Salary Prediction for Different Professors

Sheng Yuan

University of Kentucky

February 24, 2020

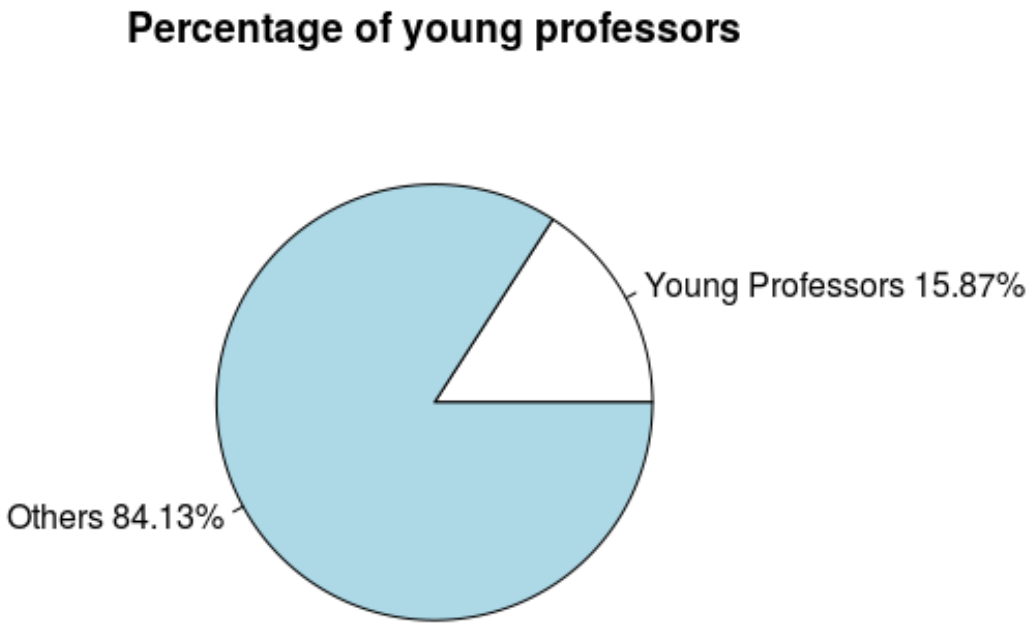
Overview

Background

- I had salaries data set from R, which had 397 cases and 6 variables, I wanted to see the the demographic information of professors and made prediction to professors' salary.
- Among the 6 variables, we had 3 categorical and 3 quantitative variables.
- I extended the prediction model to other research questions, and made some explanations on the extension.
- I applied the data analysis on R, with packages ggplot, sqldf, car, tidyverse, caret, leaps, MASS.

Young Professors

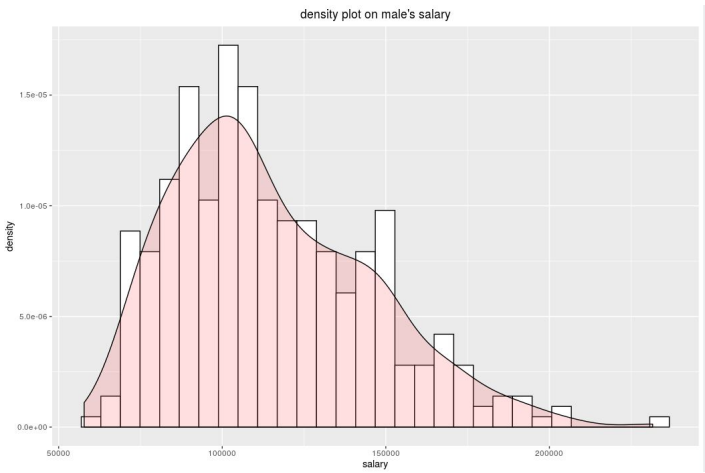
We considered assistant professor with less than 5 years of experience being young professors.

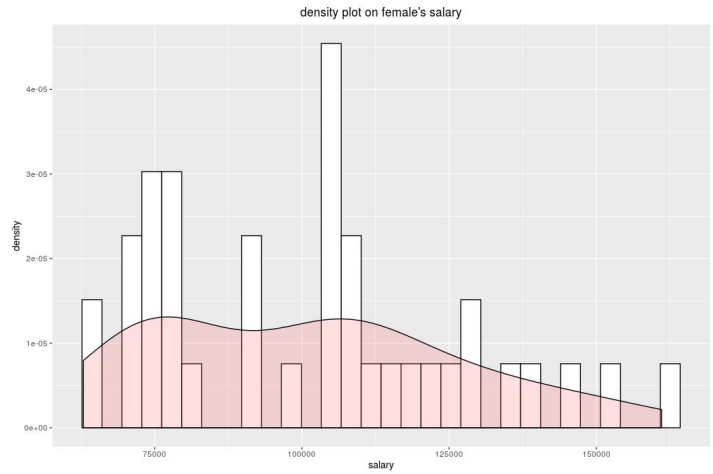


Hence 15.87% of the professors were assistant professor with less than 5 years of experience.

Salary Difference between gender

- If we want to see whether there exist salary difference we need to see the distribution of salaries based on different genders, create a side-by-side box-plot, then conducted a t-test.

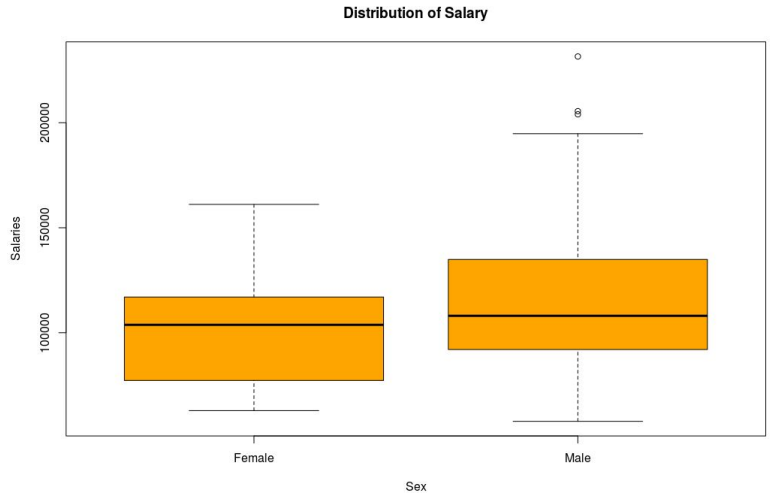




- The distribution plot for male professor's salary seemed to be normal, while the female professor's salary seemed not. However, since the sample size for female professor (39) larger than 30, we were still using t-test here to see the difference between mean salaries for different sexual group.

Salary Difference between gender

- Below was the side-by-side box-plot, there seemed to exist some difference between female and male salaries.



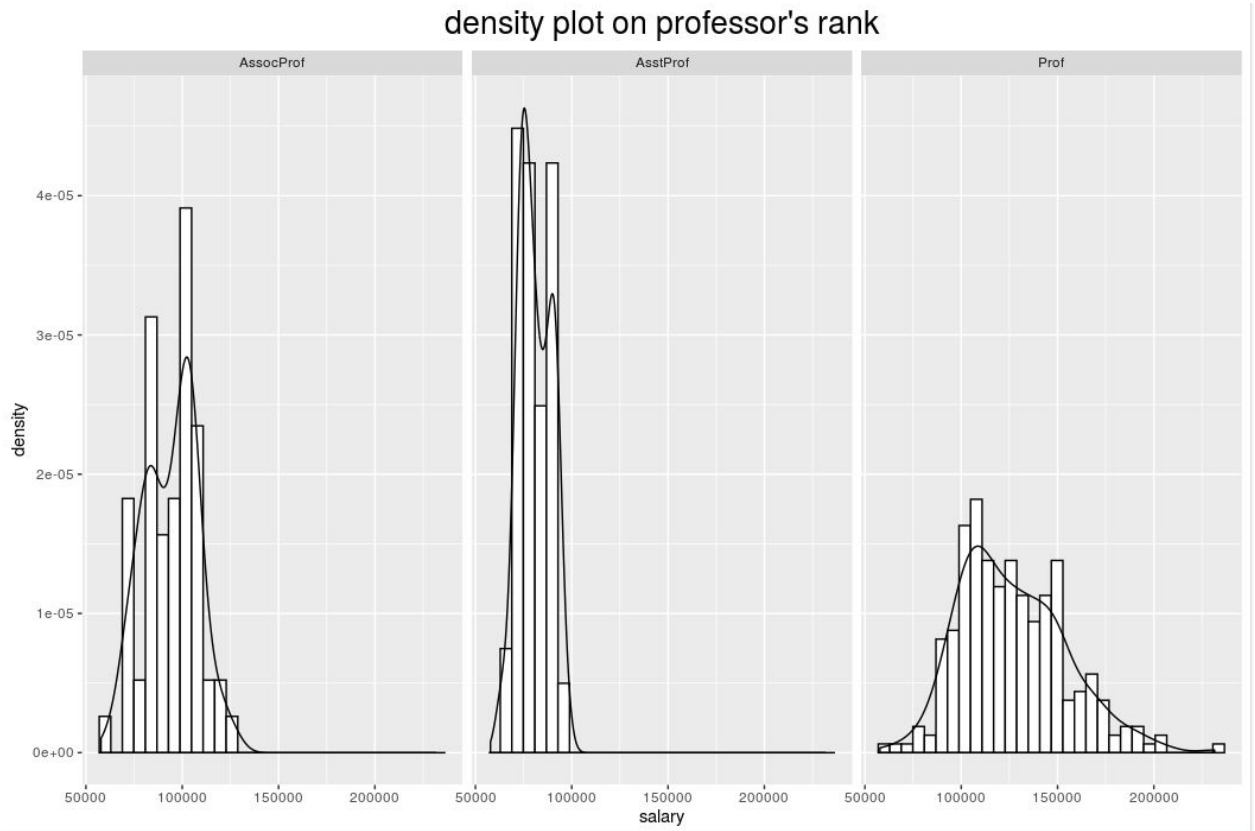
- From the test procedure below, we conducted a t-test for difference in means, got a test statistic of 3.1615 with 50.122 degrees of freedom, and reported a p-value of 0.002664, based on our common significance level $\alpha = 0.05$, p-value is smaller than α, hence we assume that there exist difference for salaries between sex.

```
        Welch Two Sample t-test

data:  male$salary and female$salary
t = 3.1615, df = 50.122, p-value = 0.002664
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5138.102 23037.916
sample estimates:
mean of x mean of y
 115090.4  101002.4
```

Distribution of Salary based on rank

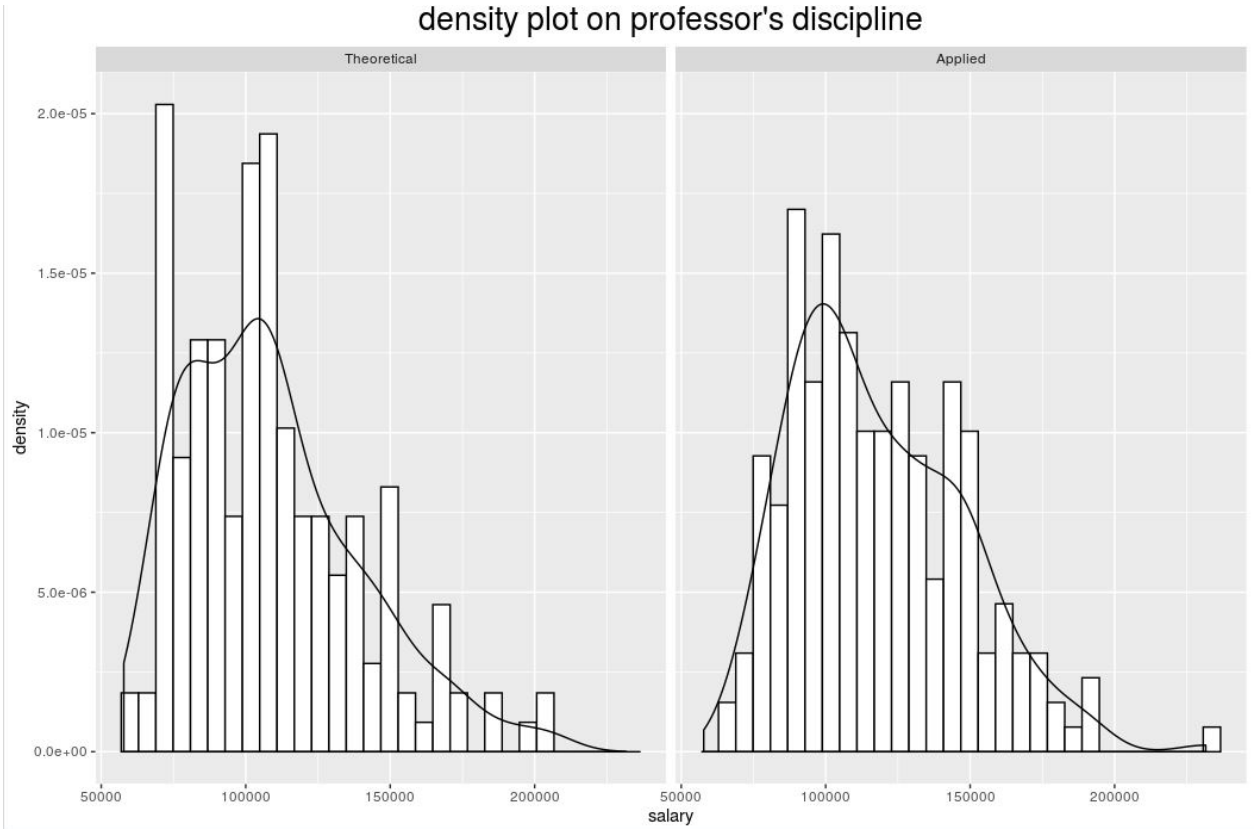
- We had the distribution of professor's salary based on rank.



- From the plot above, the salary distribution based on professor's rank seemed to be normal, with assistant professor earned the least, professor earned the most.

Distribution of Salary based on discipline

- We have the distribution of professor's salary based on discipline.



- From the plot above, the salary distribution based on professor's discipline seems to be normal, there seemed to be no difference between different discipline on professor's salary.

Dummy Coding

- For qualitative variables on regression, we need to do dummy coding on categorical variable. Here we used discipline as an example

$$X_D = \begin{cases} 0 & \text{if } \textit{discipline} \text{ is A(theoretical)} \\ 1 & \text{if } \textit{discipline} \text{ is B(applied)} \end{cases}$$

- From the binary coding above, we could use this dummy coding through our regression and classification purpose.

Data Preparation

- Because there exist categorical variables in the predictors, hence we needed to do dummy coding for them in order to perform linear regression.

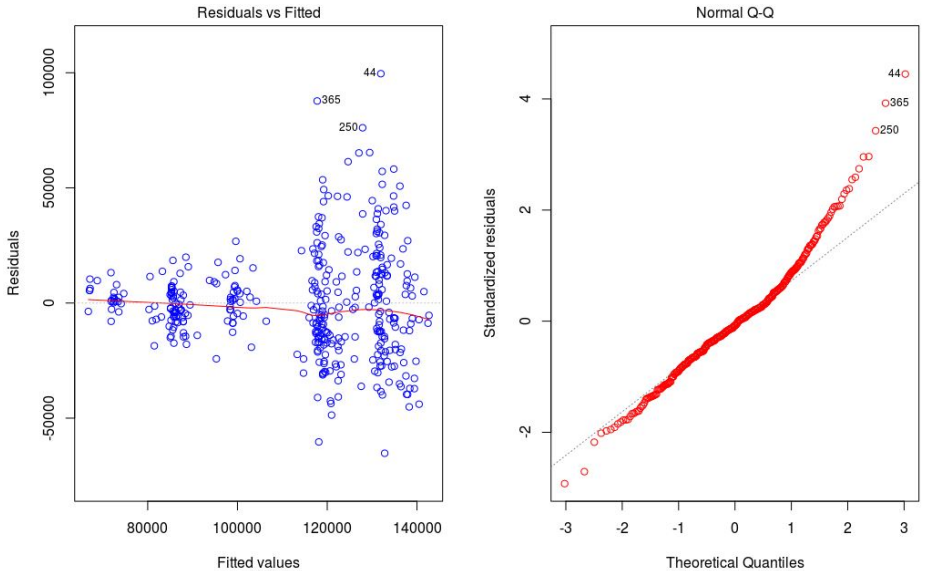
$$X_{R1} = \begin{cases} 0 & \textit{else} \\ 1 & \text{if } \textit{rank} \text{ is AssocProf} \end{cases}$$

$$X_{R2} = \begin{cases} 0 & \textit{else} \\ 1 & \text{if } \textit{rank} \text{ is Prof} \end{cases}$$

$$X_{S} = \begin{cases} 0 & \text{if } \textit{sex} \text{ is female} \\ 1 & \text{if } \textit{sex} \text{ is male} \end{cases}$$

residual analysis

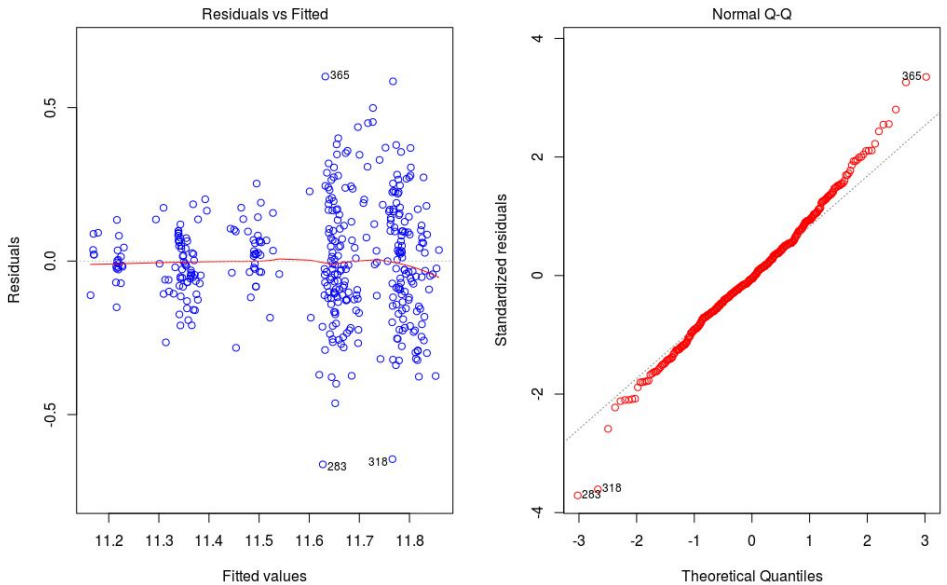
- First we wanted to see the residual analysis for plotting salaries over residuals, about whether we need transform on the response variable, salary.



- From the residual plot we can see that the consistent variance in linear regression seemed to be violated and the normal Q-Q plot there seems to be violation for normal assumption. After doing a Shapiro-Wilk normality test, we had a test-statistic 0.96857 with a p-value $1.0555 * 10^{-7}$, the normality assumption was also violated too.

residual analysis

- Then we log transformed the response variable salary.



- From the residual plot we could see that the consistent variance in linear regression seemed to be alleviated, and from the Shapiro-Wilk normality test, we reported test statistic W = 0.9915, with p-value = 0.02242, now the log transform of Y seemed a good fit for linear regression model.

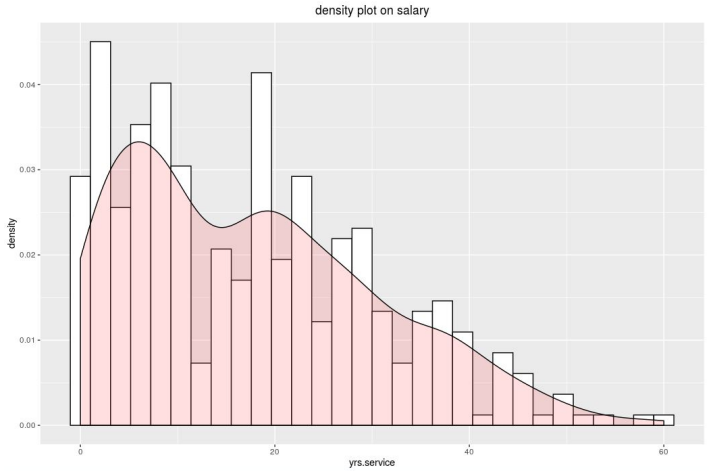
variable analysis

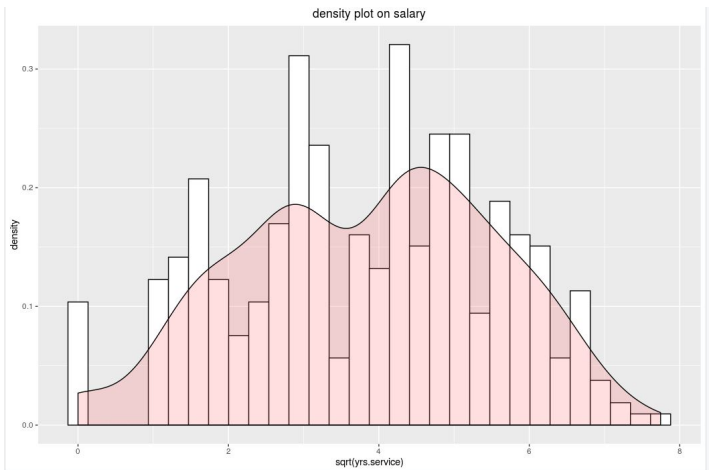
- Now we wanted to see whether the normality assumption is satisfied on the explanatory variables.
- First we saw the variable years since phd.

- We still used the normality test for the data, and reported a test statistic 0.96957 with p-value = 2.328e-07, if we do log transform, square root transform and 1/(x+0.1) transform on data, would report p-value 1.402e-14, 6.31e-06, and 2.2e-16, which did not make much difference. I will use years since phd as an explanatory variable.

variable analysis

- Then we saw the variable years of service.





- The plot on the right was square root of years of service while the plot on the left was years of service. We can see that the square root of years of service is more normal, and we conducted test on transformation, we have p-value of 0.00001379 compared with original variables p-value 2.337e-11, this meant square root of years service might be necessary.
- the other transform are log(x+0.1), 1/(x+0.1) and both reported a p-value smaller than 2.2e-16, hence we were using square root transformation on the variables.

log transform of salary

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

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 ***
rank           12907.6     4145.3   3.114  0.00198 **
rank2          45066.0     4237.5  10.635  < 2e-16 ***
discipline     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 *
sex             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
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-0.66236 -0.10813 -0.00914  0.09804  0.60107

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.164144   0.036794 303.425  < 2e-16 ***
rank           0.153787   0.033239   4.627 5.06e-06 ***
rank2          0.449463   0.033979  13.228  < 2e-16 ***
discipline     0.131869   0.018786   7.019 9.94e-12 ***
yrs.since.phd  0.003289   0.001932   1.702   0.0896 .
yrs.service   -0.003918   0.001699  -2.305   0.0217 *
sex            0.045583   0.030941   1.473   0.1415
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1807 on 390 degrees of freedom
Multiple R-squared:  0.5248,    Adjusted R-squared:  0.5175
F-statistic: 71.79 on 6 and 390 DF,  p-value: < 2.2e-16
```

- From the R output above, we could see that log transform of salary would lead to an increase of adjusted r-square about 0.07, which was a big improvement. Hence, I would adapt the log-transform of salary.

Cross-validation and Step-wise Regression

- Afterwards we needed to automatically select variables as well as higher order term of quantitative variable(second order term, sqrt(yrs.service)), interaction term(years.since.phd and yrs.service multipy rank, discipline and sex).
- We used 10-fold cross validation to present the result.

```
> step.model$results
   nvmax      RMSE  Rsquared       MAE     RMSESD RsquaredSD      MAESD
1      1 0.1962293 0.4380201 0.1594121 0.01693740 0.08277312 0.01303990
2      2 0.1861799 0.4969806 0.1456793 0.01827013 0.10165280 0.01377089
3      3 0.1809674 0.5222073 0.1382227 0.01851342 0.09308719 0.01497469
4      4 0.1814217 0.5201526 0.1391562 0.01776379 0.08711021 0.01448033
5      5 0.1814577 0.5201973 0.1394725 0.01912993 0.09524478 0.01547401
6      6 0.1813133 0.5214896 0.1400765 0.01728904 0.08690038 0.01448995
7      7 0.1787705 0.5343517 0.1372475 0.01510080 0.07674627 0.01245191
8      8 0.1786072 0.5367655 0.1377931 0.01538455 0.07140527 0.01311210
9      9 0.1796860 0.5311786 0.1381413 0.01506265 0.07051737 0.01237219
10    10 0.1801127 0.5293589 0.1386513 0.01512332 0.07029398 0.01289152
11    11 0.1798140 0.5306587 0.1383115 0.01503607 0.07021531 0.01296350
12    12 0.1796153 0.5314494 0.1378297 0.01495097 0.06973326 0.01268953
13    13 0.1795509 0.5329238 0.1380712 0.01504325 0.07080633 0.01259137
14    14 0.1793592 0.5337450 0.1377073 0.01495593 0.07055531 0.01237751
15    15 0.1792872 0.5341018 0.1375891 0.01496477 0.07049173 0.01237667
```

- Here nvmax represented how many variables in the model, RMSE represented root mean squared error, Rsquared represented the percentage of salaries explained by variables, MAE represented mean absolute error, and SD represented the standard deviation of RMSE, Rsquared and MAE respectively.

Cross-validation and Step-wise Regression

- We conducted automatic backward selection based on the minimum MSE, which is 8.
- After showing what variables in the pool, I found only rank1*years.of.service in the pool, where rank2*years.of.service is not. From the cv result plot, RMSE for 7 and 8 didn't change much. Hence, I ignored rank*years.of.service .

```
1 subsets of each size up to 8
Selection Algorithm: backward
         rank yrs.service rank2 yrs.since.phd discipline sqrt(yrs.service) sex rank:yrs.service yrs.service:rank2 rank:yrs.since.phd
1  ( 1 ) " "  " "          "*"   " "          " "        " "               " " " "              " "               " "
2  ( 1 ) " "  " "          "*"   " "          "*"        " "               " " " "              " "               " "
3  ( 1 ) "*"  " "          "*"   " "          "*"        " "               " " " "              " "               " "
4  ( 1 ) "*"  "*"          "*"   " "          "*"        " "               " " " "              " "               " "
5  ( 1 ) "*"  "*"          "*"   "*"          "*"        " "               " " " "              " "               " "
6  ( 1 ) "*"  "*"          "*"   "*"          "*"        " "               " " " "              " "               " "
7  ( 1 ) "*"  "*"          "*"   "*"          "*"        " "               " " " "              " "               " "
8  ( 1 ) "*"  "*"          "*"   "*"          "*"        " "               " " "*"              " "               " "
         yrs.service:discipline yrs.since.phd:discipline yrs.since.phd:sex yrs.service:sex
1  ( 1 ) " "                    " "                      " "               " "
2  ( 1 ) " "                    " "                      " "               " "
3  ( 1 ) " "                    " "                      " "               " "
4  ( 1 ) " "                    " "                      " "               " "
5  ( 1 ) " "                    " "                      " "               " "
6  ( 1 ) "*"                    " "                      " "               " "
7  ( 1 ) "*"                    "*"                      " "               " "
8  ( 1 ) "*"                    "*"                      " "               " "
```

Multicolinearity

- We then saw the fit of the data here. The fit is not bad.

```
Call:
lm(formula = log(salary) ~ rank + rank2 + discipline + yrs.service +
    yrs.since.phd + yrs.since.phd * discipline + yrs.service *
    discipline, data = Salaries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68119 -0.11098  0.00222  0.09882  0.64475

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              11.167567   0.033595 332.420  < 2e-16 ***
rank                      0.149081   0.032730   4.555 7.03e-06 ***
rank2                     0.445209   0.033617  13.243  < 2e-16 ***
discipline                0.184993   0.038683   4.782 2.46e-06 ***
yrs.service              -0.011240   0.002430  -4.627 5.07e-06 ***
yrs.since.phd             0.010675   0.002661   4.012 7.23e-05 ***
discipline:yrs.since.phd -0.013287   0.003397  -3.911 0.000108 ***
discipline:yrs.service    0.014004   0.003330   4.205 3.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1775 on 389 degrees of freedom
Multiple R-squared:  0.543,     Adjusted R-squared:  0.5348
F-statistic: 66.03 on 7 and 389 DF,  p-value: < 2.2e-16
```

- After checking the VIF in order to get information with respect to multicolinearity, we found multicolinearity be a big issue here because VIF value>5 or >10 is a sign of collinearity that could considerably affect results.

```
          rank          rank2     discipline    yrs.service   yrs.since.phd discipline:yrs.since.phd   discipline:yrs.service
      1.826137       3.149811       4.679147      12.555757      14.786490              25.678347              19.593919
```

Variable Selection

- From experience, we thought that yrs.service and yrs.since.phd might have multicolinearity issue, and I tried to delete them separately.

```
Call:
lm(formula = log(salary) ~ rank + rank2 + discipline + yrs.service +
    yrs.service * discipline, data = Salaries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67220 -0.10841 -0.01354  0.10139  0.62013

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             11.239111   0.029765 377.595  < 2e-16 ***
rank                     0.159744   0.032858   4.862 1.69e-06 ***
rank2                    0.470893   0.030887  15.246  < 2e-16 ***
discipline               0.089474   0.031437   2.846  0.00466 **
yrs.service             -0.002254   0.001086  -2.076  0.03855 *
discipline:yrs.service   0.002109   0.001432   1.473  0.14153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1812 on 391 degrees of freedom
Multiple R-squared:  0.5213,    Adjusted R-squared:  0.5152
F-statistic: 85.16 on 5 and 391 DF,  p-value: < 2.2e-16

> car::vif(mod=fitfinal2)
              rank              rank2         discipline       yrs.service discipline:yrs.service
          1.765950           2.551445           2.965427           2.406965           3.475931
```

```
Call:
lm(formula = log(salary) ~ rank + rank2 + discipline + yrs.since.phd +
    yrs.since.phd * discipline, data = Salaries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68922 -0.11147 -0.00497  0.09369  0.57656

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              11.2116806  0.0330693 339.036  < 2e-16 ***
rank                      0.1548790  0.0335713   4.613 5.38e-06 ***
rank2                     0.4593576  0.0343209  13.384  < 2e-16 ***
discipline                0.1321962  0.0379984   3.479  0.00056 ***
yrs.since.phd            -0.0003348  0.0012090  -0.277  0.78202
discipline:yrs.since.phd -0.0001462  0.0014672  -0.100  0.92069
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1821 on 391 degrees of freedom
Multiple R-squared:  0.5161,    Adjusted R-squared:  0.5099
F-statistic:  83.4 on 5 and 391 DF,  p-value: < 2.2e-16

> car::vif(mod=fitfinal3)
              rank              rank2         discipline       yrs.since.phd discipline:yrs.since.phd
          1.823635           3.116361           4.285871           2.897658           4.545745
```

- Top left is model with years of service and top right is model with years since phd, we could see that keep years of service will be more appropriate (smaller p-value and higher R-squared). Also, the multicolinearity issue has been settled.

Variable Selection

- Finally, we talked about square root transform of years of service previously, here we change the years of service to sqrt(years of service).

```
Call:
lm(formula = log(salary) ~ rank + rank2 + discipline + sqrt(yrs.service) +
    sqrt(yrs.service) * discipline, data = Salaries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.66825 -0.10739 -0.00980  0.09937  0.60319

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 11.251037   0.037051 303.664  < 2e-16 ***
rank                         0.163725   0.034960   4.683  3.9e-06 ***
rank2                        0.473997   0.035274  13.437  < 2e-16 ***
discipline                   0.082682   0.046326   1.785   0.0751 .
sqrt(yrs.service)           -0.014491   0.009164  -1.581   0.1146
discipline:sqrt(yrs.service)  0.011702   0.010974   1.066   0.2869
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1816 on 391 degrees of freedom
Multiple R-squared:  0.519,     Adjusted R-squared:  0.5129
F-statistic: 84.39 on 5 and 391 DF,  p-value: < 2.2e-16

> car::vif(mod=fitfinal4)
                     rank                     rank2                discipline          sqrt(yrs.service) discipline:sqrt(yrs.service)
                 1.989780                  3.312084                  6.409449                  2.913184                  6.784758
```

- Although the transformation of years of service would fit the normality assumption, but the model fitting was not as good as original years of service do, hence, we chose the model with years of services and discipline*(years of service) in the pool.

Result Interpretation

```
Call:
lm(formula = log(salary) ~ rank + rank2 + discipline + yrs.service +
    yrs.service * discipline, data = Salaries)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67220 -0.10841 -0.01354  0.10139  0.62013

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              11.239111   0.029765 377.595  < 2e-16 ***
rank                      0.159744   0.032858   4.862 1.69e-06 ***
rank2                     0.470893   0.030887  15.246  < 2e-16 ***
discipline                0.089474   0.031437   2.846  0.00466 **
yrs.service              -0.002254   0.001086  -2.076  0.03855 *
discipline:yrs.service    0.002109   0.001432   1.473  0.14153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1812 on 391 degrees of freedom
Multiple R-squared:  0.5213,    Adjusted R-squared:  0.5152
F-statistic: 85.16 on 5 and 391 DF,  p-value: < 2.2e-16

> car::vif(mod=fitfinal2)
                     rank                     rank2                discipline           yrs.service discipline:yrs.service
                 1.765950                  2.551445                  2.965427                  2.406965                  3.475931
```

- First, since we had interaction term, we could write the model as:

$$\hat{log(Salary)} = \begin{cases} 11.2391 + 0.1597 * assocprof + 0.4709 * prof \\ -0.002254 * (yrs.service) \\ \textbf{if discipline is theoretical} \\ (11.2391 + 0.0895) + 0.1597 * assocprof \\ +0.4709 * prof \\ +(0.002109 - 0.002254) * (yrs.service) \\ \textbf{if discipline is applied} \end{cases}$$

Result Interpretation

- We interpreted the result separately:
- For theoretical discipline, if an assistant professor with 0 years of service, he would be expected to earn exp(11.2391)=76046.48 dollars 9 months.
- For theoretical discipline, an associate professor with same years of service compared to others, he would be expected to earn (exp(0.1597)-1)*100% =17.35% more for 9 months.
- For theoretical discipline, a professor with same years of service compared to others, he would be expected to earn (exp(0.4709)-1)*100% = 60.14% more for 9 months.
- For theoretical discipline, Any professor with one year more of service compared to others, he would be expected to earn (1-exp(-0.002254))*100%=2.23% less for 9 months.

Result Interpretation

- We then interpreted applied discipline:
- For applied discipline, if an assistant professor with 0 years of service, he would be expected to earn exp(11.3286)=83166.51 dollars 9 months.
- For applied discipline, an associate professor with same years of service compared to others, he would be expected to earn (exp(0.1597)-1)*100% =17.35% more for 9 months.
- For applied discipline, a professor with same years of service compared to others, he would be expected to earn (exp(0.4709)-1)*100% = 60.14% more for 9 months.
- For theoretical discipline, Any professor with one year more of service compared to others, he would be expected to earn (1-exp(-0.000145))*100%=0.015% less for 9 months.

Result Interpretation

- 52.13% of the variability in salaries for professors was explained by rank, discipline and years of service.
- We wanted to see if we do a dichotomous coding for salary, how well could we predict the group for professors, what would be the difference between this two analysis method.
- We created binary salary indicator by using indicator functions, where indicator had a value of 0 if the salary is below the median and 1 otherwise.

Binary Classification

- Since we had interaction term in the pool, the best way to do classification is logistic regression, from which we could do parameter interpretation and classification.

```
Call:
glm(formula = salary ~ rank + rank2 + discipline + yrs.service +
    yrs.service * discipline, family = "binomial", data = resalaries)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9920  -0.5594   0.5439   0.6549   2.3000

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)            -19.169629 785.865120  -0.024    0.981
rank                    16.619300 785.865098   0.021    0.983
rank2                   20.126877 785.865067   0.026    0.980
discipline               0.902076   0.592675   1.522    0.128
yrs.service             -0.021096   0.014517  -1.453    0.146
discipline:yrs.service   0.009512   0.023544   0.404    0.686

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.36  on 396  degrees of freedom
Residual deviance: 337.71  on 391  degrees of freedom
AIC: 349.71

Number of Fisher Scoring iterations: 17
```

- From the result above, we could see the result is mostly not significant, however, if we use the point estimate we can interpret coefficient as follow(intercept will be ignored).

Result Interpretation

- Rank has such a large p-value, so we ignored the interpretation of rank.
- For applied discipline group, the professor would be exp(0.9021-(0.021+0.009512)*yrs.service) times likely to be in the high salary group.
- If a professor is in theoretical discipline group, an years increase of years of service would lead to (1-exp(-0.021))*100%=2.09% less likely to be in the high salary group.
- If a professor is in applied discipline group, an years increase of years of service would lead to (1-exp(0.009512-0.021))*100%=1.15% less likely to be in the high salary group.

Test Error

- Finally, we wanted to see the test error based on previous 10-fold cross validation dataset.

```
> train( salary ~ rank+rank2+discipline+yrs.service+yrs.service*discipline, data = resalaries,
+       method = "glm",
+       family="binomial",
+       trControl = train.control
+ )
Generalized Linear Model

397 samples
  4 predictor
  2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 359, 357, 357, 357, 357, 357, ...
Resampling results:

  Accuracy   Kappa
  0.7931107  0.5856595
```

- From the result above, we could see the logistic regression gave an average of accuracy of 0.7931, which means 0.2069 test error. It was better than 50% random test result, meant our model is having some effect on classification. Kappa is a statistic measure the fit of the classification, the closer to 1, the better the classification.

Extended Research Questions

- Whether there exist association between tenure (assistant professor become associate professor or higher) and professor's sex, that is, whether there exist kind of discrimination between sex in academy area.
- The reason I proposed this research topic was that the discrimination between sex are very common, we want to see whether this exist in the professors, not only in salaries but also in rank.

Extended Research Questions

- Was the proportion of professors who has salary larger than 110000 for 9 months larger than 0.5?
- Since professor is the highest rank for professor, I wanted to see the proportion of them who could earn 110000 for 9 month, is more than half of them doing this well?

Extended Research Questions

- Was there exist any difference between average amount of year service that assistant professor become associate professor and associate professor become professor?
- Usually there exist requirement for assistant professor to become associate, but there exist no requirement for associate to become professor, therefore, some associate professors might not want to be professors, wanted to see the pattern on the amount of year service.

Additional Attributes

- Add different **working zones** (like New York Zones, Bay Area, etc) where the professor worked. The degree of prosperity of different area would be an important impact factor on the salaries.
- Add **race** for each professor as an attribute because diversity between races might reflect on difference of salaries.
- Add **years of service before status** for advanced professors. You could not have a 0 year of service associate professors, so the years about when they earned the professor's title will be an important factor for their salary.
- A status of **Offers from other institutions** (Yes/No). The professor who got offers from other university might receive a salary raise in the department now.

Additional Attributes

- Add **total impact factor** for papers professor published. This is a measurement of the quality and quantity for professor's papers. The more and/or better paper get published, the higher professor more likely earned.
- Add **Previously University Experience** as an attribute, for the reason that if a professor had previous teaching experience, he or she might get paid more.
- **Rank of College** of the professor's graduate school. Especially for recent graduate professors, his or her graduate school rank would have big impact on the salary.

Sample Size Estimation

- For the question that whether there exist difference on the proportion of tenured professor based on different sex, we have the hypothesis below:

$$H_0 : p_1 = p_2 \quad v.s. \quad H_a : p_1 \neq p_2$$

- From above, p_1 is the proportion of male assistant professor, and p_2 is proportion of female assistant professor.
- We wanted to conduct a test for difference in proportion, where the power should be above 80% and the confidence level should within 5%.

$$n_i = 2 * (\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES})^2$$

where ES should be

$$\frac{|p_1 - p_2|}{\sqrt{p(1-p)}}$$

Sample Size Estimation

- Now we had $Z_{1-\alpha/2} = 1.96$, $Z_{1-\beta} = 0.84$, p_1=0.1564246, p_2=0.2820513, p=0.1687, the sample size should be 139.33≈140 for each group.
- The data on female group is very unbalanced, only 39 female in the data set while 358 cases for males. Hence, for female we can do bootstrapping from sample to enlarge our data set and male we can do random sampling, so the sample size of female and male group should be balanced.

Sample Size Estimation

- For the question that whether there proportion of professor earned more than 110000 larger than 0.5 we have the hypothesis below:

$$H_0 : p_1 \leq 0.5 \quad v.s. \quad H_a : p_1 > 0.5$$

- From above, p_1 is the proportion of professor who earned more than 110000 for nine months.
- We want to conduct a test for single proportion, where the power should be above 80% and the confidence level should within 5%.

$$n_i = 2 * (\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES})^2$$

where ES should be

$$\frac{|p_1 - p_0|}{\sqrt{p_1(1-p_1)}}$$

Sample Size Estimation

- Now we have $Z_{1-\alpha/2} = 1.96$, $Z_{1-\beta} = 0.84$, p_1=0.6654, p_0=0.5, the sample size should be 127.6≈128 professors.
- The sample size for for the rank being professor is 266, which is enough for our hypothesis test.

Sample Size Estimation

- For the question that whether there exist difference on the mean time professor goes to higher rank:

$$H_0 : \mu_1 = \mu_2 \quad v.s. \quad H_a : \mu_1 \neq \mu_2$$

- From above, μ_1 is the mean time assistant professor become associate professor, here I will use mean(years of service of associate) - mean(years of service of assistant); μ_2 is the mean time associate professor become professor, here I will use mean(years of service of professor) - mean(years of service of associate).
- We want to conduct a test for difference in proportion, where the power should be above 80% and the confidence level should within 5%.

$$n_i = 2 * (\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{ES})^2$$

where ES should be

$$\frac{|\mu_1 - \mu_2|}{\sigma}$$

Sample Size Estimation

- Now we have $Z_{1-\alpha/2} = 1.96$, $Z_{1-\beta} = 0.84$, μ_1=9.57991, μ_2=10.86266, σ=13.00602 , the sample size should be 1611.9 $\approx$ 1612 for each group.
- Since the data size not enough, we need assistant professor, associate professor and professor, each of them should have the size 1612 in order to get average years of service between rank groups.

Conclusion

- Linear regression to predict salaries for professor was doing better when we transformed salary into log of salaries.
- The same predictor for linear regression was not doing so well when we did logistic regression for classification.
- Rank was doing a positive effect on salaries, sex seemed like doing nothing on salaries, while discipline would be complicated to estimate the effect on salary. Basically, the applied discipline was earning more than theoretical discipline.
- The information from the data was not enough to conduct a very precisely prediction, hence we need more information on the data.

Reference

[1] Michel H Kutner, Christopher J. Nachtsheim., John Neter, William Li [*Applied Linear Statistical Models*]. 5th Edition, 2005.

[2] Annette J. Dobson, Adrian G. Barnett [*An Introduction to Generalized Linear Models*]. 3rd Edition, 2008.

[3] Power and Sample Size Determination
`http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Power/BS704_Power_print.html`

[4] Best Subsets Regression Essentials in R
`http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/`