Detecting Gender Salary Discrimination Using Multiple Regression

Jiayi Ji

Abstract

I assessed the relationship between gender and salary using a case from the United States District Court of Houston. Several multiple regression models were conducted and yielded convergent finding: male and female doctors were equivalently paid taking account of other productivity-related attributes. While these results should be reassuring, they would not go very far toward resolving salary discrimination disputes. The difficulties of employing statistical analyses within a politically-charged arena are discussed.

Introduction

Is gender-pay discrimination a reality or a myth? It is the purpose of the paper to addresses the question by investigating a case from the United States District Court of Houston a few years ago that arises under Title VII of the Civil Rights Act of 1964, 42 U.S.C. 200e et seq. The plaintiffs in this case were all female doctors at Houston College of Medicine who claimed that the College had engaged in a pattern and practice of discrimination against women in giving promotions and setting salaries.

Methods

Data

The dataset comprises information on 261 employees of Houston College of Medicine. The variables included in the analyses are: 1) salary in in academic year 1994 and 1995, 2) gender, 3) department, 4) Board certified or not, 5) primarily clinical/research emphasis, 6) publication rate 7) years since obtaining MD, 8) rank. Since the two repeated measures in salary are highly collinear with a correlation 0.99, I calculated the amount increase in salary from 1994 to 1995. It was then log transformed as the response variable in our analysis to render its distribution more symmetric, which is frequently used in economics. The transformation I use is also align with the general observation that annual increases in salary tend to be calculated as a percentage of current salary.[1]

Data Analysis

Similar to previous literature,[1,2] multiple regression analysis is used to establish whether females have been disadvantaged in our case. Multiple regression analysis is a methodology that allows all the relevant factors to the wage determination process to be considered simultaneously. Ideally, it allows for estimation of the relationship between gender and salary increase after accounting for other factors presumably related to wages, such as qualifications or productivity levels. For our model, the estimated parameter for gender is interpreted as the log ratio of salaries comparing males and females adjusting for other covariates; hence, a significant positive value indicates evidence of bias against females. Model fitting was evaluated by R^2 values, AIC and model diagnosis.

One major problem that can affect estimation in the employment discrimination context is multicollinearity.[1] Multicollinearity refers to the situation in which some of the factors used to explain salary are interrelated. Variance inflation factor (VIF) is used to detect multicollinearity. Generally, VIFs larger than 10 indicate a multicollinearity problem.[3] Problematic variables are isolated to reduce the effect of multicollinearity.

Results

There are 106 female doctors and 155 male doctors in our dataset. Table 1 presents descriptive statistics of continuous variables used in the empirical analysis. The statistics indicate a gender wage differential of around 60,000 US dollars.

The multiple linear regression using all the covariates shows that all the variables except for gender and publication rate are significantly associated with salary increase (Table 3). VIF is then computed for each predictor to detect collinear. In this case, VIF associated with publication rate suggests a serious multicollinearity problem. The variance of publication rate is nearly 16 times as large as it would be if there were no multicollinearity present (Table 2). Therefore, the variable publication rate is dropped. The new model without it fit better based on AIC and the diagnostics also indicate that underlying assumptions are met, and no additional terms are needed (Table 3, Figure 1 & 2). There are no further estimation problems since VIFs for all variables in the equation have been reduced to a level that is no longer of concern (Table 2). In the new model without publication rate, the coefficient for gender is found not to be significantly different from 0 (p=0.857).

Table 1: Descriptive statistics of continuous variables by gender

Gender	Male			Female		
Statistic	Mean	Median	St. Dev.	Mean	Median	St. Dev.
Publication Rate	4.646	4.000	1.938	5.350	5.250	1.886
Experience	12.1	10.0	6.703	7.491	7.000	4.166
Salary in 1994	177,339	155,006	85,930.54	118,871	108,457	56,168.01
Salary in 1995	194,914	170,967	94,902.73	130,877	119,135	62,034.51

Table 2: VIFs for full model and model without publication rate

	Physiology	Genetics	Pediatrics	Medicine	Surgery	Male
Full Model	1.607	1.629	4.283	6.380	7.216	1.444
No Prate	1.607	1.439	1.895	2.704	2.393	1.356
	Clinical	certified	Publication Rate	Experience	Associate Prof	Full Prof
Full Model	5.878	1.330	16.626	1.885	1.508	2.226
No Prate	1.659	1.328	–	1.852	1.500	2.210

Discussions

To cure multicollinearity in the full model, I dropped the insignificant variable publication rate because it is a proxy to represent productivity and experience, which have been directly measured by experience and rank in the model. The model without publication rate as a covariate fit better than the full model based on AIC and multicollinearity is no longer a problem now. Therefore, the model without publication rate is preferred and I based my conclusion on the final model.

Table 3: Multiple regression estimates and standard errors for 3 models

	Full Model	Model Without Prate	Model Without Prate and Rank
Gender:Male	-0.011 (0.038)	-0.007 (0.037)	0.041 (0.037)
Dept:Physiology	-0.175^{*} (0.055)	-0.175^{*} (0.055)	-0.159^{*} (0.057)
Dept:Genetics	0.153^{*} (0.073)	0.163^{*} (0.068)	0.163^{*} (0.072)
Dept:Pediatrics	0.260^{*} (0.101)	0.292^{*} (0.067)	0.290^{*} (0.070)
Dept:Medicine	0.509^{*} (0.085)	0.537^{*} (0.055)	0.520^{*} (0.058)
Dept:Surgery	0.947^{*} (0.116)	0.986^{*} (0.067)	0.955^{*} (0.069)
Clin:Clinical	0.248^{*} (0.077)	0.275^{*} (0.041)	0.273^{*} (0.043)
Cert:certified	0.147^{*} (0.040)	0.146^{*} (0.040)	0.135^{*} (0.041)
Publication Rate	-0.014 (0.033)	–	–
Experience	0.019^{*} (0.003)	0.019^{*} (0.003)	0.028^{*} (0.003)
Rank:Associate Prof	0.173^{*} (0.044)	0.175^{*} (0.044)	–
Rank:Full Prof	0.238^{*} (0.049)	0.240^{*} (0.049)	–
AIC	33.570	31.756	53.755
R^2	0.813	0.813	0.793

Note: $^{*}p<0.05$

Lasso could also reduce multicollinearity. In this case, the coefficient estimate for gender was shrunk to 0 with tuning parameter chosen by cross-validation, which is also consistent with my conclusion that gender is unrelated to salary discrimination. Other approaches for coping with multicollinearity proposed in the literature are principal component analysis (PCA) and partial least squares regression (PLS).[3] They are not adopted in my analysis because these methods cannot give me a direct interpretation of gender on salary.

A number of questions are not neatly answered through my regression analysis. For instance, since all covariates are only a "snapshot" of the workforce, follow-up research should be conducted. Besides, rank is often alleged to be tainted by discrimination, and its inclusion in the regression equation will underestimate the salary bias against women because it accounts for some of the variability in salary that otherwise would be associated with gender.[4] When the court deems a predictor variable to be tainted, the defendant cannot include that variable in a regression model used to rebut the plaintiffs' analysis. In our case, the same conclusion could be achieved when the variables representing academic rank were excluded (Table 3). However, when more variables are deemed tainted, the resulting regression model might give a conflicting conclusion.

Multiple regression analysis is a powerful tool for helping in an assessment of gender salary discrimination. However, its use is not in and of itself sufficient to be immune from legal challenges. Future analysis should be developed in close consultation with college representatives and lawyers possessing specific domain expertise to ensure that the statistical model captures as fully as possible the reality of the employment discrimination lawsuit.

Appendix

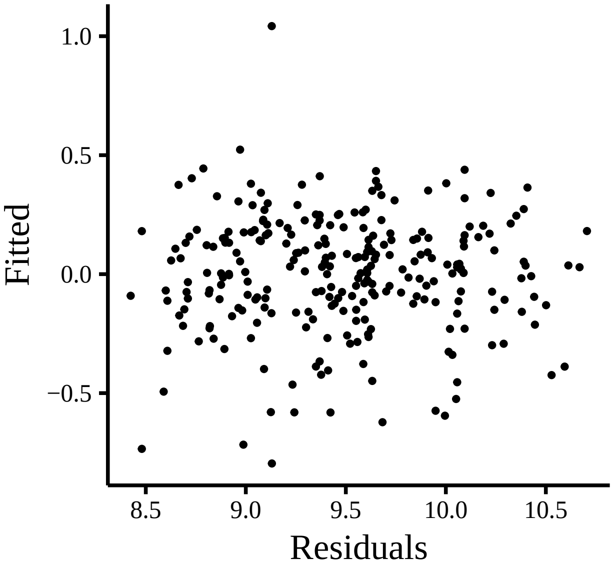

Figure 1: Residual Plot

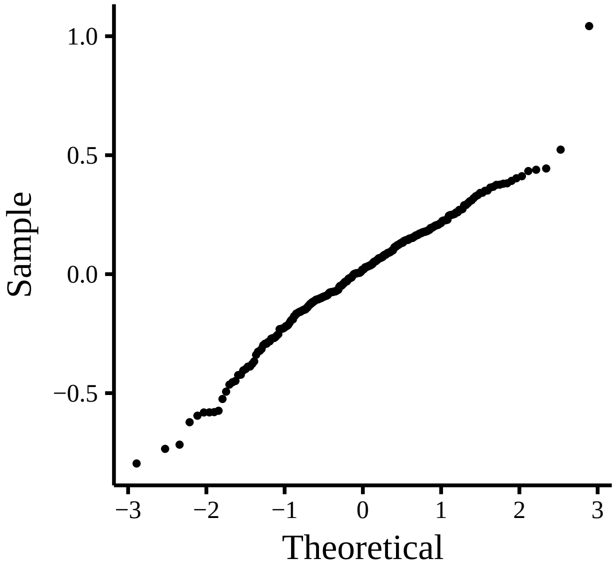

Figure 2: QQ Plot of Residuals

References

1. Paetzold RL. Multicollinearity and the use of regression analyses in discrimination litigation. Behavioral Sciences & the Law 1992;10(2):207–28.

2. Becker BE. Use of Multiple Regression for Evaluating Pay Equity : Prospects and Pitfalls. 2011;(January).

3. James G, Witten D, Hastie T, Tibishirani R. An Introduction to Statistical Learning. 2013.

4. Journal B, Law L, Gilmartin K, Gilmartintt K. Inclusion of Potentially Tainted Variables in Regression Analyses for Employment Discrimination Cases. 1991;13(1).