

# Final Project

*Oilivia Wagner | Weijia Xiong | Wurongyan Zhang | Yiling Yang*

## Contents

Abstract . . . . .	2
Introduction . . . . .	2
Exploratory data analysis . . . . .	2
Method . . . . .	7
add rank main effect + interaction . . . . .	30
on rank . . . . .	41
Conclusions . . . . .	44
Discussion . . . . .	44
Figures and tables . . . . .	45
References . . . . .	47
Appendix . . . . .	48

## Abstract

## Introduction

In our project, we examined the whether gender discrimination existed in setting salaries for people in the academia or higher education institutions.

## Exploratory data analysis

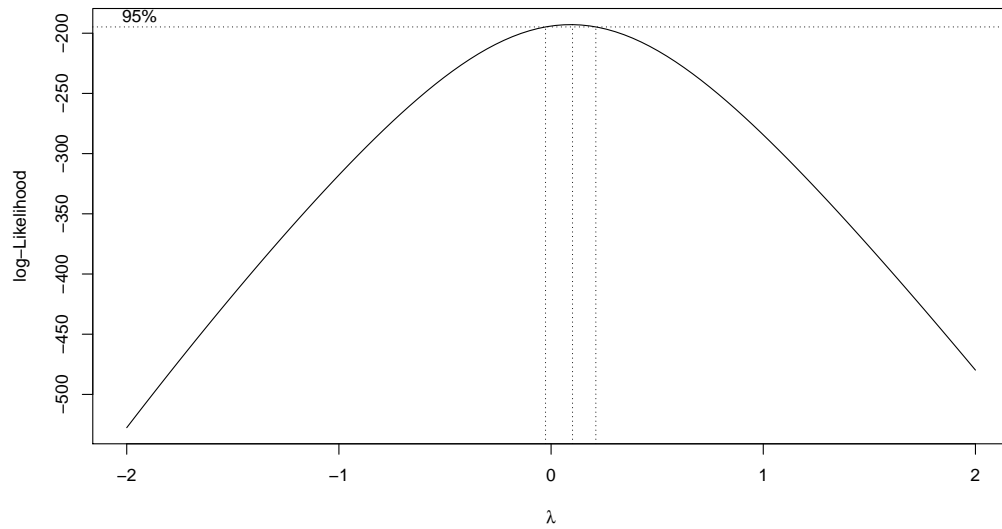
Data used in our studies were collected from 261 individuals who work in the academia or higher education institutions. The raw dataset contains following features.

- **Dept:** 1= Biochemistry/Molecular Biology 2= Physiology 3= Genetics 4= Pediatrics 5= Medicine 6= Surgery
- **Gender:** 1= Male, 0= Female
- **Clin:** 1= Primarily clinical emphasis, 0= Primarily research emphasis
- **Cert:** 1= Board certified, 0= not certified
- **Prate:** Publication rate (# publications on cv) / (# years between CV date and MD date)
- **Exper:** # years since obtaining MD
- **Rank:** 1= Assistant, 2= Associate, 3= Full professor (a proxy for productivity)
- **Sal94:** Salary in academic year 1994
- **Sal95:** Salary after increment to Sal94

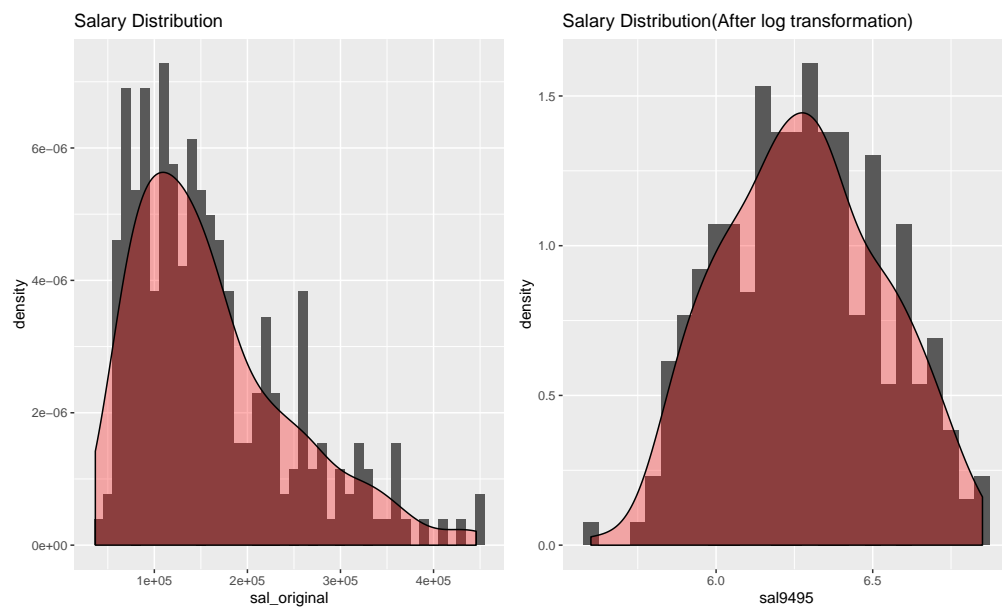
Table 1 contains the summary of variables in the dataset. Fortunately, there are no missing values in our dataset.

We then need to examine each interested variable against the main effect and main interest.

## Salary

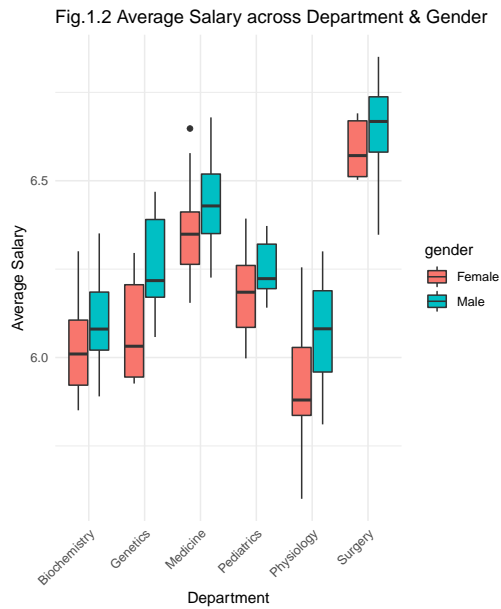
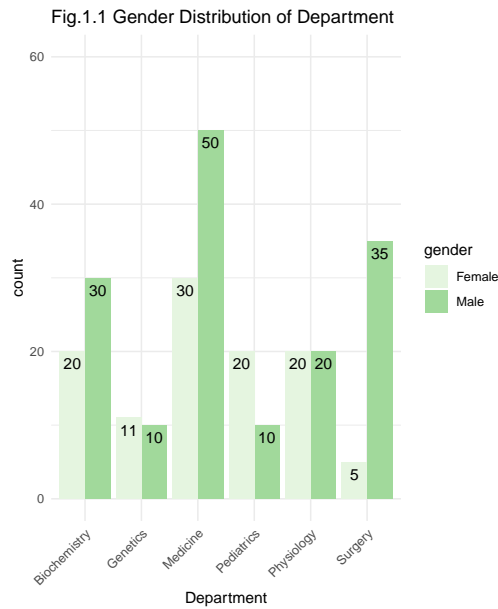


Since  $\lambda = 0$ , we use log transformation.



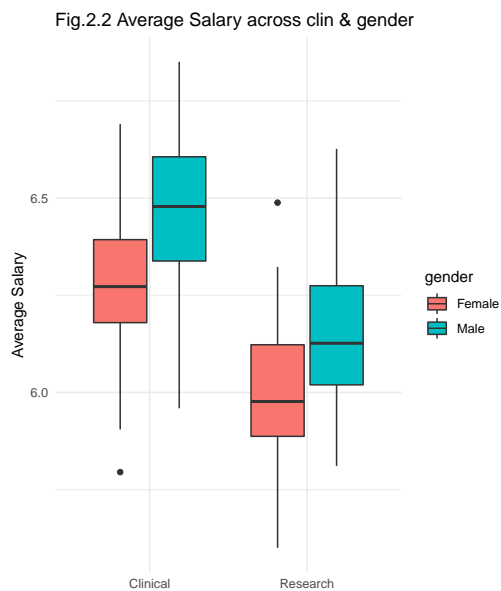
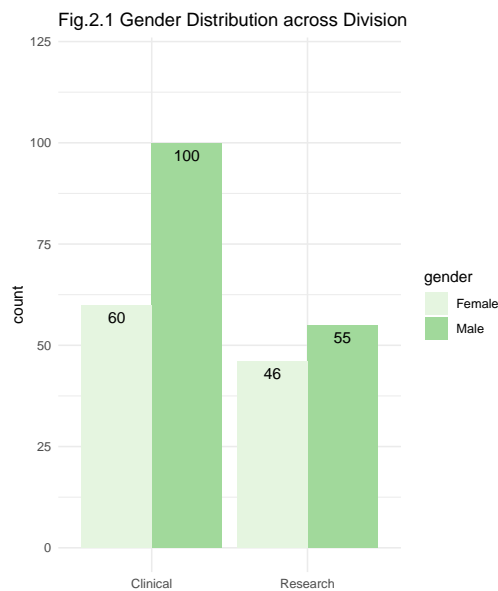
## Department

Fig 1.1. shows that the gender ratios in the department of Genetics and Physiology are very balanced, compared with department of Medicine and Surgery, which are very imbalanced. The differences in department of Biochemistry and Pediatrics are moderate. Fig 1.2 includes our main interest the salary into account. It seems that across all department, male earn more than female do. However, before further analysis, we cannot tell whether those difference are significant.



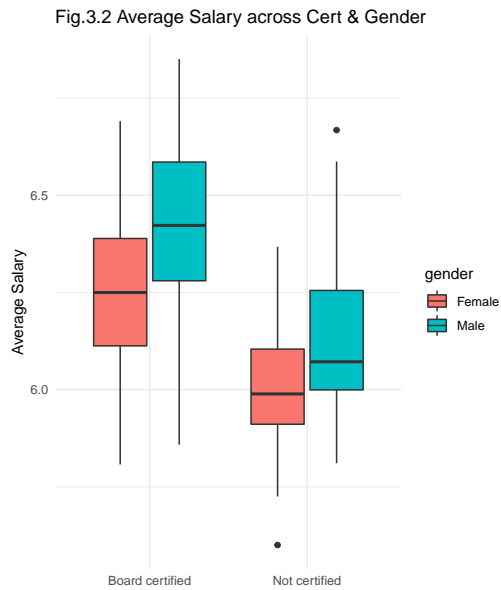
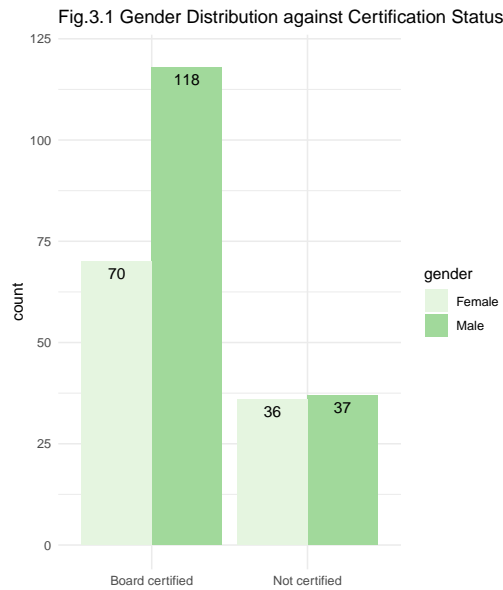
## Clinical or Research Division

Fig 2.1 shows that in either division, there are more male than female, especially the clinical division. Regarding salary, again, male in either division earned more than female do.



## Certification status

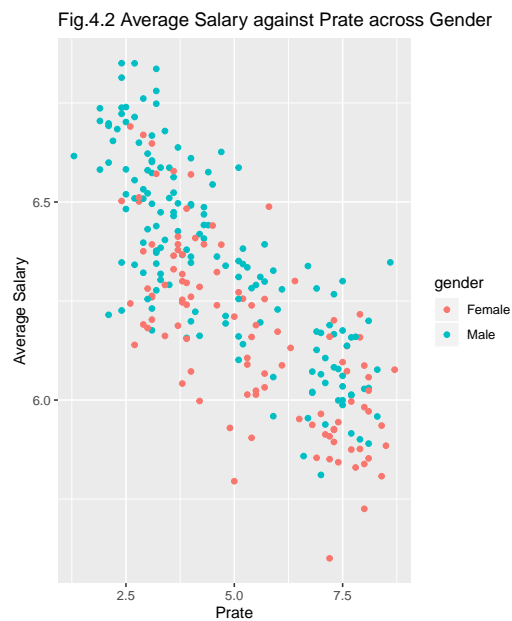
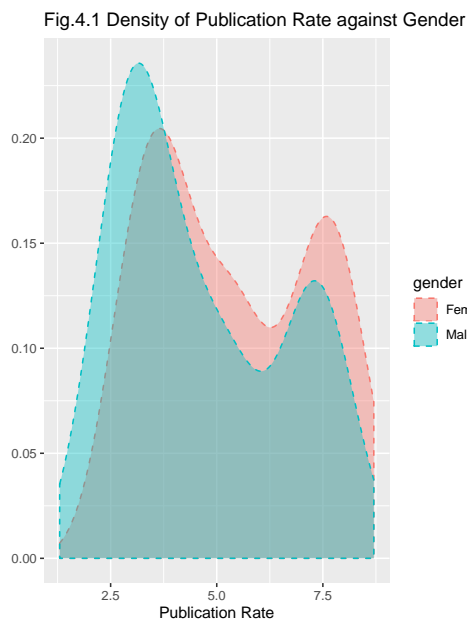
Fig.3. shows that the amount of certified male outnumbers surely the amount of certified female. However, for those without certification, those are just even. For the salary, male again earn more than female do regardless of his or her certification status.



## Prate

Fig 4.1 shows the density plot of publication rate for both male and female. No obvious difference was observed.

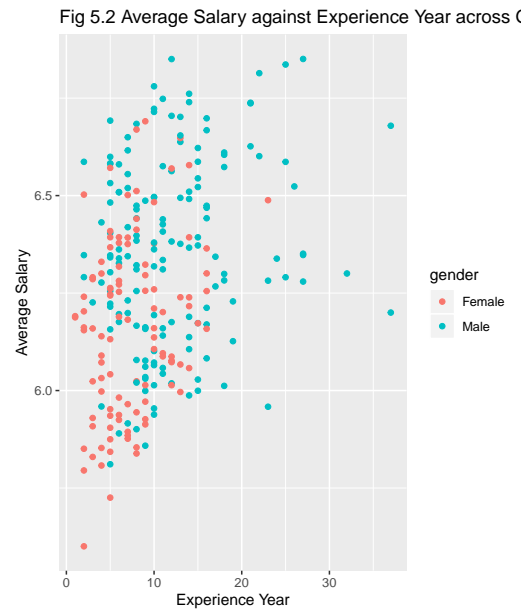
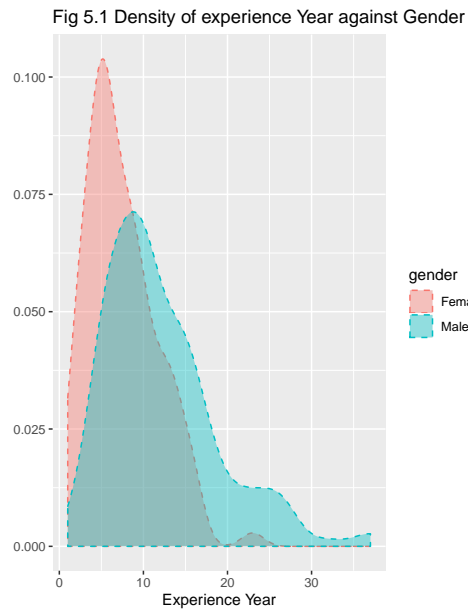
Fig 4.2 implies that there might be linear trend between average salary and publication rate. However, it's hard to tell whether there is any difference on the effect on the salary regarding the gender.



## Exper

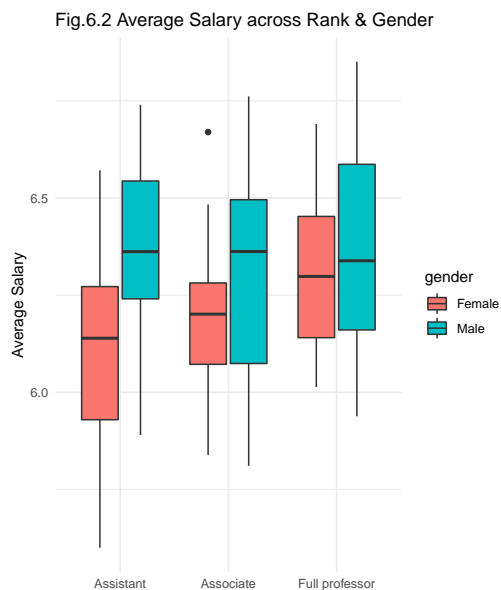
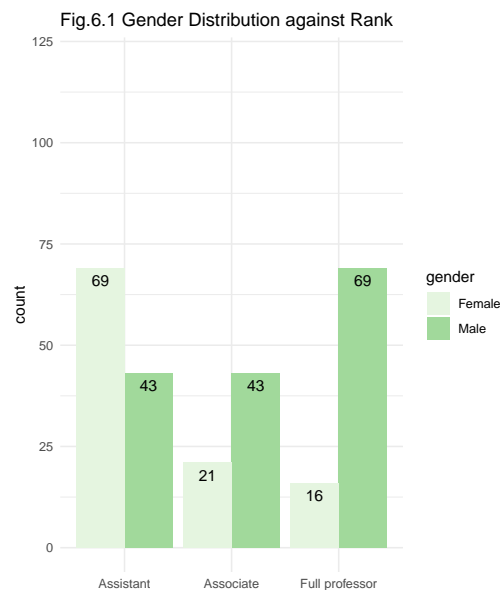
Fig.5.1 shows that the densities of experience year for both male and female are both very skewed, and seems that female have a more skewed trend. Although there are not obvious linear trend between salary and experience

year, the salary for male spread more widely than those for female and individual with high salary are dominately male.

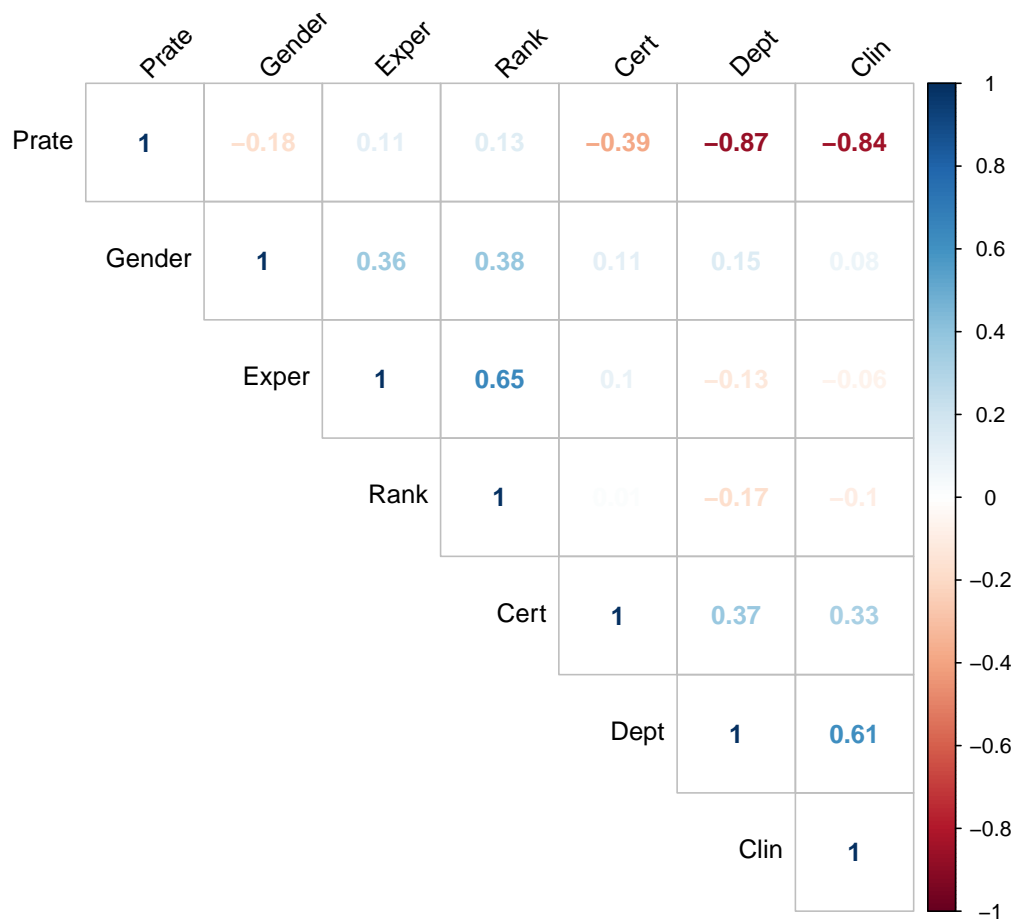


## Rank

Fig 6.1 shows that there are many female assistant professor and less associate and full professor than male. Without surprising, in either rank, male earn more than female.



Lastly, after examining the correlation matrix, we can see that there are some highly related variables ( $r > 80\%$ ). They are 1. Department (Dept) and Publish Rate (Prate) and 2. Clin and Prate. Some have morderate correlation such as 1. Experience Year (Exper) annd Cert (Certification status) and 2.Department (Dept) and Clin.



Those imply potential collinearities. As we go through those variables, there are some outliers. In later section, we will examine further about them.

## Method

```
##
## Call:
## lm(formula = sal9495 ~ gender, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5628 -0.1829 -0.0039  0.1671  0.5275
##
## Coefficients:
```

```

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.16317    0.02298 268.246 < 2e-16 ***
## genderMale    0.19265    0.02981   6.462 5.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2366 on 259 degrees of freedom
## Multiple R-squared:  0.1388, Adjusted R-squared:  0.1355
## F-statistic: 41.75 on 1 and 259 DF,  p-value: 5.103e-10
##
## Call:
## lm(formula = sal9495 ~ gender + dept, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35272 -0.09716 -0.01250  0.08171  0.30390
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.01909    0.02055 292.901 < 2e-16 ***
## genderMale      0.10260    0.01678   6.116 3.60e-09 ***
## deptGenetics    0.10046    0.03301   3.044 0.00258 **
## deptMedicine    0.32471    0.02284  14.215 < 2e-16 ***
## deptPediatrics  0.15267    0.02960   5.158 5.02e-07 ***
## deptPhysiology -0.06597    0.02693  -2.450 0.01496 *
## deptSurgery     0.53834    0.02727  19.743 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```



```

## Residual standard error: 0.1267 on 254 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:  0.752
## F-statistic: 132.4 on 6 and 254 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender
## Model 2: sal9495 ~ gender + dept
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      259 14.4927
## 2      254  4.0765   5    10.416 129.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = sal9495 ~ gender + clin, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50456 -0.12512 -0.00948  0.12282  0.49713
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.29514    0.02065  304.872 < 2e-16 ***
## genderMale     0.16859    0.02336   7.218 5.89e-12 ***
## clinResearch  -0.30410    0.02355 -12.912 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1847 on 258 degrees of freedom

```

```

## Multiple R-squared:  0.4769, Adjusted R-squared:  0.4728
## F-statistic: 117.6 on 2 and 258 DF,  p-value: < 2.2e-16

## Analysis of Variance Table

##
## Model 1: sal9495 ~ gender
## Model 2: sal9495 ~ gender + clin
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     259 14.4927
## 2     258  8.8035  1     5.6891 166.73 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = sal9495 ~ gender + cert, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55933 -0.13835 -0.00961  0.15540  0.50971
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.25144    0.02228  280.648 < 2e-16 ***
## genderMale        0.16642    0.02617   6.360 9.11e-10 ***
## certNot certified -0.25991    0.02863  -9.078 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Residual standard error: 0.2063 on 258 degrees of freedom
## Multiple R-squared:  0.3473, Adjusted R-squared:  0.3423

```

```

## F-statistic: 68.65 on 2 and 258 DF,  p-value: < 2.2e-16

## Analysis of Variance Table

##

## Model 1: sal9495 ~ gender
## Model 2: sal9495 ~ gender + cert

##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)

## 1      259 14.493

## 2      258 10.984  1     3.5087 82.416 < 2.2e-16 ***

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Call:
## lm(formula = sal9495 ~ gender + exper, data = lawsuit)

##

## Residuals:

##      Min       1Q   Median       3Q      Max

## -0.51649 -0.18686  0.02018  0.16638  0.51473

##

## Coefficients:

##              Estimate Std. Error t value Pr(>|t|)

## (Intercept)  6.100037   0.029194 208.950  < 2e-16 ***

## genderMale   0.153775   0.031384   4.900  1.7e-06 ***

## exper        0.008428   0.002480   3.399  0.000784 ***

## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2319 on 258 degrees of freedom

## Multiple R-squared:  0.1757, Adjusted R-squared:  0.1693

## F-statistic:  27.5 on 2 and 258 DF,  p-value: 1.488e-11

```

```

## Analysis of Variance Table

##

## Model 1: sal9495 ~ gender
## Model 2: sal9495 ~ gender + exper
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      259 14.493
## 2      258 13.872  1    0.62108 11.552 0.000784 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Call:
## lm(formula = sal9495 ~ gender + rank, data = lawsuit)
##

## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55399 -0.18936  0.00746  0.17699  0.51766
##

## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.154393    0.025482 241.520 < 2e-16 ***
## genderMale        0.174740    0.032210   5.425 1.34e-07 ***
## rankAssociate    -0.002612    0.038094  -0.069  0.9454
## rankFull professor  0.061583    0.036611   1.682  0.0938 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 0.2358 on 257 degrees of freedom
## Multiple R-squared:  0.151, Adjusted R-squared:  0.1411
## F-statistic: 15.23 on 3 and 257 DF,  p-value: 3.747e-09

```

```
## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender
## Model 2: sal9495 ~ gender + rank
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      259 14.493
## 2      257 14.288   2   0.20422 1.8366 0.1614
```

There is no need to consider prate since the correlation of prate and dept is -0.87. Otherwise there might be collinearity.

From linear regression results, rank is not significant. From partial F test, p is more than 0.05, which shows that model with rank is not superior.

Since the changes of estimate of genderMale are more than 10% and all p-values from partial F test are less than 0.05, dept, clin, cert and exper are all confounders. So we need to add these covariates into the model.

First we need to add department....

```
##
## Call:
## lm(formula = sal9495 ~ gender + dept, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35272 -0.09716 -0.01250  0.08171  0.30390
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.01909    0.02055 292.901 < 2e-16 ***
## genderMale     0.10260    0.01678   6.116 3.60e-09 ***
## deptGenetics   0.10046    0.03301   3.044 0.00258 **
## deptMedicine   0.32471    0.02284 14.215 < 2e-16 ***
```

```

## deptPediatrics  0.15267      0.02960    5.158 5.02e-07 ***
## deptPhysiology -0.06597      0.02693   -2.450  0.01496 *
## deptSurgery     0.53834      0.02727   19.743 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1267 on 254 degrees of freedom
## Multiple R-squared:  0.7578, Adjusted R-squared:  0.752
## F-statistic: 132.4 on 6 and 254 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = sal9495 ~ gender + dept + clin, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.315740 -0.083385 -0.003147  0.075969  0.309553
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.12952    0.02539 241.453 < 2e-16 ***
## genderMale      0.10156    0.01554   6.537 3.44e-10 ***
## deptGenetics    0.04401    0.03175   1.386 0.166977
## deptMedicine    0.23677    0.02503   9.460 < 2e-16 ***
## deptPediatrics  0.07587    0.02980   2.546 0.011481 *
## deptPhysiology -0.09789    0.02540  -3.854 0.000148 ***
## deptSurgery     0.43195    0.02999  14.401 < 2e-16 ***
## clinResearch   -0.12477    0.01898  -6.572 2.81e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```
##
## Residual standard error: 0.1173 on 253 degrees of freedom
## Multiple R-squared:  0.7931, Adjusted R-squared:  0.7874
## F-statistic: 138.5 on 7 and 253 DF,  p-value: < 2.2e-16

## Analysis of Variance Table

##

## Model 1: sal9495 ~ gender + dept
## Model 2: sal9495 ~ gender + dept + clin

##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      254 4.0765
## 2      253 3.4821   1    0.59448 43.194 2.811e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##

## Call:
## lm(formula = sal9495 ~ gender + dept + cert, data = lawsuit)

##

## Residuals:

##      Min       1Q   Median       3Q      Max
## -0.278961 -0.084532 -0.001179  0.077452  0.274284

##

## Coefficients:

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.09277    0.02079 293.077 < 2e-16 ***
## genderMale        0.09339    0.01515   6.164 2.78e-09 ***
## deptGenetics      0.11555    0.02978   3.880 0.000133 ***
## deptMedicine      0.28065    0.02133  13.156 < 2e-16 ***
## deptPediatrics    0.09115    0.02780   3.279 0.001187 **
## deptPhysiology   -0.07712    0.02428  -3.176 0.001681 **
```

```

## deptSurgery          0.47954    0.02569  18.668 < 2e-16 ***
## certNot certified -0.13629    0.01754  -7.771 1.96e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1141 on 253 degrees of freedom
## Multiple R-squared:  0.8044, Adjusted R-squared:  0.799
## F-statistic: 148.7 on 7 and 253 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept
## Model 2: sal9495 ~ gender + dept + cert
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      254 4.0765
## 2      253 3.2911  1    0.78546 60.382 1.959e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.261287 -0.066217  0.000115  0.061314  0.277094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.877078   0.018514 317.446 < 2e-16 ***
## genderMale     0.034869   0.013529   2.577  0.0105 *

```



```

## deptGenetics      0.132781    0.024959    5.320 2.29e-07 ***
## deptMedicine      0.368417    0.017480   21.076 < 2e-16 ***
## deptPediatrics    0.209135    0.022647    9.235 < 2e-16 ***
## deptPhysiology   -0.044616    0.020330   -2.195  0.0291 *
## deptSurgery       0.584001    0.020788   28.093 < 2e-16 ***
## exper             0.014613    0.001046   13.968 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09538 on 253 degrees of freedom
## Multiple R-squared:  0.8632, Adjusted R-squared:  0.8594
## F-statistic: 228.1 on 7 and 253 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept
## Model 2: sal9495 ~ gender + dept + exper
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      254 4.0765
## 2      253 2.3017  1      1.7749 195.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = sal9495 ~ gender + dept + rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.284681 -0.065757 -0.002947  0.064790  0.292831
##

```

```
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.94873    0.01773 335.554 < 2e-16 ***
## genderMale        0.03203    0.01470   2.179 0.03030 *
## deptGenetics      0.11661    0.02642   4.413 1.51e-05 ***
## deptMedicine      0.36277    0.01856  19.547 < 2e-16 ***
## deptPediatrics    0.19220    0.02390   8.043 3.43e-14 ***
## deptPhysiology    -0.06727    0.02153  -3.125 0.00199 **
## deptSurgery       0.59041    0.02225  26.531 < 2e-16 ***
## rankAssociate     0.09392    0.01662   5.650 4.33e-08 ***
## rankFull professor 0.19639    0.01628  12.061 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1013 on 252 degrees of freedom
## Multiple R-squared:  0.8464, Adjusted R-squared:  0.8415
## F-statistic: 173.6 on 8 and 252 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept
## Model 2: sal9495 ~ gender + dept + rank
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      254 4.0765
## 2      252 2.5845  2      1.492 72.737 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Result: Compared with other covariates, F value of adding exper is the largest. So we add exper into our model.

##
## Call:
```

```
## lm(formula = sal9495 ~ gender + dept + exper, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.261287 -0.066217  0.000115  0.061314  0.277094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.877078   0.018514  317.446 < 2e-16 ***
## genderMale      0.034869   0.013529   2.577  0.0105 *
## deptGenetics    0.132781   0.024959   5.320 2.29e-07 ***
## deptMedicine    0.368417   0.017480  21.076 < 2e-16 ***
## deptPediatrics  0.209135   0.022647   9.235 < 2e-16 ***
## deptPhysiology -0.044616   0.020330  -2.195  0.0291 *
## deptSurgery     0.584001   0.020788  28.093 < 2e-16 ***
## exper           0.014613   0.001046  13.968 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09538 on 253 degrees of freedom
## Multiple R-squared:  0.8632, Adjusted R-squared:  0.8594
## F-statistic: 228.1 on 7 and 253 DF,  p-value: < 2.2e-16
```

Then we try to look for next covariate.

```
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + clin, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```

## -0.23172 -0.05501 0.00399 0.05391 0.37148
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.9835666  0.0204581 292.479 < 2e-16 ***
## genderMale   0.0352073  0.0119141   2.955 0.003422 **
## deptGenetics 0.0791242  0.0228446   3.464 0.000626 ***
## deptMedicine 0.2849665  0.0181866  15.669 < 2e-16 ***
## deptPediatrics 0.1358926  0.0216792   6.268 1.57e-09 ***
## deptPhysiology -0.0750153  0.0182475  -4.111 5.33e-05 ***
## deptSurgery   0.4831773  0.0217262  22.239 < 2e-16 ***
## exper        0.0143278  0.0009219  15.542 < 2e-16 ***
## clinResearch -0.1171971  0.0136012  -8.617 7.70e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08399 on 252 degrees of freedom
## Multiple R-squared:  0.8944, Adjusted R-squared:  0.891
## F-statistic: 266.7 on 8 and 252 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper
## Model 2: sal9495 ~ gender + dept + exper + clin
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      253 2.3017
## 2      252 1.7779  1    0.52381 74.247 7.696e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##

```

```
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + cert, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22497 -0.05412  0.00002  0.05689  0.35517
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.9461021   0.0187608  316.943 < 2e-16 ***
## genderMale      0.0334723   0.0121359    2.758  0.00624 **
## deptGenetics    0.1417147   0.0224149    6.322 1.16e-09 ***
## deptMedicine    0.3306264   0.0163909   20.171 < 2e-16 ***
## deptPediatrics  0.1567611   0.0213655    7.337 2.99e-12 ***
## deptPhysiology -0.0550454   0.0182822   -3.011  0.00287 **
## deptSurgery     0.5346466   0.0196625   27.191 < 2e-16 ***
## exper          0.0133753   0.0009513   14.060 < 2e-16 ***
## certNot certified -0.1054425   0.0133379   -7.905 8.35e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08555 on 252 degrees of freedom
## Multiple R-squared:  0.8904, Adjusted R-squared:  0.8869
## F-statistic: 255.9 on 8 and 252 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper
## Model 2: sal9495 ~ gender + dept + exper + cert
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
```

```
## 1      253 2.3017
## 2      252 1.8443  1    0.45739 62.497 8.354e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.248969 -0.058604 -0.001912  0.060896  0.293323
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.880355   0.017465 336.698 < 2e-16 ***
## genderMale      0.015785   0.013060   1.209  0.22795
## deptGenetics    0.132201   0.023296   5.675 3.82e-08 ***
## deptMedicine    0.376538   0.016390  22.973 < 2e-16 ***
## deptPediatrics  0.214872   0.021166  10.152 < 2e-16 ***
## deptPhysiology -0.051669   0.019007  -2.718  0.00702 **
## deptSurgery     0.599303   0.019587  30.596 < 2e-16 ***
## exper           0.010341   0.001193   8.670 5.47e-16 ***
## rankAssociate   0.047863   0.015547   3.079  0.00231 **
## rankFull professor 0.109455   0.017475   6.264 1.62e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08902 on 251 degrees of freedom
## Multiple R-squared:  0.8818, Adjusted R-squared:  0.8776
```

```
## F-statistic: 208.1 on 9 and 251 DF, p-value: < 2.2e-16
```

```
## Analysis of Variance Table
```

```
##
```

```
## Model 1: sal9495 ~ gender + dept + exper
```

```
## Model 2: sal9495 ~ gender + dept + exper + rank
```

```
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
```

```
## 1      253 2.3017
```

```
## 2      251 1.9889  2    0.31273 19.733 1.098e-08 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result: Compared with other covariates, F value of adding clin is the largest. So we add exper into our model.

```
##
```

```
## Call:
```

```
## lm(formula = sal9495 ~ gender + dept + exper + clin, data = lawsuit)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -0.23172 -0.05501  0.00399  0.05391  0.37148
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    5.9835666  0.0204581 292.479 < 2e-16 ***
```

```
## genderMale      0.0352073  0.0119141   2.955 0.003422 **
```

```
## deptGenetics    0.0791242  0.0228446   3.464 0.000626 ***
```

```
## deptMedicine    0.2849665  0.0181866  15.669 < 2e-16 ***
```

```
## deptPediatrics  0.1358926  0.0216792   6.268 1.57e-09 ***
```

```
## deptPhysiology -0.0750153  0.0182475  -4.111 5.33e-05 ***
```

```
## deptSurgery     0.4831773  0.0217262  22.239 < 2e-16 ***
```

```
## exper           0.0143278  0.0009219  15.542 < 2e-16 ***
```

```
## clinResearch   -0.1171971  0.0136012  -8.617 7.70e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08399 on 252 degrees of freedom
## Multiple R-squared:  0.8944, Adjusted R-squared:  0.891
## F-statistic: 266.7 on 8 and 252 DF,  p-value: < 2.2e-16
```

Then we try to look for next covariate :

```
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + clin + cert, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18571 -0.05097  0.00044  0.04429  0.42863
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0314905  0.0194750 309.705 < 2e-16 ***
## genderMale      0.0339569  0.0107443   3.160 0.00177 **
## deptGenetics    0.0929245  0.0206775   4.494 1.07e-05 ***
## deptMedicine    0.2615613  0.0166801  15.681 < 2e-16 ***
## deptPediatrics  0.0986776  0.0201405   4.899 1.72e-06 ***
## deptPhysiology -0.0806445  0.0164703  -4.896 1.75e-06 ***
## deptSurgery     0.4516913  0.0200155  22.567 < 2e-16 ***
## exper          0.0132859  0.0008423  15.774 < 2e-16 ***
## clinResearch    -0.1039933  0.0123844  -8.397 3.40e-15 ***
## certNot certified -0.0915366  0.0119239  -7.677 3.63e-13 ***
## ---
```



```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07574 on 251 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9114
## F-statistic: 298.1 on 9 and 251 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + clin
## Model 2: sal9495 ~ gender + dept + exper + clin + cert
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      252 1.7779
## 2      251 1.4398  1    0.33805 58.932 3.632e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + clin + rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20615 -0.05091 -0.00325  0.04672  0.38905
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.985316   0.018728 319.585 < 2e-16 ***
## genderMale      0.015583   0.011247   1.386 0.167129
## deptGenetics    0.078800   0.020850   3.779 0.000197 ***
## deptMedicine    0.293627   0.016642  17.644 < 2e-16 ***
## deptPediatrics  0.141346   0.019833   7.127 1.10e-11 ***

```

```

## deptPhysiology      -0.081881    0.016681   -4.909 1.66e-06 ***
## deptSurgery         0.499108     0.019951   25.017 < 2e-16 ***
## exper               0.010055     0.001028    9.784 < 2e-16 ***
## clinResearch        -0.116999     0.012440   -9.405 < 2e-16 ***
## rankAssociate        0.054547     0.013408    4.068 6.35e-05 ***
## rankFull professor  0.109071     0.015049    7.248 5.26e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07666 on 250 degrees of freedom
## Multiple R-squared:  0.9127, Adjusted R-squared:  0.9092
## F-statistic: 261.4 on 10 and 250 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + clin
## Model 2: sal9495 ~ gender + dept + exper + clin + rank
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      252 1.7779
## 2      250 1.4691  2    0.30872 26.267 4.423e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Result: Compared with other covariates, F value of adding cert is the largest. So we add cert into our model.

```

##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + clin + cert, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18571 -0.05097  0.00044  0.04429  0.42863

```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0314905   0.0194750 309.705 < 2e-16 ***
## genderMale      0.0339569   0.0107443   3.160 0.00177 **
## deptGenetics    0.0929245   0.0206775   4.494 1.07e-05 ***
## deptMedicine    0.2615613   0.0166801  15.681 < 2e-16 ***
## deptPediatrics  0.0986776   0.0201405   4.899 1.72e-06 ***
## deptPhysiology -0.0806445   0.0164703  -4.896 1.75e-06 ***
## deptSurgery     0.4516913   0.0200155  22.567 < 2e-16 ***
## exper          0.0132859   0.0008423  15.774 < 2e-16 ***
## clinResearch    -0.1039933   0.0123844  -8.397 3.40e-15 ***
## certNot certified -0.0915366   0.0119239  -7.677 3.63e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07574 on 251 degrees of freedom
## Multiple R-squared:  0.9144, Adjusted R-squared:  0.9114
## F-statistic: 298.1 on 9 and 251 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + clin + cert +
##     rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17303 -0.03848 -0.00936  0.03798  0.45196
##
## Coefficients:
```

```

##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.033505   0.017199 350.801 < 2e-16 ***
## genderMale       0.012882   0.009812   1.313    0.19
## deptGenetics     0.092985   0.018250   5.095 6.90e-07 ***
## deptMedicine     0.269652   0.014758  18.272 < 2e-16 ***
## deptPediatrics   0.101673   0.017856   5.694 3.48e-08 ***
## deptPhysiology   -0.087875   0.014561  -6.035 5.73e-09 ***
## deptSurgery      0.466910   0.017767  26.280 < 2e-16 ***
## exper            0.008863   0.000906   9.783 < 2e-16 ***
## clinResearch     -0.104170   0.010943  -9.520 < 2e-16 ***
## certNot certified -0.094875   0.010622  -8.932 < 2e-16 ***
## rankAssociate     0.067332   0.011779   5.716 3.10e-08 ***
## rankFull professor 0.111107   0.013125   8.466 2.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06685 on 249 degrees of freedom
## Multiple R-squared:  0.9339, Adjusted R-squared:  0.931
## F-statistic: 319.7 on 11 and 249 DF, p-value: < 2.2e-16
## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + clin + cert
## Model 2: sal9495 ~ gender + dept + exper + clin + cert + rank
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      251 1.4398
## 2      249 1.1126  2    0.32717 36.608 1.154e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Result:

p value of partial F test is more than 0.05, which shows that the model adding cert is superior. Therefore, we need to add cert in the model.

```
##
## Call:
## lm(formula = sal9495 ~ gender + dept + exper + cert + clin +
##     rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17303 -0.03848 -0.00936  0.03798  0.45196
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.033505   0.017199  350.801 < 2e-16 ***
## genderMale        0.012882   0.009812   1.313    0.19
## deptGenetics      0.092985   0.018250   5.095 6.90e-07 ***
## deptMedicine      0.269652   0.014758  18.272 < 2e-16 ***
## deptPediatrics    0.101673   0.017856   5.694 3.48e-08 ***
## deptPhysiology    -0.087875   0.014561  -6.035 5.73e-09 ***
## deptSurgery       0.466910   0.017767  26.280 < 2e-16 ***
## exper            0.008863   0.000906   9.783 < 2e-16 ***
## certNot certified -0.094875   0.010622  -8.932 < 2e-16 ***
## clinResearch      -0.104170   0.010943  -9.520 < 2e-16 ***
## rankAssociate     0.067332   0.011779   5.716 3.10e-08 ***
## rankFull professor 0.111107   0.013125   8.466 2.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06685 on 249 degrees of freedom
```

```
## Multiple R-squared:  0.9339, Adjusted R-squared:  0.931
## F-statistic: 319.7 on 11 and 249 DF,  p-value: < 2.2e-16
```

### add rank main effect + interaction

```
##
## Call:
## lm(formula = sal9495 ~ gender * dept + exper + clin + cert +
##      rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17103 -0.03840 -0.00794  0.03623  0.45156
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.0440044   0.0200987  300.717 < 2e-16 ***
## genderMale      -0.0079465   0.0201844  -0.394 0.694151
## deptGenetics     0.0698020   0.0256629   2.720 0.006999 **
## deptMedicine     0.2621409   0.0211666  12.385 < 2e-16 ***
## deptPediatrics   0.0895477   0.0228907   3.912 0.000119 ***
## deptPhysiology  -0.1091064   0.0216163  -5.047 8.76e-07 ***
## deptSurgery      0.4526402   0.0353668  12.798 < 2e-16 ***
## exper           0.0087701   0.0009153   9.582 < 2e-16 ***
## clinResearch    -0.1030410   0.0111278  -9.260 < 2e-16 ***
## certNot certified -0.0928949   0.0107955  -8.605 9.47e-16 ***
## rankAssociate    0.0686003   0.0119119   5.759 2.54e-08 ***
## rankFull professor 0.1129723   0.0133424   8.467 2.37e-15 ***
## genderMale:deptGenetics 0.0438303   0.0359109   1.221 0.223442
## genderMale:deptMedicine 0.0153184   0.0249702   0.613 0.540139
## genderMale:deptPediatrics 0.0245812   0.0326826   0.752 0.452705
```

```

## genderMale:deptPhysiology  0.0389448  0.0289740  1.344 0.180155
## genderMale:deptSurgery      0.0251571  0.0378988  0.664 0.507446
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06717 on 244 degrees of freedom
## Multiple R-squared:  0.9346, Adjusted R-squared:  0.9303
## F-statistic: 217.9 on 16 and 244 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + cert + clin + rank
## Model 2: sal9495 ~ gender * dept + exper + clin + cert + rank
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      249 1.1126
## 2      244 1.1010  5  0.011642 0.516 0.7641

not significant

##
## Call:
## lm(formula = sal9495 ~ gender * exper + dept + clin + cert +
##   rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16065 -0.03930 -0.00493  0.03550  0.43455
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.993406    0.020845 287.519 < 2e-16 ***
## genderMale      0.064466    0.018456   3.493 0.000566 ***

```

```

## exper          0.013887    0.001773    7.834 1.38e-13 ***
## deptGenetics   0.094885    0.017913    5.297 2.60e-07 ***
## deptMedicine   0.273386    0.014522   18.825 < 2e-16 ***
## deptPediatrics 0.109302    0.017671    6.185 2.54e-09 ***
## deptPhysiology -0.082534    0.014377   -5.741 2.75e-08 ***
## deptSurgery    0.469915    0.017453   26.924 < 2e-16 ***
## clinResearch   -0.104087    0.010735   -9.696 < 2e-16 ***
## certNot certified -0.091083    0.010484   -8.688 5.09e-16 ***
## rankAssociate   0.059115    0.011824    5.000 1.09e-06 ***
## rankFull professor 0.104018    0.013056    7.967 5.90e-14 ***
## genderMale:exper -0.005864    0.001790   -3.276 0.001204 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06558 on 248 degrees of freedom
## Multiple R-squared:  0.9366, Adjusted R-squared:  0.9336
## F-statistic: 305.4 on 12 and 248 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + cert + clin + rank
## Model 2: sal9495 ~ gender * exper + dept + clin + cert + rank
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      249 1.1126
## 2      248 1.0665  1  0.046146 10.731 0.001204 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

significant

##
## Call:

```



```
## lm(formula = sal9495 ~ gender * clin + dept + exper + cert +
##      rank, data = lawsuit)
##
## Residuals:
##      Min        1Q    Median        3Q        Max
## -0.17678 -0.03949 -0.00832  0.03988  0.44659
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.0372930   0.0176148  342.740 < 2e-16 ***
## genderMale        0.0057321   0.0121564    0.472   0.638
## clinResearch     -0.1136612   0.0145087   -7.834 1.38e-13 ***
## deptGenetics      0.0960066   0.0185009    5.189 4.39e-07 ***
## deptMedicine      0.2703457   0.0147743   18.298 < 2e-16 ***
## deptPediatrics    0.1011700   0.0178633    5.664 4.09e-08 ***
## deptPhysiology   -0.0870209   0.0145866   -5.966 8.35e-09 ***
## deptSurgery       0.4693860   0.0179398   26.165 < 2e-16 ***
## exper            0.0089798   0.0009136    9.829 < 2e-16 ***
## certNot certified -0.0959499   0.0106767   -8.987 < 2e-16 ***
## rankAssociate     0.0654823   0.0119242    5.492 9.85e-08 ***
## rankFull professor 0.1077608   0.0135478    7.954 6.41e-14 ***
## genderMale:clinResearch 0.0183748  0.0184439    0.996   0.320
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06685 on 248 degrees of freedom
## Multiple R-squared:  0.9341, Adjusted R-squared:  0.931
## F-statistic: 293.2 on 12 and 248 DF,  p-value: < 2.2e-16
## Analysis of Variance Table
```

```
##
## Model 1: sal9495 ~ gender + dept + exper + cert + clin + rank
## Model 2: sal9495 ~ gender * clin + dept + exper + cert + rank
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     249 1.1126
## 2     248 1.1082  1 0.0044352 0.9925 0.3201

not significant

##
## Call:
## lm(formula = sal9495 ~ gender * cert + dept + exper + clin +
##     rank, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17395 -0.03981 -0.00970  0.03773  0.45001
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0350729   0.0177871  339.295 < 2e-16 ***
## genderMale      0.0107683   0.0114924   0.937  0.350
## certNot certified -0.0986536   0.0150536  -6.553 3.23e-10 ***
## deptGenetics     0.0937244   0.0184008   5.094 6.96e-07 ***
## deptMedicine     0.2694641   0.0147931  18.216 < 2e-16 ***
## deptPediatrics   0.1010776   0.0179658   5.626 4.96e-08 ***
## deptPhysiology  -0.0874661   0.0146322  -5.978 7.84e-09 ***
## deptSurgery      0.4671599   0.0178117  26.228 < 2e-16 ***
## exper           0.0088515   0.0009082   9.746 < 2e-16 ***
## clinResearch    -0.1045826   0.0110233  -9.487 < 2e-16 ***
## rankAssociate    0.0670573   0.0118247   5.671 3.94e-08 ***
```

```

## rankFull professor          0.1113403  0.0131641   8.458 2.37e-15 ***
## genderMale:certNot certified 0.0069040  0.0194548   0.355   0.723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06696 on 248 degrees of freedom
## Multiple R-squared:  0.9339, Adjusted R-squared:  0.9307
## F-statistic: 292.1 on 12 and 248 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + cert + clin + rank
## Model 2: sal9495 ~ gender * cert + dept + exper + clin + rank
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     249 1.1126
## 2     248 1.1121  1 0.00056472 0.1259  0.723

not significant

##
## Call:
## lm(formula = sal9495 ~ gender * rank + dept + exper + clin +
##     cert, data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16333 -0.04040 -0.00537  0.03823  0.43343
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.0203633   0.0179001  336.331 < 2e-16 ***
## genderMale      0.0372396   0.0137840   2.702  0.00738 **

```

```

## rankAssociate          0.0865708  0.0169521   5.107 6.55e-07 ***
## rankFull professor    0.1411406  0.0197970   7.129 1.11e-11 ***
## deptGenetics          0.0922859  0.0181030   5.098 6.84e-07 ***
## deptMedicine          0.2716021  0.0146819  18.499 < 2e-16 ***
## deptPediatrics        0.1042339  0.0177642   5.868 1.41e-08 ***
## deptPhysiology        -0.0877718  0.0144357  -6.080 4.53e-09 ***
## deptSurgery           0.4656938  0.0176337  26.409 < 2e-16 ***
## exper                 0.0090856  0.0009028  10.064 < 2e-16 ***
## clinResearch          -0.0985156  0.0110875  -8.885 < 2e-16 ***
## certNot certified     -0.0956067  0.0106813  -8.951 < 2e-16 ***
## genderMale:rankAssociate -0.0414716  0.0223749  -1.853 0.06501 .
## genderMale:rankFull professor -0.0526354  0.0233270  -2.256 0.02492 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06627 on 247 degrees of freedom
## Multiple R-squared:  0.9355, Adjusted R-squared:  0.9322
## F-statistic: 275.8 on 13 and 247 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Model 1: sal9495 ~ gender + dept + exper + cert + clin + rank
## Model 2: sal9495 ~ gender * rank + dept + exper + clin + cert
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1      249 1.1126
## 2      247 1.0847  2  0.027927 3.1796 0.04331 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
significant.

```

### Final Model with Interaction:

```
##
## Call:
## lm(formula = sal9495 ~ gender * rank + dept + clin + cert + rank,
##     data = lawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19136 -0.05284 -0.00350  0.04668  0.41145
##
## Coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.09474    0.01932  315.457 < 2e-16 ***
## genderMale        0.04073    0.01633   2.494 0.013278 *
## rankAssociate     0.11798    0.01975   5.975 7.94e-09 ***
## rankFull professor 0.20597    0.02218   9.285 < 2e-16 ***
## deptGenetics      0.08033    0.02141   3.752 0.000218 ***
## deptMedicine      0.25273    0.01726  14.645 < 2e-16 ***
## deptPediatrics    0.07521    0.02077   3.620 0.000356 ***
## deptPhysiology    -0.10289    0.01701  -6.047 5.39e-09 ***
## deptSurgery       0.45045    0.02082  21.636 < 2e-16 ***
## clinResearch      -0.10166    0.01313  -7.740 2.52e-13 ***
## certNot certified -0.11131    0.01252  -8.889 < 2e-16 ***
## genderMale:rankAssociate -0.02136    0.02641  -0.809 0.419483
## genderMale:rankFull professor -0.03481    0.02756  -1.263 0.207822
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07853 on 248 degrees of freedom
```

```
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9047
```

```
## F-statistic: 206.7 on 12 and 248 DF,  p-value: < 2.2e-16
```

### Remove outlier

```
## Potentially influential observations of
```

```
##  lm(formula = sal9495 ~ gender * rank + dept + clin + cert + rank,      data = lawsuit) :
```

```
##
```

```
##      dfb.1_ dfb.gndM dfb.rnkA dfb.rnFp dfb.dptG dfb.dptM dfb.dptPd
```

```
## 6   -0.17  -0.04    0.01   -0.02    0.14    0.22    0.17
```

```
## 19   0.11   0.05   -0.01   -0.01   -0.17   -0.19   -0.15
```

```
## 39   0.00   0.00    0.00    0.00    0.00    0.00    0.00
```

```
## 58  -0.08   0.09   -0.03    0.02    0.11    0.02   -0.01
```

```
## 82   0.01   0.12   0.15    0.14   -0.01   -0.08   -0.06
```

```
## 86   0.00   0.00    0.00    0.00    0.00    0.00    0.00
```

```
## 101  0.00   0.00    0.00    0.08    0.06    0.00    0.00
```

```
## 109 -0.01   0.00    0.03    0.00    0.03    0.01    0.00
```

```
## 127  0.00   0.00    0.00   -0.04    0.00    0.00   -0.02
```

```
## 133  0.00   0.00    0.00    0.06    0.00    0.00    0.03
```

```
## 184 -0.81   0.88    0.11    0.12    0.12    0.83    0.49
```

```
## 208  0.23  -0.26   -0.15   -0.16   -0.03    0.09   -0.10
```

```
## 216 -0.05   0.01    0.01    0.13    0.02    0.07    0.02
```

```
##      dfb.dptPh dfb.dptS dfb.clnR dfb.crNc dfb.gM:A dfb.gM:p dffit  cov.r
```

```
## 6    0.22     0.20   -0.07    0.16    0.01   -0.05   -0.46   0.79_*
```

```
## 19  -0.20    -0.16    0.01    0.13    0.14   -0.02    0.50   0.82_*
```

```
## 39   0.00     0.00    0.00    0.00    0.00    0.00    0.00   1.17_*
```

```
## 58   0.49    -0.01    0.24   -0.37    0.27   -0.10   0.92_*  0.56_*
```

```
## 82  -0.30    -0.11   -0.10   -0.18   -0.08   -0.09   -0.51   0.83_*
```

```
## 86   0.00     0.00    0.00    0.00    0.00    0.00    0.00   1.16_*
```

```
## 101 -0.01     0.00   -0.02    0.02    0.00   -0.06    0.12   1.17_*
```

```
## 109  0.00     0.01    0.02   -0.02   -0.02    0.00    0.05   1.18_*
```

```

## 127  0.00      0.00      0.00      0.00      0.00      0.03      -0.05      1.16_*
## 133  0.00      0.00      0.00      0.00      0.00     -0.04      0.07      1.16_*
## 184  0.19      0.40      0.82      0.67     -0.64     -0.52      1.57_*  0.22_*
## 208 -0.03     -0.04     -0.10     -0.08      0.14      0.15      0.47      0.70_*
## 216  0.01      0.04      0.09     -0.04     -0.01     -0.11      0.17      1.17_*

##      cook.d hat
## 6      0.02  0.03
## 19     0.02  0.04
## 39     0.00  0.10
## 58     0.06  0.06
## 82     0.02  0.04
## 86     0.00  0.09
## 101    0.00  0.11
## 109    0.00  0.11
## 127    0.00  0.09
## 133    0.00  0.09
## 184    0.17  0.07
## 208    0.02  0.03
## 216    0.00  0.11

## # A tibble: 13 x 18
##   .rownames  dfb.1_ dfb.gndM dfb.rnkA dfb.rnFp dfb.dptG dfb.dptM
##   <chr>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 6          -1.66e-1 -4.07e-2  1.31e-2 -1.76e-2  1.42e-1  2.18e-1
## 2 19          1.14e-1  4.59e-2 -1.07e-2 -8.10e-3 -1.72e-1 -1.95e-1
## 3 39          -4.67e-4  2.83e-5  3.41e-5 -6.04e-4  3.00e-4  4.70e-4
## 4 58          -8.21e-2  8.72e-2 -3.49e-2  1.85e-2  1.13e-1  2.18e-2
## 5 82          1.14e-2  1.23e-1  1.48e-1  1.37e-1 -6.21e-3 -8.09e-2
## 6 86          1.17e-4  2.87e-5 -1.22e-4  3.01e-3 -5.34e-4 -3.28e-4
## 7 101         -1.65e-3 -1.98e-3 -3.64e-3  7.67e-2  5.67e-2 -3.41e-3

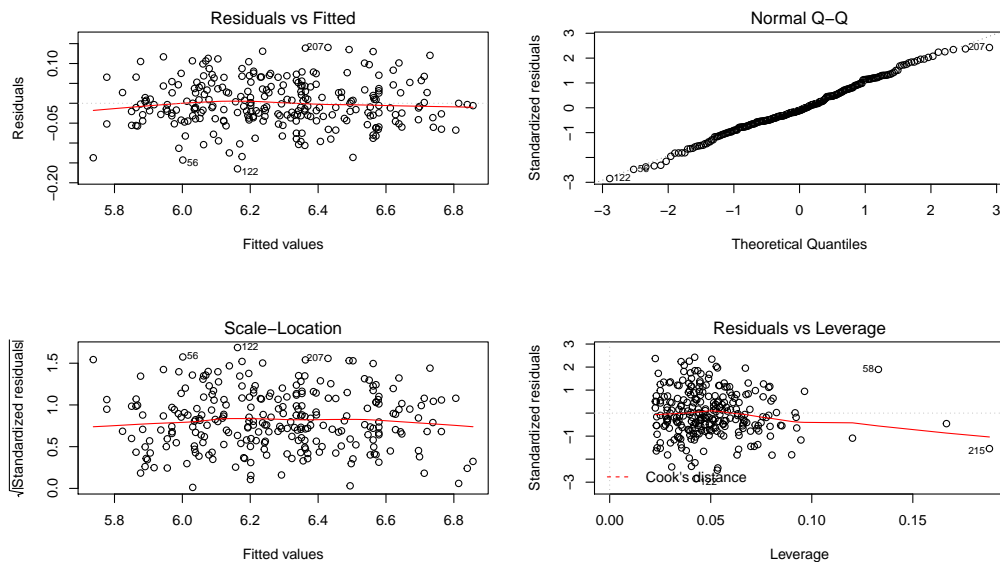
```

```
## 8 109      -6.94e-3  9.20e-4  2.58e-2  2.22e-3  2.88e-2  6.23e-3
## 9 127      2.78e-3 -2.62e-3 -3.93e-4 -3.71e-2  9.91e-4  5.76e-4
## 10 133     -4.16e-3  3.91e-3  5.87e-4  5.54e-2 -1.48e-3 -8.60e-4
## 11 184     -8.07e-1  8.79e-1  1.09e-1  1.24e-1  1.19e-1  8.27e-1
## 12 208      2.28e-1 -2.58e-1 -1.50e-1 -1.56e-1 -2.90e-2  9.06e-2
## 13 216     -5.40e-2  5.81e-3  9.72e-3  1.33e-1  2.24e-2  7.20e-2
## # ... with 11 more variables: dfb.dptPd <dbl>, dfb.dptPh <dbl>,
## #   dfb.dptS <dbl>, dfb.clnR <dbl>, dfb.crNc <dbl>, dfb.gM.A <dbl>,
## #   dfb.gM.p <dbl>, dffit <dbl>, cov.r <dbl>, cook.d <dbl>, hat <dbl>

##
## Call:
## lm(formula = sal9495 ~ gender * exper + rank + dept + clin +
##     cert, data = newlawsuit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.164921 -0.036222 -0.006406  0.040199  0.140506
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.009705   0.018933  317.415 < 2e-16 ***
## genderMale      0.049018   0.016779   2.921  0.00381 **
## exper          0.013290   0.001601   8.298 6.91e-15 ***
## rankAssociate   0.065807   0.010707   6.146 3.16e-09 ***
## rankFull professor 0.108273   0.011795   9.179 < 2e-16 ***
## deptGenetics    0.091972   0.016170   5.688 3.62e-08 ***
## deptMedicine    0.259894   0.013225  19.652 < 2e-16 ***
## deptPediatrics  0.099663   0.015997   6.230 2.00e-09 ***
## deptPhysiology -0.086228   0.012983  -6.641 1.96e-10 ***
```



```
## deptSurgery      0.461249    0.015791   29.209   < 2e-16 ***
## clinResearch     -0.113043    0.009759  -11.584   < 2e-16 ***
## certNot certified -0.099511    0.009526  -10.446   < 2e-16 ***
## genderMale:exper -0.004838    0.001621   -2.984    0.00313 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05918 on 247 degrees of freedom
## Multiple R-squared:  0.9483, Adjusted R-squared:  0.9458
## F-statistic: 377.6 on 12 and 247 DF,  p-value: < 2.2e-16
```



Stratification:

**on rank**

```
##
## Call:
## lm(formula = sal9495 ~ gender + dept + clin + cert + exper, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.134447 -0.030859 0.004221 0.034784 0.093496
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.103475   0.028568 213.650 < 2e-16 ***
## genderMale     -0.006639   0.015506  -0.428 0.670252
## deptGenetics    0.085294   0.027870   3.060 0.003439 **
## deptMedicine    0.253549   0.025026  10.132 4.30e-14 ***
## deptPediatrics  0.105034   0.027960   3.757 0.000424 ***
## deptPhysiology -0.094671   0.021741  -4.355 5.99e-05 ***
## deptSurgery     0.465950   0.028550  16.321 < 2e-16 ***
## clinResearch   -0.110123   0.018853  -5.841 3.06e-07 ***
## certNot certified -0.100244   0.015902  -6.304 5.53e-08 ***
## exper          0.010756   0.001309   8.214 4.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05164 on 54 degrees of freedom
## Multiple R-squared:  0.9621, Adjusted R-squared:  0.9558
## F-statistic: 152.2 on 9 and 54 DF, p-value: < 2.2e-16
##
## Call:
## lm(formula = sal9495 ~ gender + dept + clin + cert + exper, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15497 -0.04615 -0.00685  0.03846  0.39427
##
## Coefficients:
```

```

##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.975397   0.037982 157.323 < 2e-16 ***
## genderMale        0.041328   0.017673   2.338 0.021316 *
## deptGenetics       0.071716   0.034593   2.073 0.040681 *
## deptMedicine       0.300403   0.030516   9.844 < 2e-16 ***
## deptPediatrics     0.127710   0.033358   3.829 0.000223 ***
## deptPhysiology    -0.100657   0.031689  -3.176 0.001973 **
## deptSurgery        0.471541   0.035022  13.464 < 2e-16 ***
## clinResearch      -0.089536   0.021083  -4.247 4.80e-05 ***
## certNot certified -0.059914   0.020445  -2.931 0.004176 **
## exper             0.012368   0.002677   4.620 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07707 on 102 degrees of freedom
## Multiple R-squared:  0.9126, Adjusted R-squared:  0.9049
## F-statistic: 118.3 on 9 and 102 DF,  p-value: < 2.2e-16
##
## Call:
## lm(formula = sal9495 ~ gender + dept + clin + cert + exper, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.139999 -0.039721 -0.000282  0.037863  0.126949
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.189109   0.028450 217.541 < 2e-16 ***
## genderMale     -0.020206   0.018108  -1.116  0.26805

```

```

## deptGenetics      0.124412    0.027602    4.507 2.38e-05 ***
## deptMedicine      0.264095    0.019605   13.471 < 2e-16 ***
## deptPediatrics    0.083397    0.033185    2.513 0.01412 *
## deptPhysiology    -0.063930    0.019417   -3.292 0.00152 **
## deptSurgery       0.474317    0.025165   18.848 < 2e-16 ***
## clinResearch      -0.089652    0.016706   -5.367 8.61e-07 ***
## certNot certified -0.129005    0.016995   -7.591 7.14e-11 ***
## exper             0.007436    0.001126    6.601 5.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05823 on 75 degrees of freedom
## Multiple R-squared:  0.9511, Adjusted R-squared:  0.9452
## F-statistic: 161.9 on 9 and 75 DF,  p-value: < 2.2e-16

```

Gender is only significant in assistant professors' salary.

## Conclusions

## Discussion

## Figures and tables

	Female (N=106)	Male (N=155)	Total (N=261)	p value
Department				< 0.001
- Biochemistry	20 (18.9%)	30 (19.4%)	50 (19.2%)	
- Genetics	11 (10.4%)	10 (6.5%)	21 (8.0%)	
- Medicine	30 (28.3%)	50 (32.3%)	80 (30.7%)	
- Pediatrics	20 (18.9%)	10 (6.5%)	30 (11.5%)	
- Physiology	20 (18.9%)	20 (12.9%)	40 (15.3%)	
- Surgery	5 (4.7%)	35 (22.6%)	40 (15.3%)	
- Missing	0	0	0	
Clinical				0.197
- Clinical	60 (56.6%)	100 (64.5%)	160 (61.3%)	
- Research	46 (43.4%)	55 (35.5%)	101 (38.7%)	
- Missing	0	0	0	
Certified				0.074
- Board certified	70 (66.0%)	118 (76.1%)	188 (72.0%)	
- Not certified	36 (34.0%)	37 (23.9%)	73 (28.0%)	
- Missing	0	0	0	
Publication Rate				0.004
- Mean (SD)	5.350 (1.886)	4.646 (1.938)	4.932 (1.944)	
- Median (IQR)	5.250 (3.725, 7.275)	4.000 (3.100, 6.700)	4.400 (3.200, 6.900)	
- Min - Max	2.400 - 8.700	1.300 - 8.600	1.300 - 8.700	
- Missing	0	0	0	
Years since obtaining				< 0.001
MD				
- Mean (SD)	7.491 (4.166)	12.103 (6.704)	10.230 (6.227)	
- Median (IQR)	7.000 (5.000, 10.000)	10.000 (7.000, 15.000)	9.000 (6.000, 14.000)	
- Min - Max	1.000 - 23.000	2.000 - 37.000	1.000 - 37.000	
- Missing	0	0	0	
Rank				< 0.001

(continued)

	Female (N=106)	Male (N=155)	Total (N=261)	p value
- Assistant	69 (65.1%)	43 (27.7%)	112 (42.9%)	
- Associate	21 (19.8%)	43 (27.7%)	64 (24.5%)	
- Full professor	16 (15.1%)	69 (44.5%)	85 (32.6%)	
- Missing	0	0	0	
Salary in academic year 1994				< 0.001
- Mean (SD)	118871.274 (56168.006)	177338.761 (85930.540)	153593.345 (80469.667)	
- Median (IQR)	108457.000 (75774.500, 143096.000)	155006.000 (109687.000, 231501.500)	133284.000 (90771.000, 200543.000)	
- Min - Max	34514.000 - 308081.000	52582.000 - 428876.000	34514.000 - 428876.000	
- Missing	0	0	0	
Salary after increment to Sal94				< 0.001
- Mean (SD)	130876.915 (62034.507)	194914.090 (94902.728)	168906.655 (88778.425)	
- Median (IQR)	119135.000 (82345.250, 154170.500)	170967.000 (119952.500, 257163.000)	148117.000 (99972.000, 218955.000)	
- Min - Max	38675.000 - 339664.000	58923.000 - 472589.000	38675.000 - 472589.000	
- Missing	0	0	0	

## References

## Appendix