Bhumika H Yogesh

Contents

Analysis on the salary difference between male and female faculty members

The salaries dataset and variables description

The datasets that we'll use is the 'Salaries' dataset within the 'carData' package. The dataset consists of nine-month salaries collected from 397 collegiate professors in the U.S. during 2008 to 2009. In addition to salaries, the professor's rank, sex, discipline, years since Ph.D., and years of service was also collected. Thus, there is a total of 6 variables, which are described below.

rank: Factor variable composed by te following (AssocProf,AsstProf,Prof).

discipline: Factor variable with levels A ("theoretical" departments) or B ("applied" departments).

yrs.since.phd: Number of years since the professor has obtained their PhD.

yrs.service: Number of years the professor has served university.

sex: Factor variable with levels (Female, Male)

salary: Nine-month salary, in dollars.

Figure 1: Credit: ABA Journal

Load the dataset

Descriptive summary

Our project's goal is to identify the analysis on the salary difference between male and female faculty members. In order to do this, we are going to analyze which of the variables in the dataset are important and should be considered through regression analysis.

At first we will spell out the rank variables and rename discipline variables to its meaningful name ensuring both rank and discipline are factors

```
##                  rank discipline yrs.since.phd yrs.service  sex salary
## 1           Professor    Applied            19          18 Male 139750
## 2           Professor    Applied            20          16 Male 173200
## 3 Assistant Professor    Applied             4           3 Male  79750
## 4           Professor    Applied            45          39 Male 115000
## 5           Professor    Applied            40          41 Male 141500
## 6 Associate Professor    Applied             6           6 Male  97000
```

```
##                    rank           discipline  yrs.since.phd    yrs.service
##  Assistant Professor: 67   Applied    :216   Min.   : 1.00   Min.   : 0.00
##  Associate Professor: 64   Theoretical:181   1st Qu.:12.00   1st Qu.: 7.00
##  Professor          :266                     Median :21.00   Median :16.00
##                                              Mean   :22.31   Mean   :17.61
##                                              3rd Qu.:32.00   3rd Qu.:27.00
##                                              Max.   :56.00   Max.   :60.00
##      sex          salary
```

```
##  Female: 39   Min.   : 57800
##  Male  :358   1st Qu.: 91000
##               Median :107300
##               Mean   :113706
##               3rd Qu.:134185
##               Max.   :231545
```

The salaries data has three categorical variables including sex and three numerical variables including salary which is a response variable. The professor's salary in this sample range from 57800 to 231545 dollars. The mean of the salaries is 113706 dollars, which means the average amount a professor earn in nine months is 113706 dollars.
We can see that the sex data is highly unbalanced, that means the size of the male faculty (358) is a lot more than the female faculty (only 39).

Contingency table

```
##                       discipline
## rank                   Applied Theoretical
##   Assistant Professor      43          24
##   Associate Professor      38          26
##   Professor               135         131
```

We have three different ranks - Assistant Prof, Associate Prof and the Professor. Similarly, we have two different discipline - Applied and Theoretical.
From the contingency table, we can say that there are 62.5% professor's in the applied department and 72% of Professor's in the theoretical department which is slightly higher. We can also see that there are bit more faculties in the Applied discipline than the Theoretical discipline.

Normalized form

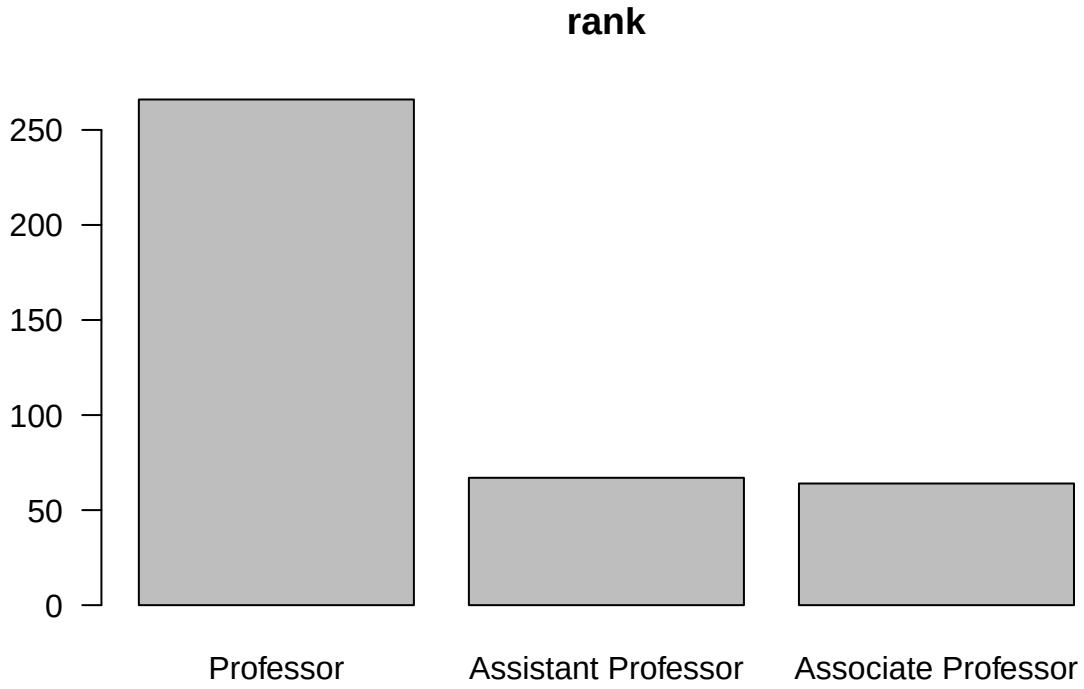
```
##                       discipline
## rank                     Applied Theoretical
##   Assistant Professor 0.1990741   0.1325967
##   Associate Professor 0.2099448   0.1203704
##   Professor           0.6250000   0.7237569
```

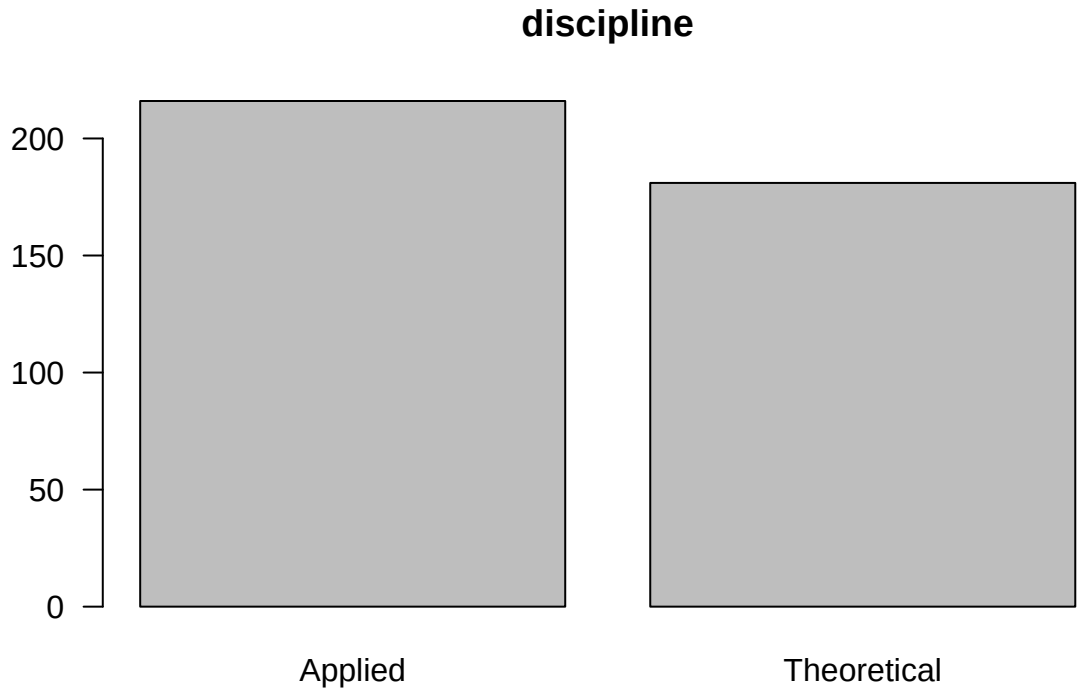
Numerical and graphical summary

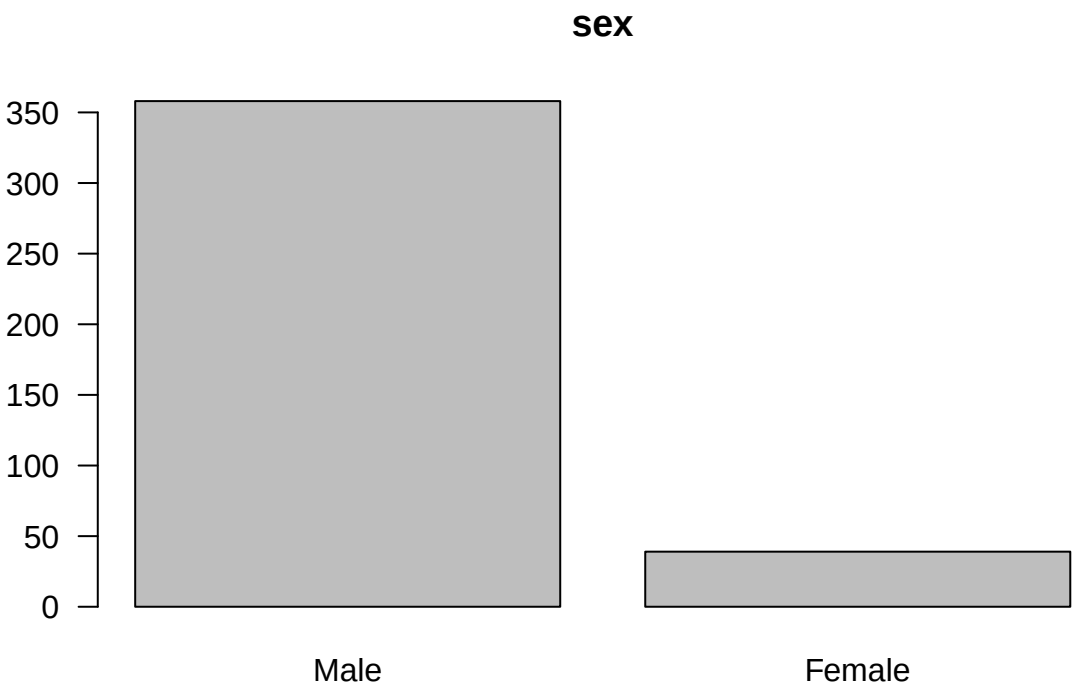
Here, we will first extract the categorical and numerical variables. 'rank', 'discipline', 'sex' are the categorical variables and 'years since phd', 'years service', 'salary' are the numerical variables.
Bar chart
(1) Professor rank is quite higher than the other two ranks
(2) More faculty in the Applied department than the Theoretical department. In terms of proportions we are reasonably close 54% vs 46%.
(3) Size of the male faculty is higher than the female faculty

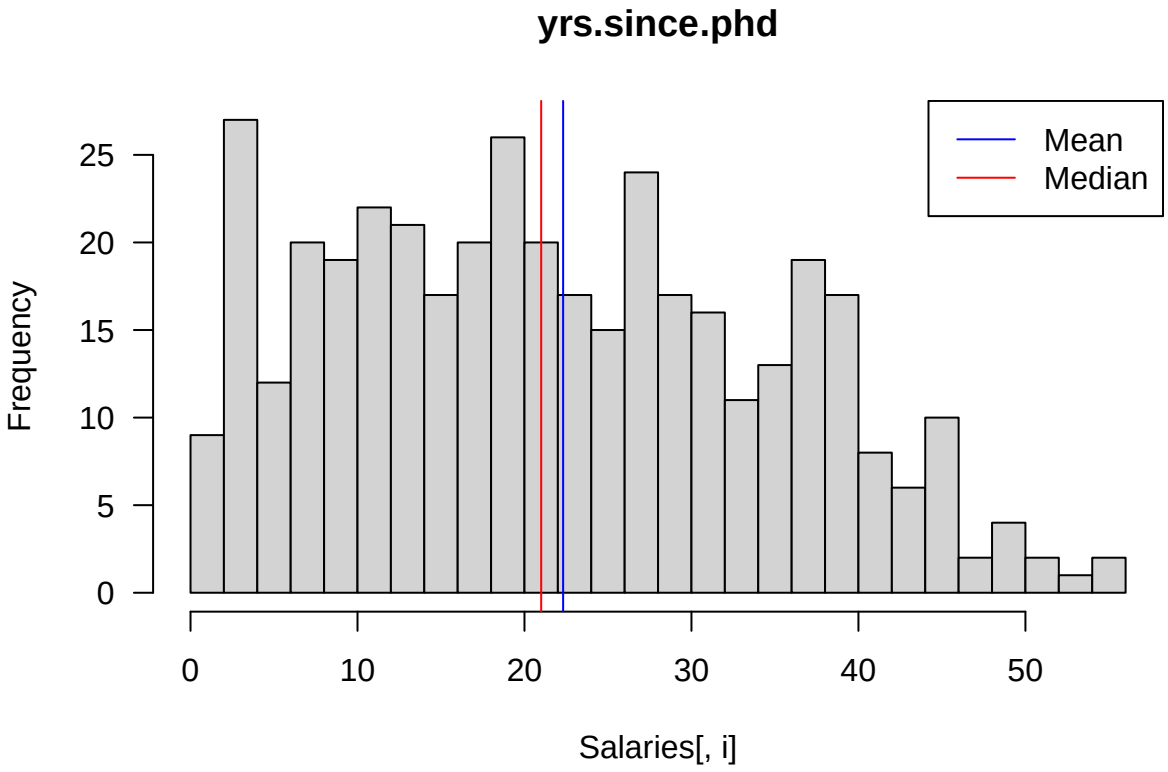
Histogram
(1)The years since PhD is skewed right. The mean and median values are reasonably close
(2) The years of service is little bit skewed to the right right. More people could have joined recently and you can see some faculty have stayed more than 60 years. Here we can see that mean value is slightly larger than the median value
(3)Salary distribution is skewed right but a bit closer to symmetric. We can see that the mean is slightly higher than the median.

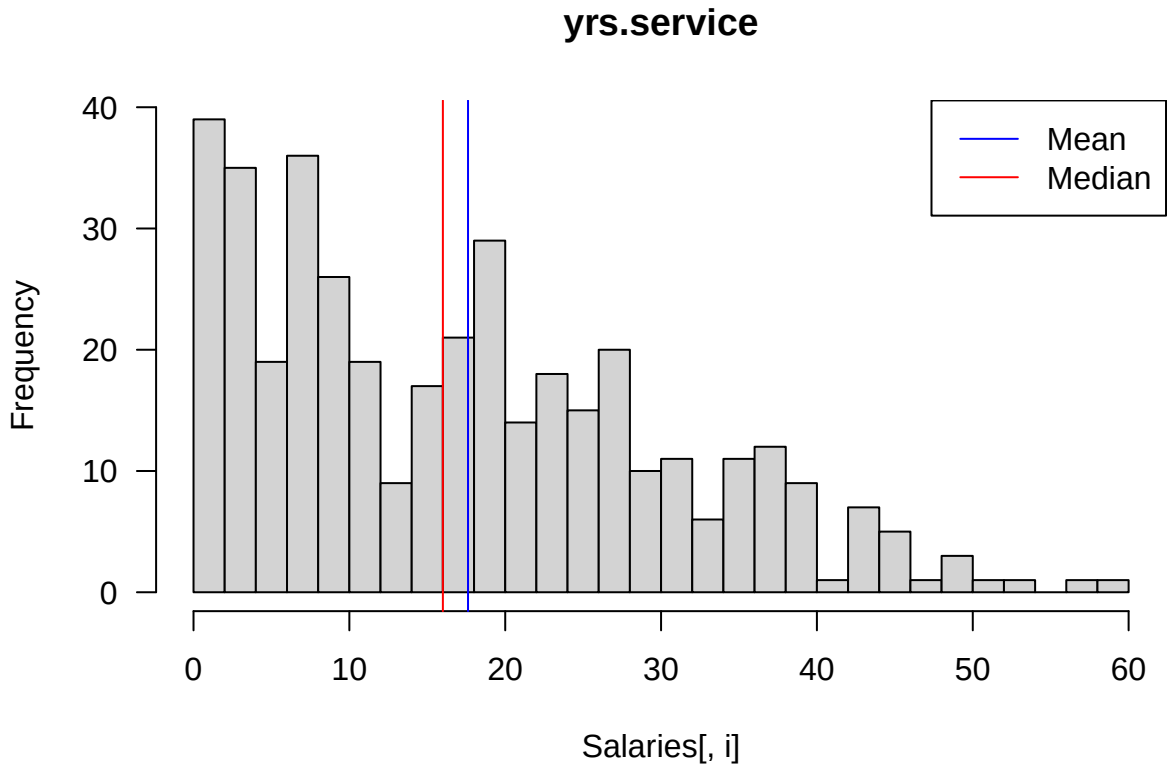
There are more professors (~2/3) than associate and assistant professors combined (~1/3). The disciplines are relatively close to equal.There are way more male than female professors.

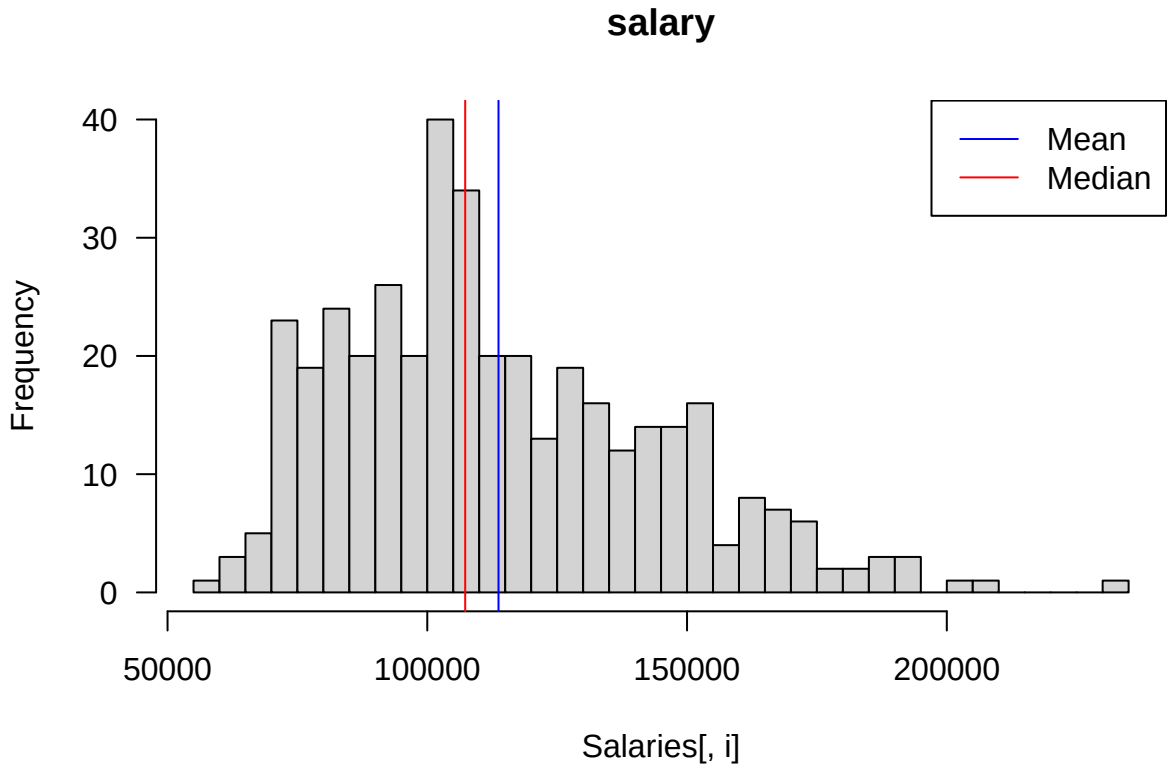




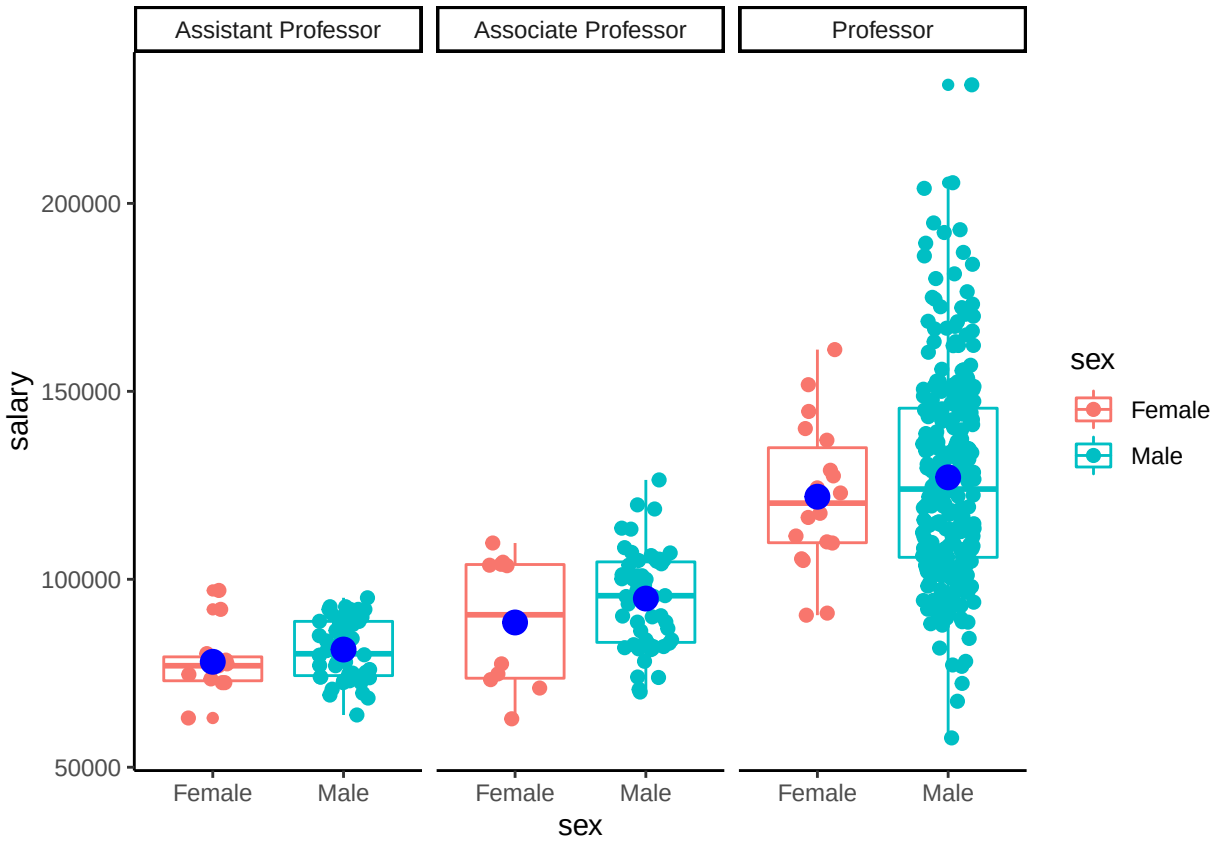

sex
350
300
250
200
150
100
50
0
Male
Female


yrs.since.phd
Mean
Median
Frequency
25
20
15
10
5
0
0
10
20
30
40
50
Salaries[, i]



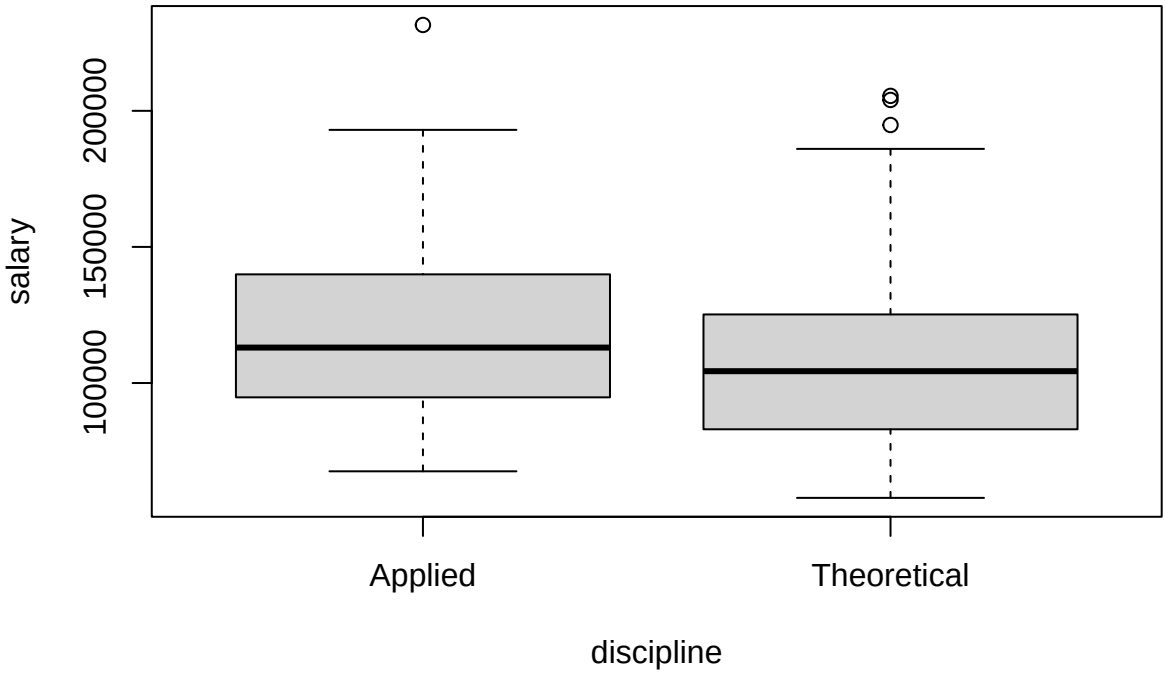


Boxplot From the boxplot, we can observe that as the rank increases the salary also increases. The mean salary (blue dot) for Male is comparatively higher as compared to female. It also suggests that associate professors earn lower salaries compare to assistant professors, and professors.



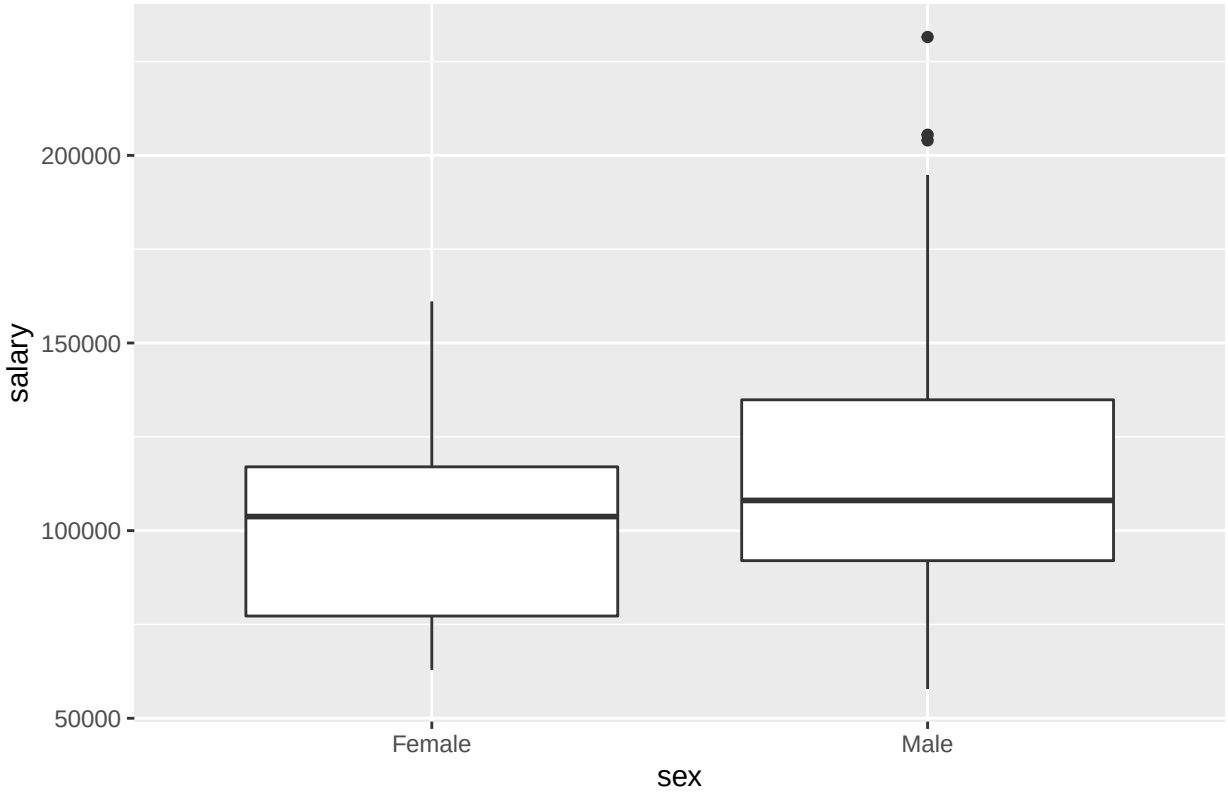
Box plot - Salary vs Discipline

The box plot for the categorical variables "discipline" suggests that, salaries differences by discipline where applied departments professors seem to receive significant more salaries either when they have lower or higher salaries.



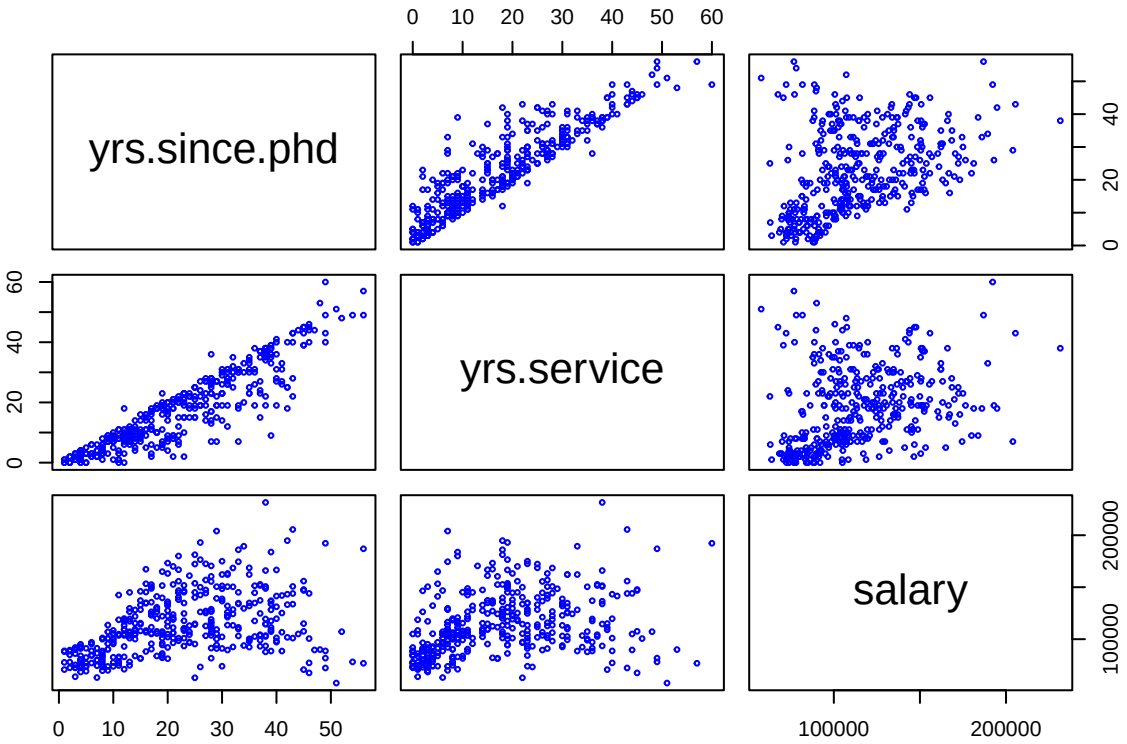
According to the below boxplot (Salar Vs Sex), there are three outliers in the male salary.

Figure 1: Female and Male Professors' Salaries



Pairwise Scatterplot

- From the pairwise matrix scatter plot, we can see that there is, as expected, a strong positive linear relationship with 'yrs.service' and 'yrs.since.phd'. This suggest that multicolinearity will probably be an issue with these two columns as they are numerical variables. So, we will need to further investigate that in the future models.
- There is a moderate positive linear relationship between 'salary' and 'yrs.since.phd'.
- And a even weaker positive linear relationship between 'salary' and 'yrs.service'.



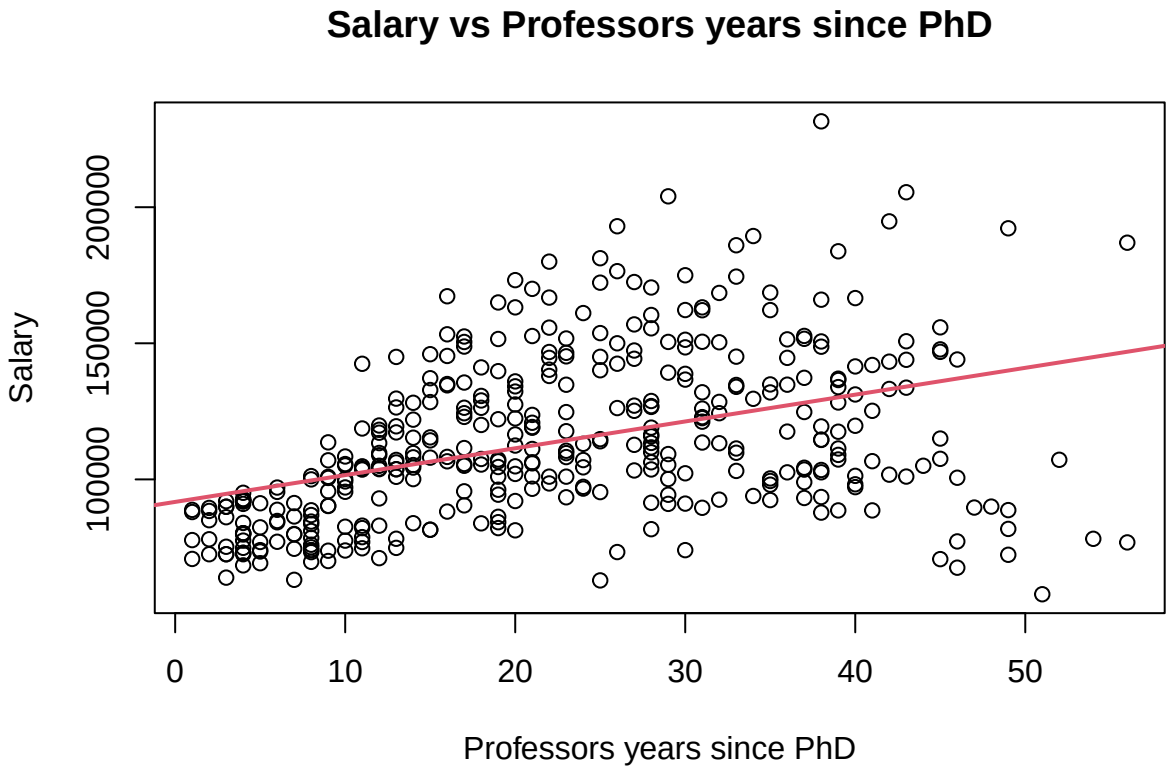

```
##                yrs.since.phd yrs.service    salary
## yrs.since.phd      1.0000000   0.9096491 0.4192311
## yrs.service        0.9096491   1.0000000 0.3347447
## salary             0.4192311   0.3347447 1.0000000
```

Regression

Simple Linear Regression

The first fitted model we decided to fit was salary Vs. years since PhD.

```
##
## Call:
## lm(formula = salary ~ yrs.since.phd, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -84171 -19432  -2858  16086 102383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd    985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27530 on 395 degrees of freedom
## Multiple R-squared:  0.1758, Adjusted R-squared:  0.1737
## F-statistic: 84.23 on 1 and 395 DF,  p-value: < 2.2e-16
```



The linear regression model for Salary vs Years since Phd is as follows:
Salary=91718.7+985.3∗yrs.since.phd

We can see that the summary statistics show that 17.58% of the variability in salaries can be explained by the fitted linear regression model and the model overall seems to be valid. The plot suggests that there is a positive relationship between these variables. Although, this relationship is not considered to be strong we are going to keep exploring this model to see if it is a good fit for salaries analysis.

However, salary cannot be explained by years since Phd alone - we can extend the model by including additional explanatory variables - we will estimate the multiple regression model

Full Model (Multiple Linear Regression)

Let's fit a multiple linear regression model by supplying all independent variables except the dependent variable (salary).

Here we can observe that a person gets an average salary of 65955.2 dollars. The associate professor level is set to the reference level. You can interpret that as ranking increases i.e., from assistant to associate to the professor, the average salary also increases. let's interpret a continuous variable to say "years of service". As years of service increases by 1 year, the average salary drops by 489.5 dollars holding all other variables constant. It also shows that 45.47% of the variability can be explained by the fitted linear regression model.

Similarly, here the discipline Theoretical dept is the reference category. The Applied discipline is significantly associated with an average increase of 14417.6 dollars in salary compared to theoretical departments holding other variables at constant.

In this section we decided to further our analysis examining if salary is affected for more than one variable. For this part we start considering all variables. The full model is as follows:

```
##
## 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)               80372.9     4372.3  18.382  < 2e-16 ***
## rankAssociate Professor   12907.6     4145.3   3.114  0.00198 **
## rankProfessor             45066.0     4237.5  10.635  < 2e-16 ***
## disciplineTheoretical    -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
```

Stepwise Regression

As per our initial assumption of multicolinearity (correlation among independent variables) and filter out essential variables/features from a large set of variables, a stepwise regression is usually performed. The process starts with initially fitting all the variables and after that, with each iteration, it starts eliminating variables one by one if the variable does not improve the model fit. The AIC metric is used for checking model fit improvement.

```
## Start:  AIC=7965.19
## salary ~ rank + discipline + yrs.since.phd + yrs.service + sex
##
##                 Df  Sum of Sq        RSS    AIC
## - sex            1 7.8068e+08 1.9890e+11 7964.8
## <none>                        1.9812e+11 7965.2
## - yrs.since.phd  1 2.5041e+09 2.0062e+11 7968.2
## - yrs.service    1 2.7100e+09 2.0083e+11 7968.6
## - discipline     1 1.9237e+10 2.1735e+11 8000.0
## - rank           2 6.9508e+10 2.6762e+11 8080.6
##
## Step:  AIC=7964.75
## salary ~ rank + discipline + yrs.since.phd + yrs.service
##
##                 Df  Sum of Sq        RSS    AIC
## <none>                        1.9890e+11 7964.8
## - yrs.since.phd  1 2.5001e+09 2.0140e+11 7967.7
## - yrs.service    1 2.5763e+09 2.0147e+11 7967.9
## - discipline     1 1.9489e+10 2.1839e+11 7999.9
## - rank           2 7.0679e+10 2.6958e+11 8081.5

##
## Call:
```

```
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = Salaries)
##
## Coefficients:
##             (Intercept)  rankAssociate Professor           rankProfessor
##                 84374.2                  12831.5                45287.7
##   disciplineTheoretical           yrs.since.phd             yrs.service
##                -14505.2                    534.6                 -476.7
```

Here, as we can see that it eliminated the 'sex' variable from the full model but ir hardly caused any improvement in the AIC value.

Fitting the improved model

Now, let's refit the full model with the best model variables suggested by the stepwise process above.

We note that from our improved model, having **more** experience **lowers** the salary but the more time since PhD - the higher the salary.

```
##
## Call:
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65244 -13498  -1455   9638  99682
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               84374.2     2951.4  28.588  < 2e-16 ***
## rankAssociate Professor   12831.5     4147.7   3.094  0.00212 **
## rankProfessor             45287.7     4236.7  10.689  < 2e-16 ***
## disciplineTheoretical    -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
```

Hence these explanatory variables are tightly connected (eg: linear relationship), we will use VIF to measure the multicollinearity.

The Pearson correlation between years since Ph.D. and years service is 0.9096491 which is almost 1. This means that both the variables hold a strong positive linear relationship. Therefore, we should avoid considering both the variables in our fitted regression model since they are not independent and would affect the our prediction results. The VIF test for this model shows that 2 out of the 5 variables are indeed pretty high, these two variables are years since Ph.D. with a VIF of 7.518936 a years of service with a VIF of 5.923038.Therefore, we can say that the coefficients in our sample were poorly estimated and the variables years since PhD and years service should be further analyzed.

```
##                   GVIF Df GVIF^(1/(2*Df))
## rank          2.013193  2        1.191163
## discipline    1.064105  1        1.031555
## yrs.since.phd 7.518936  1        2.742068
## yrs.service   5.923038  1        2.433729
## sex           1.030805  1        1.015285
```

In order to address these issues we are going to compare the full model with a reduce model. Essentially, we will consider all predictor variables except sex and years since PhD because it has a higher VIF. The output for the reduced model is below:

```
##
## Call:
## lm(formula = salary ~ rank + discipline + yrs.service, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64198 -14040  -1299  10724  99253
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              85814.96    2893.28  29.660  < 2e-16 ***
## rankAssociate Professor  14483.23    4100.53   3.532 0.000461 ***
## rankProfessor            49377.50    3832.90  12.883  < 2e-16 ***
## disciplineTheoretical   -13561.43    2315.91  -5.856 1.01e-08 ***
## yrs.service                -76.33     111.25  -0.686 0.493039
## ---
## 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
```

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

Reducing the model seems to partially address the issues with col-linearity. The VIF test shows that all values are below 5 which is a good indicator for no col-linearity problems. However, the summary of the model shows that excluding sex and years service from the model does not completely fix collinearity. As we can see years service still has a negative coefficient.

Subset wise selection

We are going to focus on estimating which variables should be included in the fitted multiple linear model. In this case, we applied subset wise selection.

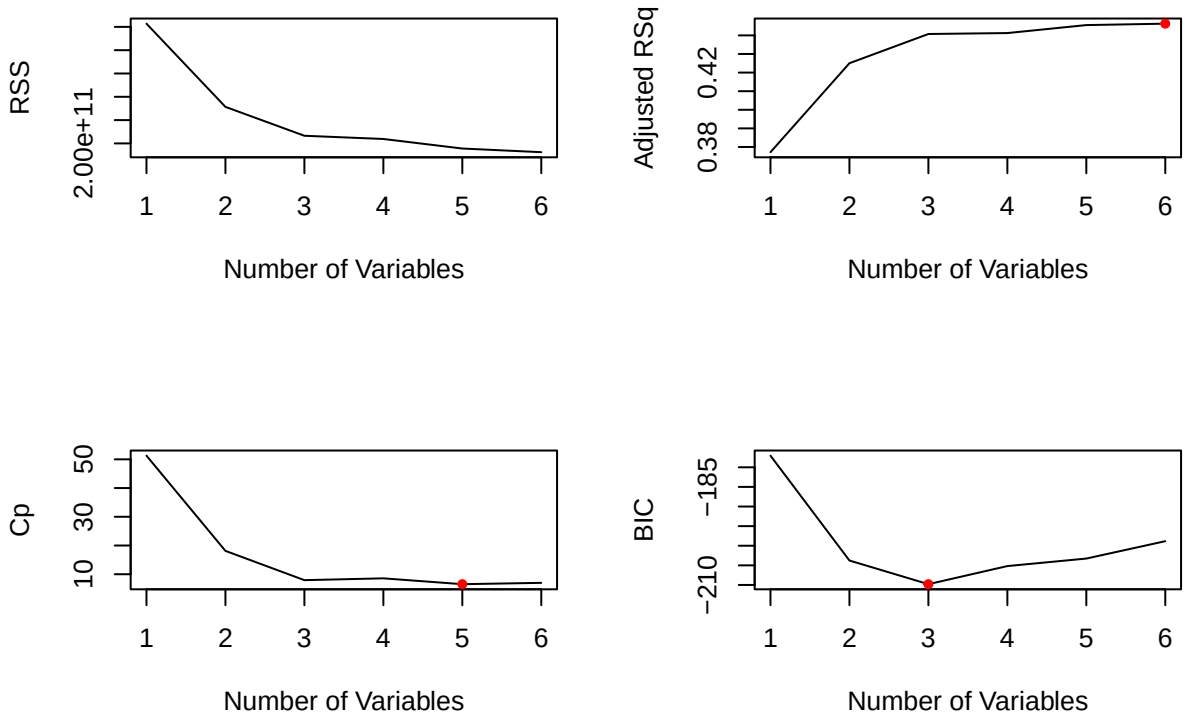
```
##      Adj.R2        Cp       BIC
## 1 0.3772158 51.273074 -177.0373
## 2 0.4250270 18.127814 -203.7709
## 3 0.4407437  7.934407 -209.7988
```

```
## 4 0.4412470  8.568101 -205.1838
## 5 0.4455269  6.536793 -203.2665
## 6 0.4462870  7.000000 -198.8438
```

```
## [1] 6
```

```
## [1] 5
```

```
## [1] 3
```



As we can see that subset wise selection suggests third model according to BIC criteria. Therefore we are going to fit a model using BIC suggestion such as 'rank' and 'discipline'.
The fitted model is shown below:

```
##
## Call:
## lm(formula = salary ~ rank + discipline, data = Salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65990 -14049  -1288  10760  97996
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 85705       2887  29.687  < 2e-16 ***
## rankAssociate Professor     13762       3961   3.475 0.000569 ***
## rankProfessor               47844       3112  15.376  < 2e-16 ***
## disciplineTheoretical      -13761       2296  -5.993 4.65e-09 ***
## ---
## 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
```

T-TEST of Significance Difference

```
##
##  Welch Two Sample t-test
##
## data:  c1 and c2
## 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
```

```
##
##  Welch Two Sample t-test
##
## data:  c1 and c2
## t = 1.0391, df = 22.451, p-value = 0.3098
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5119.769 15426.192
## sample estimates:
## mean of x mean of y
##  127120.8  121967.6
```

Conclusion: 44.07% of the variability can be explained by the fitted multiple linear regression model. Finally we can conclude that discipline and rank are significant in the analysis of Salary than the **sex** variable. Therefore, we should consider having them in the model since they have steady positive relation with salary.

The above t-test yields a p-value=0.002664 (significance level=0.05), which means that the salaries of male and female faculty are statistically different. To probe a bit further, we wonder whether the difference might be due to the different numbers of male and female, at the different ranks. So to eliminate the effect of rank, let us do a t-test on the salaries of male and female full professors (rank='Prof'). This test indicates the salaries of the two groups are NOT statistically different when the variable rank is fixed(p-value = 0.3098 with significance level=0.05).