Y_0]$.[25] Also, according to Wooldridge (2010), when the set of control variables is richer, CIA has more chance to hold. Our dataset provides multidimensional productivity measures on publication information and other educational and demographic information in up to 80 categories, so we believe CIA could hold in our dataset. Still, we know we cannot guarantee the variables in this dataset are plenty enough to represent or be able to proxy any unobserved information that affects the determination on salary.[26]

Gender We investigated the salary differences between young male and young female economics professors holding personal, educational, and productivity information unchanged. The gender difference in salaries has been proven to be affected by many factors. We assume that as long as we control for enough productivity and background information, the ATE of gender is identified by estimating both CEF

[23] Wooldridge (2010) Assumption ATE.1 and a weaker version ATE.1'

[24] Wooldridge (2010) Assumption ATE.2

[25] Wooldridge (2010) Proposition 21.1

[26] According to Wooldridge (2010), section 21.3, unobservables are allowed to be correlated with D only when the unobservables are not correlated with Y_0 and Y_1.

models (2.4.1) and (2.4.3) consistently. We assume

$$(Y_1, Y_0) \perp\!\!\!\perp Gender \mid \mathbf{X}$$

here $\mathbf{X}$ includes all the other information except "Male" and "Salary" in Table 2.3. However, as mentioned before, it is reasonable to doubt that the CIA holds in our dataset. For example, we do not have teaching information (teaching load, teaching ability, etc.) in our dataset. Teaching time has been proven to negatively affect time faculty spent on research, inhibiting publications per faculty (Graves, Marchand, and Thompson, 1982) and leading to lower salaries (Perna, 2001). People tend to believe female professors have higher teaching ability and are therefore paid more (Siegfried and White, 1978), which biases the salary gender gap to be smaller than it actually is. Even though it is preferable to have additional teaching information under our control, it is not a huge concern for us because we are focusing on public research universities, and we presume those schools assign similar teaching loads to young economics professors regardless of gender. Also, we do not have information on reputation included in our dataset, which as stated early, has an influential effect on salary as well. That also does not present a big problem because we are looking at young economics professors, most of whom have not yet built up solid reputation in the field. Nevertheless, we believe our dataset contains valuable research and productivity information that is richer than that found in other studies to date. This in-depth dataset is essential to examine our research problem.

Graduation School Rank Researchers may ask whether PhD graduation school rank has a causal effect on young economics professors' salaries and how the effect evolves with more working experience gained. Specifically, when estimating the grad-

uate school rank effect, we use "PhD Graduation School Rank (IDEAS_2013)" exclusively and omit "PhD Graduation School Rank (USnews_2013)" and "PhD graduation School Score(USnews_2013)" from the dataset and put other background and productivity measures in the control set $\mathbf{X}$. "PhD graduation school rank" is a continuous variable, so only a partial linear structural model (2.5.1) is estimable by DML. For the purpose of being comparable with partial linear models, we create a binary variable for our structural model called "Rank Dummy," taking on a value of 1 if the school is ranked in the top half of the ranking list and 0 if ranked in the second half of ranking list. Again, we argue that the CIA

$$\textit{Graduate School Rank} \perp\!\!\!\perp U \mid \mathbf{X}$$

is satisfied in our dataset because the rank of the graduation school is mainly related to the research and teaching ability of the PhD receiving school, which may influence the young professor's research ability, and we have controlled for the young faculties research ability.

Undergraduate Major Undergraduate education is an important factor for a person's career path so we want to find out how professors' salaries are affected by their undergraduate major. Undergraduate major may influence salary through many paths, indirectly through graduate school rank, indirectly through human capital productivity, etc. As we control this information in our model, we are able to check on causality between undergraduate majors and salary. We investigate the impact of economics as well as STEM undergraduate majors on salary. We examine how these majors affect salary differently and whether they interact with each other.[27]

[27]the two variables are not exclusive from each other.

We undertake this analysis to find out if an economics faculty's research ability and productivity is related to a science education background. Employers might have the impression that people who have better math, statistics and\or other science background are productive in their later on research life even if they do not have a difference in actually paper productivity with others (Bonzi, 1992). In order to test this, when estimating the average treatment effect (ATE) of undergraduates being economics majors on salary we put everything else in the control set $\mathbf{X}$; and when estimating the effect of being an undergraduate STEM major we switch the position of "Undergraduate major in economics" and "Undergraduate major in STEM." For each structural model (2.5.1) and (2.5.2), we report both "Economics" results and "STEM" results. Because we controlled for "STEM" in $\mathbf{X}$, we believe that CIA

$$(Y_1, Y_0) \perp\!\!\!\perp \textit{Undergraduate major in economics} \mid \mathbf{X} \tag{2.5.3}$$

holds and when controlled for "Undergraduate major in economics" in $\mathbf{X}$

$$(Y_1, Y_0) \perp\!\!\!\perp \textit{Undergraduate in STEM} \mid \mathbf{X} \tag{2.5.4}$$

holds. Actually, (2.5.3) and (2.5.4) seem easier to be justified than the previous two. Undergraduate major choice could be correlated with such factors as years of education and research ability. Thus, undergraduate major may affect salary through many paths. Once we controlled for demographic and research information, we considered the choice of undergraduate major as independent with the potential outcomes.

We believe that within each experience year the observations are independent and identically distributed (iid). That is because individuals having same work experience

year of a sub-sample are collected from year 2008/2009 to 2014/2015 who work at research public universities in Table 2.1. We separately estimate causal effects with respect to different experience years. For each model, we are using cross sectional data with iid observations. This iid assumption rules out the situations where the treatment on one individual affects others' outcomes (Wooldridge, 2010). Again, it is never a bad thing to be critical. Because here iid is a very strong assumption,[28] people may doubt its validity. For example, for newly recruited young economics professors within one department, their salaries might be close to one other. Also, when promotion and\or salary increase decisions are made, fundings levels between departments might be different and for those with limited funding, they might have only some qualified (but not all) professors' salaries increased. This means that young faculty's salaries from the same department may correlates with each other. But at least we can assume iid holds between departments. Since we do not have too many individuals coming from one department in one year, that is not a huge concern. However, if many people are drawn from same department, then we realized the potential problem it could cause. Another reason we separately estimated the effects by experience year is that people can come and leave the population pool of research public universities; also, we were concerned about average effect on people who actually worked at those research public universities instead of focusing on a fixed group of people, so we do not specifically track anyone's career path. We were not concerned with who was in the pool or for how long they were in the pool but we wanted to make sure that within each experience year pool we had iid samples.

[28]In Wooldridge (2010) chapter 21, the author makes a stronger assumption of iid sample than what is needed Stable Unit Treatment Value Assumption (SUTVA) in treatment literature. Because random sampling implies SUTVA.

Table 2.4: Means and Standard Deviations of Salaries

		Exp0	Exp1	Exp2	Exp3	Exp4
Mean		112,987	119,930	116,417	110,126	110,911
SD		16,014	22,372	18,951	15,460	19,916
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
Mean	111,102	118,783	127,603	133,948	138,424	145,388
SD	23,630	36,333	46,649	54,620	51,163	54,557

Note: Means are in year 2008 dollar value.

Because we separated the data by different experience years, each year's data contains a relatively large number of variables (p) compared to the number of observations (n).[29] Also, considering the nature of the DML sample splitting procedure[30], we took one more step to carry out principle component analysis on the control variables each time before DML estimation and treated all principal components as control variables in the model. This procedure helps to reduce multicollinearity without losing any information. In DML, we used this PCA pre-processed data in the estimation accompanied by machine learning methods random forest, neural network, and decision tree to build a DML+PCA model, which you can find in the report tables. Also, we used the original data in DML estimation accompanied by Post Lasso[31], which is a method more robust to $p > n$ cases as another set of models.

Table 2.5: Gender effects on log salary (DML+PCA)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		0.010	0.045	0.027	0.011	0.108
se(median)		(0.084)	(0.244)	(0.071)	(0.048)	(0.079)
se		(0.072)	(0.207)	(0.071)	(0.038)	(0.067)
B. Partial Linear Model						
ATE		0.035	0.018	0.048	0.037	0.091
se(median)		(0.023)	(0.037)	(0.026)	(0.022)	(0.03)
se		(0.022)	(0.034)	(0.026)	(0.022)	(0.03)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.046	0.068	0.161	0.093	0.085	0.005
se(median)	(0.060)	(0.094)	(0.074)	(0.127)	(0.162)	(0.151)
se	(0.057)	(0.075)	(0.064)	(0.113)	(0.162)	(0.124)
B. Partial Linear Model						
ATE	0.111	0.080	0.161	0.066	0.094	-0.063
se(median)	(0.034)	(0.041)	(0.061)	(0.077)	(0.091)	(0.136)
se	(0.032)	(0.038)	(0.054)	(0.075)	(0.087)	(0.096)

Note: "Exp" is short for experience years. "ATE" reports "best" median treatment effect estimations across splits, here "best" is among Trees, Random Forest and Neural network methods. How "best" are calculated is refered to Chernozhukov et al. (2018). "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. "Splits" means how many times we randomly separate the data into different (pre-setted) folds.

2.6 Results

Table 2.4 provides the average and standard deviation of the CPI-adjusted salaries. The average salary is gradually increasing by experience year. Tables 2.5 to 2.8 show DML with PCA pre-processed datasets estimation results. Tables 2.9 to 2.12 show DML along with the post Lasso method using raw data results. Specifically, in DML we set fold=2, which means treating half of the sample as the auxiliary sample to estimate $g(.)$ and $m(.)$ and treating the other half as the main sample to estimate the treatment\causal effect. Then, we swapped the main and auxiliary sample and

[29]More than 80 variables; approximately 100 observations for each experience year's data set.

[30]In DML, only a partial of the data are used for causal parameter estimation after sample splitting.

[31]Post Lasso is a method originated by Belloni and Chernozhukov (2013). Lasso is not applicable at the same time with PCA, because after PCA the principle components are already orthogonal with each other.

Table 2.6: Graduate school rank effects on log salary (DML+PCA)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	-0.166	-0.048	0.018	0.004	0.053
se(median)	(0.170)	(0.148)	(0.087)	(0.046)	(0.095)
se	(0.170)	(0.148)	(0.085)	(0.045)	(0.087)
B. Partial Linear Model					
ATE	0.070	0.084	0.043	0.039	0.165
se(median)	(0.057)	(0.088)	(0.064)	(0.049)	(0.083)
se	(0.051)	(0.074)	(0.059)	(0.048)	(0.076)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.023	0.043	0.151	0.118	0.141	0.092
se(median)	(0.074)	(0.079)	(0.080)	(0.082)	(0.112)	(0.125)
se	(0.074)	(0.064)	(0.072)	(0.078)	(0.105)	(0.125)
B. Partial Linear Model						
ATE	0.132	0.210	0.294	0.041	0.19	0.289
se(median)	(0.099)	(0.131)	(0.187)	(0.187)	(0.147)	(0.211)
se	(0.092)	(0.128)	(0.173)	(0.187)	(0.134)	(0.206)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Estimation on interactive model effects are obtained by creating a dummy variable which is denoted 1 when rank is less or equal to the rank median; 0 if rank is greater than the rank median. Estimation on partial linear model effects are obtained by creating a transformed rank variable, which is calculated by $1 - rank/max(rank)$. In this way, the rank range is between 0 and 1. Better ranked schools are assigned values close to 1, worse ranked schools are assigned values close to 0.

Table 2.7: Undergraduate ECON effects on log salary(DML+PCA)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		-0.023	-0.039	-0.06	-0.049	-0.030
se(median)		(0.095)	(0.114)	(0.081)	(0.086)	(0.11)
se		(0.092)	(0.107)	(0.081)	(0.067)	(0.080)
B. Partially Linear Model						
ATE		0.004	-0.023	-0.057	-0.036	-0.030
se(median)		(0.035)	(0.037)	(0.035)	(0.031)	(0.040)
se		(0.030)	(0.035)	(0.034)	(0.031)	(0.036)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	-0.058	-0.024	-0.117	-0.17	-0.106	-0.159
se(median)	(0.103)	(0.077)	(0.087)	(0.100)	(0.138)	(0.266)
se	(0.085)	(0.062)	(0.082)	(0.091)	(0.135)	(0.246)
B. Partially Linear Model						
ATE	-0.027	-0.021	-0.127	-0.129	-0.137	-0.098
se(median)	(0.038)	(0.044)	(0.058)	(0.099)	(0.092)	(0.133)
se	(0.038)	(0.043)	(0.055)	(0.087)	(0.091)	(0.121)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp8", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

Table 2.8: Undergraduate STEM effects on log salary (DML+PCA)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		0.106	-0.062	-0.042	0.014	0.124
se(median)		(0.175)	(0.247)	(0.077)	(0.048)	(0.117)
se		(0.131)	(0.130)	(0.074)	(0.048)	(0.099)
B. Partially Linear Model						
ATE		-0.020	-0.015	0.004	0.031	0.098
se(median)		(0.028)	(0.044)	(0.032)	(0.029)	(0.033)
se		(0.027)	(0.035)	(0.032)	(0.028)	(0.033)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.091	0.138	0.142	0.096	0.101	0.177
se(median)	(0.131)	(0.153)	(0.113)	(0.168)	(0.197)	(0.172)
se	(0.105)	(0.114)	(0.113)	(0.162)	(0.178)	(0.157)
B. Partially Linear Model						
ATE	0.054	0.097	0.089	0.080	0.116	0.119
se(median)	(0.040)	(0.053)	(0.063)	(0.101)	(0.109)	(0.101)
se	(0.039)	(0.048)	(0.058)	(0.098)	(0.106)	(0.093)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp6", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

Table 2.9: Gender effects on log salary (DML with PostLasso)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE(PostLasso)		0.038	0.022	0.006	0.040	0.071
se(median)		(0.042)	(0.124)	(0.255)	(0.025)	(0.076)
se		(0.032)	(0.104)	(0.255)	(0.025)	(0.066)
B. Partial Linear Model						
ATE(PostLasso)		0.029	0.009	0.045	0.024	0.060
se(median)		(0.024)	(0.034)	(0.028)	(0.026)	(0.030)
se		(0.022)	(0.031)	(0.026)	(0.020)	(0.028)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.117	0.111	0.047	0.168	0.112	0.353
se(median)	(0.117)	(0.080)	(0.262)	(0.167)	(0.260)	(14.50)
se	(0.099)	(0.072)	(0.230)	(0.115)	(0.178)	(7.663)
B. Partial Linear Model						
ATE	0.064	0.084	0.041	0.074	-0.017	-0.035
se(median)	(0.048)	(0.037)	(0.062)	(0.088)	(0.072)	(0.099)
se	(0.027)	(0.036)	(0.049)	(0.059)	(0.070)	(0.097)

Note: "Exp" is short for experience years. "ATE" reports median treatment effect estimations across splits. "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. "Splits" means how many times we randomly separate the data into different (pre-setted) folds.

Table 2.10: Graduate school rank effects on log salary (DML with PostLasso)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		0.022	0.011	-0.029	-0.032	-0.046
se(median)		(0.042)	(0.049)	(0.149)	(0.083)	(0.124)
se		(0.037)	(0.048)	(0.122)	(0.064)	(0.123)
B. Partial Linear Model						
ATE		0.004	0.016	0.002	0.004	0.027
se(median)		(0.052)	(0.110)	(0.090)	(0.065)	(0.100)
se		(0.048)	(0.083)	(0.049)	(0.051)	(0.076)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	-0.035	0.048	-0.190	-0.939	-0.220	-0.033
se(median)	(0.118)	(0.178)	(0.292)	(1.034)	(0.308)	(0.483)
se	(0.085)	(0.134)	(0.292)	(0.811)	(0.307)	(0.434)
B. Partial Linear Model						
ATE	0.069	0.056	0.087	-0.109	0.009	0.134
se(median)	(0.087)	(0.083)	(0.174)	(0.195)	(0.136)	(0.204)
se	(0.079)	(0.08)	(0.156)	(0.131)	(0.136)	(0.163)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Estimation on interactive model effects are obtained by creating a dummy variable which is denoted 1 when rank is less or equal to the rank median; 0 if rank is greater than the rank median. Estimation on partial linear model effects are obtained by creating a transformed rank variable, which is calculated by $1 - rank/max(rank)$. In this way, the rank range is between 0 and 1. Better ranked schools are assigned values close to 1, worse ranked schools are assigned values close to 0.

Table 2.11: Undergraduate ECON effects on log salary(DML with PostLasso)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	-0.053	-0.086	-0.052	0.013	-0.006
se(median)	(0.065)	(0.075)	(0.084)	(0.054)	(0.055)
se	(0.052)	(0.056)	(0.076)	(0.043)	(0.052)
B. Partially Linear Model					
ATE	0.024	0.011	-0.05	-0.005	0.034
se(median)	(0.033)	(0.037)	(0.033)	(0.028)	(0.042)
se	(0.028)	(0.035)	(0.029)	(0.027)	(0.031)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.011	0.038	-0.031	-0.45	-0.112	-0.797
se(median)	(0.094)	(0.267)	(0.273)	(0.398)	(0.655)	(2.567)
se	(0.094)	(0.218)	(0.27)	(0.314)	(0.358)	(1.316)
B. Partially Linear Model						
ATE	0.019	0.015	-0.021	-0.135	-0.085	-0.129
se(median)	(0.035)	(0.053)	(0.061)	(0.068)	(0.076)	(0.112)
se	(0.035)	(0.040)	(0.046)	(0.067)	(0.069)	(0.112)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp8", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

Table 2.12: Undergraduate STEM effects on log salary (DML with PostLasso)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	0.009	-0.021	0.009	0.016	0.098
se(median)	(0.083)	(0.162)	(0.057)	(0.052)	(0.228)
se	(0.065)	(0.093)	(0.048)	(0.046)	(0.135)
B. Partially Linear Model					
ATE	0.001	-0.018	-0.008	0.005	0.044
se(median)	(0.026)	(0.034)	(0.03)	(0.03)	(0.03)
se	(0.025)	(0.034)	(0.028)	(0.027)	(0.030)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.093	0.008	-0.085	0.218	0.308	0.564
se(median)	(0.268)	(0.122)	(0.191)	(1.030)	(0.935)	(1.304)
se	(0.114)	(0.093)	(0.137)	(0.951)	(0.569)	(0.857)
B. Partially Linear Model						
ATE	0.031	0.057	-0.011	0.080	-0.101	0.030
se(median)	(0.039)	(0.044)	(0.069)	(0.076)	(0.094)	(0.108)
se	(0.032)	(0.041)	(0.064)	(0.075)	(0.083)	(0.088)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp6", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

repeated this procedure and calculated the average of the two (i.e., one time of DML estimation). In order to make the estimator more robust to outliers, by setting split=5, we ran the DML estimation procedure five times and take the median. We follow Chernozhukov et al. (2018) result reporting style and take the "best"[32] column as our reported ATE estimators in Tables 2.5 to 2.8. We also include a more robust version of standard errors, defined as median methods adjusted standard error[33]. The coefficients are point estimates of causal effects on the log CPI-adjusted salary. Also, it is worth noting that all of the the effects are based on people in the population pool, which is young economics professors who work at the top 50 research university economics departments that are on tenure track.[34]

Gender The gender effect estimators of both interactive and partial linear models in Table 2.5 are in a rough range of 0.01 to 0.1 across all experience years. This means male economics professors are estimated to earn 1% to 10% more on average than female economics professors during the early years of their career when holding research productivity and demographic variables constant. However, most of the gender effect estimates are neither statistically nor economically significant. When experience is less than four years, in both partial linear and interactive models the point estimates are less than 4% without statistical significance. Specifically, in experience year zero, the more general interactive model gives point estimates of 1%, which is neither statistically significant nor economically meaningful. There is not enough evidence to show that gender causally make a difference in salaries once we

[32] A result getting from choosing best methods for $g(.)$ and $m(.)$ separately.

[33] Chernozhukov et al. (2018) Definition 3.3.

[34] We are not interested in people who are not in this pool, for example people who has been offered a position from any of these universities but not taking it and people who leave out of the tenure track or move to private school after a few years working in the system.

control for research and productivity measures and other demographic variables.

The coefficients of gender effects go up and then down with most of them non statistically and economically significant. When professors' working experience goes close to near promotion dates (experience year 6 or 7), the salary difference caused by gender becomes severe. In years 6 to 9, the effects are around 10%, which is a significant amount and is higher than the previous few years, peaking in year 7. In particular, in experience year 7, the gender effects are strong and both statistically and economically significant. Also in Table 2.9, there exists a similar trend with 5 to 7 years' effects being around 10%, much greater than the other years.

Overall, the positive male and female salary gap is consistent with other gender salary gap studies in the literature such as Perna (2001), Toutkoushian and Conley (2005). However, the lack of statistical significance of most of our coefficients makes our result different and the values are also smaller. Toutkoushian and Conley (2005) reported an overall 26% salary difference between full-time male and female faculty in four-year institutions. There might be problems associated with data collection and the sample size is limited, which makes standard errors of the estimates large. But according to our study, there is no gender causal effect economically or statistically on young economics professors salary in their first few career years (before five experience years) and a slight\moderate gender gap in experience years 5 to 7, once controlling for research productivity and other demographic variables. Also, our outcome is different from the conclusion of Altonji and Pierret (2001), which asserts that by doing linear regressions on wages, coefficients of easy-to-observe variables that are highly related to productivity decline with years of experience.

The reasons for causality come from many sources. Statistical discrimination

is a highly discussed topic (Altonji and Pierret, 2001; Altonji, 2005). Even with similar productivity, employers might still have the stereotype impression that young female researchers have less potential to be productive as compared to males. In this case, females are provided less salary even though they have the same productivity record at the same point in time as their male counterparts. One explanation for the different outcome in our study is that there have been more protections for women from being discriminated against in recent years. This idea is in accord with our empirical results: in the first few experience years there is no gender discrimination between males and females. However, in experience year 5 to 7. which are also near most faculty's promotion date, the gender difference or discrimination is obvious (being either economically or statistically significant). Another source is negotiation power differences between men and women. Some research shows that women have less power to negotiate their salaries than men because decision-makers believe women do not have the incentive to move if their salary demands are not fully satisfied. It has also been suggested that women negotiate less frequently than men (Gerhart and Rynes, 1991; Claypool et al., 2017). So it seems promising to run a work council or collective bargaining for balancing the gender negotiation power (Oberfichtner, Schnabel, and Töpfer, 2020).

Moreover, female usually take on more family burdens like child care and other household duties, which make them less likely to move and search for a new higher paying job once settled into a community (Blackaby, Booth, and Frank, 2005).

The reasons mentioned above help explain the gender difference in salary in experience years 5 to 7. This result is consistent with the notion that men have more bargaining power because they have more options by being in a stronger position

to seek outside offers and having more of an incentive to move if obtained better compensation. In contrast, women may be less likely to move if they receive other offers.

PhD graduation school rank The original graduate school rank variable is continuous and only applicable to a partial linear model. In order to make the model results easier to interpret and comparable with each other, we undertook two transformations of it. First, we created a dummy variable for graduate school rank. The rank dummy is assigned 1 when the actual PhD graduation school rank is less or equal to the rank median over all individuals in that experience year; it is assigned 0 if rank is greater than the rank median. Panel A in Table 2.6 and Table 2.10 reports interactive model results estimated with the rank dummy. Second, we transfered the original rank variable to be in a range of 0 to 1 values, letting the smallest number represent the least-ranked school and the highest number represent the highest ranked school. The transferred continuous rank variable was calculated by $1 - rank/max(rankofthatyear)$. Panel B in Table 2.6 and Table 2.10 reported the partial linear model results estimated with a transferred continuous rank variable.

There are explained differences between the two models. Take the effect of experience in year 5 as an example: in the interactive model, the estimate of ATE 0.023 means a change from a worse than median ranked school to a better than median ranked school would increase the average salary of young economics professors with five years of work experience by 2.3%; in the partial linear model, the school rank increased by 10 percentile the average salary of young economics faculty with 5 years of work experience would increase by 1.32%. The partial linear model provides more insights on the quantiles, but it has been restricted to a linear effect. The interactive

model is more general in terms of the function form, but it conveys less information because it only separate all schools into two categories, one is above the median rank while the other below the median. Obviously, the highest ranked schools would have much difference from the lowest ranked one even within the same category.

Overall, the partial linear "rank effect" model gives larger estimates (in magnitude) across all experience years than those in the interactive model. In the interactive model, rank effect estimates range roughly from 0.01 to 0.15, without statistical significance. In the partial linear model, the estimates range from approximately 0.07 to 0.2, which are not statistically significant.[35] Except for the observed negative signs on the first two experience year estimates in the interactive model, most of the rank effect estimates are positive. That means graduating from a higher ranked school could help a young economics professor earn more money, even with other conditions (productivity, other background) held constant. This result is in agreement with Stock and Alston (2000), who state that there is a positive correlation between program rank and initial salaries, even after controlling for information on qualifications, such as research and teaching. The effects become larger from year 6 to 9. However, we notice that even though in years such as 6 and 9 the point estimates are large, there is not enough evidence to show those are significantly different from zero. In some years, the estimates fall on the upper bound of its 95% confidence interval but that includes zero. Therefore, it is hard to say there is PhD graduation school rank casual effect on young economics professors' salary in most of the early working experience years. However, we must be careful with rank effects in year 7, 8, and 9, even if those are not statistically significant; they may express certain practical meanings,

[35]We would prefer to focus on Table 2.6 results because Table 2.10 gives too many negative point estimates with very large standard errors which is less informative than Table 2.6.

and being economically significant, we just don't have a big enough sample to prove it .

Reasonably, employers might have discriminated against them in terms of salary based on PhD graduation rank during those years because those are the years approaching the time of promotion. The discrimination could also be associated with the fact that people who graduate from higher-ranked universities also have higher potential human capital (Claypool et al., 2017). Moreover, perhaps coming from a higher-ranked PhD program is helpful in getting positive letters for promotion and tenure, which usually corresponds to a salary increase.

Undergraduate Major Tables 2.7 and 2.8, and Tables 2.11 and 2.12 show how two related and interesting factors, undergraduate major in economics and undergraduate major in a STEM field, affect young economics professors salary. They are not exclusive to each other, because there exists cases where one person owns a dual degree with one of them economics and the other a STEM major. When estimating economics major effect we take the STEM under control and vice versa.

The economics undergraduate major negatively influences salary according to our findings. The range of the effects is approximately from -0.17 to -0.02. That is a 2% to 17% decrease in salary. Based on this data, economics as an undergraduate major may not be a good choice if you want to earn a higher salary as an economics professor. And similarly, the estimates of economics undergrad are quite small and not statistically significant in the first few experience years (0 to 6), such that there was no difference in salary caused by whether one was an undergraduate economics major in these years, once controlled for research productivity and demographic information. However, we should pay attention to the fact that in both the interactive and partial

linear models, the effect estimates become larger from experience year 7. Again, since those point estimates are not statistically significant, even with large point estimates, we cannot exclude them from zero effects.

Other than on the first few experience years (0 to 2), where the estimates take negative sign, the STEM undergraduate majors seems positively affected by salary with controls. Similar to the economics undergraduate major effects, the STEM undergraduate major effects are not statistically significant either. We cannot confidently distinguish them from zero. The point estimates take a range of approximately -0.04 to 0.14. According to our finding, STEM undergraduate majors would tend to positively affects salary compared to economics undergraduate major. However, we should be careful that we cannot exclude them from zero effect because of the absence of statistical significance.

Overall, holding an economics undergraduate degree may not be as helpful as holding a STEM major undergraduate degree to achieve a better salary, even with an otherwise similar research and demographic background. As mentioned in Section 2.2, observable economics talent is a key determinant in the PhD admissions process. Among the factors considered, quantitative ability is an important consideration because quantitative ability has been proved to be highly correlated with research productivity (Grove and Wu, 2007). Also, as mentioned earlier, Zhang (2005) showed that social science, business, and some other art undergraduate majors have negative coefficients on the probability of graduate enrollment whereas math, biological, and other science majors have positive coefficients on the probability of graduate enrollment. The findings of these other studies help to explain what we observed in the estimation.

Post Lasso Selected Variables Figures B.1 to B.9 give an idea on the variables selected by the post lasso method in "DML with PostLasso" interactive models. This part is trying to show how DML applied machine learning. In interactive models, DML estimates $g(1, \mathbf{X})$ and $g(0, \mathbf{X})$ separately so we can see there are two different series representing each. The vertical axis shows the frequency of each variable being selected over all (2 folds) $\times$ (5 splits) $= 10$ times possible for a given experience year model. Some variables are selected more than once and by both functions; some are selected only by one function with very few times. Since the reported ATE are aggregated over all folds and splits, we consider them as "selected" as long as they have ever been selected in the DML estimation process. The frequency difference might be caused by sample size, which leaves more randomness to the results. One might wonder why some variables are not selected at all. This is because the Lasso mechanism pushes hard on some of the variable coefficients to be exactly zero, causing them to be not "selected" by the model. That is not to say they are not suitable for being controlled. That is exactly where the difference between traditional methods and DML comes from. The L1 regularization terms in Lasso (Tibshirani, 1996) give exactly zero solutions on regression coefficients for the purpose of seeking minimum MSE under bias and variance trade-off. As mentioned in Section 2.4, the regularization terms in machine learning methods could cause regularization bias, even so, DML can reduce the regularization bias and give consistent causal estimators. But if you simply put everything into an OLS model, the estimation would crash because of the singularity problem, let alone finding causal estimators. We also notice that those variables selected by different models are quite similar to each other. They all select year, school code, and research productivity by journal rank tier variables as a

subset of control.

The graphs provide insights on how DML works, also provoking more thinking. The good thing is that we know which variables are important in the estimation of the $g(.)$ and $m(.)$ functions for prediction purposes; those variables convince us that it is worth controlling for detailed research productivity measures as well as background information. The negative side is that because of the randomness, we hardly see a variable selected with high frequency (more than 5/10).

Discussion Since economics departments have a hard time determining the true research ability of their young faculty members, we suspect that in determining young economics professors' salaries, employers not only rely on publication information but also on other easier-to-observe information that they may think relate to research ability. With our novel dataset, we were able to control for more detailed research productivity measures beyond the estimation of causal effects of gender, PhD graduation school rank, and undergraduate majors on salary. We questioned if there were some other reasons other than research ability that affected the determination in salary. But as said by Ginther and Hayes (2003), we cannot make a definitive conclusion that the difference only stems from discrimination or any other single source. For some of the early career year, we can not say there is an effect from all three aspects (gender, PhD school rank and undergrad major) when control for research productivity and other information; for some later year in career, the gender causal effect are strong and significant and may come from multiple reasons. Statistical discrimination based on PhD graduation school rank and undergraduate majors could have happened during later experience years, such as 6, 7 or 8. But we must be careful interpreting this because those estimates are not statistically significant in DML models. Overall, we

Table 2.13: Robustness Check: Gender Effects (DML+PCA)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	0.059	0.052	0.057	0.056	0.105
se(median)	(0.035)	(0.056)	(0.047)	(0.044)	(0.057)
se	(0.031)	(0.05)	(0.045)	(0.032)	(0.054)
B. Partial Linear Model					
ATE	0.078	0.064	0.055	0.036	0.094
se(median)	(0.027)	(0.038)	(0.032)	(0.024)	(0.033)
se	(0.026)	(0.037)	(0.032)	(0.023)	(0.032)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.11	0.117	0.089	0.203	0.061	NA
se(median)	(0.059)	(0.083)	(0.097)	(0.086)	(0.131)	(NA)
se	(0.048)	(0.073)	(0.094)	(0.086)	(0.129)	(NA)
B. Partial Linear Model						
ATE	0.092	0.109	0.126	0.161	0.066	-0.129
se(median)	(0.035)	(0.047)	(0.07)	(0.073)	(0.09)	(0.092)
se	(0.033)	(0.046)	(0.065)	(0.068)	(0.087)	(0.09)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. NA shows up in the whole column of Exp10, meaning interactive model does not work on this specific dataset and no result reported.

noticed a jump on the point estimates in almost all models from experience year 6 to 8. That could be evidence supporting that during those years (near promotion dates), young economics professors may experience discrimination based on gender, PhD background, as well as undergraduate major.

2.7 Robustness Check

In order to test the model specification in DML estimation, we do another set of models as a robustness check. They are different from the original DML models in the control variables we used. The robustness check models will do exactly as the four previous DML models in Tables 2.5 to 2.8 except we change the control variables in X. The robustness check models include next years' current and cumulative productivity

Table 2.14: Robustness Check: Graduate School Rank Effects (DML+PCA)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		0	0.076	0.008	0.014	0.036
se(median)		(0.057)	(0.072)	(0.046)	(0.041)	(0.044)
se		(0.046)	(0.056)	(0.044)	(0.041)	(0.044)
B. Partial Linear Model						
ATE		0.111	0.171	0.063	0.068	0.149
se(median)		(0.057)	(0.09)	(0.057)	(0.06)	(0.084)
se		(0.056)	(0.075)	(0.056)	(0.055)	(0.081)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.064	0.145	0.187	0.128	0.079	0.114
se(median)	(0.056)	(0.056)	(0.08)	(0.133)	(0.149)	(0.143)
se	(0.046)	(0.056)	(0.075)	(0.132)	(0.136)	(0.141)
B. Partial Linear Model						
ATE	0.113	0.26	0.216	0.143	0.338	0.035
se(median)	(0.061)	(0.133)	(0.221)	(0.147)	(0.221)	(0.279)
se	(0.059)	(0.124)	(0.208)	(0.142)	(0.206)	(0.253)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Estimation on interactive model effects are obtained by creating a dummy variable which is denoted 1 when rank is less or equal to the rank median; 0 if rank is greater than the rank median. Estimation on partial linear model effects are obtained by creating a transformed rank variable, which is calculated by $1 - rank/max(rank)$. In this way, the rank range is between 0 and 1. Better ranked schools are assigned values close to 1, worse ranked schools are assigned values close to 0.

Table 2.15: Robustness Check: Undergraduate Econ Effects (DML+PCA)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		-0.016	-0.002	-0.069	-0.034	-0.007
se(median)		(0.072)	(0.101)	(0.064)	(0.056)	(0.067)
se		(0.068)	(0.099)	(0.064)	(0.056)	(0.065)
B. Partially Linear Model						
ATE		-0.001	-0.025	-0.068	-0.048	-0.016
se(median)		(0.039)	(0.047)	(0.047)	(0.036)	(0.038)
se		(0.038)	(0.045)	(0.045)	(0.035)	(0.036)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	-0.037	-0.11	-0.159	-0.163	-0.307	NA
se(median)	(0.068)	(0.1)	(0.073)	(0.116)	(0.218)	(NA)
se	(0.062)	(0.084)	(0.072)	(0.112)	(0.215)	(NA)
B. Partially Linear Model						
ATE	-0.025	-0.066	-0.17	-0.172	-0.292	-0.148
se(median)	(0.042)	(0.054)	(0.067)	(0.109)	(0.127)	(0.222)
se	(0.041)	(0.05)	(0.065)	(0.105)	(0.119)	(0.2)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. NA shows up in the whole column of Exp10, meaning interactive model does not work on this specific dataset and no result reported.

Table 2.16: Robustness Check: Undergraduate STEM Effects (DML+PCA)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	-0.029	-0.018	0.038	0.046	0.106
se(median)	(0.063)	(0.059)	(0.058)	(0.042)	(0.066)
se	(0.06)	(0.058)	(0.055)	(0.042)	(0.057)
B. Partially Linear Model					
ATE	-0.019	-0.008	0.043	0.049	0.089
se(median)	(0.035)	(0.043)	(0.039)	(0.03)	(0.037)
se	(0.035)	(0.042)	(0.037)	(0.029)	(0.036)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.089	0.111	0.116	0.146	0.168	0.068
se(median)	(0.067)	(0.137)	(0.105)	(0.134)	(0.204)	(0.22)
se	(0.061)	(0.128)	(0.101)	(0.133)	(0.202)	(0.19)
B. Partially Linear Model						
ATE	0.053	0.08	0.104	0.138	0.174	0.01
se(median)	(0.044)	(0.062)	(0.072)	(0.112)	(0.122)	(0.198)
se	(0.042)	(0.059)	(0.068)	(0.11)	(0.118)	(0.131)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits.

measures, current year current productivity measures, and all the other demographic measures.[36] As has been mentioned previously, because economics papers usually take a while to be officially published after being accepted, people may update their resumes by a "forthcoming" paper in some year if the paper was officially published the year after. Employers can observe that changing information as time goes by, whereas our dataset is not able to catch "forthcoming" information. Therefore, we want to see how it would look like to include next year's productivity information and whether this change is robust to the model performance or not. Tables 2.13 to 2.16 reports the robustness check model results.[37] There is not a big difference in point estimates and overall trend between the main DML models and robustness check

[36]We do not include current year cumulative productivity information, because we believe that next year's cumulative productivity measures have already contain the current year cumulative information. But current productivity is not representable by next years' information.

[37]In the result tables, if NA shows up in the whole column, meaning this model does not work on this specific dataset and no result reported; if NA shows up in se(median), meaning no "best" result reported and we report a median value among all ML estimators and its corresponding se.

models. One may notice that the point estimate range changes a little bit, but overall they are fairly similar to the main DML models. For example, the range of gender effects in the robust check models is approximately 0.03 to 0.16, slightly shifting higher compared to the main DML model, and the estimates are mostly statistically insignificant; the range of graduate school rank effects in the robust check interactive model is about 0.01 to 0.18, in the partial linear model it is about 0.03 to 0.2; those are about the same as in main DML model.

2.8 Conclusion

This paper makes the following contributions. First, we created a novel dataset that has never been studied before in the literature. This dataset collects multi-dimensional and time-varying research productivity measures of young economics faculties and can be helpful in answering many research questions. Second, we applied a newly boarded method to consistently estimate causal effects with high-dimensional control variables. Last but not least, we looked into how easy-to-observe variables that are correlated with productivity influence a young economics professor's salary once productivity measurements are held constant.

The paper's results raise numerous questions. Why are the larger-effect estimates not statistically significant? Is it because the true effect is indeed zero or because we do not have enough data to reduce the randomness? Because of the limit of sample size, the noise from the data might affect the results significantly.

In the future, it would be useful to expand the current work in several directions, one way would be to expand the current dataset and try to obtain more precise

confidence intervals for point estimates; another way would be to try to dig into the DML model even further to make it more visible, showing the estimation process and results and making it more flexible to combine with other deep learning methods; besides, we can also try other nonparametric method such as matching to serve as model validity check.

Chapter 3

Academic Paper Publication Value and Gender Bias Based on Text Analysis

3.1 Introduction

In the academic world, finding good measures of research ability and properly evaluating it has been broadly discussed over the past few decades ((Moed, 2008; Adams, 2009; Taylor, 2011)). Paper publications and citation rankings are considered important measures for showing one's research ability ((Frey and Rost, 2010; Moed, 2008; Gibbons and Fish, 1991)). However, that information is not available until the papers are published in a journal, a process that could take months, even years. Before knowing the publication information, what can people learn from a finished paper itself rather than just objectively judging it (as discussed by Adams (2009))? This question motivated us to build a prediction model based on text information of a

paper to quantify its publication quality potential.

We use journal H-index as the outcome measure. H-index evaluates both the productivity and citation impact of a publication (Bornmann and Daniel, 2007), and it reflects the (potential) quality of a paper (Hirsch, 2007). Usually, higher quality papers are more likely to be published in higher H-indexed journals.

However, there are issues with using journal H-index as a measure of a paper's quality that must be carefully considered. First, papers published in the same journal are considered to be at the same level. But even within the same journal, some papers get many citations and some may get only a few; thus, even in the same journal, there could exist a wide range of paper quality. Second, the H-index could partially reflect subfield information (Hirsch, 2007). Even if we take economics papers[1] as one big group, many subfields exist under it, and different subfields may have different citations. For example, a game theory paper with very good quality may get lower citation rates with a lower H-index relative to an education or finance paper, where the number of citations are typically larger. Third, the H-index could even reflect gender bias (Tamblyn, Girard, Qian, and Hanley, 2018; Witteman, Hendricks, Straus, and Tannenbaum, 2019). Studies have shown that being female could be negatively correlated to publishing refereed papers and total publications (Sax et al., 2002; Yang and Webber, 2015). There is also a correlation between the percentage of women on journal editorial boards and the percentage of papers written by female authors in some fields (McElhinny, Hols, Holtzkener, Unger, and Hicks, 2003). Potentially, people tend to believe that male authors dominate in authoring high-quality papers since they author more papers than women (McElhinny et al., 2003; Nakhaie, 2008).

[1]Here economics papers refer to the papers we include in our dataset, defined in Section 3.2.1.

It is therefore natural to ask whether there is gender bias on H-indices. Or more specifically, whether higher H-indexed papers are more likely to be written by male authors or the opposite. In this paper, we address these questions and give plausible explanations.

We first create a novel dataset containing the most recent papers published in a large group of economics journals as well as authors' gender information. We apply Natural Language Processing (NLP) to transform the text into numerical values[2] and discuss how to use the text data to train machine learning (ML) models. We then take advantage of NLP extract features, and trace the relations between paper text content and any pattern it may contain. The features extracted are "keywords" representations, which serve as regressors. The numerical values in each feature are quantified information of word importance, word relations (quantified word associations), and potential patterns (preferred wording habits can be captured by classification of the features). This information conveys the main idea of a journal paper. Then we build models for both prediction and causal inference purpose.

There has not been any research using text information to detect a journal paper's potential quality, but there are related studies of text analysis in the literature. There are studies doing feature extraction and classification from short text content like movie reviews, Twitter posts, and text messages (Wang, Liu, Sun, Wang, and Wang, 2015; Tang, Wei, Yang, Zhou, Liu, and Qin, 2014; dos Santos and Gatti, 2014). Many focus on semantic analysis (Wang et al., 2015; Tang et al., 2014; dos Santos and Gatti, 2014; Collobert and Weston, 2008) and applications for predicting the next word or

[2]We use NLTK Python library to conduct the NLP practice with basic preprocess from the raw data; and Gensim, a popular Python package, to get value representations of words from different mechanisms.

sentence in a specific context (Mikolov, Karafiát, Burget, Cernocký, and Khudanpur, 2010), including in finance and accounting (Lang and Stice-Lawrence, 2015; Loughran and McDonald, 2016) and for translation purposes (Hu, Lu, Li, and Chen, 2014). Westergaard, Starfeldt, Tønsberg, Jensen, and Brunak (2018) use text analysis to grasp information from scientific papers but do not link the text to any performance outcomes.

Text data are quite different from other economics data. Text are usually high-dimensional, so effectively reducing the feature dimension yet producing reliable model predictions leads to another important regimen in the text mining world. We use term frequency–inverse document frequency (tf–idf), a simple yet efficient way to represent keywords (Gentzkow, Kelly, and Taddy, 2019). Very low and very high frequently shown words in a document might have similar tf–idf values, could both be small. By choosing different tf–idf cutoff values for each document, we can control the keyword frequency we want to keep, drop those very high frequency words such as "the", "very", "every" which do not convey too much information as well as those very low frequency words in each document which may not be contributing much to the main idea. Then, we build prediction models and discuss gender bias based on the vectorized data.

We find that it is useful and informative to use paper text data to detect its potential H-index, although the prediction power is not very high. Different gender may have different emphasis on paper topics, but the gender bias on H-indices is not obvious when controlling for paper text.

This paper contributes to the literature in the following ways. First, we create a novel dataset, implement an NLP technique, and use the paper text data to construct

models for outcome prediction and gender effects. Second, we visualize the keywords and discuss the patterns used by different gender groups. Third, we compare ordinary least squares (OLS) and double machine learning (DML) gender causal effect estimates to provide creative insights on how gender relates to paper outcomes, that is, H-indices.

In the rest of our paper, we present how the data are collected and preprocessed in Section 3.2; in Section 3.3, we discuss paper text prediction models and gender bias causality; in Section 3.4, we discuss the results; and finally we conclude in Section 3.5.

3.2 Data

In this section, we first introduce the data collection procedure and then show how we preprocess the data.

3.2.1 Data Collection

We first organized a text corpus through papers from a pre-made list of economics journals, and we collected each paper from an open application programming interface (API) provided by Elsevier[3] using web crawling Python libraries.[4] Elsevier provides a comprehensive yet easy-to-use API for directly downloading paper content and its relevant information.

We then made a search criterion to control the pool of papers we wanted to collect, which can be simply passed to a search query to get the link and access all journal

[3]Website: https://api.elsevier.com/

[4]Including json, requests, and Beautiful Soup.

papers satisfying that criterion. To construct our dataset, we set the criterion to collect papers under the domain of *Economics, Econometrics and Finance (General)*, between January 2015 and March 2020. Through the API, we put the content of papers satisfying the search criterion in .txt files, with titles as the filenames. Other than the text content itself, we also collected and saved author names and publication dates.

Elsevier's API does not provide authors' gender information, so we collected gender information separately. We created a numeric variable of values from 0 to 1 to represent the fraction of male authors of each paper. For example, if a paper was written by only a male author(s), then the value is 1; if the paper was written by only a female author(s), then the value is 0. If the paper was written by mixed-gender authors, then we set the value to the weighted average of the authors' genders.[5] Based on this numeric variable, we also create four gender dummies; details are in Section 3.4.2.

The H-indices were collected from the SJR website[6] for year 2018 and are used as the outcome variable.

In summary, our dataset contains the paper name, journal name, author name(s), author gender(s), publication date, and H-index, as well as the paper's text in a .txt file. The paper text contains the main text content, excluding abstract, references, figures, equations, and tables.

[5]The fraction of male authors.

[6]`https://www.scimagojr.com/index.php`

3.2.2 Data Cleaning

Before applying the analytical models, we must process the text data into numerical values. NLTK Python library provides many useful tools for preprocessing the text data. After examining a number of sentences using sentence tokenization tools provided by the NLTK, we find that the papers we collected contained an average of 500 sentences, with few containing more than 1,000 sentences. From this perspective, papers could contain more than 10,000 words very easily, and when transforming these papers into document matrices, the size of the matrices could cause the computer's memory to overflow very easily. Wei, Qin, Ye, and Zhao (2019) provide methods for down-sizing legal documents for deep learning (DL) models. Some of the useful methods including picking out sentences randomly and reorganizing the sentences based on their original sequence in the text. However, the randomization might lose important information; for example, by picking out sentences at random, the core methodology content of the papers could only be left with very few sentences, and majority of the content left could provide very little help in associating the contents to the value of the paper.

We then needed to use a method that could downsample the paper text yet grasp the most important part of the content. To do this, we used Gensim, a Python library. Gensim uses a tool called the summarizer, which summarizes text using text rank and can summarize and downsample a text based on the percentage of the content one would like to preserve. The summarizer uses a text rank algorithm (Barrios, López, Argerich, and Wachenchauzer, 2016), which originated from the page rank algorithm (Page, Brin, Motwani, and Winograd, 1999) used by Google for determining which web pages are more important based on people's searches of them. Text rank assigns

importance scores for each sentence based on a similarity graph that is built for the sentences using a similarity matrix created between the sentences. Sentences with higher similarity are connected with each other in the graph, and sentences containing more connections with other sentences are considered to be more important. For each paper, we keep the sentences with the most such connections, specifically the top 50%.

After the text content was downsampled, the documents were cleaned using the Regex python library. Regex can further help clean the text data, including removing digits, non-word symbols, and extra spaces. It works by formatting those parts of a document we would like to remove into a pattern represented by regular expressions and then replacing them with a single space.

Last, representing the words in a numerical way within each downsampled and cleaned document is a crucial step for feeding text data into prediction models. We use the tf–idf technique (Ramos et al., 2003; Wu, Luk, Wong, and Kwok, 2008), which transforms words into numbers. The score is calculated as $tf(t,d) \times idf(t)$, where t is the text term (a word) and d is a document (a specific journal paper in our study). $tf(t,d)$ is the term frequency of term t in document d, simply calculated as the raw count of the term within document d.[7] $idf(t)$ is defined as $\log(\frac{N+1}{n_t+1})$, the inverse document frequency of the term t, where N is the total number of the documents and n_t is the number of documents containing term t. Thus, the tf–idf score increases with the number of times a word shows up in one paper, but it is offset by the number of times this word occurs across all other papers. It takes account of a word's raw count while also measuring how much information this word provides within the specific corpus. Words that appear most frequently (over all text documents) are

[7]Sometimes it is divided by the max value among all frequencies so that the highest frequency would be normalized to 1.

called keywords.

Scikit-learn is a Python library dedicated to ML algorithms and provides a very useful vectorizer that can turn text documents into tf–idf matrices. When using the vectorizer, several parameters must be pre-determined. The range of the N-grams, which is the number of consecutive words combined and treated as one word unit, is set as 1 to 3, meaning that all the combinations of 1, 2, and 3 consecutive words were treated as a word or phrase. The max feature has been set as 15,000; that is, the tf–idf contains only 15,000 of the most frequent keywords. However, if a word appears in more than 98% of documents, then we ignore it because it is highly likely that the word is a "stop word" like "a," "an," and "the," which does not contribute to the algorithms as meaningful information.

In all, after the cleaning process, 2,295 journal papers have been saved in the text corpus. The vectorizer turned the corpus into a tf–idf matrix with a dimension of 2,295 by 15,000. Each time a prediction model is run, all data will be split randomly into two datasets for training and validation purposes. Eighty-five percent of the data, 1,951 documents, were assigned for training purposes, while 15%, 344 documents, were assigned for validation purposes.

3.3 Modeling

In this section, we first introduce how we implement the H-index prediction model by using the text data we just preprocessed, and then we build models discuss gender bias.

3.3.1 Prediction Model

We use random forest (RF) (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) as the ML methods in our prediction model. We choose ML instead of other DL techniques due to the sample size limit of our data. When training DL models, usually stochastic gradient descend (Bottou, 2010) is used. For example, in constructing RNN and CNN models (Sak, Senior, and Beaufays, 2014; Laurent, Pereyra, Brakel, Zhang, and Bengio, 2016; Goyal, Dollár, Girshick, Noordhuis, Wesolowski, Kyrola, Tulloch, Jia, and He, 2017), usually a mini-batch is used in each step of the gradient descent to reduce the noise caused by data redundancy. Even if stochastic gradient descent is not used, DL models are "deeper" and multiple hidden layers are needed. Many more parameters need to be estimated than traditional ML models, which needs quite a large sample size to support the computation. Therefore, considering that our dataset has only 2,295 observations, using ML models would be more practical.

RF is made up of decision tree models, and it aggregates many small trees through the bagging method (Breiman, 1996), with each tree built at each split using a random subset of features. When predicting using data, the predictions from each decision tree are averaged as the final prediction. RF has the same accuracy as bagged trees in the sense that the average of many identically distributed trees have the same expectation as each individual tree, whereas RF with random setting greatly reduces the correlation between each tree so as to reduce the variance.

XGBoost (Chen and Guestrin, 2016) is a new workhorse in the ML applications; it originated from stochastic gradient boosting (Friedman, 2002) but incorporates a regularization to reduce overfitting. It involves a second-order Taylor approximation of the loss function so to make sure it works better than regular boosting method.

Because we have only 2,295 documents, the number of the features (keywords) is much larger than the sample size. Therefore, for the purpose of dimension reduction, we first standardize the tf–idf scores to have mean 0 and standard deviation 1. We then perform truncated singular value decomposition (Frank and Buhmann, 2011) using Scikit-learn's TruncatedSVD tool, keeping only a subset of feature columns. The inputs we end up using are 20, 200, and 500 truncated SVD components, and each input is transferred and kept from 3,000 to 15,000 keywords.

Finally, the hyperparameters[8] of the prediction models are tuned through a grid search using a set of pre-defined parameter grids and five cross-validation (CV) folds of the training dataset. The hyperparameters are chosen from a selected set of hyperparameters that can obtain highest accuracy performance on the CV folds in average. The selected hyperparameters are then used to build a final prediction model using all the data points in the training dataset. The performance was then evaluated using the testing dataset.

3.3.2 Gender Differences

For the purpose of detecting gender causality on journal paper H-indices, we start from the simplest linear population model:

$$Y = \alpha + D\theta + U, \tag{3.3.1}$$

[8]Hyperparameters in RF are the following: "*max_depth*" the maximum depth of the decision trees; "*min_samples_split*" the minimum number of samples needed when creating a split in the decision trees; "*n_estimators*" total number of decision trees in the RF model. Hyperparameters in XGBoost are the following: "*max_depth*" maximum depth of a tree; "*learning_rate*" a weighting factor that deciding how much of a new tree added to the existing tree; "*n_estimators*" total number of decision trees in XGBoost; "*subsample*" the fraction of observations to be randomly sampled before growing each tree.

where Y is the H-index, α is the intercept, D is the gender variable, and U is the error term. It would be more proper to interpret the estimated θ here as the statistical association of gender with H-indices because there is omitted information in U that is suspected to be correlated with gender yet also affects Y, such as research ability or research topic choice. Due to this concern, we add more control variables to (3.3.1):

$$Y = \alpha + D\theta + \mathbf{X}\boldsymbol{\beta} + U$$

where $\mathbf{X}$ denotes the vector of control variables, here the (top 20, 200, or 500) truncated SVD components from the tf–idf matrix. We believe that a paper's text is a good source to be controlled because it can reflect different subfields and topics that male and female authors may concentrate on differently, and it can reflect the potential paper quality associated with research ability. Those are the information contained in U that could cause bias on the causality of gender effect estimator. Adding the tf–idf SVD components can help us reduce the data dimension yet also keep as much information as possible.

However, bias is a potential concern in a linear model that contains hundreds of explanatory variables, and the model may not be a good choice in terms of the best approximation of the true conditional expectation function (CEF). Moreover, the true CEF could be a flexible one, with the truncated SVD components being free to interact with each other:

$$Y = D\theta + g(\mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0, \tag{3.3.2}$$

where $\mathbf{X}$ denotes the truncated SVD components vectors, and they join the model

by a nonparametric function $g(\mathbf{X})$, or even more generally as

$$Y = g(D, \mathbf{X}) + U, \quad \mathbb{E}[U \mid \mathbf{X}, D] = 0. \tag{3.3.3}$$

DML (Chernozhukov et al., 2018) can be used to estimate (3.3.2) and (3.3.3). Under the DML framework, gender is allowed to be correlated with $\mathbf{X}$, here the truncated SVD components of paper text, through another CEF:

$$D = m(\mathbf{X}) + V, \quad \mathbb{E}[V \mid \mathbf{X}] = 0, \tag{3.3.4}$$

where $m(.)$ is another non-parametric function.

When conditional independence assumption (CIA)[9] and overlap assumption[10] hold, DML can provide consistent estimators of gender causal effects.[11] As mentioned in the beginning of this section, the gender variable could be affected by many other issues, for example, research ability. We believe that once we control for the paper text's content, gender D would not be associated with U, if not though the paper text could help capture the most of the omitted variable bias. Then, we could identify the average causal effect of gender or at least make the estimation as close as possible to the true causal effect.

[9] Wooldridge (2010) Assumption ATE.1 and a weaker version ATE.1

[10] Wooldridge (2010) Assumption ATE.2

[11] In either estimation of (3.3.2) or (3.3.3), (3.3.4) is estimated at the same time in DML, for the purposes of constructing Neyman orthogonality. See Chernozhukov et al. (2018) for more details.

3.4 Results

3.4.1 Prediction

Tables 3.1 and 3.2 show the prediction results from basic linear regression models to highly trained ML models (RF and XGBoost). In regression models, we use the square root of mean squared error as the evaluation measure. In classification models, we use a categorized H-index as outcome variable, with each observation assigned automatically to one of the four quartile categories (which are separated by 25%, 50%, and 75% quartiles of H-index across the data). We report the percentage prediction accuracy as evaluation measure in the classification model. According to Sokolova, Japkowicz, and Szpakowicz (2006), instead of using other empirical evaluation measures in ML applications such as precision, recall and F1 score, using accuracy is because our data are separated evenly into 4 categories and we care about correct predictions on all 4 categories.

Hyperparameters determines the performance of ML models. In RF, "max_depth" is tuned from 3 to 9, "min_samples_split" is tuned from 3 to 7, and "n_estimators" is tuned from 100 to 1,000 with 100 as one step. In XGBoost, "max_depth" is tuned from 3 to 9; "learning_rate" is tuned from 0.0001 to 1; "n_estimators" is tuned from 100 to 1,000 with 100 as one step; and "subsample" is tuned from 0.6, 0.7, and 0.8. The best prediction result is obtained from a fivefold cross-validation in tuning the above hyperparameters on the training sample. We also use the corresponding hyperparameter set to calculate the test errors as the final result reported for each SVD/keywords scenario.

For regression models (Table 3.1), XGBoost performs better than RF and both of

them performs better than OLS. The first row shows a benchmark which shows the result of using only a constant (mean value) to predict. For the *OLS* column, the training set's root mean squared error (rmse) decreases with more regressors (SVDs) included; however, there is no such trend on testing set. The lowest test rmse comes from the scenario with 200 SVD components transferred from 3,000 keywords, which is 43.3; after that, increasing SVD components or keywords in the OLS models does not help. Too many variables in OLS (500 SVDs from 15,000 Keywords) make the test rmse even worse than the benchmark. In *Random Forest*, the training set's rmse ranges from 27 to 30, which is a lot better than the benchmark or OLS, while the RF test rmse even though are larger than the train rmse, still got improved from its same level OLS model. XGBoost performs similar to RF on the training set, except in some 3,000 and 6,000 keyword scenarios, the training set rmse drops below 10, where XGBoost works much better. Looking at the rmse of XGBoost's testing set overall, the improvement on test rmse compared to OLS or RF is less apparent. The best test result from XGBoost is 43.5, from the scenario where 6,000 keywords were transferred for 20 SVD components. Looking at each SVD and Keywords senerio horizontally, XGBoost beats other models. There could be overfitting problems but not definitely. Because overall we observe the ML train rmse being smaller than the OLS train rmse and both of them being smaller than the OLS and the ML test rmse. According to the model results, people should expect an rmse average of 44 in the H-index differences up and down.

For classification models in Table 3.2, XGBoost is the winner again. The benchmark showed in the first row is 25% for random guessing. RF and XGBoost provide substantial higher accuracy in the training set compared to Logistic models no matter

how many SVDs are used. On the testing sets, the two ML methods still performs better but the improvement from logistic models are not as big. The best results are from the 200 SVD and 3000 scenario, where the test set accuracy is 58.7%, meaning when predicting the H-index categories, there is a 59% chance the prediction is right.

In summary, for both the regression and classification models, ML methods can always beats the benchmark regardless of SVD/keywords, whereas OLS can actually do worse than the benchmark for some SVD/keywords scenarios.

However, even though XGBoost shows higher prediction power compared to the others, its test set's performances are worse than its train set's. Apart from overfitting that I just mentioned, the reason could be more likely attributed to the data we use, as the text data only conveys a paper's partial information, and the excluded tables, graphs, and formulas are informative as well. On the other hand, the outcome variable H-index might convey multiple meanings. As mentioned in the introduction, it not only about paper quality, but also paper subfield, gender differences and so on. Moreover, it worth mentioning that we are using paper text to predict. In the scientific paper writing world, novelty is a very important criteria for publication outcomes (Veugelers and Wang, 2019, Wang, Veugelers, and Stephan, 2017), so using already published paper information to build prediction models might mean losing some part of the novelty information of a paper. Even though we have taken this into account and only keep the most recent papers, we are not able to exclude noise caused by new information that comes along with a paper. So for all the above reasons, we consider the prediction errors of our models acceptable and the predictions quite informative. When using text information to predict a paper's H-index, the regression model predicts an rmse of around 44, and the classification model predicts which tier

Table 3.1: Regression Prediction Models (RMSE)

		OLS		*Random Forest*		*XGBoost*	
SVD	Keywords	Train	Test	Train	Test	Train	Test
0	0	46.3	50.7	-	-	-	-
20	3000	43.0	47.1	30.0	46.0	23.7	44.1
20	6000	42.8	47.0	29.4	44.7	23.0	43.5
20	9000	42.9	47.7	28.3	45.3	5.7	43.9
20	12000	43.0	48.2	29.1	46.1	22.1	44.3
20	15000	42.8	49.0	29.1	47.2	5.8	44.7
200	3000	37.5	43.3	29.4	47.3	2.3	43.8
200	6000	36.1	43.5	28.4	46.4	2.1	44.1
200	9000	36.2	44.8	26.8	46.3	21.4	44.4
200	12000	36.3	47.4	27.6	47.8	21.2	45.3
200	15000	36.6	52.7	26.4	47.6	21.4	45.5
500	3000	32.4	44.5	30.9	48.4	1.9	45.3
500	6000	31.7	44.0	29.1	47.3	1.7	44.4
500	9000	32.4	45.7	30.8	47.9	18.9	45.0
500	12000	31.7	59.9	28.1	49.6	19.7	47.0
500	15000	31.6	72.3	27.6	49.2	19.8	46.1

Note: Truncated SVD components are used as model inputs. H-indices are used as model output.

the paper would be in with around a 60% correct possibility.

In Appendix C.1, we also give prediction results which exclude outliers in the data. We also use test rmse as evaluation measure. The test rmse shows without outliers the linear regression estimated by OLS works better than ML models and only OLS can beat the benchmark whereas ML models some time do worse.

3.4.2 Gender differences

Descriptive Statistics

Figure 3.1 reports the H-index box plots by gender, which are distributed similarly among different gender values. Figure 3.1a shows all numeric gender values' corresponding H-indices, and Figure 3.1b focuses on a subset of them. When gender=0.8,

Table 3.2: Classification Prediction Models (Accuracy in Percentage)

		Multinomial Logit		*Random Forest*		*XGBoost*	
SVD	Keywords	Train	Test	Train	Test	Train	Test
0	0	25	25	-	-	-	-
20	3000	49.1	46.5	90.1	48.0	94.1	49.4
20	6000	48.5	45.9	88.7	50.3	100.0	52.3
20	9000	49.9	46.5	86.9	50.3	100.0	51.7
20	12000	50.2	45.1	87.2	48.0	100.0	50.6
20	15000	49.8	45.9	88.1	48.6	100.0	52.6
200	3000	67.8	50.0	98.5	50.3	100.0	58.7
200	6000	69.8	52.0	98.3	53.8	100.0	55.2
200	9000	68.5	52.9	99.0	53.8	100.0	54.1
200	12000	69.5	50.6	98.9	50.0	100.0	53.8
200	15000	69.9	51.5	98.5	51.2	100.0	52.3
500	3000	96.4	49.4	99.3	48.0	100.0	54.7
500	6000	98.2	49.1	98.9	50.9	100.0	52.0
500	9000	98.9	47.4	99.5	52.3	99.9	51.7
500	12000	99.6	50.6	99.5	50.0	99.9	50.9
500	15000	99.8	47.1	99.2	50.3	100.0	54.1

Note: Truncated SVD components are used as model inputs. 4 quartile separated H-index categories are used as model output.

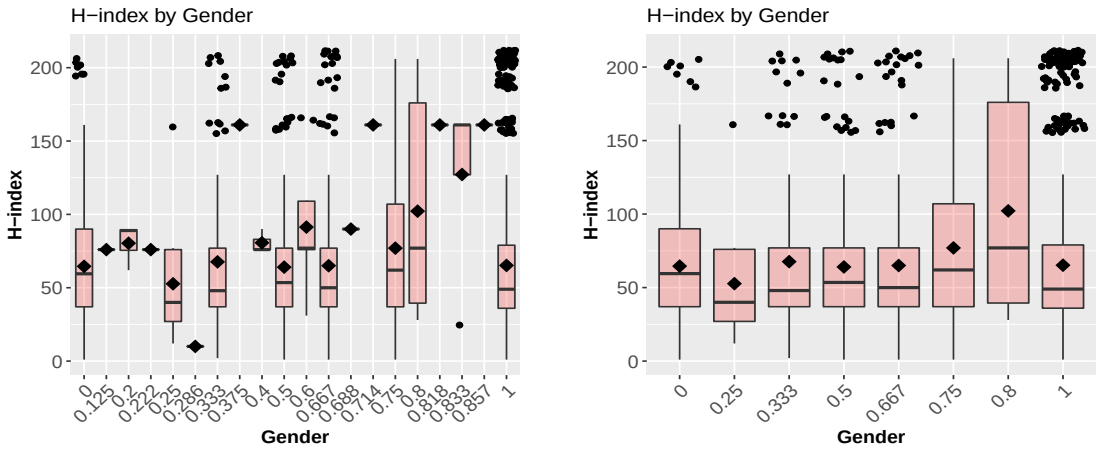
the span of H-indices are more spread out compared to others, but actually the count of cases when gender=0.8 is quite small. Figure 3.2 shows the data counts by gender values. Male-only author(s) make up nearly half of the whole dataset, whereas female-only author(s) make up only about 10%.

Black dots on the box plots in Figure 3.1 are outliers corresponding to each gender value. The outliers are defined as more than 1.5 IQR[12] below the first quartile or above the third quartile of H-indices corresponding to each specific gender value. Nearly all of the outliers come form the upper side of the boxes, which makes the distributions right skewed such that within most of the gender values, the H-index mean is greater than the median.[13] Among them, the largest number of outliers comes from when

[12]The distance between the third quartile and first quartile.

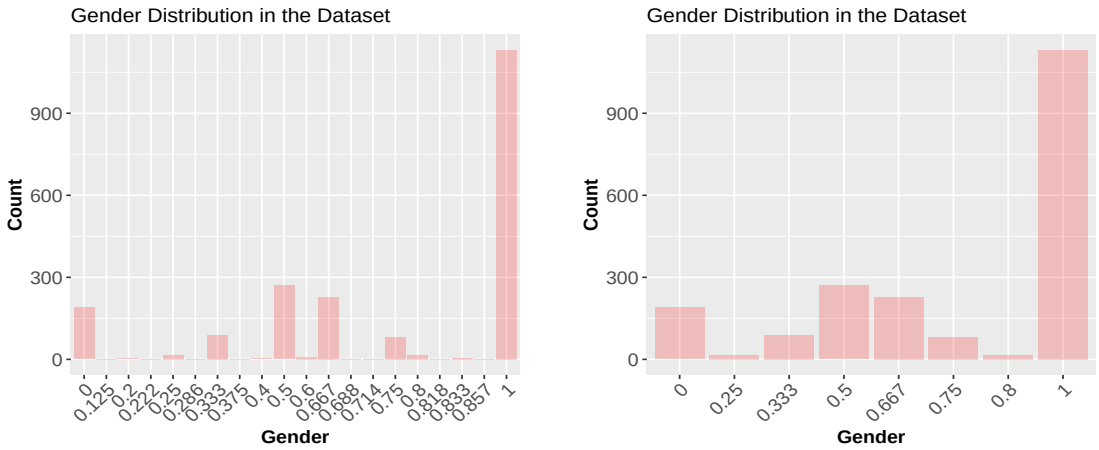
[13]Each black diamond on the box plot represents the mean value of all data points within that group; each bar represents the median value.

gender=1, where there are 93 outliers. The group of gender=1 box also has largest number of cases compared to other gender values, as shown by the highest column in Figure 3.2. It is therefore not surprising that this group has the most outliers.



(a) (b)

Figure 3.1: H-index box plots by Gender



(a) (b)

Figure 3.2: Gender distribution in the data

(a) (b)

Figure 3.3: H-index Density

The top ten keywords in Table 3.3 show some interesting facts and patterns. First, the word "students" is in first place in only the female author(s) (Gender=0) group, while it is ninth place in the gender≤0.5 group, and then drops out of the top ten in the other two groups. The word "food," which is in second place in the Gender=0 group, shows a similar pattern as "students." Second, we find that female authors use the word "policy" quite often, in third place in the Gender=0 group, and the word "model" is in seventh place. Conversely, in the other groups the word "model" shows up in the first two or three rows and "policy" in seventh place. It seems that the higher proportion of male-authored papers place more emphasis on techniques, whereas female authors place emphasis more on the policy itself when expressing their analysis. Third, the rank of "risk" increases when the male author proportion increases. Last, it is quite provoking to note that only female-authored papers seem more focused on life-related social topics such as "water," "school," and "housing," whereas with papers with more male authors, the top keywords are related

Table 3.3: Top 10 keywords

	Gender=0	Gender≤0.5	Gender>0.5	Gender=1
1	students	market	market	market
2	food	firms	model	model
3	policy	model	firms	price
4	market	time	price	risk
5	water	food	risk	firms
6	school	risk	time	time
7	model	policy	policy	policy
8	housing	rate	firm	stock
9	social	students	results	firm
10	inflation	price	stock	results

Note: Ranked descending by aggregated tf–idf score.

to finance and business subfields. In summary, papers written by female author(s) or by over half female authors focus on a more broad range of social topics such as education, health, and environment, whereas in more male-authored papers, financial and business topics are covered more and more emphasis is placed on the technical side ("model") rather than observable facts ("policy") itself.

Even though the ranked differences of the top ten words between gender groups seems trivial, one cannot deny men and women use different keywords and there is a concentration bias in regard to research topics between men and women researchers. There could be several reasons for this, such as men and women having different research interests. We can observe such differences when gender proportions of published works are approximately the same as that in the field (Lutz, 1990; McElhinny et al., 2003). However, McElhinny et al. (2003) also show that some journals with a larger number of women on editorial boards are more likely to publish female-authored papers, which could be another reason for the differences we see among different gender groups.

(a) Gender=0 (b) Gender=1

(c) Gender≤0.5 (d) Gender>0.5

Figure 3.4: Word Cloud

Causality

We create two types of gender variables to analyze gender bias in paper H-indices. One is a numeric gender variable, defined as the proportion of male authors. For example, if a paper is written by five people and three of them are male, then the gender variable is 3/5 = 0.6. The interpretation of this variable is that the higher the proportion of male authors in a paper, the larger the value on the gender variable

and the more it gets closer to one. From this perspective, the potential causality could be explained as H-index differences caused by a change in the male proportion of authors in a paper. The other type is a binary variable, where we create four different dummies based on the numeric gender variable. They differ from each other in the cut-off values, that is, the four dummies in Table 3.4. Accordingly, the causality explanation changes. For example, the causality caused by dummy 1 can be explained as the H-index changes when gender constitution of a paper changes from all female (Gender=0) to male (Gender>0) contributed. We want to explore how the H-index distributed by gender, the difference of paper publications between genders. Since we collect paper text data, those can be used as controls when estimating causal effects. Because we believe that text could reflect paper quality as well as research ability, even if the text only reflects subfields or topics, it is still important to condition on when evaluating gender differences.

Table 3.4 reports subgroup averages and average differences of the four dummy variables. We find that male author(s) contributed (group of Gender>0) papers have a higher H-index than only female author(s) (group of Gender=0) by 1.810, on average; single-gender authored papers (group of Gender=1,0) have a lower averaged H-index than mixed-gender (group of 0<Gender<1) authored papers by 3.099. Papers written by only male author(s) (group of Gender=1) have a lower average H-index than female contributed (group of Gender<1) papers by 2.258, and more male contributed (group of Gender>0.5) papers have a higher average H-index than more female contributed (group of Gender$\leq$0.5) papers by 2.025. Tables 3.5 to 3.9 report OLS and DML estimations of gender effects, where 20, 200, and 500 truncated SVD components (transferred from 3,000 to 15,000 keywords) are used as the control

Table 3.4: Average (Mean) H-Index by Gender Groups

	Dummy 1	Dummy 2	Dummy 3	Dummy 4
1	Gender>0	Gender=1,0	Gender=1	Gender>0.5
	66.399	65.128	65.218	66.801
0	Gender=0	0<Gender<1	Gender<1	Gender≤0.5
	64.589	68.227	67.476	64.776
Diff.	1.810	-3.099	-2.258	2.025

Note: Gender is defined as 1 if the paper is written by all male authors; 0 if by all female authors; and weighted average if by mixed gender authors with the weights correspond to the number of authors within each gender type.

variables in models.

In Table 3.5, the OLS estimates increase as the number of SVD components (control variables) increases. Even if those coefficients are not statistically significant, the estimates from larger SVD(s) and larger keywords may be more reliable. This is because when more information (variables) is controlled, the coefficient is more likely to be causal (Wooldridge, 2010). Moreover, the increase in adjusted R^2 justifies that adding more SVD components from larger keywords to the model is more helpful in explaining the dependent variable H-index. Therefore, it seems that the larger coefficients like 3.792 and 3.770 are more close to the true causal effect of gender than the coefficient 0.639 from the simplest linear model. However, the DML model results contradict the OLS results; they increase and decrease around 0 when the number of controls are increased. The reported DML standard errors[14] are all within 95% confidence intervals, and all of them include 0, and therefore they are indistinguishable from each other or zero statistically.

In Tables 3.6 to 3.9, the gender dummies share similar upward trend on the OLS coefficients and similar DML performances from which no significant effects can be drawn. The OLS coefficients of dummy 1 (gender>0) are around zero but with

[14]SE (median) is a more robust version of standard errors, defined as median methods adjusted standard error for DML estimates according to Chernozhukov et al. (2018), Definition 3.3.

obvious increases on the last three estimates, and up to close to the value of simplest linear regression coefficient, that is, the average group difference. The DML model results of dummy 1 varies within a small range, and most of the estimates are above 1 and close to the group difference. Neither OLS nor the DML model shows statistical significance on the coefficients.

Dummy 2 separates papers written by single-gender author(s) and mixed-gender authors. The average group difference is -3.009, meaning, on average, the mixed-gender authored papers have a higher H-index compared to single-gender authored ones. The OLS point estimate of 20 truncated SVD from 3,000 keywords is equal to -2.872, very close to the average group difference, and then shows an upward trend when more control variables are added. However, the DML estimates are all pretty close to the group difference. Both OLS and the DML have similar standard errors and are not statistically significant.

Dummy 3 separates only male-authored papers and the others; dummy 4 separates more male contributed papers and less male contributed ones. Like the first two dummies, even though the threshold value and point estimates are different, the model outcomes are similar in that in the OLS model, the dummy coefficients increase when the control variables increases and in the DML model the estimates are more concentrated with each other and are closer to the group difference.

Overall, the results in Tables 3.5 to 3.9 show that the DML gender effect estimates are closer than OLS to the simplest regression coefficients, which are also the group differences of the dummies. In regard to the numeric gender variables, dummy 2 and dummy 3, the OLS coefficients increases from a value close to the simplest regression coefficient to a larger one as the SVD controls changes from 20 up to 500. On the

contrary, the OLS coefficients of dummy 1 and dummy 4 change from a smaller value to one close to the simplest regression coefficient. The differences are quite trivial, and no solid causal conclusion can be made. Simply looking at the standard errors, it seems that the DML does not perform better than OLS. But instead of always going upward like OLS, the DML's estimates are more concentrated and consistent across different SVD scenarios. More specifically, no one model is more successful than the other. The DML allows interactions between the control and treatment/causal variables, which makes the model quite flexible, open to any form, and actually includes linear model as one of its possibilities. It also seems to be able to estimate any functions. But in the process of obtaining consistent causal estimators, the DML needs to split the data at least into two folds and use one fold to estimate nuisance parameters and the rest to obtain the causal estimators (Chernozhukov et al., 2018), requiring a relatively large sample size to get good performances.

Therefore, when dealing with the same sample as OLS, the DML just uses partial data to obtain the causal parameter. For OLS, the upward trends of coefficients might come from model mis-specification considering the large control variable set in a linear regression. Sometimes, one only takes OLS regression results and believe it reflects true causal meanings once the essential conditions are satisfied. But in reality we do not know what the true function looks like, whether the covariates interact with other, and how.

Nevertheless, we cannot conclude that there is a gender causal effect on papers' H-indices. Both the OLS and DML estimates are small in magnitude and not statistically significant, and most come with a 95% confidence interval that includes other estimates within the same model and cannot exclude zero. It is hard to say there is a

virtual difference. When gender changes from 0 (female) to 1 (male) or from mixed authors to single author(s), the H-index changes only about 1 to 3, which is trivial compared to the H-index range from 1 to 206.[15]

Therefore, there are several mechanisms that could be responsible for what we observed from data, and they could be working together to offset each other and even the true gender effect, if there is one. On the one hand, papers coming from the same journals have the same H-indices; however there might be a paper quality range even within the same one journal. For example, *European Economics Review* accepts papers from all areas of economics, both theoretical and empirical, and may publish some papers that get cited a lot of and some that do not. Those papers may differ in actual quality but could end up having the same H-index. Consequently, there could be quality overlapping between different H-indexed papers so that two identical papers published into different journals may have a different H-index.

On the other hand, in Section 3.4.2 we observed the different keywords used by male and female authors and discussed the different topics and subfields they publish in, which could make another path of the mechanism. For example, the *Journal of Financial Economics* publishes financial topics, which mostly male authors focus on, and the *Journal of Health Economics* publishes topics frequently focused on by female authors. Even if two identical papers were submitted to the same journal, they would have the same accept/rejection decision regardless of their authors' genders, since there are more finance papers than health papers due to most research economists being male. So naturally, a finance paper would get more citations and get a higher h-index, whereas health papers would get a relatively lower h-index, increasing the

[15]Shown in Table C.9.

causal effect.

Moreover, other subfield differences such as a good quality game theory paper may have a lower citation rate compared to a similar quality finance paper due to the lower number of game theory researchers. For example, when comparing theoretical game theory papers with a low H-index that are written by male authors that are similar in quality to higher H-index empirical health economics papers mostly written by female authors, the causal effect will be driven down.

Also, the process when a journal accepts papers could be disturbed by gender discrimination. As presented by McElhinny et al. (2003), in their study among the five journals in the sociology and sociocultural anthropology field, two journals with more women in its editorial boards tend to have more female-authored work published.

In Appendix C.2, we discuss the gender effect using data exclude outliers and display the results in Tables C.2 to C.6. Without outliers, male authors do not contribute to increasing the H-index. In Appendix C.3, we also include a median regression for gender effect, because we consider that regression to the median would be more robust to outliers. The results are showed in Table C.8, they are in accordance with our main result, indicating no gender effect. In Appendix C.4, we show results with controlling for number of authors and if single author.

3.5 Conclusion

We use a novel dataset that collects recently published economics papers, and we build ML and simple linear models to predict the paper outcomes measured by H-index as well as discuss the gender bias associate with them. We find that using

Table 3.5: Model Results Using Numeric Gender Variable

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	0.639	3.239	0.000	-	-
20	3000	-0.174	3.268	0.056	1.404	3.089
20	6000	0.088	3.256	0.061	0.989	3.042
20	9000	0.523	3.261	0.059	1.361	2.999
20	12000	0.469	3.261	0.057	1.876	3.240
20	15000	0.433	3.261	0.056	2.309	3.217
200	3000	0.698	3.123	0.254	0.087	3.370
200	6000	1.106	3.111	0.258	-1.504	3.134
200	9000	1.236	3.107	0.257	0.610	3.159
200	12000	1.183	3.106	0.259	-2.220	3.262
200	15000	0.950	3.117	0.257	-0.660	3.137
500	3000	0.908	3.350	0.305	-0.029	3.046
500	6000	3.889	3.305	0.318	0.420	2.998
500	9000	3.673	3.279	0.322	1.274	3.071
500	12000	3.792	3.269	0.323	0.207	3.043
500	15000	3.770	3.267	0.325	0.711	3.158

Table 3.6: Model Results Using Dummy 1 (Gender>0)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	1.809	3.646	0.000	-	-
20	3000	-0.411	3.645	0.056	1.366	3.226
20	6000	0.372	3.629	0.061	1.511	3.138
20	9000	0.624	3.636	0.059	0.590	3.204
20	12000	0.560	3.637	0.057	1.647	3.226
20	15000	0.649	3.635	0.056	2.197	3.257
200	3000	-1.054	3.428	0.254	0.759	3.282
200	6000	-0.179	3.411	0.258	1.026	3.209
200	9000	0.015	3.417	0.257	0.714	3.219
200	12000	0.081	3.414	0.259	0.892	3.233
200	15000	0.027	3.423	0.257	1.314	3.243
500	3000	-2.241	3.653	0.306	1.184	3.292
500	6000	0.809	3.615	0.317	1.537	3.281
500	9000	1.157	3.600	0.321	1.229	3.277
500	12000	1.566	3.592	0.323	1.181	3.305
500	15000	1.359	3.587	0.325	1.435	3.372

Note: Dummy 1 is set to 1 when numeric Gender is greater than 0; 0 when it is equal to 0 (only female authors).

Table 3.7: Model Results Using Dummy 2 (Gender=1 or Gender=0)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-3.099	2.207	0.000	-	-
20	3000	-2.872	2.169	0.057	-2.790	2.214
20	6000	-2.989	2.162	0.619	-3.062	2.195
20	9000	-2.832	2.164	0.060	-3.181	2.236
20	12000	-2.823	2.167	0.058	-2.894	2.219
20	15000	-2.891	2.168	0.057	-2.864	2.217
200	3000	-0.322	2.063	0.254	-2.928	2.221
200	6000	-0.802	2.052	0.258	-3.031	2.203
200	9000	-0.934	2.054	0.257	-2.890	2.223
200	12000	-1.026	2.054	0.259	-2.943	2.212
200	15000	-1.180	2.058	0.257	-3.120	2.231
500	3000	1.216	2.194	0.306	-2.977	2.208
500	6000	1.621	2.181	0.318	-2.891	2.212
500	9000	1.235	2.172	0.321	-2.084	2.235
500	12000	1.041	2.116	0.323	-3.074	2.210
500	15000	1.199	2.162	0.325	-3.049	2.198

Note: Dummy 2 is set to 1 when numeric Gender is equal to 0 or 1; 0 otherwise.

Table 3.8: Model Results Using Dummy 3 (Gender=1)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-2.258	2.125	0.000	-	-
20	3000	-2.750	2.117	0.056	-2.809	2.243
20	6000	-2.724	2.111	0.062	-2.943	2.227
20	9000	-2.488	2.113	0.060	-2.812	2.182
20	12000	-2.499	2.115	0.058	-2.698	2.161
20	15000	-2.532	2.115	0.057	-2.344	2.243
200	3000	-0.682	2.030	0.254	-2.534	2.111
200	6000	-0.843	2.023	0.258	-2.593	2.142
200	9000	-0.899	2.022	0.257	-2.652	2.170
200	12000	-0.967	2.022	0.259	-2.389	2.106
200	15000	-1.136	2.027	0.257	-2.487	2.112
500	3000	0.398	2.168	0.184	-2.355	2.102
500	6000	1.855	2.146	0.318	-2.530	2.126
500	9000	1.594	2.131	0.321	-2.455	2.102
500	12000	1.549	2.125	0.323	-2.333	2.109
500	15000	1.631	2.122	0.325	-2.507	2.127

Note: Dummy 3 is set to 1 when numeric gender is equal to 1 (only male authors); 0 otherwise.

Table 3.9: Model Results Using Dummy 4 (Gender>0.5)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	2.024	2.351	0.000	-	-
20	3000	1.771	2.354	0.056	1.486	2.457
20	6000	1.950	2.344	0.061	1.733	2.377
20	9000	2.267	2.348	0.060	1.486	2.502
20	12000	2.206	2.348	0.057	2.032	2.414
20	15000	2.139	2.348	0.057	1.727	2.342
200	3000	1.101	2.246	0.254	1.847	2.324
200	6000	1.292	2.234	0.258	1.563	2.330
200	9000	1.521	2.229	0.258	1.612	2.296
200	12000	1.546	2.227	0.259	1.699	2.363
200	15000	1.416	2.235	0.257	1.748	2.282
500	3000	0.572	2.386	0.305	1.779	2.280
500	6000	2.282	2.372	0.318	1.811	2.335
500	9000	1.996	2.353	0.322	1.618	2.328
500	12000	2.172	2.348	0.323	1.778	2.316
500	15000	2.103	2.346	0.325	1.829	2.266

Note: Dummy 4 is set to 1 when continuous Gender is greater than 0.5; 0 otherwise.

text information to predict paper H-index is informative but only in a limited sense. It can provide a rough idea about the range of the paper's potential H-index (with regression rmse 44 and 4-category classification accuracy 60%), although the H-index does not signify its absolute quality. We also find that when controlling for the paper text information, gender does not cause a change in its H-index. It would be meaningful to further extend this research by expanding the data pool by including other fields' papers and, more importantly, to explore other possibilities in information transformation, including extracting keywords, transforming words to vectors, finding ways to transfer tables and graphs to numeric values, and exploring other model-building techniques.

Appendix A

Double Machine Learning Discussion and Application to the Causal Effect of the *California Math* Curriculum

In this part, I will summarize how DML gives a consistent estimator as showed in Chernozhukov et al., 2018, section 1.

A.1 Partially Linear Model Consistency

A simple idea of estimating θ_0 in (1.2.1) is to subtract an estimator of $g_0(\mathbf{X})$ from Y and apply OLS procedure afterwards.[1] A naive estimator of θ_0 is then given by

[1]Because of the high dimensional nuisance parameter space, usually machine learning methods are used in estimating g_0, you can think this as an ML estimator.

equation (1.3) in Chernozhukov et al., 2018

$$\hat{\theta}_0 = (\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g_0}(X_i))$$

In (1.2.1), $D\theta + g_0(\mathbf{X})$ is a conditional expectation function (CEF). Functional form of g_0 is unknown and unrestricted. Maybe it's nonlinear and complicated. $D\theta_0$ part is a linear restriction (may not be correct to do so). But it's hard to know what a true CEF look like. With a more flexible $g_0(\mathbf{X})$, the whole function could be close enough to the true CEF as much as possible.

However, this naive way of θ_0 estimation cannot provide a properly converged estimator when machine learning is used in the estimation of $g_0(\mathbf{X})$. (Chernozhukov et al., 2018) Restrictions in Lasso and Ridge regression, penalty in Neural Nets and other penalty forms would increase the estimation bias in order to control variances. Unavoidably, regularization bias is produced. It comes along with the trade off between bias and variance in the estimations. The regularization keeps variance small but increase bias. The scaled decomposed estimation error of a naive estimator of partially linear model is given by

$$\begin{aligned}\sqrt{n}(\hat{\theta_0} - \theta_0) = &(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i \\ &+ (\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g_0}(X_i))\end{aligned} \tag{A.1.1}$$

the second term of (A.1.1) is indeed

$$(\frac{1}{n}\sum_{i\in I} D_i^2)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} m_0(\mathbf{X_i})(g_0(X_i) - \hat{g_0}(X_i)) + o_P(1)$$

(see page 3 on Chernozhukov et al., 2018, the equations right below (1.4)). Because of the existence of bias, the sum of n terms of $m_0(\mathbf{X_i})(g_0(X_i) - \hat{g}_0(X_i))$ do not have a mean zero. If we say the converge rate of an estimator converging in a root mean squared error sense is $1/2$. The converge rate of $\hat{g}_0$ to g_0 is slower than $1/2$. We denote the converge rate of $\hat{g}_0$ by φ_g, $\varphi_g < 1/2$. Thus, the performance of $\hat{\theta}_0$ is poor. When the sample size is relatively small, because of the slow converge rate, the estimator may deviate from the true parameter too much.

A general moment condition such as (Chernozhukov et al., 2018) equation (2.9) is,

$$\mathbb{E}(\psi(D, \mathbf{X}; \theta_0, \eta_0)) = 0 \tag{A.1.2}$$

where ψ is a vector of score functions, it could be in any form, a maximum likelihood score function, a GMM moment function and so on; η_0 denote the true value of nuisance parameters included in g_0 and m_0, $\eta_0 \in \tau$ where τ is the nuisance parameter space. Besides, the score function must satisfy an additional condition which is its Gateaux derivative $D_r[\eta - \eta_0]$ exists and non-sensitive to to the change of nuisance parameters η towards any direction. The Gateaux derivative is

$$D_r[\eta - \eta_0] := \partial_r\{\mathbb{E}[\psi(D, \mathbf{X}; \theta_0, \eta_0 + r(\eta - \eta_0)]\}, \quad \eta \in \tau$$

where $r \in [0, 1)$. The Neyman orthogonality condition is defined as (Chernozhukov et al., 2018) definition (2.1), $D_r[\eta - \eta_0]$ exists for all $r \in [0, 1)$ and $\eta \in \tau$ and at $r = 0$,

$$\partial_\eta \, \mathbb{E} \, \psi(D, \mathbf{X}; \theta_0, \eta_0)[\eta - \eta_0] = 0 \tag{A.1.3}$$

The above two equations are (Chernozhukov et al., 2018) equation (2.1) and the

equation below (2.1). The two conditions (A.1.2) and (A.1.3) together constitute the orthogonality of DML estimation.

Now let's see how DML makes the estimator of true θ_0 consistent in a partially linear model. Let $\check{\theta}_0$ denote the consistent DML estimator. The inference on θ_0 relays on the score function defined as in (Chernozhukov et al., 2018) equation (4.3):

$$\psi(D, \mathbf{X}; \theta, \eta) := (Y - D\theta - g(\mathbf{X}))(D - m(\mathbf{X})) \tag{A.1.4}$$

which satisfies the moment condition $\mathbb{E}(VU) = 0$ and the orthogonality condition (A.1.3). After some algebra, a DML estimator of $\check{\theta}_0$ is given by,

$$\check{\theta}_0 = (\frac{1}{n}\sum_{i\in I}\hat{V}_i D_i)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{g}_0(X_i)) \tag{A.1.5}$$

where $\hat{V}$ is from ML estimation $\hat{V} = D - \hat{m}_0(X)$. Note that $\hat{m}_0$ and $\hat{g}_0$ are obtained by auxiliary sample and $\hat{\theta}_0$ is obtained by main sample. The scaled decomposed estimation error is then,

$$\begin{aligned}\sqrt{n}(\check{\theta}_0 - \theta_0) = (\mathbb{E}\,V^2)^{-1}&\frac{1}{\sqrt{n}}\sum_{i\in I}V_i U_i \\ &+ (\mathbb{E}\,V^2)^{-1}(\frac{1}{\sqrt{n}}\sum_{i\in I}(m_0(X_i) - \hat{m}_0(X_i))(g_0(X_i) - \hat{g}_0(X_i))) \\ &+ o_p(1)\end{aligned} \tag{A.1.6}$$

(can be found in (Chernozhukov et al., 2018) page 3.) Now, this equation can be

separated into three parts

$$a^* = (\mathbb{E}\, V^2)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i$$

$$b^* = (\mathbb{E}\, V^2)^{-1} (\frac{1}{\sqrt{n}} \sum_{i \in I} (m_0(X_i) - \hat{m_0}(X_i))(g_0(X_i) - \hat{g_0}(X_i)))$$

$$c^* = o_P(1)$$

The first term converges to a normal distribution under mild condition, $a^* \rightsquigarrow N(0, \Sigma)$. The second term b^* now is determined by estimation errors of both $m_0(\mathbf{X})$ and $g_0(\mathbf{X})$. It contains regularization bias from both of them. The converge rate depends on specific Machine Learning methods used, usually is slower than a square root rate. Biau (2012) prove that the converge rate of Random Forest estimators depends on its strong features, the rate order has a form of $n^{\frac{-0.75}{Slog2+0.75}}$ where S is a subset of features. Chen (2007) gives convergence properties of least square regression under L_2 norm.[2] The converge rate is $n^{\frac{-p}{2p+d}}$, where d is the dimension of the raw explanatory variables, and p is the assumed degree of smoothness of the CEF (like number of derivatives). We can control the bound by choosing proper p's for any given d. We now know $m_0(\mathbf{X})$ and $g_0(\mathbf{X})$ are estimated with a slower converge rate to their true value, but the product of the two makes the whole term converge within a vanishing upper bound. This upper bound is $\sqrt{n} n^{-(\varphi_m + \varphi_g)}$, where φ_m is the converge rate of $\hat{m}_0$ to m_0 and φ_g is the converge rate of $\hat{g}_0$ to g_0. Thus b^* vanishes eventually if $\varphi_m + \varphi_g > 1/2$.[3]

[2] See proposition (3.6) (Chen, 2007).

[3] Chernozhukov et al. (2018) made this claim in their paper. They prove if the specific machine learning method used in the model has this property, then b^* will converge as well as proved. They also show good simulation results. In practice it's hard to find theoretical justifications for how each method converge, but DML has been shown outperforms just arbitrary picking some variables and running a simple regression.

Another requirement for $\check{\theta}_0$ to be consistent is to control the remainder items in c^* and make sure $c^* = o_P(1)$. In partially linear model, terms like

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i(\hat{g}_0(\mathbf{X}_i) - g_0(X_i)) \tag{A.1.7}$$

are included in c^*. Without sample splitting, model error terms V_i and estimation errors $\hat{g}_0(\mathbf{X}_i) - g_0(\mathbf{X}_i)$ are generally related. The reason is that in estimating $\hat{g}_0$ information contained in observation i has already been used, whereas V_i also has information from observation i, the relation between them will cause poor performance of c^*. Conditional on the auxiliary sample and with $\mathbb{E}(V_i \mid \mathbf{X}_i) = 0$, (A.1.7) has mean zero and variance with order $\frac{1}{n} \sum_{i \in I} (\hat{g}_0(\mathbf{X}_i) - g_0(\mathbf{X}_i) \to_P 0$.

A.2 A General Nonparametric DML Score Function

For estimating ATE, a score function for example in (Chernozhukov et al., 2018) (5.3),

$$\psi(\mathbf{W}; \theta, \eta) := (g(1, \mathbf{X}) - g(0, \mathbf{X})) + \frac{D(Y - g(1, \mathbf{X}))}{m(\mathbf{X})} - \frac{(1 - D)(Y - g(0, \mathbf{X}))}{1 - m(\mathbf{X})} - \theta \tag{A.2.1}$$

is needed, which must satisfy moment condition (A.1.2), as well as the orthogonality condition (A.1.3).

Appendix B

Are Young Economics Professors' Salaries Affected by their Background?

B.1 Supplementary Results

People may be curious about how OLS performs. Would they be different or similar with DML estimation? Tables B.1 to B.7 show the OLS estimation results. They are conducted using the same data as Tables 2.5 to 2.8. We report three kinds of models. The first one is a simple OLS model with only the treatment variable itself, as reported in panel A; the second one is an OLS model with only background information as control variables and report with 5 principal components and with 10 principal component results as showed in panel B; the third one, we include all (background and productivity) information as control variables and also report 5 prin-

cipal and 10 principal components results.[1] DML results are thus more comparable with OLS panel C. We notice that when adding more information to the control the coefficients are decreasing towards zero, and within each panel, when include more principal components the coefficients are also decreasing. These are consistent with our perception, because people would suspect there is a positive bias due to the unobservant in the simple OLS model. And again, with the thoery from (Wooldridge, 2010) saying that with more controls in the model the treatment coefficient is more likely to have causal meaning, people would believe more in the OLS results with more information controlled. In Gender effect models, compared with OLS model, DML (interactive) is a totally functional form relaxed model, gives near zero estimates in first few experience years estimates and greater estimates in experience year 6 to 7 than the OLS. In graduate school rank effect models, OLS panel C gives overall smaller point estimates than DML and even several negative point estimates. We are afraid that if it is OLS that put too much restriction onto the strcuture that make the estimates towards zero. In undergraduate major models, if we compare panel C of Table B.5 and Table B.7 and DML results, we may find that the overall signs are the same, undergraduate ECON major positive and undergraduate STEM major negative. But OLS results might still be more close to zero than DML results.

[1]Like how we process the data in DML estimation, we do principal component analysis on all the control variables we want to include into the OLS model, and put the principal components together with treatment variables into OLS model. 5 principal components are the top 5 ones explaining the variance of the original variable set; 10 principal components are the top 10 ones explaining the variance of the original variable set.

Table B.1: Gender effects on log salary (OLS)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. only gender dummy					
coefficient	0.053	0.027	0.051	0.042	0.098
se	(0.024)	(0.038)	(0.029)	(0.023)	(0.032)
B. add controls W/O productivity					
(5 principal components)					
coefficient	0.064	0.038	0.07	0.068	0.103
se	(0.026)	(0.04)	(0.027)	(0.02)	(0.032)
(10 principal component)					
coefficient	0.068	0.022	0.062	0.043	0.047
se	(0.028)	(0.037)	(0.027)	(0.022)	(0.028)
C. add all controls					
(5 principal components)					
coefficient	0.055	0.024	0.044	0.037	0.075
se	(0.026)	(0.041)	(0.029)	(0.022)	(0.029)
(10 principal component)					
coefficient	0.067	0.018	0.056	0.055	0.082
se	(0.029)	(0.042)	(0.03)	(0.023)	(0.03)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. only gender dummy						
coefficient	0.116	0.138	0.168	0.135	0.082	0.067
se	(0.035)	(0.044)	(0.06)	(0.077)	(0.098)	(0.106)
B. add controls W/O productivity						
(5 principal component)						
coefficient	0.114	0.153	0.132	0.111	0.006	0.141
se	(0.037)	(0.04)	(0.06)	(0.069)	(0.01)	(0.144)
(10 principal component)						
coefficient	0.074	0.055	0.066	0.061	0.067	0.006
se	(0.033)	(0.038)	(0.052)	(0.078)	(0.105)	(0.143)
C. add all controls						
(5 principal component)						
coefficient	0.082	0.07	0.055	0.048	-0.03	-0.061
se	(0.031)	(0.038)	(0.045)	(0.063)	(0.08)	(0.092)
(10 principal component)						
coefficient	0.08	0.078	0.085	0.038	-0.029	-0.093
se	(0.031)	(0.038)	(0.049)	(0.056)	(0.076)	(0.094)

Note: Panel A shows simple regression of salary on gender dummy. Panel B includes demographic, educational and year fix effect, school fix effect but not productivity measures. Panel C includes all control variables. All OLS result tables report heteroscedasticity robust standard errors.

Table B.2: Graduate school rank effects on log salary (OLS)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. only school rank					
coefficient	0.102	0.172	0.078	0.07	0.191
se	(0.059)	(0.101)	(0.072)	(0.057)	(0.092)
B. add controls W/O productivity					
(5 principal component)					
coefficient	0.056	0.175	0.022	0.064	0.184
se	(0.055)	(0.105)	(0.089)	(0.072)	(0.092)
(10 principal component)					
coefficient	0.014	-0.142	-0.027	-0.008	0.012
se	(0.077)	(0.137)	(0.097)	(0.082)	(0.112)
C. add all controls					
(5 principal component)					
coefficient	0.082	0.184	0.024	0.059	0.09
se	(0.052)	(0.082)	(0.073)	(0.054)	(0.075)
(10 principal component)					
coefficient	0.032	0.161	-0.043	0.033	0.101
se	(0.073)	(0.104)	(0.097)	(0.063)	(0.071)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. only school rank						
coefficient	0.21	0.293	0.288	0.014	0.175	0.279
se	(0.094)	(0.151)	(0.254)	(0.249)	(0.141)	(0.389)
B. add control W/O productivity						
(5 principal component)						
coefficient	0.136	0.17	0.191	-0.041	0.14	0.315
se	(0.085)	(0.13)	(0.214)	(0.229)	(0.147)	(0.39)
(10 principal component)						
coefficient	-0.026	0.026	0.07	-0.136	0.031	0.174
se	(0.1)	(0.107)	(0.138)	(0.152)	(0.128)	(0.25)
C. add all controls						
(5 principal component)						
coefficient	0.108	0.058	0.125	-0.118	-0.009	0.291
se	(0.081)	(0.08)	(0.144)	(0.166)	(0.158)	(0.285)
(10 principal component)						
coefficient	0.078	0.037	0.142	-0.073	-0.016	0.308
se	(0.076)	(0.086)	(0.175)	(0.145)	(0.167)	(0.292)

Note: Transformed rank variable is being used, calculated by $1 - rank/max(rank)$. Panel A shows simple regression of salary on rank. Panel B includes demographic, educational and year fix effect, school fix effect. Panel C includes all.

Table B.3: Graduate school rank (dummy) effects on log salary (OLS)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. only school rank					
coefficient	0.033	0.049	0.019	0.016	0.046
se	(0.029)	(0.034)	(0.03)	(0.027)	(0.033)
B. add controls W/O productivity					
(5 principal component)					
coefficient	0.017	0.043	0.015	0.007	0.041
se	(0.031)	(0.037)	(0.032)	(0.027)	(0.036)
(10 principal component)					
coefficient	0.013	0.009	0.004	-0.023	-0.027
se	(0.035)	(0.038)	(0.027)	(0.027)	(0.032)
C. add all controls					
(5 principal component)					
coefficient	0.018	0.043	0.024	0.003	0.01
se	(0.032)	(0.038)	(0.031)	(0.26)	(0.031)
(10 principal component)					
coefficient	0.014	0.037	0.007	-0.012	0.005
se	(0.037)	(0.04)	(0.034)	(0.027)	(0.033)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. only school rank						
coefficient	0.053	0.111	0.186	0.14	0.105	0.109
se	(0.037)	(0.045)	(0.057)	(0.077)	(0.084)	(0.103)
B. add control W/O productivity						
(5 principal component)						
coefficient	0.04	0.085	0.151	0.042	0.068	0.161
se	(0.038)	(0.046)	(0.054)	(0.079)	(0.099)	(0.098)
(10 principal component)						
coefficient	-0.017	0.01	0.028	-0.048	0.037	0.058
se	(0.033)	(0.046)	(0.058)	(0.068)	(0.08)	(0.087)
C. add all controls						
(5 principal component)						
coefficient	0.023	0.038	0.079	0.014	-0.036	0.076
se	(0.034)	(0.041)	(0.045)	(0.06)	(0.067)	(0.08)
(10 principal component)						
coefficient	0.016	0.026	0.09	-0.043	-0.049	0.083
se	(0.034)	(0.043)	(0.051)	(0.054)	(0.076)	(0.067)

Note: Dummy rank variable is being used, equal to 1 when rank is less or equal to the rank median, 0 if rank is greater than the rank median. Panel A shows simple regression of salary on dummy rank. Panel B includes demographic, educational and year fix effect, school fix effect. Panel C includes all.

Table B.4: Undergraduate major effects on log salary (OLS coefficients)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. only dummy					
Ugrad_ECON	0.005	-0.015	-0.071	-0.038	-0.038
se	(0.033)	(0.04)	(0.037)	(0.032)	(0.039)
Ugrad_STEM	-0.024	-0.035	0.017	0.03	0.102
se	(0.03)	(0.039)	(0.036)	(0.03)	(0.034)
B. add controls W/O productivity					
Ugrad_ECON	0.004	-0.023	-0.019	-0.013	0.014
se	(0.034)	(0.041)	(0.034)	(0.036)	(0.047)
Ugrad_STEM	0.023	-0.021	0.067	0.047	0.111
se	(0.054)	(0.069)	(0.07)	(0.06)	(0.064)
Ugrad_ECON*Ugrad_STEM	-0.069	0.034	-0.097	-0.039	-0.032
se	(0.06)	(0.088)	(0.081)	(0.066)	(0.078)
C. add all controls					
Ugrad_ECON	0.018	0	-0.046	-0.033	0.002
se	(0.035)	(0.045)	(0.037)	(0.038)	(0.053)
Ugrad_STEM	0.027	0.009	0.06	0.032	0.095
se	(0.053)	(0.08)	(0.072)	(0.059)	(0.063)
Ugrad_ECON*Ugrad_STEM	-0.077	-0.001	-0.099	-0.007	-0.028
se	(0.062)	(0.088)	(0.081)	(0.065)	(0.072)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. only dummy						
Ugrad_ECON	-0.021	-0.051	-0.133	-0.174	-0.15	-0.192
se	(0.042)	(0.048)	(0.06)	(0.094)	(0.103)	(0.169)
Ugrad_STEM	0.054	0.107	0.109	0.096	0.136	0.18
se	(0.042)	(0.053)	(0.063)	(0.11)	(0.115)	(0.122)
B. add controls W/O productivity						
Ugrad_ECON	-0.032	-0.073	-0.15	-0.223	-0.233	0.008
se	(0.062)	(0.074)	(0.096)	(0.089)	(0.114)	(0.289)
Ugrad_STEM	0.001	-0.034	-0.026	-0.065	-0.061	0.439
se	(0.081)	(0.085)	(0.103)	(0.17)	(0.227)	(0.319)
Ugrad_ECON*Ugrad_STEM	0.048	0.208	0.169	0.134	0.209	-0.342
se	(0.096)	(0.112)	(0.56)	(0.229)	(0.282)	(0.37)
C. add all controls						
Ugrad_ECON	-0.013	-0.044	-0.147	-0.196	-0.194	0.028
se	(0.061)	(0.048)	(0.06)	(0.078)	(0.095)	(0.103)
Ugrad_STEM	0.03	-0.055	-0.084	0.038	-0.146	0.282
se	(0.077)	(0.068)	(0.074)	(0.163)	(0.162)	(0.194)
Ugrad_ECON*Ugrad_STEM	0.017	0.171	0.192	0.074	0.216	-0.258
se	(0.092)	(0.09)	(0.107)	(0.191)	(0.2)	(0.254)

Note: Panel A shows simple regression of log salary on undergraduate major dummy separately. Panel B controls demographic, educational and year fix effect, school fix effect as being principle components. Panel C controls all information as being principle components. Both B and C are using 5 principal components as control variables.

Table B.5: Undergraduate major effects on log salary (OLS ATE)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. only dummy					
Ugrad_ECON	0.005	-0.015	-0.071	-0.038	-0.038
se	(0.033)	(0.04)	(0.037)	(0.032)	(0.039)
Ugrad_STEM	-0.024	-0.035	0.017	0.03	0.102
se	(0.03)	(0.039)	(0.036)	(0.03)	(0.034)
B. add controls W/O productivity					
ATE (Ugrad_ECON)	-0.02	-0.013	-0.046	-0.024	0.004
se	(0.031)	(0.038)	(0.033)	(0.031)	(0.037)
ATE (Ugrad_STEM)	-0.027	0.004	-0.005	0.019	0.088
se	(0.03)	(0.043)	(0.032)	(0.027)	(0.037)
C. add all controls					
ATE (Ugrad_ECON)	-0.009	0	-0.074	-0.035	-0.007
se	(0.032)	(0.04)	(0.033)	(0.032)	(0.04)
ATE (Ugrad_STEM)	-0.028	0.008	-0.013	0.027	0.074
se	(0.028)	(0.04)	(0.031)	(0.025)	(0.033)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. only dummy						
Ugrad_ECON	-0.021	-0.051	-0.133	-0.174	-0.15	-0.192
se	(0.042)	(0.048)	(0.06)	(0.094)	(0.103)	(0.169)
Ugrad_STEM	0.054	0.107	0.109	0.096	0.136	0.18
se	(0.042)	(0.053)	(0.063)	(0.11)	(0.115)	(0.122)
B. add control W/O productivity						
ATE (Ugrad_ECON)	-0.017	-0.016	-0.103	-0.19	-0.184	-0.091
se	(0.05)	(0.061)	(0.074)	(0.086)	(0.098)	(0.208)
ATE (Ugrad_STEM)	0.036	0.114	0.084	0.026	0.095	0.158
se	(0.047)	(0.059)	(0.074)	(0.107)	(0.119)	(0.139)
C. add all controls						
ATE (Ugrad_ECON)	-0.008	0.003	-0.093	-0.177	-0.143	-0.046
se	(0.048)	(0.042)	(0.05)	(0.078)	(0.082)	(0.065)
ATE (Ugrad_STEM)	0.043	0.068	0.041	0.088	0.015	0.07
se	(0.044)	(0.05)	(0.058)	(0.105)	(0.093)	(0.087)

Note: Panel A shows simple regression of log salary on undergraduate major dummy separately. Panel B controls demographic, educational and year fix effect, school fix effect as being principle components. Panel C controls all information as being principle components. Both B and C are using 5 principal components as control variables.

Table B.6: Undergraduate major effects on log salary (OLS coefficients) II

	Exp0	Exp1	Exp2	Exp3	Exp4
B. add controls W/O productivity					
Ugrad_ECON	0.011	0.007	-0.041	-0.037	0.023
se	(0.04)	(0.036)	(0.038)	(0.035)	(0.053)
Ugrad_STEM	0.052	-0.007	0.027	0.002	0.056
se	(0.054)	(0.077)	(0.06)	(0.058)	(0.06)
Ugrad_ECON*Ugrad_STEM	-0.096	0.009	-0.067	-0.001	-0.006
se	(0.062)	(0.089)	(0.071)	(0.063)	(0.072)
C. add all controls					
Ugrad_ECON	0.002	-0.002	-0.038	-0.043	0.03
se	(0.039)	(0.047)	(0.037)	(0.039)	(0.054)
Ugrad_STEM	0.013	0.012	0.067	0.022	0.125
se	(0.057)	(0.083)	(0.073)	(0.063)	(0.064)
Ugrad_ECON*Ugrad_STEM	-0.052	0.001	-0.104	0.003	-0.07
se	(0.061)	(0.09)	(0.085)	(0.068)	(0.075)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
B. add control W/O productivity						
Ugrad_ECON	-0.006	-0.044	-0.102	-0.153	-0.108	0.062
se	(0.057)	(0.057)	(0.088)	(0.101)	(0.134)	(0.103)
Ugrad_STEM	0.044	-0.011	-0.079	-0.04	0.038	0.385
se	(0.071)	(0.064)	(0.094)	(0.172)	(0.207)	(0.197)
Ugrad_ECON*Ugrad_STEM	-0.032	0.139	0.209	0.178	0.108	-0.189
se	(0.09)	(0.09)	(0.126)	(0.21)	(0.261)	(0.272)
C. add all controls						
Ugrad_ECON	-0.003	-0.05	-0.164	-0.161	-0.19	0.041
se	(0.064)	(0.055)	(0.071)	(0.09)	(0.119)	(0.132)
Ugrad_STEM	0.022	-0.07	-0.114	0.031	-0.194	0.338
se	(0.079)	(0.072)	(0.086)	(0.173)	(0.2)	(0.227)
Ugrad_ECON*Ugrad_STEM	0.006	0.185	0.199	0.037	0.241	-0.363
se	(0.092)	(0.089)	(0.118)	(0.204)	(0.248)	(0.291)

Note: Panel B controls demographic, educational and year fix effect, school fix effect as being principle components. Panel C controls all information as being principle components. Both B and C are using 10 principal components as control variables.

Table B.7: Undergraduate major effects on log salary (OLS ATE) II

		Exp0	Exp1	Exp2	Exp3	Exp4
B. add controls no productivity						
ATE (Ugrad_ECON)		-0.022	0.01	-0.06	-0.037	0.021
se		(0.033)	(0.034)	(0.033)	(0.029)	(0.039)
ATE (Ugrad_STEM)		-0.017	-0.001	-0.022	0.002	0.051
se		(0.031)	(0.041)	(0.028)	(0.029)	(0.033)
C. add all controls						
ATE (Ugrad_ECON)		-0.016	-0.003	-0.067	-0.042	0.008
se		(0.034)	(0.04)	(0.035)	(0.033)	(0.04)
ATE (Ugrad_STEM)		-0.025	0.013	-0.01	0.024	0.073
se		(0.032)	(0.043)	(0.031)	(0.03)	(0.034)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
B. add control no productivity						
ATE (Ugrad_ECON)	-0.016	-0.006	-0.043	-0.109	-0.083	0.008
se	(0.044)	(0.046)	(0.069)	(0.093)	(0.121)	(0.091)
ATE (Ugrad_STEM)	0.02	0.088	0.056	0.081	0.118	0.229
se	(0.043)	(0.05)	(0.062)	(0.095)	(0.138)	(0.102)
C. add all controls						
ATE (Ugrad_ECON)	-0.001	0.001	-0.108	-0.152	-0.133	-0.064
se	(0.05)	(0.048)	(0.057)	(0.079)	(0.093)	(0.091)
ATE (Ugrad_STEM)	0.027	0.062	0.015	0.057	-0.015	0.039
se	(0.043)	(0.051)	(0.06)	(0.1)	(0.096)	(0.095)

Note: Panel B controls demographic, educational and year fix effect, school fix effect as being principle components. Panel C controls all information as being principle components. Both B and C are using 10 principal components as control variables.

Table B.8: Gender effects on standardized salary (DML)

		Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model						
ATE		0.355	0.145	0.364	0.327	0.511
se(median)		(0.564)	(0.228)	(0.235)	(0.298)	(0.296)
se		(0.251)	(0.216)	(0.2)	(0.279)	(0.216)
B. Partial Linear Model						
ATE		0.385	0.015	0.35	0.416	0.527
se(median)		(0.181)	(0.231)	(0.174)	(0.182)	(0.168)
se		(0.166)	(0.203)	(0.163)	(0.162)	(0.162)
	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.518	0.458	0.563	0.406	0.179	0.225
se(median)	(0.337)	(0.253)	(NA)	(0.253)	(0.347)	(0.344)
se	(0.296)	(0.252)	(0.198)	(0.232)	(0.314)	(0.343)
B. Partial Linear Model						
ATE	0.524	0.413	0.386	0.321	0.168	0.22
se(median)	(0.168)	(0.165)	(0.199)	(0.217)	(0.285)	(0.293)
se	(0.144)	(0.155)	(0.186)	(0.202)	(0.256)	(0.277)

Note: "Exp" is short for experience years. "ATE" means median treatment effect estimations across splits, here it reports "best" estimation results among Trees, Random Forest and Neural network methods. how "best" are calculated is refered to Chernozhukov et al. (2018). "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. "Splits" means how many times we randomly separate the data into different (pre-setted) folds.

Table B.9: Graduate school rank effects on standardized salary (DML)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	0.164	0.219	0.137	-0.003	0.263
se(median)	(0.371)	(0.209)	(0.541)	(0.317)	(0.205)
se	(0.221)	(0.179)	(0.388)	(0.317)	(0.204)
B. Partial Linear Model					
ATE	0.684	0.546	0.529	0.421	1.068
se(median)	(0.465)	(0.476)	(0.381)	(0.495)	(0.463)
se	(0.43)	(0.404)	(0.38)	(0.356)	(0.444)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.203	0.152	0.465	0.363	0.333	0.155
se(median)	(0.196)	(0.28)	(0.245)	(0.292)	(0.274)	(0.357)
se	(0.187)	(0.26)	(0.211)	(0.29)	(0.263)	(0.355)
B. Partial Linear Model						
ATE	0.923	0.612	0.688	0.214	0.528	0.686
se(median)	(0.506)	(0.438)	(0.554)	(0.553)	(0.406)	(0.593)
se	(0.422)	(0.41)	(0.55)	(0.519)	(0.354)	(0.56)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Estimation on interactive model effects are obtained by creating a dummy variable which is denoted 1 when rank is less or equal to the rank median; 0 if rank is greater than the rank median. Estimation on partial linear model effects are obtained by creating a transformed rank variable, which is calculated by $1 - rank/max(rank)$. In this way, the rank range is between 0 and 1. Better ranked schools are assigned values close to 1, worse ranked schools are assigned values close to 0.

Table B.10: Undergraduate ECON effects on standardized salary(DML)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	-0.3	-0.113	-0.475	-0.231	0.196
se(median)	(0.791)	(0.327)	(0.279)	(0.27)	(0.377)
se	(0.791)	(0.235)	(0.246)	(0.258)	(0.304)
B. Partially Linear Model					
ATE	-0.007	0.049	-0.411	-0.338	0.132
se(median)	(0.235)	(0.243)	(0.253)	(0.235)	(0.224)
se	(0.233)	(0.211)	(0.238)	(0.232)	(0.21)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	-0.173	-0.132	-0.286	-0.41	-0.431	NA
se(median)	(0.254)	(0.274)	(0.29)	(NA)	(0.543)	(NA)
se	(0.235)	(0.272)	(0.244)	(0.31)	(0.532)	(NA)
B. Partially Linear Model						
ATE	-0.056	-0.075	-0.218	-0.489	-0.333	-0.465
se(median)	(0.23)	(0.177)	(0.22)	(0.29)	(0.307)	(0.492)
se	(0.208)	(0.176)	(0.217)	(0.286)	(0.307)	(0.469)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp8", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

Table B.11: Undergraduate STEM effects on standardized salary (DML)

	Exp0	Exp1	Exp2	Exp3	Exp4
A. Interactive Model					
ATE	-0.079	-0.031	0.062	0.198	0.547
se(median)	(0.429)	(0.311)	(0.308)	(0.261)	(0.231)
se	(0.429)	(0.308)	(0.269)	(0.25)	(0.212)
B. Partially Linear Model					
ATE	-0.08	-0.035	0.106	0.239	0.549
se(median)	(0.217)	(0.209)	(0.227)	(0.215)	(0.215)
se	(0.216)	(0.209)	(0.227)	(0.207)	(0.191)

	Exp5	Exp6	Exp7	Exp8	Exp9	Exp10
A. Interactive Model						
ATE	0.407	0.366	0.166	0.311	0.249	NA
se(median)	(0.546)	(0.437)	(0.343)	(0.566)	(0.888)	(NA)
se	(0.544)	(0.34)	(0.326)	(0.566)	(0.668)	(NA)
B. Partially Linear Model						
ATE	0.278	0.253	-0.009	0.304	0.321	0.316
se(median)	(0.209)	(0.19)	(0.229)	(0.345)	(0.335)	(0.28)
se	(0.207)	(0.187)	(0.195)	(0.341)	(0.329)	(0.254)

Note: "Exp" is short for experience years. "ATE" reports median average treatment effect across splits; "se(median)" reports median methods adjusted standard errors; "se" reports median standard error across splits. Under "Exp6", because there is not a "best" result reported, what listed is chosen from a medium value of coefficients among all reported machine learning methods. No results reported under "Exp10".

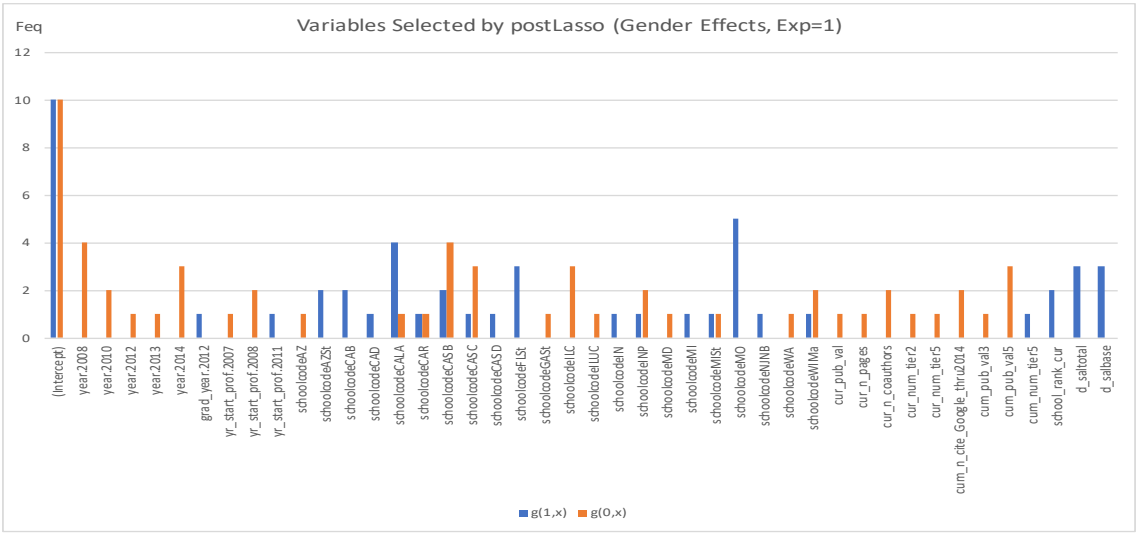


Figure B.1: Post Lasso Selected variables in interactive model: Gender Effect, EXP=1

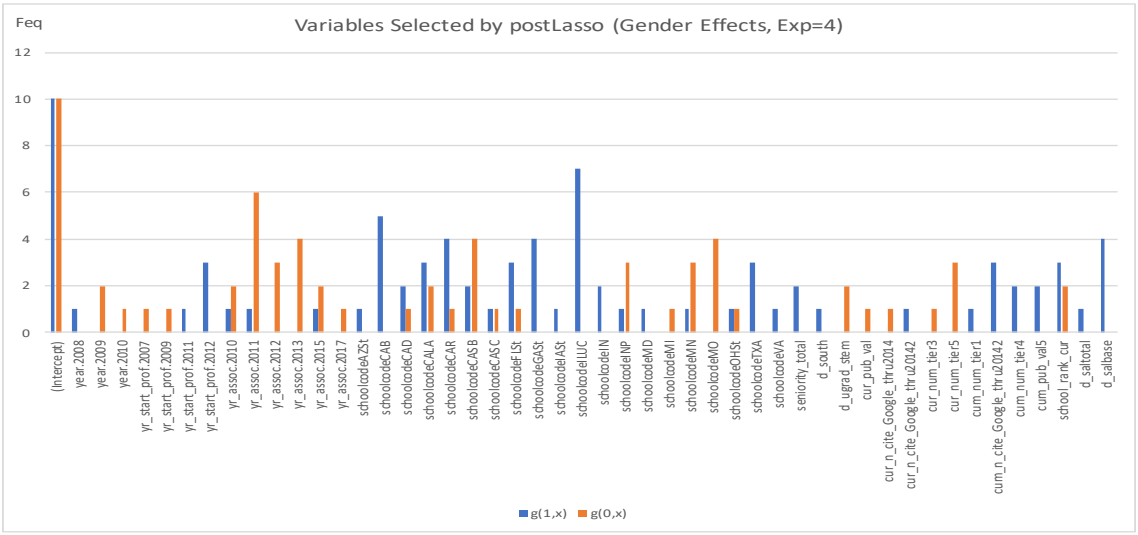


Figure B.2: Post Lasso Selected variables in interactive model: Gender Effect, EXP=4

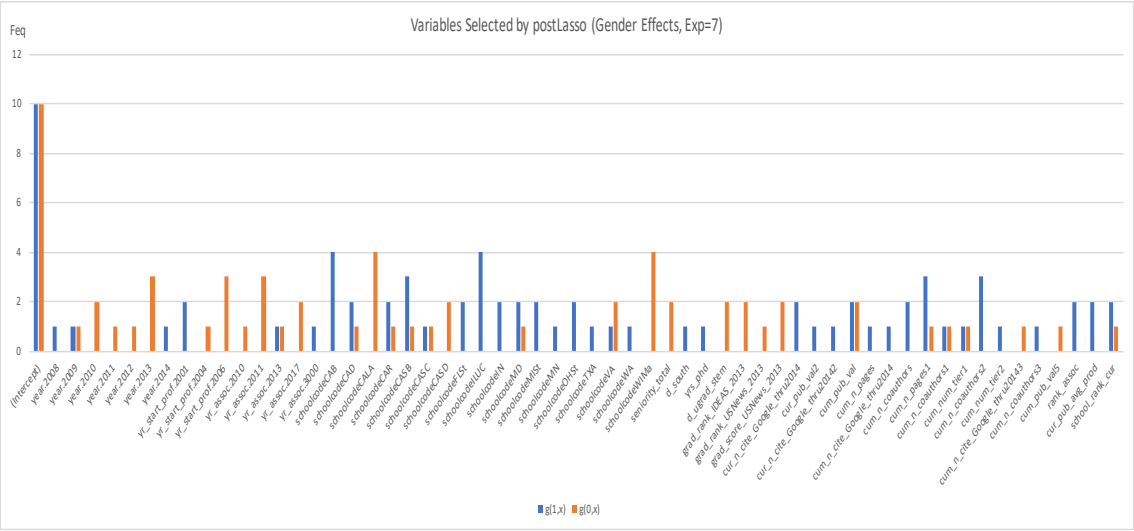


Figure B.3: Post Lasso Selected variables in interactive model: Gender Effect, EXP=7

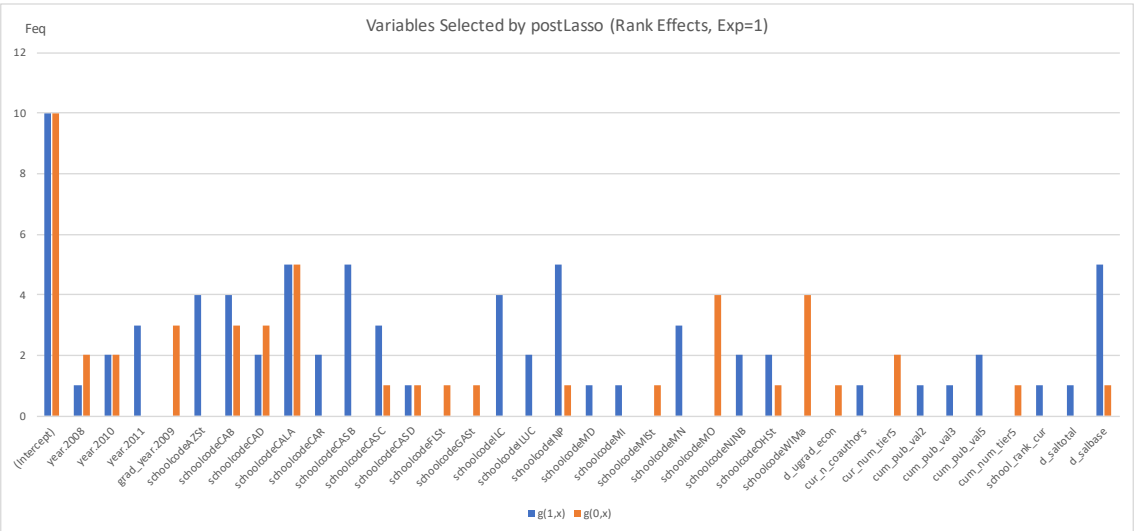


Figure B.4: Post Lasso Selected variables in interactive model: Rank Effect, EXP=1

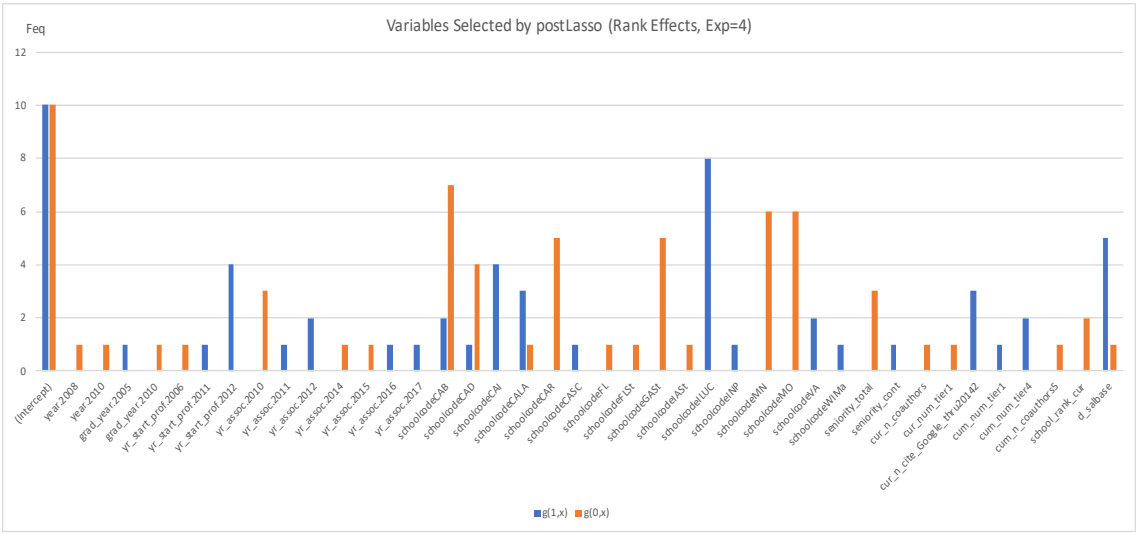


Figure B.5: Post Lasso Selected variables in interactive model: Rank Effect, EXP=4

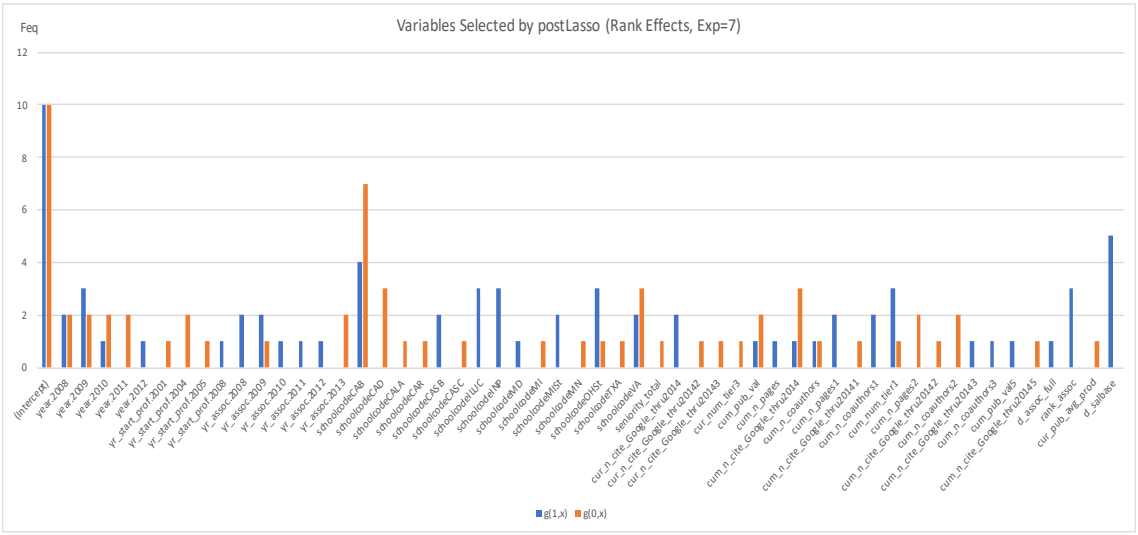


Figure B.6: Post Lasso Selected variables in interactive model: Rank Effect, EXP=7

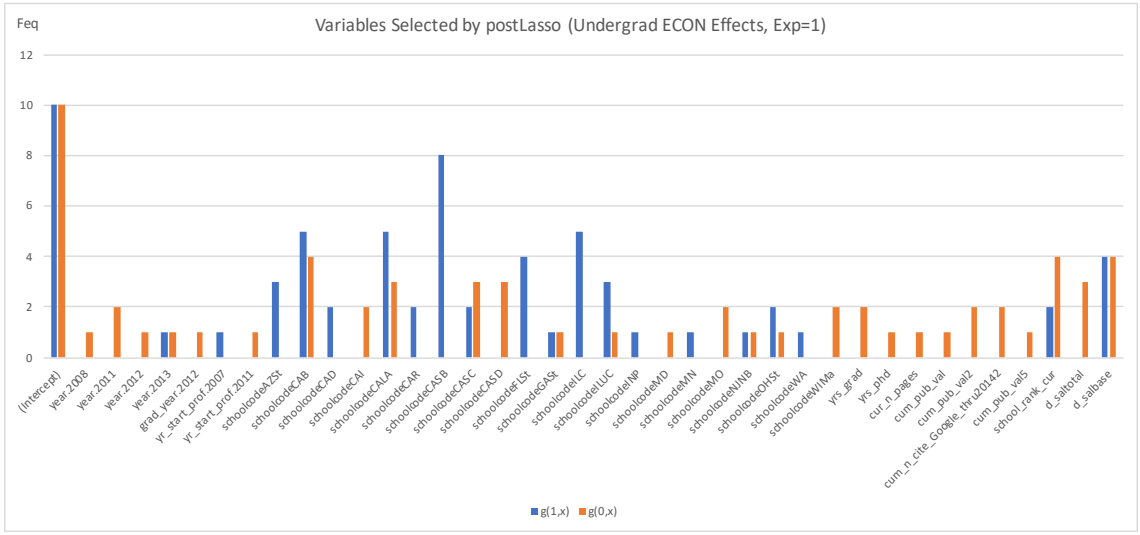


Figure B.7: Post Lasso Selected variables in interactive model: Undergraduate ECON Effect, EXP=1

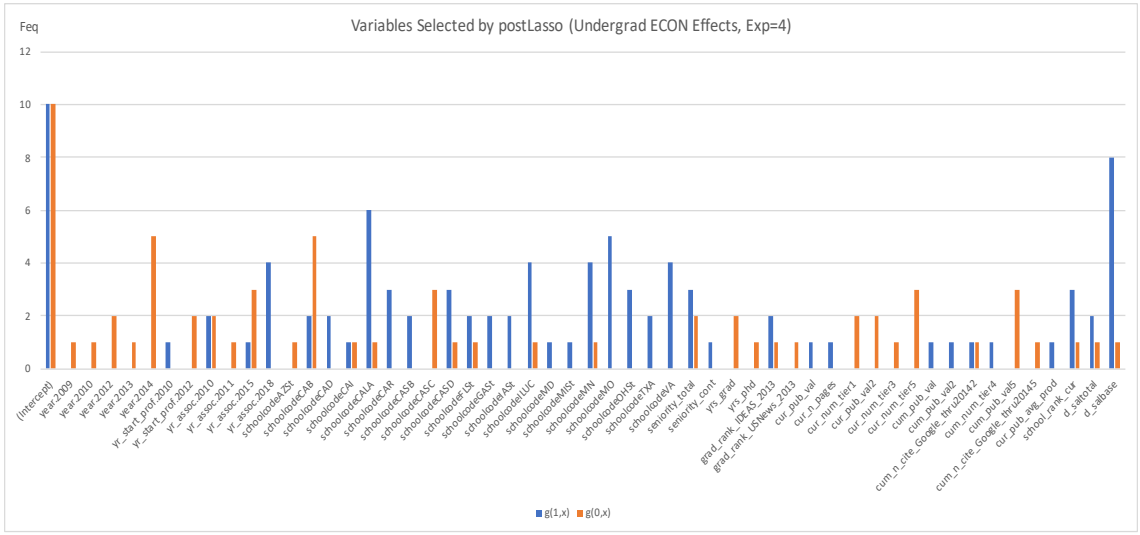


Figure B.8: Post Lasso Selected variables in interactive model: Undergraduate ECON Effect, EXP=4

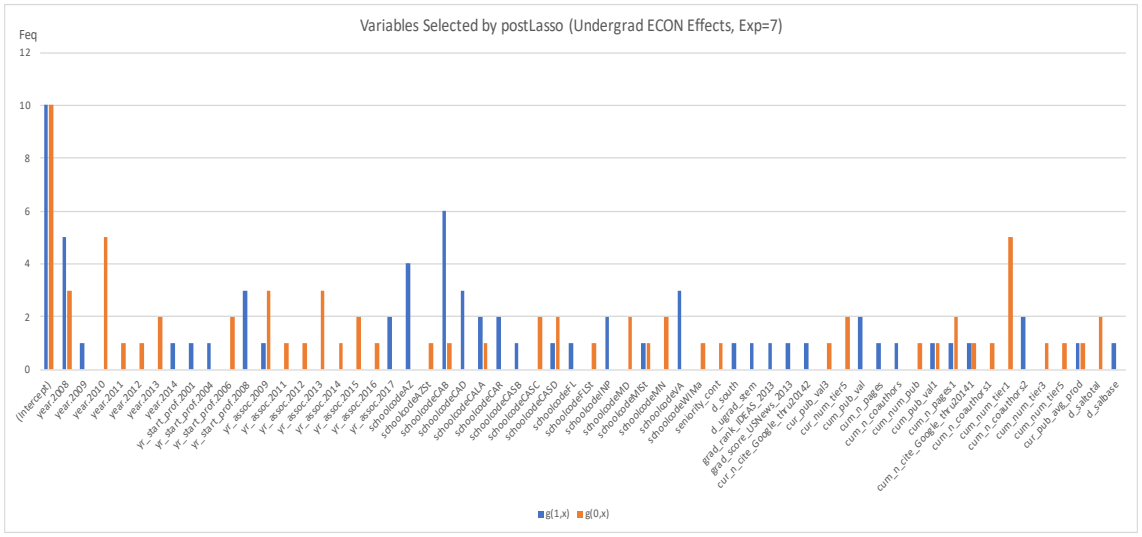


Figure B.9: Post Lasso Selected variables in interactive model: Undergraduate ECON Effect, EXP=7

Appendix C

Academic Paper Publication Value and Gender Bias Based on Text Analysis

C.1 Supplementary I: Prediction (Exclude outliers)

Table C.1 shows prediction results excluding outliers.

C.2 Supplementary II: Gender Effect (Excluding Outliers)

Considering it is usually quite hard and only a few can get their papers published into very high H-index journals, studying the rest of the papers' model performance could be insightful. From Figure 3.1, we find that no matter what the author's gender, there are some high-valued H-indices that skew the H-index distribution. The outliers that

Table C.1: H_index Prediction (exclude outliers)

SVD	Keywords	*Mean Predictor*	*Linear Regression*	*Random Forest*	*XGBoost*
200	3000	26.152	24.158	40.588	29.664
200	6000	26.547	24.639	29.964	29.779
200	9000	25.731	23.587	27.029	29.889
200	12000	25.034	22.903	25.218	28.750
200	15000	27.067	25.302	27.399	29.952
500	3000	25.174	22.879	39.801	28.318
500	6000	26.352	23.780	29.413	30.192
500	9000	26.572	25.159	28.602	29.455
500	12000	28.143	26.314	28.709	31.515
500	15000	26.444	25.020	25.969	29.971
1000	3000	25.930	24.373	42.021	29.658
1000	6000	25.371	24.312	29.922	28.434
1000	9000	27.137	25.206	28.488	30.106
1000	12000	27.121	25.124	27.214	29.426
1000	15000	26.248	24.745	26.982	28.581
1500	3000	26.339	24.449	38.064	29.674
1500	6000	26.141	23.551	30.234	29.632
1500	9000	26.419	24.839	26.851	29.422
1500	12000	26.367	24.906	27.978	28.727
1500	15000	24.029	22.363	25.828	26.910
2000	3000	25.712	23.487	40.532	29.374
2000	6000	24.919	22.817	27.503	28.135
2000	9000	24.616	22.709	25.445	28.199
2000	12000	25.471	22.692	25.573	29.859
2000	15000	26.997	25.160	27.308	29.235

show up in Figure 3.1 are all above H-index 150, and the rest are below 150. So for the rest of the papers, we re-run the gender models and show the results in Tables C.2 to C.6.

The model estimates in Tables C.2 and C.3 are mainly negative, contrary to Tables 3.5 and 3.6, which means a paper having more male authors pulls down the H-index if the extremely good cases (outliers) were removed. Even though those estimates are not statistically significant, the negative trend is quite thought-provoking. Results from dummy 2 and dummy 3 in Tables C.4 and C.5 further confirm that if the extremely good H-index cases are excluded, the mixed-gender authored papers can actually obtain higher H-indices and female contribution in a paper actually improves its H-index.

The most obvious different results are observed from dummy 3, which separates only male author(s) and female contributed author(s), those two categories. The negative OLS estimates also show an upward trend from -5.8 to -3.0 with an increase in SVD components, whereas the DML estimates varies only in a small range of -5.1 to -5.7, all statistically significant. This means that if a paper changes from female contributed author(s) to male author(s), the average H-index would decrease by (approximately) -5.3 when controlling for 15,000 keywords with 500 SVD truncated components.

C.3 Supplementary III: Gender Effect (Median)

Regression to the median is more robust to extreme values compared to regression to the mean. The median regression results complement our previous conclusion that

Table C.2: Model Results Using Numeric Gender Variable (exclude outliers)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-3.040	2.333	0.000	-	-
20	3000	-3.988	2.320	0.087	-2.912	2.237
20	6000	-3.859	2.313	0.090	-2.905	2.217
20	9000	-3.810	2.312	0.090	-2.977	2.227
20	12000	-3.850	2.310	0.090	-2.622	2.241
20	15000	-3.849	2.311	0.089	-2.632	2.230
200	3000	-2.019	2.283	0.244	-2.835	2.169
200	6000	-1.302	2.270	0.251	-2.682	2.268
200	9000	-1.230	2.264	0.253	-2.283	2.205
200	12000	-0.994	2.261	0.256	-2.954	2.200
200	15000	-1.055	2.269	0.254	-2.388	2.327
500	3000	-0.444	2.467	0.295	-2.446	2.227
500	6000	0.732	2.444	0.308	-2.267	2.274
500	9000	0.732	2.420	0.317	-2.318	2.253
500	12000	0.538	2.408	0.320	-2.459	2.308
500	15000	0.779	2.408	0.323	-2.645	2.196

Table C.3: Model Results Using Dummy 1 (Gender>0, exclude outliers)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-3.548	2.598	0.000	-	-
20	3000	-4.589	2.560	0.087	-3.434	2.732
20	6000	-4.363	2.551	0.090	-3.374	2.708
20	9000	-4.326	2.554	0.090	-3.230	2.680
20	12000	-4.385	2.553	0.090	-2.814	2.774
20	15000	-4.297	2.553	0.089	-3.194	2.735
200	3000	-3.017	2.478	0.244	-3.411	2.607
200	6000	-1.781	2.464	0.251	-3.244	2.678
200	9000	-1.882	2.464	0.253	-2.978	2.721
200	12000	-1.680	2.464	0.256	-2.888	2.786
200	15000	-1.797	2.468	0.254	-2.907	2.643
500	3000	-2.503	2.662	0.296	-3.252	2.682
500	6000	-0.525	2.637	0.308	-3.438	2.653
500	9000	-0.395	2.605	0.318	-3.336	2.678
500	12000	-0.533	2.599	0.320	-3.406	2.750
500	15000	-0.325	2.595	0.323	-3.592	2.793

Note: Dummy 1 is set to 1 when numeric Gender is greater than 0; 0 when it is equal to 0 (only female authors).

Table C.4: Model Results Using Dummy 2 (Gender=1 or Gender=0, exclude outliers)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-4.011	1.601	0.003	-	-
20	3000	-4.391	1.550	0.090	-4.025	1.721
20	6000	-4.455	1.545	0.093	-4.105	1.865
20	9000	-4.487	1.544	0.093	-4.061	1.707
20	12000	-4.469	1.545	0.092	-4.167	1.765
20	15000	-4.512	1.545	0.092	-4.140	1.703
200	3000	-3.091	1.519	0.245	-4.049	1.676
200	6000	-3.617	1.507	0.253	-3.924	1.697
200	9000	-3.552	1.505	0.255	-3.997	1.685
200	12000	-3.480	1.504	0.259	-3.959	1.660
200	15000	-3.457	1.507	0.256	-4.015	1.690
500	3000	-2.230	1.636	0.296	-3.974	1.665
500	6000	-2.744	1.612	0.310	-3.960	1.696
500	9000	-2.982	1.601	0.319	-3.876	1.716
500	12000	-3.044	1.599	0.322	-3.931	1.712
500	15000	-2.958	1.599	0.324	-4.123	1.657

Note: Dummy 2 is set to 1 when numeric Gender is equal to 0 or 1; 0 otherwise.

Table C.5: Model Results Using Dummy 3 (Gender=1, exclude outliers)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	-4.960	1.538	0.001	-	-
20	3000	-5.779	1.508	0.093	-5.733	1.577
20	6000	-5.765	1.505	0.096	-5.562	1.659
20	9000	-5.775	1.503	0.096	-5.436	1.614
20	12000	-5.776	1.503	0.095	-5.472	1.666
20	15000	-5.789	1.504	0.095	-5.689	1.640
200	3000	-4.092	1.493	0.247	-5.394	1.609
200	6000	-4.165	1.485	0.254	-5.245	1.599
200	9000	-4.123	1.480	0.256	-5.342	1.589
200	12000	-3.975	1.479	0.259	-5.219	1.566
200	15000	-4.002	1.463	0.257	-5.234	1.565
500	3000	-3.098	1.615	0.297	-5.310	1.559
500	6000	-2.887	1.597	0.310	-5.129	1.566
500	9000	-3.073	1.586	0.320	-5.267	1.573
500	12000	-3.167	1.579	0.322	-5.248	1.585
500	15000	-3.009	1.580	0.324	-5.261	1.563

Note: Dummy 3 is set to 1 when numeric gender is equal to 1 (only male authors); 0 otherwise.

Table C.6: Model Results Using Dummy 4 (Gender>0.5, exclude outliers)

		OLS			*DML*	
SVD	Keywords	Gender	SE	Adjusted R^2	Gender	SE(Median)
0	0	0.636	1.703	0.000	-	-
20	3000	0.388	1.681	0.086	-0.046	2.057
20	6000	0.462	1.676	0.089	-0.023	1.929
20	9000	0.524	1.675	0.089	0.020	2.195
20	12000	0.498	1.674	0.088	0.278	2.135
20	15000	0.464	1.674	0.087	-0.111	2.368
200	3000	0.879	1.652	0.244	0.689	1.695
200	6000	1.360	1.639	0.251	0.466	1.778
200	9000	1.383	1.635	0.253	0.546	1.860
200	12000	1.538	1.632	0.257	0.625	1.730
200	15000	1.510	1.636	0.254	0.464	1.780
500	3000	1.984	1.767	0.296	0.237	1.950
500	6000	2.330	1.760	0.309	0.384	1.743
500	9000	2.550	1.743	0.319	0.453	1.728
500	12000	2.567	1.737	0.321	0.677	1.766
500	15000	2.711	1.737	0.324	0.486	1.761

Note: Dummy 4 is set to 1 when continuous Gender is greater than 0.5; 0 otherwise.

gender has no causal influence on a paper's H-index when controlling for paper text. We find that among all the gender variables, numeric gender has a coefficient equal to zero in a median regression model with no control variables; dummy 2 and dummy 4 have zero group median differences. Comparing Table C.8 with the original gender models in Tables 3.5 to 3.9, the estimates of all these five variables from median regression models are overall more toward zero in magnitude and even with smaller standard errors are not statistically significant.

C.4 Supplementary IV: Gender Effect (controlling for number of authors and if single author)

Table C.7: Median H-Index by Gender Groups

	Dummy 1	Dummy 2	Dummy 3	Dummy 4
1	Gender>0	Gender=1,0	Gender=1	Gender>0.5
	50	50	49	50
0	Gender=0	0<Gender<1	Gender<1	Gender≤0.5
	59.5	50	53.5	50
Diff.	-9.5	0	-4.5	0

Table C.8: Quantile Regression (Median)

		Gender		*Dummy 1*		*Dummy 2*		*Dummy 3*		*Dummy 4*	
SVD	Keywords	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
0	0	0.000	4.599	-7.000	4.308	0.000	2.963	-1.000	3.006	0.000	3.167
20	3000	-0.280	1.493	0.662	1.836	-1.028	1.090	-0.861	0.844	0.283	1.106
20	6000	-0.170	1.623	0.753	1.379	-1.357	0.847	-0.994	1.037	0.713	1.056
20	9000	0.258	1.722	1.134	1.993	-1.607	0.965	-1.237	1.165	0.738	1.137
20	12000	0.247	1.385	1.097	1.993	-1.824	1.137	-1.325	1.088	0.747	1.022
20	15000	-0.115	1.546	1.067	2.129	-2.140	1.048	-1.610	1.007	0.617	1.085
200	3000	-0.072	2.342	-0.873	2.742	0.342	1.585	-0.104	1.521	0.608	1.711
200	6000	0.535	2.197	-0.232	2.389	-0.438	1.411	-0.619	1.471	1.151	1.565
200	9000	-0.323	2.050	-0.270	2.382	-1.138	1.536	-1.144	1.584	0.466	1.574
200	12000	0.495	1.908	0.084	2.048	-0.464	1.387	-0.515	1.379	0.644	1.492
200	15000	-1.076	2.386	-0.562	2.318	-1.204	1.595	-1.210	1.509	0.000	1.638
500	3000	1.486	4.772	-0.880	4.818	0.913	2.848	0.707	3.028	1.491	3.128
500	6000	2.712	3.727	-0.177	4.542	1.594	2.642	1.723	2.702	2.747	3.181
500	9000	2.471	4.568	0.686	5.041	1.777	2.937	1.573	2.848	2.402	3.348
500	12000	2.719	4.007	0.316	4.894	1.337	2.875	1.297	2.727	2.279	3.117
500	15000	2.648	4.917	0.331	5.721	1.861	2.708	1.551	3.145	2.240	3.557

Note: This median regression is conducted by using all data (keeping outliers).

Table C.9: Journals and H-index

Journal Name:	H-index	Journal Name:	H-index
Journal of Financial Economics	206	Economics & Human Biology	46
Research Policy	191	Information Economics and Policy	43
Ecological Economics	161	Journal of Economics and Business	43
International Journal of Production Economics	141	Pacific-Basin Finance Journal	43
World Development	140	Journal of International Financial Markets, Institutions and Money	42
Physica A: Statistical Mechanics and its Applications	133	Journal of Housing Economics	41
Journal of Econometrics	127	Socio-Economic Planning Sciences	41
Journal of Banking & Finance	126	Journal of Policy Modeling	40
Journal of Accounting and Economics	122	Emerging Markets Review	39
Journal of Development Economics	115	International Review of Economics & Finance	38
Journal of Public Economics	115	International Review of Financial Analysis	38
Journal of International Economics	113	Structural Change and Economic Dynamics	38
European Economic Review	110	International Review of Law and Economics	37
Energy Economics	109	Journal of Asian Economics	37
Journal of Monetary Economics	107	Journal of Macroeconomics	37
Journal of Health Economics	103	Explorations in Economic History	36
Journal of Environmental Economics and Management	101	Journal of Mathematical Economics	35
Journal of Economic Behavior & Organization	94	Journal of the Japanese and International Economies	34
Geoforum	90	Journal of Multinational Financial Management	34
Journal of Urban Economics	89	Review of Financial Economics	32
Journal of Economic Theory	84	Mathematical Social Sciences	31
Games and Economic Behavior	79	Research in Transportation Economics	30
Economics Letters	77	Economic Systems	29
Journal of Corporate Finance	77	Japan and the World Economy	28
Journal of Economic Psychology	77	Global Finance Journal	26
Journal of International Money and Finance	77	Environmental Innovation and Societal Transitions	24
Food Policy	76	Review of Radical Political Economics	23
Journal of Economic Dynamics and Control	75	Research in Economics	22
International Journal of Industrial Organization	71	Finance Research Letters	21
Journal of Comparative Economics	69	Advances in Accounting	20
Economics of Education Review	65	Economic Analysis and Policy	18
European Journal of Political Economy	65	Journal of Choice Modelling	17
Agricultural Economics	63	City, Culture and Society	16
Journal of Empirical Finance	63	Journal of Applied Economics	16
Journal of Financial Intermediation	63	Journal of Rail Transport Planning & Management	14
Regional Science and Urban Economics	63	International Review of Economics Education	13
Insurance: Mathematics and Economics	62	Economics of Transportation	12
Labour Economics	60	Journal of Contemporary Accounting & Economics	12
Resource and Energy Economics	57	Value in Health Regional Issues	11
China Economic Review	56	Water Resources and Economics	10
Review of Economic Dynamics	52	Review of Development Finance	9
Economic Modelling	50	Borsa Istanbul Review	7
Resources Policy	50	The Journal of the Economics of Ageing	7
Journal of Financial Markets	49	Development Engineering	3
Journal of Behavioral Economics	48	Econometrics and Statistics	2
Journal of Behavioral and Experimental Economics	48	Russian Journal of Economics	1

Note: Journal H-index is collected from SJR under year 2018 criterion. Website: https://www.scimagojr.com/index.php

Table C.10: Gender Effects Using Numeric Gender Variable

		OLS			*DML*	
SVD	Keywords	Gender	#Authors	Single Author	Gender	Authors
0	0	0.639	-	-	-	-
0	0	0.883	6.539***	2.551	-	-
20	3000	-0.174	-	-	1.404	-
20	3000	-0.019	5.534***	1.016	1.647	Y
20	6000	0.088	-	-	0.989	-
20	6000	0.228	5.527***	1.079	0.563	Y
20	9000	0.523	-	-	1.361	-
20	9000	0.699	5.611***	1.392	1.343	Y
20	12000	0.469	-	-	1.876	-
20	12000	0.632	5.627***	1.462	1.912	Y
20	15000	0.433	-	-	2.309	-
20	15000	0.574	5.601***	1.375	2.292	Y
200	3000	0.698	-	-	0.087	-
200	3000	0.246	2.995**	-2.890	-0.713	Y
200	6000	1.106	-	-	-1.504	-
200	6000	0.714	3.534***	-2.105	-1.031	Y
200	9000	1.236	-	-	0.610	-
200	9000	0.866	3.789***	-1.652	-1.349	Y
200	12000	1.183	-	-	-2.220	-
200	12000	0.799	3.792***	-1.342	0.221	Y
200	15000	0.950	-	-	-0.660	-
200	15000	0.545	3.777***	-1.410	-1.871	Y
500	3000	0.908	-	-	-0.029	-
500	3000	0.195	2.625*	-2.963	0.598	Y
500	6000	3.889	-	-	0.420	-
500	6000	3.189	2.975**	-2.416	1.006	Y
500	9000	3.673	-	-	1.274	-
500	9000	2.971	3.015**	-2.455	1.028	Y
500	12000	3.792	-	-	0.207	-
500	12000	3.165	3.190***	-1.912	0.731	Y
500	15000	3.770	-	-	0.711	-
500	15000	3.074	2.954**	-2.557	0.176	Y

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. "-" means not included. "Y" means yes, control for number of authors and if single author.

Table C.11: Gender Effects Using Dummy 1

		OLS			*DML*	
SVD	Keywords	Gender	#Authors	Single Author	Gender	Authors
0	0	1.809	-	-	-	-
0	0	-4.027	6.645***	1.837	-	-
20	3000	-0.411	-	-	1.366	-
20	3000	-5.732	5.722***	0.132	-2.314	Y
20	6000	0.372	-	-	1.511	-
20	6000	-5.249	5.703***	0.276	-2.058	Y
20	9000	0.624	-	-	0.590	-
20	9000	-4.891	-4.891***	0.637	-1.416	Y
20	12000	0.560	-	-	1.647	-
20	12000	-4.950	5.789***	0.698	-1.125	Y
20	15000	0.649	-	-	2.197	-
20	15000	-4.892	5.764***	0.620	-3.642	Y
200	3000	-1.054	-	-	0.759	-
200	3000	-6.372*	3.223**	-3.965	-1.953	Y
200	6000	-0.179	-	-	1.026	-
200	6000	-5.572	3.726***	-3.082	-0.671	Y
200	9000	0.015	-	-	0.714	-
200	9000	-5.372	3.972***	-2.600	-0.373	Y
200	12000	0.081	-	-	0.892	-
200	12000	-5.159	3.967***	-2.262	-0.044	Y
200	15000	0.027	-	-	1.314	-
200	15000	-5.272	3.959***	-2.341	0.090	Y
500	3000	-2.241	-	-	1.184	-
500	3000	-7.639*	2.953**	-4.243	0.767	Y
500	6000	0.809	-	-	1.537	-
500	6000	-4.284	3.135**	-3.383	-0.260	Y
500	9000	1.157	-	-	1.229	-
500	9000	-3.892	3.151**	-3.362	0.358	Y
500	12000	1.566	-	-	1.181	-
500	12000	-3.281	3.297**	-2.731	-0.006	Y
500	15000	1.359	-	-	1.435	-
500	15000	-3.638	3.071**	-3.460	-0.865	Y

Note: Dummy 1 is set to 1 when numeric Gender is greater than 0; 0 when it is equal to 0 (only female authors). *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. "-" means not included. "Y" means yes, control for number of authors and if single author.

Table C.12: Gender Effects Using Dummy 2

		OLS			*DML*		
SVD	Keywords	Gender	#Authors	Single Author	Gender	Authors	
0	0	-3.099	-	-	-	-	-
0	0	3.236	7.042***	1.948	-	-	-
20	3000	-2.872	-	-	-2.790	-	-
20	3000	2.782	5.973***	0.502	-5.703	Y	Y
20	6000	-2.989	-	-	-3.062	-	-
20	6000	2.607	5.939***	0.598	-5.358	Y	Y
20	9000	-2.832	-	-	-3.181	-	-
20	9000	2.792	6.049***	0.870	-1.834	Y	Y
20	12000	-2.823	-	-	-2.894	-	-
20	12000	2.781	6.064***	0.944	-1.869	Y	Y
20	15000	-2.891	-	-	-2.864	-	-
20	15000	2.706	6.028***	0.875	-1.608	Y	Y
200	3000	-0.322	-	-	-2.928	-	-
200	3000	4.273*	3.668**	-3.593	-1.784	Y	Y
200	6000	-0.802	-	-	-3.031	-	-
200	6000	4.054*	4.186***	-2.774	1.900	Y	Y
200	9000	-0.934	-	-	-2.890	-	-
200	9000	3.974*	4.419***	-2.332	5.194	Y	Y
200	12000	-1.026	-	-	-2.943	-	-
200	12000	3.698	4.377***	-1.967	-1.188	Y	Y
200	15000	-1.180	-	-	-3.120	-	-
200	15000	3.525	4.339***	-1.989	-3.607	Y	Y
500	3000	1.216	-	-	-2.977	-	-
500	3000	5.701**	3.568**	-3.759	-1.371	Y	Y
500	6000	1.621	-	-	-2.891	-	-
500	6000	6.398***	4.013***	-3.562	3.377	Y	Y
500	9000	1.235	-	-	-2.084	-	-
500	9000	5.895**	3.969***	-3.516	-0.108	Y	Y
500	12000	1.041	-	-	-3.074	-	-
500	12000	5.633**	4.107***	-2.942	-1.012	Y	Y
500	15000	1.199	-	-	-3.049	-	-
500	15000	5.780**	3.900***	-3.596	2.682	Y	Y

Note: Dummy 2 is set to 1 when numeric Gender is equal to 0 or 1; 0 otherwise. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. "-" means not included. "Y" means yes, control for number of authors and if single author.

Table C.13: Gender Effects Using Dummy 3

		OLS			*DML*	
SVD	Keywords	Gender	#Authors	Single Author	Gender	Authors
0	0	-2.258	-	-	-	-
0	0	1.210	6.682***	2.501	-	-
20	3000	-2.750	-	-	-2.809	-
20	3000	0.364	5.580***	1.005	-2.388	Y
20	6000	-2.724	-	-	-2.943	-
20	6000	0.376	5.572***	1.063	-2.718	Y
20	9000	-2.488	-	-	-2.812	-
20	9000	0.644	5.685***	1.355	-2.065	Y
20	12000	-2.499	-	-	-2.698	-
20	12000	0.615	5.698***	1.428	-2.093	Y
20	15000	-2.532	-	-	-2.344	-
20	15000	0.572	5.668***	1.344	-2.444	Y
200	3000	-0.682	-	-	-2.534	-
200	3000	1.427	3.166**	-2.897	-2.666	Y
200	6000	-0.843	-	-	-2.593	-
200	6000	1.510	3.718***	-2.128	-2.756	Y
200	9000	-0.899	-	-	-2.652	-
200	9000	1.508	3.970***	-1.689	-2.392	Y
200	12000	-0.967	-	-	-2.389	-
200	12000	1.357	3.954**	-1.373	-2.437	Y
200	15000	-1.136	-	-	-2.487	-
200	15000	1.182	3.920**	-1.426	-2.54	Y
500	3000	0.398	-	-	-2.355	-
500	3000	2.263	2.901**	-2.913	-2.314	Y
500	6000	1.855	-	-	-2.530	-
500	6000	3.897*	3.440**	-2.531	-2.435	Y
500	9000	1.594	-	-	-2.455	-
500	9000	3.599*	3.446**	-2.556	-2.221	Y
500	12000	1.549	-	-	-2.333	-
500	12000	3.580	3.623**	-2.022	-2.465	Y
500	15000	1.631	-	-	-2.507	-
500	15000	3.601*	3.392**	-2.649	-2.583	Y

Note: Dummy 3 is set to 1 when numeric gender is equal to 1 (only male authors); 0 otherwise. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. "-" means not included. "Y" means yes, control for number of authors and if single author.

Table C.14: Gender Effects Using Dummy 4

		OLS			*DML*	
SVD	Keywords	Gender	#Authors	Single Author	Gender	Authors
0	0	2.024	-	-	-	-
0	0	-	-	-	-	-
20	3000	1.771	-	-	1.486	-
20	3000	0.797	5.468***	0.872	1.369	Y
20	6000	1.950	-	-	1.733	-
20	6000	0.976	5.443***	0.897	1.387	Y
20	9000	2.267	-	-	1.486	-
20	9000	1.291	5.497***	1.140	1.863	Y
20	12000	2.206	-	-	2.032	-
20	12000	1.218	5.519***	1.225	1.660	Y
20	15000	2.139	-	-	1.727	-
20	15000	1.144	5.501***	1.154	1.745	Y
200	3000	1.101	-	-	1.847	-
200	3000	0.387	2.960**	-2.959	1.743	Y
200	6000	1.292	-	-	1.563	-
200	6000	0.505	3.486**	-2.215	1.445	Y
200	9000	1.521	-	-	1.612	-
200	9000	0.696	3.725***	-1.798	2.279	Y
200	12000	1.546	-	-	1.699	-
200	12000	0.701	3.728***	-1.483	1.955	Y
200	15000	1.416	-	-	1.748	-
200	15000	0.552	3.728***	-1.517	2.183	Y
500	3000	0.572	-	-	1.779	-
500	3000	-0.284	2.648*	-2.943	1.834	Y
500	6000	2.282	-	-	1.811	-
500	6000	1.382	2.841**	-2.797	1.688	Y
500	9000	1.996	-	-	1.618	-
500	9000	1.123	2.903**	-2.798	1.813	Y
500	12000	2.172	-	-	1.778	-
500	12000	1.306	3.059**	-2.296	1.903	Y
500	15000	2.103	-	-	1.829	-
500	15000	1.259	2.825*	-2.936	1.920	Y

Note: Dummy 4 is set to 1 when continuous Gender is greater than 0.5; 0 otherwise. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. "-" means not included. "Y" means yes, control for number of authors and if single author.

Bibliography

Adams, Jonathan. 2009. "The use of bibliometrics to measure research quality in UK higher education institutions." *Archivum Immunologiae et Therapiae Experimentalis* 57 (9).

Altonji, Joseph G. 2005. "Employer learning, statistical discrimination and occupational attainment." *American Economic Review* 95 (2):112–117.

Altonji, Joseph G and Charles R Pierret. 2001. "Employer Learning and Statistical Discrimination." *Quarterly Journal of Economics* 116 (1):313–350.

Angrist, Joshua D. and Jörn Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Economics Books: Princeton University Press.

Barbezat, Debra A. and Michael R. Donihue. 1998. "Do faculty salaries rise with job seniority?" *Economics Letters* 58 (2):239–244.

Barrios, Federico, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. "Variations of the Similarity Function of TextRank for Automated Summarization." *CoRR* URL `http://arxiv.org/abs/1602.03606`.

Belloni, Alexandre and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19 (2):521–547.

Biau, Gerard. 2012. "Analysis of a Random Forests Model." *Journal of Machine Learning Research* 2012 (13):1063–1095.

Blackaby, David, Alison L. Booth, and Jeff Frank. 2005. "Outside Offers and the Gender Pay Gap: Empirical Evidence from the UK Academic Labour Market." *The Economic Journal* 115 (501):F81–F107. URL `https://www.jstor.org/stable/3590464`.

Bonzi, Susan. 1992. "Trends in research productivity among senior faculty." *Information Processing & Management* 28:111–120.

Bornmann, Lutz and Hans-Dieter Daniel. 2007. "What do we know about the h index?" *Journal of the American Society for Information Science and Technology* 58 (9):1381–1385. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20609`.

Bottou, Léon. 2010. "Large-Scale Machine Learning with Stochastic Gradient Descent." In *Proceedings of COMPSTAT'2010*, edited by Yves Lechevallier and Gilbert Saporta. Heidelberg: Physica-Verlag HD, 177–186.

Bratsberg, Bernt, James F Ragan, Jr., and John T Warren. 2003. "Negative returns to seniority: New evidence in academic markets." *Industrial and Labor Relations Review* 56 (2):306–323.

Breiman, Leo. 1996. "Bagging predictors." *Machine Learning* 24 (2):123–140.

———. 2001. "Random Forests." *Machine Learning* 45 (1):5–32.

Carlin, Paul S., Michael P. Kidd, Patrick M. Rooney, and Brian Denton. 2013. "Academic Wage Structure by Gender: The Roles of Peer Review, Performance, and Market Forces." *Southern Economic Journal* 80 (1):127–146. URL `https://onlinelibrary.wiley.com/doi/abs/10.4284/0038-4038-2010.267`.

Chen, Jihui, Myongjin Kim, and Qihong Liu. 2008. "Gender Gap in Tenure & Promotion: Evidence from the Economics Ph. D. Class of 2008." *D. Class of* .

Chen, Tianqi and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *CoRR* abs/1603.02754. URL `http://arxiv.org/abs/1603.02754`.

Chen, Xiaohong. 2007. "Large sample sieve estimation of semi-nonparametric models." *Handbook of Econometrics* 6B:5549–5632.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review* 107 (5):261–65.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21 (1):C1–C68. URL `https://arxiv.org/abs/1608.00060`. ArXiv:1608.00060.

Claypool, Vicki Hesli, Brian David Janssen, Dongkyu Kim, and Sara McLaughlin Mitchell. 2017. "Determinants of Salary Dispersion among Political Science Faculty: The Differential Effects of Where You Work (Institutional Characteristics) and

What You Do (Negotiate and Publish)." *Political Science and Politics* 50 (1):146–156.

Collobert, Ronan and Jason Weston. 2008. "A unified architecture for natural language processing: Deep neural networks with multitask learning." *Proceedings of the 25th International Conference on Machine Learning* :160–167.

dos Santos, Cícero and Maíra Gatti. 2014. "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 69–78. URL `https://www.aclweb.org/anthology/C14-1008`.

Ehrenberg, Ronald G. 2002. "Studying ourselves: The academic labor market." Tech. rep., National Bureau of Economic Research.

Ehrenberg, Ronald G, Paul J Pieper, and Rachel A Willis. 1998. "Do economics departments with lower tenure probabilities pay higher faculty salaries?" *Review of Economics and Statistics* 80 (4):503–512.

Formby, John P and Gary A Hoover. 2002. "Salary Determinants of Entry-Level Academic Economists and the Charactertics of Those Hired on the Tenure Track." *Eastern Economic Journal* 28 (4):509–522.

Frank, Mario and Joachim M. Buhmann. 2011. "Selecting the rank of SVD by Maximum Approximation Capacity." *CoRR* abs/1102.3176. URL `http://arxiv.org/abs/1102.3176`.

Frey, Bruno S. and Katja Rost. 2010. "Do Rankings Reflect Research Quality?" *Journal of Applied Economics* 13 (1):1–38.

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38:67–378.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as Data." *Journal of Economic Literature* 57 (3):535–74.

Gerhart, Barry A. and Sara L. Rynes. 1991. "Determinants and consequences of salary negotiations by male and female MBA graduates." *Journal of Applied Psychology* 76 (2):256–262.

Ghosh, Pallab and Zexuan Liu. 2020. "Coauthorship and the gender gap in top economics journal publications." *Applied Economics Letters* 27 (7):580–590. URL `https://doi.org/10.1080/13504851.2019.1644420`.

Gibbons, Jean D. and Mary Fish. 1991. "Rankings of Economics Faculties and Representation on Editorial Boards of Top Journals." *The Journal of Economic Education* 22 (4):361–372.

Gibson, John, David L. Anderson, and John Tressler. 2017. "Citations or Journal Quality: Which is Rewarded More in the Academic Labor Market?" *Economic Inquiry* 55 (4):1945–1965.

Ginther, Donna K. and Kathy J. Hayes. 2003. "Gender Differences in Salary and Promotion for Faculty in the Humanities 1977-95." *The Journal of Human Resources* 38 (1):34–73.

Goyal, Priya, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour." *arXiv e-prints* :arXiv:1706.02677.

Graves, Philip E., James R. Marchand, and Randall Thompson. 1982. "Economics Departmental Rankings: Research Incentives, Constraints and Efficiency." *The American Economic Review* 72 (5):1131–1141. URL `https://www.jstor.org/stable/1812028`.

Grove, Wayne A. and Stephen Wu. 2007. "The Search for Economics Talent: doctoral Completion and Research Productivity." *American Economic Review* 97 (2).

Hadar, Linor L. 2017. "Opportunities to learn: Mathematics textbooks and students' achievements." *Studies in Educational Evaluation* 55:153–166. URL `https://www.sciencedirect.com/science/article/pii/S0191491X17300949`.

Hamermesh, Daniel S, George E Johnson, and Burton A Weisbrod. 1982. "Scholarship, citations and salaries: Economic rewards in economics." *Southern Economic Journal* 49 (2):472–481.

Hilmer, Michael J., Michael R. Ransom, and Christiana E. Hilmer. 2015. "Fame and the fortune of academic economists: How the market rewards influential research in economics." *Southern Economic Journal* 82 (2):430–452.

Hirsch, J. E. 2007. "Does the h index have predictive power?" *Proceedings of the National Academy of Sciences* 104 (49):19193–19198. URL `https://www.pnas.org/content/104/49/19193`.

Hu, Baotian, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. "Convolutional Neural Network Architectures for Matching Natural Language Sentences." In *Advances in Neural Information Processing Systems 27*. 2042–2050. URL http://papers.nips.cc/paper/5550-convolutional-neural-network-architectures-for-matching-natural-language-sentences.pdf.

Johnson, George E and Frank P Stafford. 1974. "The earnings and promotion of women faculty." *American Economic Review* 64 (6):888–903.

Jones, Adam, Peter Schuhmann, Daniel Soques, and Allison Witman. 2020. "So you want to go to graduate school? Factors that influence admissions to economics PhD programs." *The Journal of Economic Education* 51 (2):177–190. URL https://doi.org/10.1080/00220485.2020.1731385.

Kahn, Lisa B and Fabian Lange. 2014. "Employer learning, productivity, and the earnings distribution: Evidence from performance measures." *Review of Economic Studies* 81 (4):1575–1613.

Katz, David A. 1973. "Faculty salaries, promotions, and productivity at a large university." *American Economic Review* 63 (3):469–477.

Kidd, Michael P., Nigel O'Leary, and Peter Sloane. 2017. "The impact of mobility on early career earnings: A quantile regression approach for UK graduates." *Economic Modelling* 62:90 – 102. URL http://www.sciencedirect.com/science/article/pii/S0264999317300858.

Koedel, Cory, Diyi Li, Morgan Polikoff, Tenice Hardaway, and Stephani Wrabel. 2017.

"Mathematics Curriculum Effects on Student Achievement in California." *AERA Open* 3 (1):1–22.

Lang, Mark and Lorien Stice-Lawrence. 2015. "Textual analysis and international financial reporting: Large sample evidence." *Journal of Accounting and Economics* 60 (2):110–135. URL https://www.sciencedirect.com/science/article/pii/S0165410115000658.

Laurent, C., G. Pereyra, P. Brakel, Y. Zhang, and Y. Bengio. 2016. "Batch normalized recurrent neural networks." In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2657–2661.

Li, Diyi and Cory Koedel. 2017. "Representation and Salary Gaps by Race-Ethnicity and Gender at Selective Public Universities." *Educational Researcher* 46 (7):343–354.

Liebowitz, S. J. and J. P. Palmer. 1984. "Assessing the Relative Impacts of Economics Journals." *Journal of Economic Literature* 22 (1):77–88.

Loughran, Tim and Bill McDonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (4):1187–1230. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12123.

Lutz, Catherine. 1990. "The erasure of women's writing in sociocultural anthropology." *Journal of the American Ethnological Society* 17:611–627.

McElhinny, BONNIE, MARIJKE Hols, JEFF Holtzkener, SUSANNE Unger, and CLAIRE Hicks. 2003. "Gender, publication and citation in sociolinguistics and lin-

guistic anthropology: The construction of a scholarly canon." *Language in Society* 32 (3):299–328.

Mikolov, Tomas, Martin Karafiát, Lukás Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. "Recurrent neural network based language model." *Interspeech* :1045–1048URL `http://dblp.uni-trier.de/db/conf/interspeech/interspeech2010.html#MikolovKBCK10`.

Mitchell, Sara McLaughlin and Vicki L. Hesli. 2013. "Women don't ask? Women don't say no? Bargaining and Service in the Political Science Profession." *Political Science & Politics* 46 (2013):355–369.

Moed, Henk F. 2008. "UK Research Assessment Exercises: Informed judgments on research quality or quantity?" *Scientometrics* :153–161.

Moore, William J, Robert J Newman, and Geoffrey K Turnbull. 1998. "Do academic salaries decline with seniority?" *Journal of Labor Economics* 16 (2):352–366.

Nakhaie, M. Reza. 2008. "Gender Differences in Publication among University Professors in Canada." *Canadian Review of Sociology* 39:151–179.

Oberfichtner, Michael, Claus Schnabel, and Marina Töpfer. 2020. "Do unions and works councils really dampen the gender pay gap? Discordant evidence from Germany." *Economics Letters* 196:109509. URL `http://www.sciencedirect.com/science/article/pii/S0165176520303116`.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. "The PageRank Citation Ranking: Bringing Order to the Web." Technical Report 1999-

66, Stanford InfoLab. URL http://ilpubs.stanford.edu:8090/422/. Previous number = SIDL-WP-1999-0120.

Perna, Laura W. 2001. "Sex Differences in Faculty Salaries: A Cohort Analysis." *The Review of Higher Education* 24 (3):283–307.

Ramos, Juan et al. 2003. "Using tf-idf to determine word relevance in document queries." In *Proceedings of the first instructional conference on machine learning*, vol. 242. Citeseer, 29–48.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1):41–55. URL http://www.jstor.org/stable/2335942.

Sak, H., Andrew Senior, and F. Beaufays. 2014. "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* :338–342.

Sax, Linda J., Linda Serra Hagedorn, Marisol Arredondo, and Frank A. Dicrisi. 2002. "Faculty Research Productivity: Exploring the Role of Gender and Family-Related Factors." *Research in Higher Education* 43 (4):423–446.

Siegfried, John J. and Kenneth J. White. 1978. "Teaching Ability as a Determinant of Faculty Salaries." *The Journal of Economic Education* 9 (2):130–132. URL http://www.jstor.org/stable/1182107.

Smart, John C. 1991. "Gender Equity in Academic Rank and Salary." *The Review of Higher Education* 14 (4):511–525.

Sokolova, Marina, Nathalie Japkowicz, and Stan Szpakowicz. 2006. "Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation." In *AI 2006: Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 1015–1021.

Stock, Wendy A. and Richard M. Alston. 2000. "Effect of Graduate Program Rank on Success in the Job Market." *The Journal of Economic Education* 31 (4):389–401.

Tamblyn, Robyn, Nadyne Girard, Christina J Qian, and James Hanley. 2018. "Assessment of potential bias in research grant peer review in Canada." *CMAJ* 190 (16):E489–E499.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1555–1565. URL `https://www.aclweb.org/anthology/P14-1146`.

Taylor, Jim. 2011. "The Assessment of Research Quality in UK Universities: Peer Review or Metrics?" *British Journal of management* 22:202–217.

Tibshirani, Robert J. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58 (1):267–288.

Toutkoushian, Robert K., Marcia L. Bellas, and John V. Moore. 2007. "The Interaction Effects of Gender, Race, and Marital Status on Faculty Salaries." *The Journal of Higher Education* 78 (5):572–601.

Toutkoushian, Robert K. and Valerie Martin Conley. 2005. "Progress for Women in Academe, Yet Inequities Persist: Evidence from NSOPF: 99." *Research in Higher Education* 46 (1):1–28. URL http://www.jstor.org/stable/40197383.

Toutkoushian, Robert K. and Emily P. Hoffman. 2002. "Alternatives for Measuring the Unexplained Wage Gap." *New Directions for Institutional Research* :71–89.

van den Ham, Ann-Katrin and Aiso Heinze. 2018. "Does the textbook matter? Longitudinal effects of textbook choice on primary school students' achievement in mathematics." *Studies in Educational Evaluation* 59:133–140. URL https://www.sciencedirect.com/science/article/pii/S0191491X18301007.

Veugelers, Reinhilde and Jian Wang. 2019. "Scientific novelty and technological impact." *Research Policy* 48 (6):1362–1372. URL https://www.sciencedirect.com/science/article/pii/S0048733319300459.

Wang, Jian, Reinhilde Veugelers, and Paula Stephan. 2017. "Bias against novelty in science: A cautionary tale for users of bibliometric indicators." *Research Policy* 46 (8):1416–1436. URL https://www.sciencedirect.com/science/article/pii/S0048733317301038.

Wang, Xin, Yuanchao Liu, Chengjie Sun, Baoxun Wang, and Xiaolong Wang. 2015. "Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1343–1353. URL http://aclweb.org/anthology/P/P15/P15-1130.pdf.

Wei, Fusheng, Han Qin, Shi Ye, and Haozhen Zhao. 2019. "Empirical Study of Deep Learning for Text Classification in Legal Document Review." *CoRR* abs/1904.01723. URL http://arxiv.org/abs/1904.01723.

Weisshaar, Katherine. 2017. "Publish and Perish? An Assessment of Gender Gaps in Promotion to Tenure in Academia." *Social Forces* 96 (2):529–560. URL https://doi.org/10.1093/sf/sox052.

Westergaard, David, Hans-Henrik Starfeldt, Christian Tønsberg, Lars Juhl Jensen, and Søren Brunak. 2018. "A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts." *PLOS Computational Biology* 14 (2):1–16. URL https://doi.org/10.1371/journal.pcbi.1005962.

Witteman, Holly O, Michael Hendricks, Sharon Straus, and Cara Tannenbaum. 2019. "Gender bias in CIHR Foundation grant awarding." *The Lancet* 394:E41–E42.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2nd ed.

Wu, Ho Chung, Robert Luk, Kam-Fai Wong, and Kui-Lam Kwok. 2008. "Interpreting TF-IDF term weights as making relevance decisions." *ACM Trans. Inf. Syst.* 26.

Yang, Lijing and Karen L. Webber. 2015. "A decade beyond the doctorate: the influence of a US postdoctoral appointment on faculty career, productivity, and salary." *Higher Education* 70 (4):667–687.

Zhang, Liang. 2005. "Advance to Graduate Education: the Effect of College Quality and Undergraduate Majors." *The Review of Higher Education* 28 (3):313–338.

Zhang, Terrence Y. 2018. "Does a 'Gender Wage Gap' Exist at the University of Florida?" URL http://dx.doi.org/10.2139/ssrn.3305473.

VITA

I was born in Inner Mongolia, China. In 2009, I was admitted to Central University of Finance and Economics, major in statistics. After 4 years' college time at Beijing, in 2003, I went to Hong Kong to pursue my Master's Degree in Economics at the Chinese University of Hong Kong. In 2015, I went to University of Missouri–Columbia for pursuing my PhD degree, started my doctoral research under Dr. David Kaplan's supervision in 2017 and got my PhD degree in 2021.