**Gender Based Discrimination Medical Professionals’ Salary:**

**A Case Study**

Abstract:

This report uses statistical analysis to find evidence for (or against) discrimination against females in salary. A multiple linear regression model was fitted based on statistical principles, using data of sample size N=261 from a lawsuit against the Houston College of Medicine, in the mid 1990’s. Based on rigorous analysis, we believe there is a strong possibility of discrimination against women based on the data, provided that some assumptions we made hold.

By Kang-Li Cheng

**Introduction:**

Gender discrimination has been a long-standing problem in the United States, despite a history of progress starting from suffrage and leading to laws such as Equal Employment Opportunity Act and other new legislation. But more subtle forms of discrimination have persisted, for example the discrepancy in pay between genders, which is claimed to be widespread but harder to quantify given the complicated factors that go into deciding compensation for an individual. This report will examine a particular case of possible gender discrimination in salary and promotion, using data from a legal case involving Houston College of Medicine and several all-female plaintiffs.

**Methods:**

a) Examining the Data

The dataset contains observations of 9 variables from N=261 individuals, of which 106 are females and 155 are males. There is salary information for the years 1994 and 1995, academic department, publication rate per year, experience in years, board certified (yes/no), clinical or research, and rank of professorship. The variables are summarized in Table 1. After checking data quality, (no problems found) we proceeded with exploratory analysis. By plotting the salary in 1994 against Rank and Department respectively while grouping by gender, it seems clear that the male group tends to have higher medians across the board.

Figure 1)

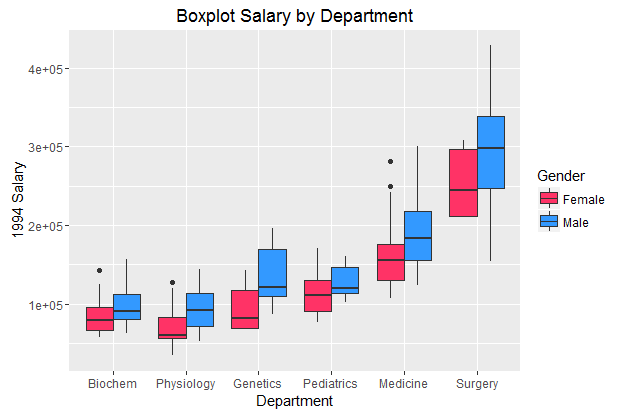
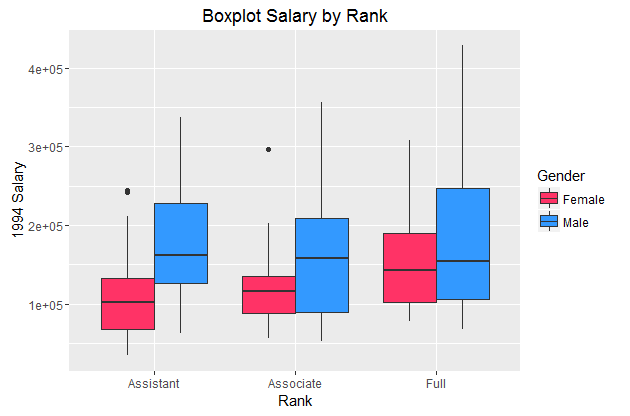


Table 1a) Summary Statistics for Continuous Covariates (Males), SD in parenthesis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | N | Experience | Publication Rate | Avg. Salary 1994 |
| Assistant | 43 | 6.42 (3.04) | 3.56 (1.44) | 188874.7 (73207.46) |
| Associate | 43 | 11.79 (6.03) | 4.80 (2.01) | 171058.8 (84797.4) |
| Full | 69 | 15.84 (6.24) | 1.90 (1.90) | 193803.7 (102699.10) |

Table 1b) Summary Statistics for Continuous Covariates (Females), SD in parenthesis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Rank | N | Experience | Publication rate | Avg. Salary 1994 |
| Assistant | 69 | 5.68 (3.07) | 5.31 (1.89) | 114303 (51443.45) |
| Associate | 21 | 9.33 (3.18) | 5.64 (1.99) | 128217 (57922.2) |
| Full | 16 | 12.88 (3.81) | 5.14 (1.79) | 166073 (75036.57) |

Fitting Regression Models

We started with regressing the log average of the 1995 and 1994 salaries against all possible covariates. In this model, the multiple F-test showed that all covariates except “gender” and “prate” were significant. Because “gender” is the main variable of interest, we remove only “prate” and refit the model. Again “gender” came up as not significant by its high pvalue.

The next step we took was to test for possible interaction terms. In the original model, we added the interaction term “gender\*Dept”. Then the process was repeated to test for interaction between clinical and gender, Prate and gender, and etc. Finally, we found that the interaction term for Rank and Gender was significant. Prate was not included in the final model because it is highly correlated with experience and also negatively correlated with salary. The explanation for this is likely because our data contains both clinical and non-clinical individuals, so using the publication rate as a predictor of salary is not very meaningful, since clinical individuals tend not to publish much. Experience is probably a better predictor of salary for both clinical and non-clinical individuals.

The final mode is:

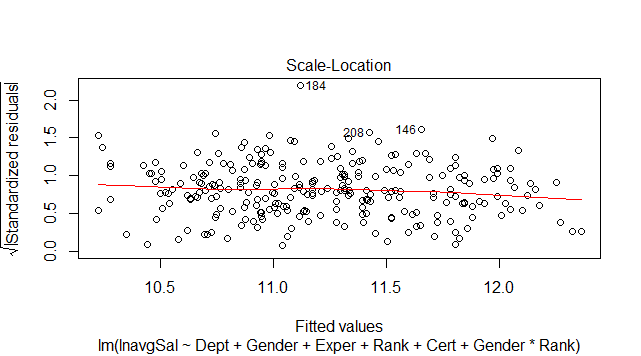
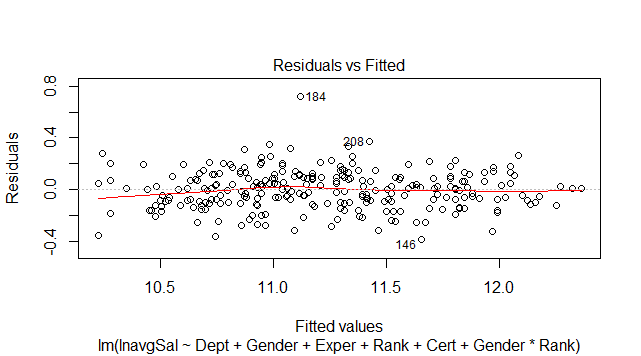
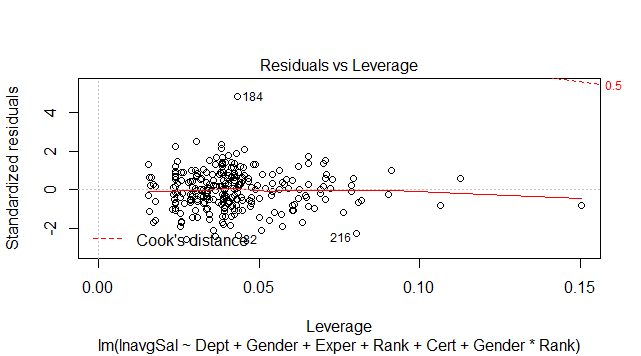
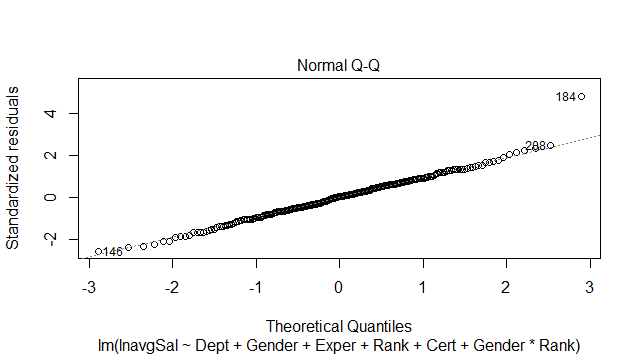
Table 2) Summary of Final Model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Covariate** | **Estimate** | **Std. Error** | **T value** | **P value** |
| Dept2 | -0.1277 | 0.0327 | -3.913 | 0.000117 |
| Dept3 | 0.274059 | 0.040015 | 6.849 | 5.75e-11 |
| Dept4 | 0.320546 | 0.038148 | 8.403 | 3.32e-15 |
| Dept5 | 0.670665 | 0.029377 | 22.829 | <2e-16 |
| Dept6 | 1.0837 | 0.35698 | 30.359 | <2e-16 |
| Gender | 0.194348 | 0.049477 | 3.928 | 0.000111 |
| Exper | 0.018563 | 0.002069 | 8.97 | <2e-16 |
| Rank | 0.170596 | 0.021390 | 7.975 | 5.45e-14 |
| Cert | 0.12932 | 0.023842 | 9.199 | 2e-16 |
| Gender:Rank | -0.096312 | 0.024186 | -3.824 | 0.000166 |
| Intercept | 10.141161 | 0.041315 | 245.46 | <2e-16 |

We then proceeded to see if being male would predict having a higher salary. Using random 25% subset of the data, we trained the full model on that data and predicted it on two test sets, one male only test set and another female only test set. The model predicted the log salary for males to be 11.37, and predicted 11.056 for females, which translates to an average of 2.9% higher salary for males. This finding is in line with the interpretation of the final model.

**Discussion**

Figure 2) Standard diagnostic plots

****

From the QQ-Plot, we see that the residuals follow a fairly straight line, which is a good sign. The residuals vs leverage however, show many outlying cases near the end of the left hand side. The residuals vs fitted plot seems to support that there is no strong non-linear relationship present, so a regression is valid. Finally, the scale-location plot shows that the residuals are fairly randomly spread out, so we are not too worried about homoscedasticity.

**Conclusion:**

Based on the statistical analysis, we can say that gender is a significant factor in determining salary amongst professionals at Houston College of Medicine. From the initial exploratory plots to the more rigorous regression modeling, it seems that there is strong evidence that there is discrimination based on gender. The regression model adjusted for each of the covariates still showed that gender is significant. If we wanted to show that discrimination was not at play, then we would expect that gender is not significant in the model, despite the appearance of a correlation when plotting the raw data. The effect of being male is positive, meaning one would expect males to earn a higher salary, holding all else equal. This is also assuming that the data set includes nearly all the relevant employees at the College, but if this is not so, the analysis should be re-performed since a more complete sample should give a more accurate picture.

**References:**

  James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer. Print

"Women's Rights." *Wikipedia*. Wikimedia Foundation, 04 May 2017. Web. 10 May 2017.