STA 141A
Fundamentals of Statistical Data Science Final Project

Instructor: Vaidotas Characiejus

Group member:

Wenfeng Chang (wfechang@ucdavis.edu)

Yeping Dong(yepdong@ucdavis.edu)
Yinglin Luo(lylluo@ucdavis.edu)
Xuecheng Zhang(hcczhang@ucdavis.edu)

Date: Friday, December 9, 2019

Introduction

Now more and more undergraduates choose to go to graduate school in order to obtain a higher salary after graduation, according to the integrated postsecondary education data system (IPEDS): “Between 2000 and 2017, total post baccalaureate enrollment increased by 39 percent (from 2.2 million to 3.0 million students)”. Our data was collected from a college in the U.S. that contains information on the 2008-2009 nine-month academic salaries of Assistant Professors, Associate Professors, and Professors. There are 397 observations on the following six variables, including 3 qualitative variables: Rank(Professor or Assocprofessor or Asstprofessor), Discipline(theoretical departments(A) or applied departments(B)), Sex (Male or Female), and Years since Ph.D.; Years of service, Salary(nine-month salary, in dollars). So we are going to study what factors affect the salary of professors.

Methods We will Use

Data exploration and Visualization

The first step is processing the data set. Data exploration and visualization is a good way for us to see the distribution of our variables so that we can identify and resolve data error.

Single Linear Regression

single linear regression is a tool for us to find the relationship between the predictor variables and response variables. It is very important for us to do interpretation for our final model.

Multiple Linear Regression

multiple linear regression helps us to find the best model fit all of the data and it can predict a quantitative response. We can pick up the best model by comparing R squared and adjusted R squared and coefficients of the different models.

Model diagnostics

We can draw Residual Versus Fitted Values Plot and Residual QQ-plot to check model assumptions including normality, equal variance. Because F-test and related procedures are pretty robust to the normality and equal variance assumptions. Once we find non-euqal variance and the QQ-plot is not linear, we can transform the response variable to stabilize the variance, which makes the distribution closer to normal. What's more, we also need to remove the potential high influential points since those points can also influence the model diagnostics.

Results of Data Visualization

Nine-month Salary:

Form the salary histogram, It seems to be skewed right. After we summary the data, we find a minimum salary of is $57,800 and the maximum salary is $231,545. The median is $107,300, and the mean is 113706. The first quarter is $91000, and the third quarter is 134185. (As shown in Appendix 1- Figure 1)

Years since graduating from Ph.D:

Form the histogram, it seems to be a normal distribution. After we summary our data, we find the median is 21 years, and the mean is 22.31 years. The minimum is 1 year and the maximum is 56 years. It means half of the faculty members received their Ph.D. between 1 and 21 years ago, and half of the faculty members received their Ph.D. between 22 and 56 years ago. (As shown in Appendix 1- Figure 2)

Years of Service:

Form the histogram, it seems to be slightly skewed right. After we summary our data, we find that the minimize years of service is 0 years, and the maximum years of service are 60 years. The median is 16 years and the mean is 17.61 years. (As shown in Appendix 1- Figure 3)

Relationship between Years since Ph.D. and Salary and Rank

From our plot, we can find Years since Ph.D. is positively correlated with salary. Moreover, we can clearly find the salary of Prof is higher than AsstProf and AssocProf, and the counts of Prof also are the largest. In general, the salary of the Ph.D. will increase when the number of years since the Ph.D. increase. However, we also can find the ranks are positively correlated with Years since Ph.D. and salary as well. AsstProf has the lowest salary by between $57800-$10000, AssocProf has middle salary by

between $50000-$125000, and Prof has the largest salary is $231545. (As shown in Appendix 1- Figure 4)

Relationship between Years since Ph.D. and Salary and Discipline

From our plot, we find that there is not a strong relationship between salary and discipline, or between discipline and Years since Ph.D. However, we find the data points of faculty members in applied departments(B) are larger than theoretical departments(A). (As shown in Appendix 1- Figure 5)

Relationship between Years of Service and Salary and Rank

From our plot, it seems does not appear a strong relationship between salary and years of service, but there is a relationship between years of service and rank. Group Prof has the highest years of service so they can get a higher salary. (As shown in Appendix 1- Figure 6)

Relationship between Years since Ph.D. and Salary and Sex

From our plot, it does not appear any relationship between salary and sex, at years of service and sex. After we summary, the counts of male (358) is larger than female (39). It seems to have some impact on our results. (As shown in Appendix 1- Figure 7)

Results in Single Regression Analysis Model

Salary ~ Rank Model:

The model is $y = 80776 + 13100x_1 + 45996x_2$

X1 is an associate professor, X2 is a professor. First of all, since the p-value of X1, X2, and Intercept are smaller than 0.05, so we reject the null hypothesis at the significance level at 0.05, and we conclude that rank is a factor that affects the salary. Moreover, the Multiple R-squared is 0.3943, it means 39.43% of data can be explained by Rank.

The average salary that an assistant professor at this school will get in nine months is $80776 (intercept). The associate professor also makes an additional $13100, and the professor makes an additional $45996 compared with the assistant professor.

Salary ~ Sex:

The model is $y = 101002 + 14088x$

X is a male faculty. Since the p-value of X and Intercept are smaller than 0.05, we reject the null hypothesis at the significance level at 0.05 and conclude that sex is a factor that affects the salary. However, the Multiple R-squared is 0.01921, it means only 1.921% of data can be explained by Sex.

The average salary that a female faculty in this school will get in nine months is $101002 (intercept). The male faculty makes an additional $14088 compared with the female faculty.

Salary ~ Discipline

The model is $y = 108548 + 9480x$

X is a faculty member in applied departments. Similar to the results above for sex, since the p-value of X and Intercept are smaller than 0.05, so we reject the null hypothesis at the significance level at 0.05. So discipline is a factor that affects the salary, but the Multiple R-squared is 0.02436, so only 2.436% of data can be explained by Discipline.
The average salary that faculty members in theoretical departments in this school will get in nine months is $108548 (intercept). The faculty members in applied departments make an additional $9480 compared with the female faculty.

Salary ~ Years Since Ph.D.

The model is $y = 91718.7 + 985.3x$

X is the year since the Ph.D. Since the p-value of X and Intercept are smaller than 0.05, we reject the null hypothesis at the significance level at 0.05, and conclude that years since Ph.D. is a factor that affects the salary. Moreover, the Multiple R-squared is 0.1758, it means 17.58% of data can be explained by Years since Ph.D.
Since receiving from Ph.D., a faculty member's salary estimate increases $985.3 for each additional year.

Salary ~ Years of Service

The model is $y = 99974.7 + 779.6x$

X is the year of service. Since the p-value of X and Intercept are smaller than 0.05. We reject the null hypothesis at the significance level at 0.05 and conclude years of service is a factor that affects the salary. Moreover, the Multiple R-squared is 0.1121, it means 11.21% of data can be explained by years of service.
For each additional year since receiving their Ph.D., a faculty member's salary estimate increases by an additional $779.6.

Results of Multiple Linear Regression

Full model with salary regressed on all predictors

Since the p-value of rank, discipline, years since Ph.D., years service and Intercept are smaller than 0.05, we reject the null hypothesis at the significance level at 0.05 and conclude that rank, discipline, years since Ph.D., and years service are the factors that affect the salary. Moreover, sex males' p-value is larger than 0.05, so we fail to reject the null hypothesis. However, for years of service, the coefficient was 779.60 whereas in the full model here it is -489.50. This could be a sign of collinearity. So we think sex may have a less effective salary or not. We decide to take out the predictor Sex and test again. (As shown in Appendix 1- Figure 8)

Take Out the Predictor Sex

Since the p-value of rank, discipline, years since Ph.D., years service and Intercept are smaller than 0.05, we also reject the null hypothesis. The adjusted R-squared is 0.4455,

and the full model is 0.4463. Compare with the full model, new R-squared does not change much after removing the sex factor. But the coefficient of the years of service is still a negative number, so we need to check it is sign collinearity with other variations? (As shown in Appendix 1- Figure 9) We decided to use VIF method for the model: The years since Ph.D.'s GVIF is 7.51892 > 5, it is the largest one. .So it is a sign of collinearity that can considerably affect results. We need to take out the years since Ph.D. and test the model again.(As shown in Appendix 1- Figure 10)

Take Out the Predictor Sex and Years since PhD

In this model, after taking out years since Ph.D and sex. The multiple R-squared is 0.44, and the full model is 0.4463. Comparing with the full model, new R-squared does not change much. Even though it solved collinearity. (As shown in Appendix 1- Figure 11)
The VIF estimate for years of service went down to 1.61. However, the sign of the coefficient for years of service is still negative. (As shown in Appendix 1- Figure 12)

Take Out the Predictor Sex, and Rank

Let's try moving rank from the model instead of years since Ph.D. And see whether it is a good way to solve our problem. However, from the R output, removing rank from the model did not solve the problem-the sign of years of service is still negative. As for collinearity, years of service both have a VIF>5 which means that this is not a good way to solve our problem. (As shown in Appendix 1- Figure 13, Figure 14)

Take Out the Predictor Sex and Years of Service

In this model, the coefficient of the years since Ph.D. is a positive number, but the p-value is larger than 0.05, so we fail to reject the null hypothesis. Taking out years of service did not change the adjusted R-squared value much compared to the full model and none of the remaining variables have a high VIF. So we decided to choose discipline and rank to be a predictor of the full model.(As shown in Appendix 1- Figure 15, Figure 16)

Final Model

In our final model, the discipline and rank can explain 44.5% of the data, we think it is justified. The years since Ph.D. and years of service may not be well understood yet. However, it should be noted that the residual plot is not free of patterns. Thus, there is still work to do in order to identify the best model. (As shown in Appendix 1- Figure 17)

Results of Model Diagnostics

Draw Residual Versus Fitted Values Plot and Residual QQ-plot

The normal QQ plot (As shown in Appendix 1- Figure 18) and residuals plots (As shown in Appendix 1- Figure 19) are shown above. The normal QQ plot indicates the slightly heavy-tailed of the residuals' distribution; while the residuals versus fitted values plots show the sign for unequal variance. Hence we need to do some transformation to the response variable.

Use Log Transformation to the Response Variable

After log transformation to the response variable, we can see that the R-squared and Adjusted R-squared both increase. However, the residuals versus fitted values plots still show the sign for unequal variance and the normal QQ plot indicates the slightly heavy-tailed of the residuals' distribution; Hence we still need to do some transformation to the response variable. (As shown in Appendix 1- Figure 20, Figure 21, Figure22)

Use the Square Root Transformation

After square root transformation, the R-squared and Adjusted R-squared compared with the final model don't change a lot. What's more, the residuals versus fitted values plots still show the sign for unequal variance and the normal QQ plot indicates the slightly heavy-tailed of the residuals' distribution. It means that the square root transformation isn't a good solution. Therefore we still need to do some transformation to the response variable. (As shown in Appendix 1- Figure 23, Figure 24, Figure25)

Use the Inverse Transformation

After inverse transformation, the R-squared and Adjusted R-squared compared with the previous model increase a lot. However the residuals versus fitted values plots show a little unequal variance. The normal QQ plot looks not like a straight line. Maybe it's because there are so many outliers that the residuals plots don't look very good. (As shown in Appendix 1- Figure 26, Figure 27, Figure28)

Identify the Outliers and Remove Them

We use Cook's distance to determine whether the outlying cases (in Y and/or in X) are influential in determining the fitted regression function The Cook's distance measures the aggregate influence on all fitted values that are made by the omission of a single case in the fitting process. We use Di >4/(n-p) is as an indicator of being a potential influential case. We use red line indicates h=4/(n-p), then we can find that there are many points that are above the line, we can drop them out and fit the model again. After dropping the potential outliers we can find that the R-squared and Adjusted R-squared increases a lot. The R-squared increases from 0.5605 to 0.6157. The residuals versus fitted values plots show almost equal variance. The normal QQ plot is a straight line. which means that the error terms are normally distributed. Therefore, we can use this

model as our improved final model. (As shown in Appendix 1- Figure 29,Figure 31, Figure32)

Conclusion

In conclusion, the final model is:

$$\frac{1}{Salary} = 1.341\times 10^{-5} - 1.694\times 10^{-6}\times rankAsscoprof - 4.374\times 10^{-6}\times rankprof - 1.173\times 10^{-6}\times disciplineB$$

From the improved final function we can calculate that:
The mean salary of rank is Assistant professor and discipline is A is

$$\frac{10^5}{1.314} = 76103.7$$

the mean salary of rank is Assistant pr fessor and discipline is B is
o

$$\frac{10^5}{1.1967} = 83563.13.$$

the mean salary of rank is Associate professor and discipline is A is

$$\frac{10^5}{1.1446} = 87366.77.$$

the mean salary of rank is Associate professor and discipline is B is

$$\frac{10^5}{1.0273} = 97342.55.$$

the mean salary of rank is professor and discipline is A is

$$\frac{10^5}{0.8766} = 114077.1$$

the mean salary of rank is professor and discipline is B is

$$\frac{10^5}{0.7593} = 131700.3$$

In our improved final model, the discipline and rank can explain 61.57% of the data, we think this is justified. Through calculate the mean salary of different factor level we can find that the result is consistent with the boxplot. From the boxplot, we can find that the salary of the professor is greater than the salary of an Associate professor and the salary of the Associate professor is greater than the Assistant professor. On the other hand, the salary discipline B is greater than the salary of discipline A, which means that the salary of the “applied department” is greater than the salary of the “theoretical department”. This result is consistent with our common sense. Therefore, the best

model explained the previous steps as we visualized the whole dataset. (As shown in Appendix 1- Figure 30, Figure33)

Appendix 1: Figures and Tables

Data Visualization

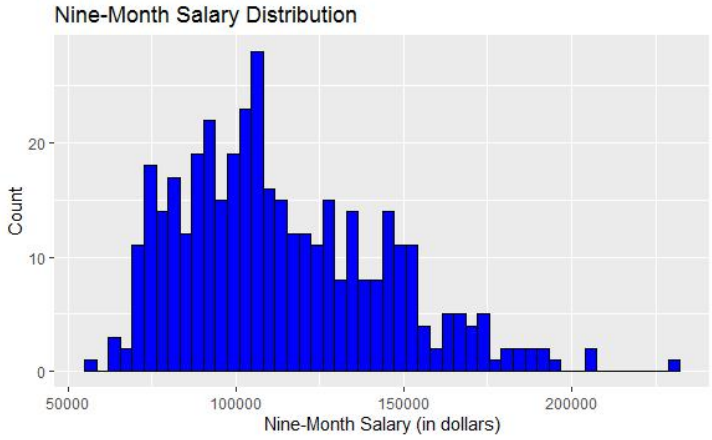

Figure 1: Histogram of Nine-month Salary

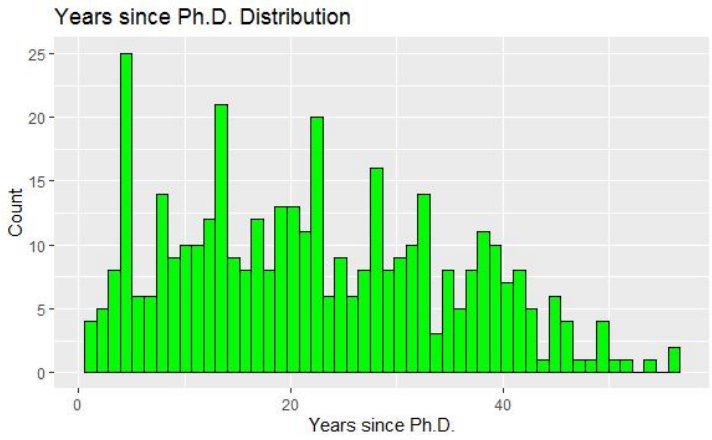

Figure 2: Histogram of Years since graduating from Ph.D.

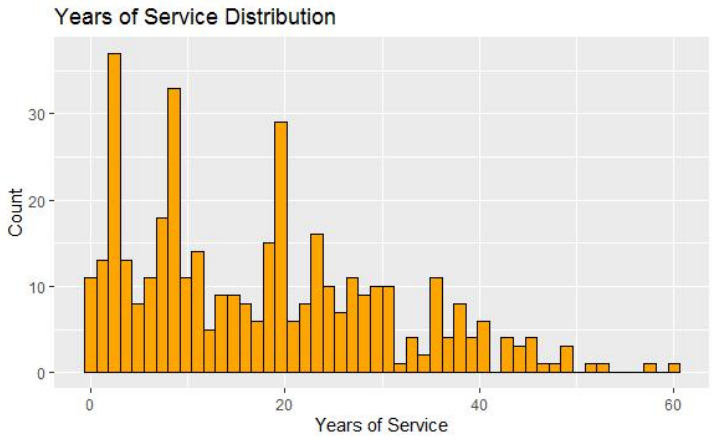

Figure 3: Histogram of Years of service

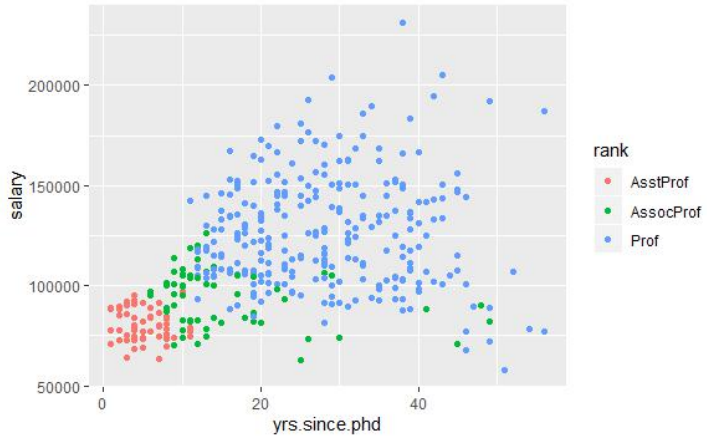

Figure 4: scatter plot of relationship between Years since Ph.D., Salary and Rank

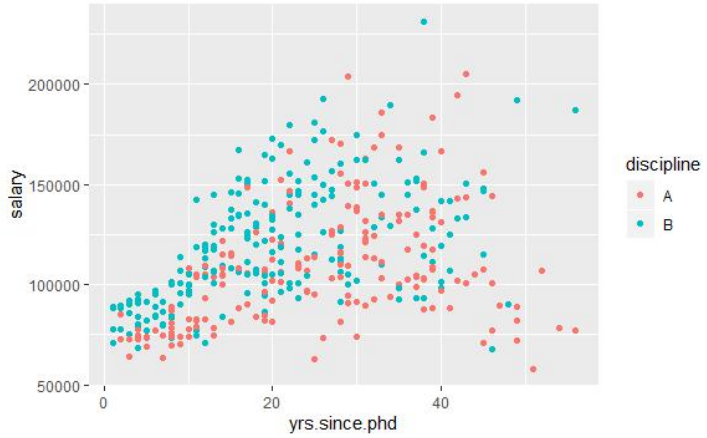

Figure 5: scatter plot of relationship between Years since Ph.D., Salary and Discipline

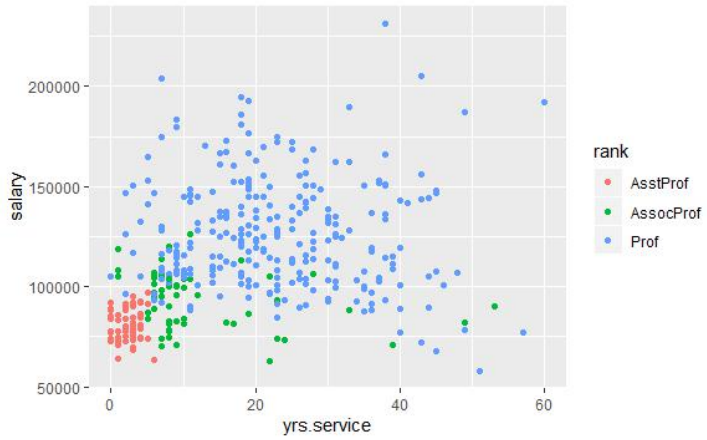

Figure 6: scatter plot of relationship between Years of service, Salary and Rank

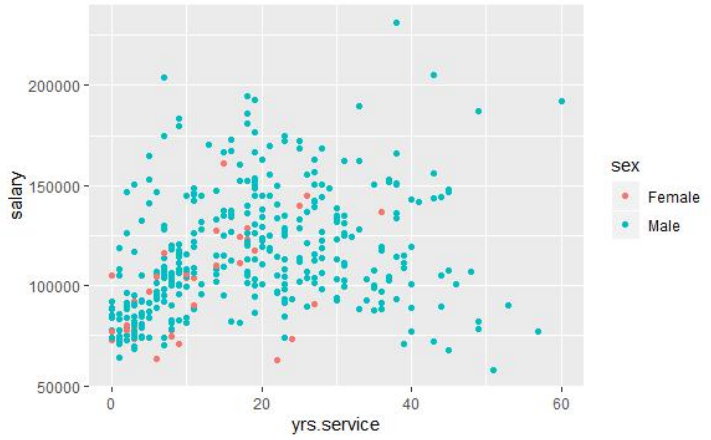

Figure 7: scatter plot of relationship between Years since Ph.D.,Salary and Sex

```
Call:
lm(formula = salary ~ ., data = df)

Residuals:
   Min     1Q Median     3Q    Max
-65248 -13211  -1775  10384  99592

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
rankAssocProf  12907.6     4145.3   3.114  0.00198 **
rankProf       45066.0     4237.5  10.635  < 2e-16 ***
disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
yrs.since.phd    535.1      241.0   2.220  0.02698 *
yrs.service     -489.5      211.9  -2.310  0.02143 *
sexMale         4783.5     3858.7   1.240  0.21584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22540 on 390 degrees of freedom
Multiple R-squared:  0.4547,   Adjusted R-squared:  0.4463
F-statistic:  54.2 on 6 and 390 DF,  p-value: < 2.2e-16
```

Figure 8: The summary of the full model with salary regressed on all predictors

```
Call:
lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
    data = df)

Residuals:
   Min     1Q Median     3Q    Max
-65244 -13498  -1455   9638  99682

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    69869.0     3332.1  20.968  < 2e-16 ***
rankAssocProf  12831.5     4147.7   3.094  0.00212 **
```

```
rankProf       45287.7     4236.7  10.689  < 2e-16 ***
disciplineB    14505.2     2343.4   6.190 1.52e-09 ***
yrs.since.phd    534.6      241.2   2.217  0.02720 *
yrs.service     -476.7      211.8  -2.250  0.02497 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22550 on 391 degrees of freedom
Multiple R-squared:  0.4525,   Adjusted R-squared:  0.4455
F-statistic: 64.64 on 5 and 391 DF,  p-value: < 2.2e-16
```

Figure 9: The summary of the model without predictor Sex

```
                GVIF Df GVIF^(1/(2*Df))
rank          2.003562  2        1.189736
discipline    1.063139  1        1.031086
yrs.since.phd 7.518920  1        2.742065
yrs.service   5.908984  1        2.430840
```

Figure 10: the VIF table of the model without predictor Sex

```
Call:
lm(formula = salary ~ discipline + yrs.service + rank, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-64198 -14040  -1299  10724  99253

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   72253.53    3169.48  22.797  < 2e-16 ***
disciplineB   13561.43    2315.91   5.856 1.01e-08 ***
yrs.service     -76.33     111.25  -0.686 0.493039
rankAssocProf 14483.23    4100.53   3.532 0.000461 ***
rankProf      49377.50    3832.90  12.883  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22670 on 392 degrees of freedom
Multiple R-squared:  0.4456,   Adjusted R-squared:   0.44
F-statistic: 78.78 on 4 and 392 DF,  p-value: < 2.2e-16
```

Figure 11: The summary of the model without predictor Sex and Years since Ph.D.

```
               GVIF Df GVIF^(1/(2*Df))
discipline  1.028057  1        1.013932
yrs.service 1.613750  1        1.270335
rank        1.588631  2        1.122679
```

Figure 12: the VIF table of the model without predictor Sex and Years since Ph.D

```
Call:
```

```
lm(formula = salary ~ discipline + yrs.since.phd + yrs.service,
    data = df)

Residuals:
   Min     1Q Median     3Q    Max
-76047 -17197  -4709  15904  97194

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    77486.4     3405.5  22.754  < 2e-16 ***
disciplineB    16480.7     2713.7   6.073 2.96e-09 ***
yrs.since.phd   1815.6      249.4   7.281 1.82e-12 ***
yrs.service     -752.8      244.5  -3.079  0.00222 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26190 on 393 degrees of freedom
Multiple R-squared:  0.258,    Adjusted R-squared:  0.2523
F-statistic: 45.55 on 3 and 393 DF,  p-value: < 2.2e-16
```

Figure 13: The summary of the model without predictor Sex and Rank

```
   discipline yrs.since.phd   yrs.service
     1.057281      5.961779      5.836347
```

Figure 14: the VIF table of the model without predictor Sex and Rank

```
Call:
lm(formula = salary ~ discipline + yrs.since.phd + rank, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-67395 -13480  -1536  10416  97166

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   71405.40    3278.32  21.781  < 2e-16 ***
disciplineB   14028.68    2345.90   5.980 5.03e-09 ***
yrs.since.phd    71.92     126.68   0.568   0.5706
rankAssocProf 13030.16    4168.17   3.126   0.0019 **
rankProf      46211.57    4238.52  10.903  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22670 on 392 degrees of freedom
Multiple R-squared:  0.4454,   Adjusted R-squared:  0.4398
F-statistic: 78.72 on 4 and 392 DF,  p-value: < 2.2e-16
```

Figure 15: The summary of the model without predictor Sex and Years of Service

```
                GVIF Df GVIF^(1/(2*Df))
discipline    1.054461  1        1.026869
yrs.since.phd 2.053425  1        1.432978
```

```
rank          1.978933  2        1.186063
```

Figure 16: the VIF table of the model without predictor Sex and Years of Service

```
Call:
lm(formula = salary ~ discipline + rank, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-65990 -14049  -1288  10760  97996

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     71944       3135  22.948  < 2e-16 ***
disciplineB     13761       2296   5.993 4.65e-09 ***
rankAssocProf   13762       3961   3.475 0.000569 ***
rankProf        47844       3112  15.376  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22650 on 393 degrees of freedom
Multiple R-squared:  0.445,    Adjusted R-squared:  0.4407
F-statistic:   105 on 3 and 393 DF,  p-value: < 2.2e-16
```

Figure 17: The summary of final model

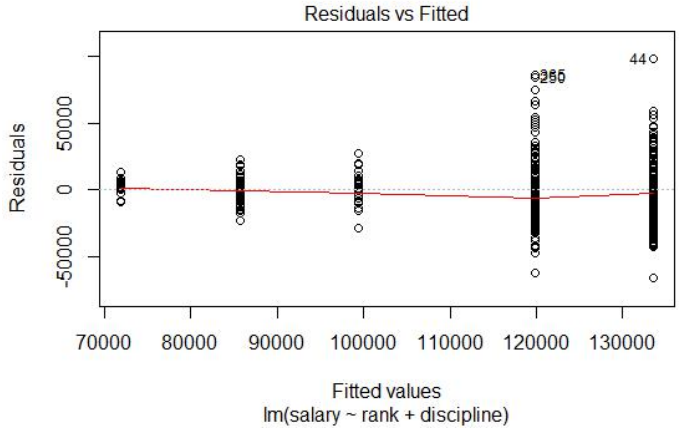

Figure 18: Residuals vs Fitted plot of final model

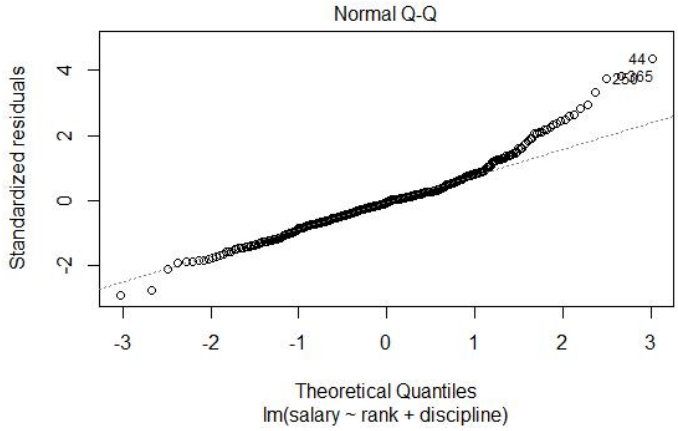

Figure 19: QQ plot of final model

```
Call:
lm(formula = salary1 ~ rank + discipline)

Residuals:
     Min       1Q   Median       3Q      Max
-0.69579 -0.11128 -0.00407  0.09423  0.57267

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.21068    0.02515 445.720  < 2e-16 ***
rankAssocProf  0.15058    0.03177   4.739 3.01e-06 ***
rankProf       0.44986    0.02496  18.021  < 2e-16 ***
disciplineB    0.13040    0.01842   7.079 6.72e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1817 on 393 degrees of freedom
Multiple R-squared:  0.5159,  Adjusted R-squared:  0.5122
F-statistic: 139.6 on 3 and 393 DF,  p-value: < 2.2e-16
```

Figure 20: the summary of the log transformation model

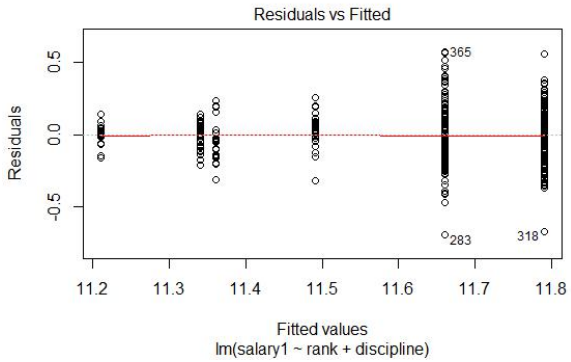

Figure 21: residuals vs fitted plots of log transformation model

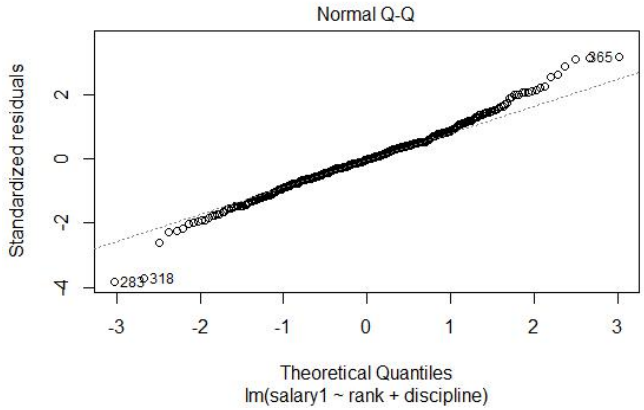

Figure 22: QQ plot of log transformation model

```
Call:
lm(formula = salary2 ~ rank + discipline)

Residuals:
     Min       1Q   Median       3Q      Max
-104.420  -19.355   -1.084   15.989  116.850

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    270.343      4.391  61.572  < 2e-16 ***
rankAssocProf   22.715      5.547   4.095 5.12e-05 ***
rankProf        72.949      4.358  16.741  < 2e-16 ***
disciplineB     21.049      3.216   6.546 1.85e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.72 on 393 degrees of freedom
Multiple R-squared:  0.483,    Adjusted R-squared:  0.4791
F-statistic: 122.4 on 3 and 393 DF,  p-value: < 2.2e-16
```

Figure 23: the summary of square root transformation model

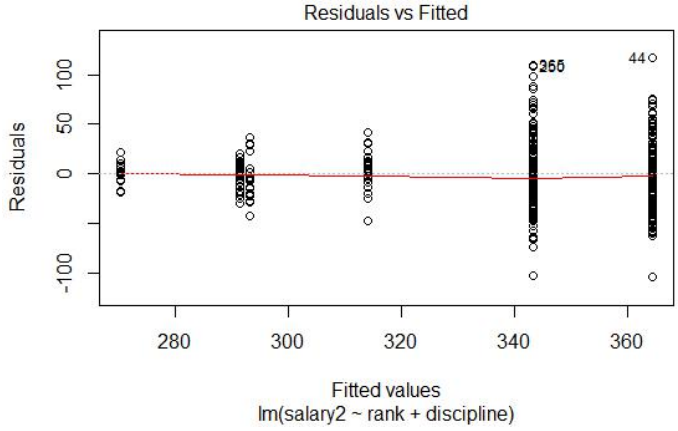

Figure 24: residuals vs fitted plots of square root transformation model

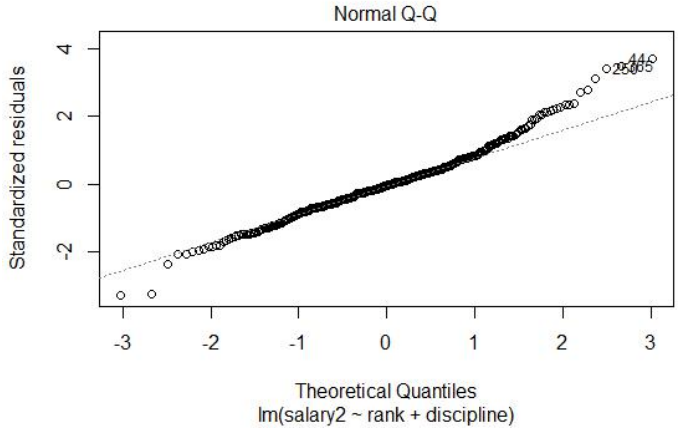

Figure 25: QQ plot of square root transformation model

```
Call:
lm(formula = salary3 ~ rank + discipline)

Residuals:
       Min         1Q     Median         3Q        Max
-4.060e-06 -1.009e-06 -8.010e-08  8.281e-07  8.375e-06

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.334e-05  2.224e-07  59.971  < 2e-16 ***
rankAssocProf -1.674e-06  2.810e-07  -5.956 5.73e-09 ***
rankProf      -4.413e-06  2.207e-07 -19.989  < 2e-16 ***
disciplineB   -1.297e-06  1.629e-07  -7.963 1.83e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.607e-06 on 393 degrees of freedom
Multiple R-squared:  0.5605,   Adjusted R-squared:  0.5571
F-statistic: 167.1 on 3 and 393 DF,  p-value: < 2.2e-16
```

Figure 26: the summary of inverse transformation model

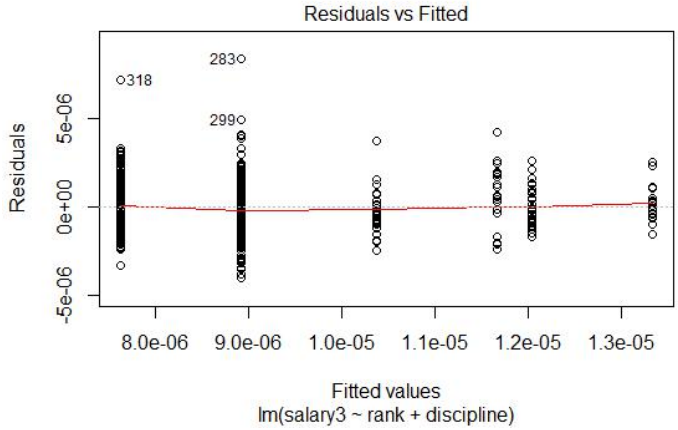

Figure 27: residuals vs fitted plots of inverse transformation model

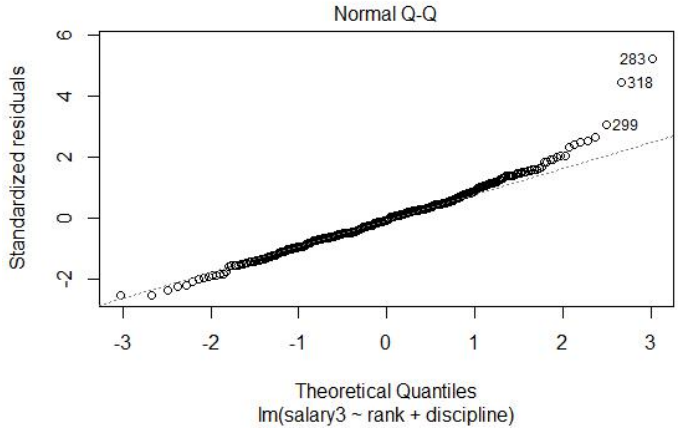

Figure 28: QQ plot of inverse transformation model

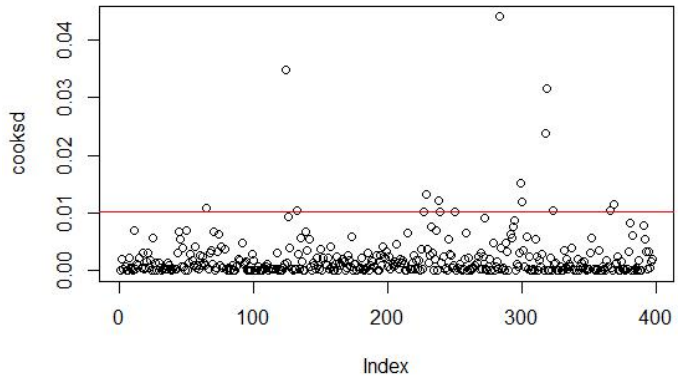

Figure 29: Cook's distance plot

```
Call:
lm(formula = salary4 ~ Salaries1$rank + Salaries1$discipline)

Residuals:
       Min         1Q     Median         3Q        Max
-3.866e-06 -9.444e-07 -1.700e-09  8.612e-07  4.026e-06

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.314e-05  1.990e-07  66.037  < 2e-16 ***
Salaries1$rankAssocProf -1.694e-06  2.536e-07  -6.681 8.51e-11 ***
Salaries1$rankProf      -4.374e-06  1.962e-07 -22.294  < 2e-16 ***
Salaries1$disciplineB   -1.173e-06  1.450e-07  -8.091 8.17e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.398e-06 on 378 degrees of freedom
Multiple R-squared:  0.6157,   Adjusted R-squared:  0.6126
```

```
F-statistic: 201.8 on 3 and 378 DF,  p-value: < 2.2e-16
```

Figure 30: the summary of model without outliers

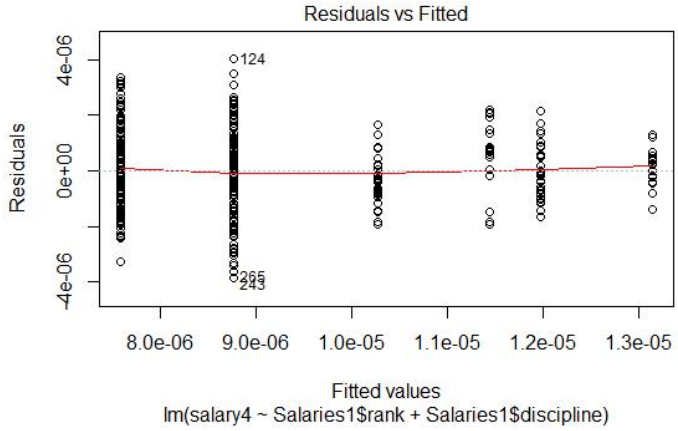

Figure 31: residuals vs fitted plots of model without outliers

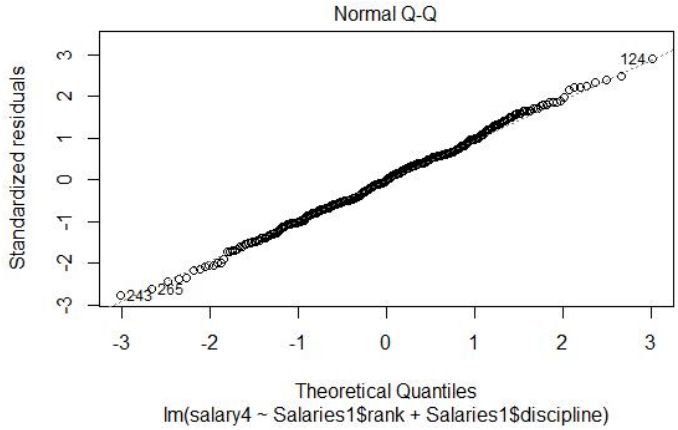

Figure 32: QQ plot of model without outliers

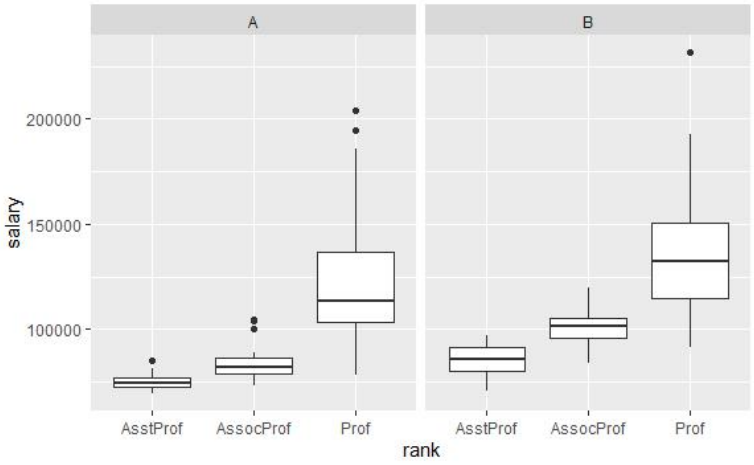

Figure 33: boxplot of final model

Appendix 2: R code

1.descriptive statistics
1.1 library package and read data

```
library(dplyr)
library(car)
df<-Salaries
nrow(df)
colnames(df)
```

1.2 check the sum of the qualitative variables

```
rank_sum <- summary(df$rank)
discpl_sum <- summary(df$discipline)
sex_sum <- summary(df$sex)
rank_sum
discpl_sum
sex_sum
```

1.3 plot histograms for the quantitative variable3 and check their summary statistic
1.3.1 salary

```
library(ggplot2)
salary_hist <- df %>% ggplot() +
geom_histogram(aes(x = salary), color = "black", fill = "blue", bins = 50) +
labs(x = "Nine-Month Salary (in dollars)", y = "Count", title = "Nine-Month Salary
Distribution")
salary_hist
summary(df$salary)
```

1.3.2 years since Ph.D

```
phd_hist <- df %>% ggplot() +
geom_histogram(aes(x = yrs.since.phd), color = "black", fill = "green", bins = 50)
+labs(x = "Years since Ph.D.", y = "Count", title = "Years since Ph.D.
Distribution")
phd_hist
summary(df$yrs.since.phd)
```

1.3.3 years of service

```
service_hist <- df %>% ggplot() +
geom_histogram(aes(x = yrs.service), color = "black", fill = "orange", bins = 50)
+labs(x = "Years of Service", y = "Count",title = "Years of Service Distribution")
service_hist
summary(df$yrs.service)
```

2 Assess the relationship between the quantitative variable and qualitative variable

2.1 Assessing years since Ph.D,salary,and rank

```
library(ggplot2)
attach(Salaries)
ggplot(mapping=aes(x=yrs.since.phd,y=salary)) +
geom_point(mapping=aes(colour=rank))
```

2.2 Assessing years since Ph.D, salary and discipline

```
library(ggplot2)
ggplot(mapping=aes(x=yrs.since.phd,y=salary)) +
geom_point(mapping=aes(colour=discipline))
```

2.3 Assessing Salary, years of service and rank

```
library(ggplot2)
ggplot(mapping=aes(x=yrs.service,y=salary)) +
```

```
geom_point(mapping=aes(colour=rank))
```

2.4 Assessing salary, sex, years of service

```
library(ggplot2)
ggplot(mapping=aes(x=yrs.service,y=salary)) +
geom_point(mapping=aes(colour=sex))
```

3. single linear regression analysis

3.1 check outlier

```
boxplot(salary)
boxplot(yrs.service)
boxplot(yrs.since.phd)
```

from the boxplot, we can find that there is no obvious outlier in salary, years of service and years since Ph.D

3.2 salary~rank

```
mod_rank_slr <- lm(salary~rank, data = df)
summary(mod_rank_slr)
```

3.3 salary~sex

```
mod_sex_slr <- lm(salary~sex, data =df)
summary(mod_sex_slr)
```

3.4 salary~discipline

```
mod_discpl_slr <- lm(salary~discipline, data = df)
summary(mod_discpl_slr)
```

3.5 salary~years since Ph.D

```
mod_phd_slr <- lm(salary~yrs.since.phd, data = df)
summary(mod_phd_slr)
```

3.6 salary~years of service

```
mod_service_slr <- lm(salary~yrs.service,data=df)
summary(mod_service_slr)
```

4 Multivariate regression analysis
4.1 full model

```
full_mod <- lm(salary~., data = df)
summary(full_mod)
```

4.2 model without sex

```
mod_no_sex <- lm(salary~rank+discipline+yrs.since.phd+yrs.service, data=df)
summary(mod_no_sex)
```

VIF

```
car::vif(mod = mod_no_sex)
```

4.3 model without sex, years since Ph.D, years of service

```
mod_no_phd <- lm(salary~discipline+yrs.service+rank, data = df)
summary(mod_no_phd)
```

VIF

```
car::vif(mod_no_phd)`
```

4.4 model without rank and sex

```
mod_no_rank <- lm(salary~discipline+yrs.since.phd+yrs.service,data = df)
summary(mod_no_rank)
car::vif(mod_no_rank)
```

4.5 model without sex and years of service

```
mod_no_service <-lm(salary~discipline+yrs.since.phd+rank,data = df)
summary(mod_no_service)
car::vif(mod_no_service)
```

4.6 final model

```
mod_final <- lm(salary~discipline+rank,data=df)
summary(mod_final)
```

5. model diagnostic

```
attach(Salaries)
reg1<-lm(salary~rank+discipline)
summary(reg1)
plot(reg1)
salary1<-log(salary)
reg2<-lm(salary1~rank+discipline)
summary(reg2)
plot(reg2)
salary2<-sqrt(salary)
reg3<-lm(salary2~rank+discipline)
summary(reg3)
plot(reg3)
salary3<-1/salary
reg4<-lm(salary3~rank+discipline)
summary(reg4)
plot(reg4)
cooksd=cooks.distance(reg4)
```

```
n<-length(salary)
p<-4
m<-4/(n-p)
t<-which(cooksd>m)
t
{plot(cooksd)
abline(h =m, col="red")}
Salaries1<-Salaries[-t,]
salary4<-1/Salaries1$salary
reg5<-lm(salary4~Salaries1$rank+Salaries1$discipline)
summary(reg5)
plot(reg5)
```

7. estimated factor level means

```
ggplot(data=Salaries1)+
geom_boxplot(mapping=aes(x=rank,y=salary))+
facet_wrap(~discipline)
```