JSC370 Final Report

Rui Miao

2022-04-22

A. Introduction

There are many different factors that affects the salary level of employers working in a company, such as degree, experience, number of years working in the company, gender, and so on. In this research, we are interested in the factors affecting the salary level of Data Scientists. According to a journal Research administrator salary: association with education, experience, credentials, and gender[1], we expect a positive relationship between experience and salary level, and male tends to have a higher salary than female.

a. Purpose of the Research:

The purpose of the research comes from the fact that as statistics students, we are curious about careers related and their prospects. This will provide us with some insights into my future career path and what to expect. The importance of the research is that most third-year students are preparing to find an internship in the near future. And salary level is one of the most essential factors that we will consider to select a job. Thus, we are interested in studying the factors which may affect the salary level.

b. Research Question:

How does salary relate to years of experience, years at the company, and gender?

We are interested in this specific research question since answering this question can give us a preliminary understanding of our possible future development working in a company. Besides, we are also interested in whether there is gender discrimination when it comes to the salaries of male data scientists and female data scientists.

Our goal for this research is to build a statistical model which is best for explaining the relationship between salary, and other possible predictors. And we may use our model to predict the salary that we may receive in out future career.

c. Background of the Dataset

This dataset[2] is found on the website *Kaggle* containing salary records from top companies. There are in total 62642 observations and 29 variables in our original dataset. We are interested in the question *How does salary relate to years of experience, years at the company, and gender?*. Thus, the variables that we are interested in are *salary*, *yearatcompany*, *yearsofexperience* and *gender*.

B. Methods

a. How and When the data was acquired

The data was scraped off levels.fyi and then cleaned by the author. The license[3] of this data set is including in the *Reference* section. According to the terms and conditions of the website, people may use the data for personal, non-commercial purposes (mentioned in the last sentence in the "License" section).

b. Cleaning and Wrangling the Data

There are in total 62642 observations and 29 variables in our original dataset. After filtering out unrelated titles and leaving only observations with Data Scientist as title, observations with missing values in gender and "Other" in gender are also filtered out. As one of the purposes of our analysis is to explore whether there is gender discrimination in salaries. The size of the dataset shrinks to 1718 observations. We are only interested in exploring 12 of the 29 variables including *salary*, *yearatcompany*, *yearsofexperience gender, Bachelors_Degree, Masters_Degree, Doctorate_Degree, Some_College, Highschool, Race_Asian, Race_White,* and *Race_Black*. We mutated the gender to a numeric values, where male is 0, and female is 1 so that we could interpret the gender data more easily.

As shown, total yearly compensation ranges from 10000 to 900000 which is a large interval. And we noticed that the median is slightly lower than the mean, implying that the distribution of salary might be left-skewed. And the table illustrates that the variance of salary is much larger than the variances of our predictors, which coincides with our large-interval finding of salary in Table 1.

Also the mean of year of experience is longer than the mean of year at company, which may indicates that employees usually have worked in different companies. And according to the values of *skew*, we noticed that all of the three numeric variables are right-skewed in different level.

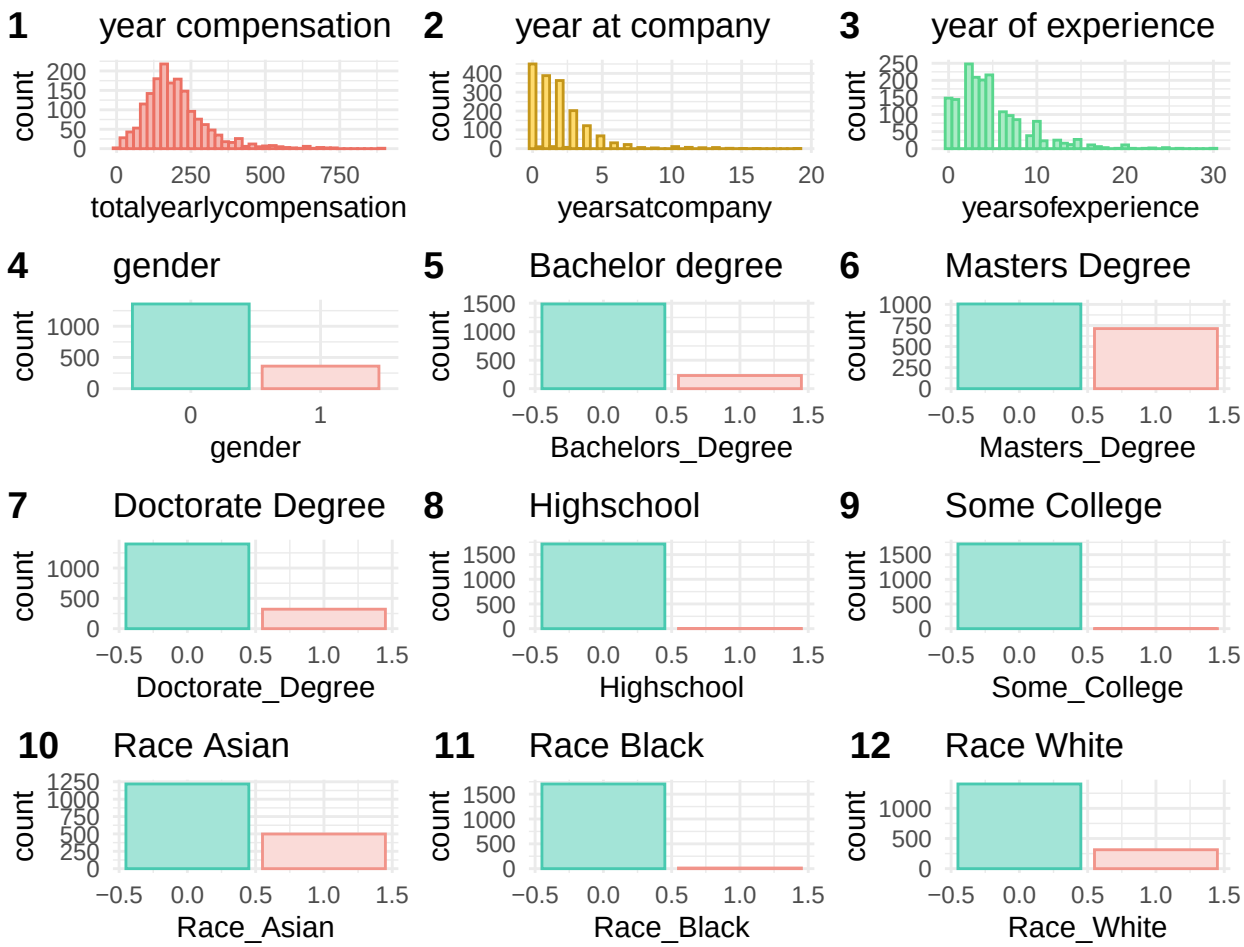
Based on the mean of Bachelor's degree,there are about 13% of the population (excluding the na) that have obtained a bachelor's degrees, while there are about 42% of the population (excluding the na) that have obtained a master's degree, and there are about 19% of the population (excluding the na) have obtained a doctor's degree. Also, Asian are the most popular race in this dataset.

Figure 1 showing the distribution of salary indicates a left-skewed trend with some outliers on the right, which have salaries over 5*105dollars per year. Most values concentrate in the interval between 10000 and 4*105. There are 35 observations that have a salary above 5*105. Since they are likely to affect the result of our linear model, we may remove them when fitting a linear model.

Figure 2 (years of experience) and Figure 3 (years at company) display a similar trend as Figure 1 (salary). They are even more left-skewed than salary and have a few outliers as well, for instance, the 41 people with more than 15 years of experience and the 44 people with more than 7 years at the same company.

Figure 4 (gender) illustrates a strong gender imbalance. There are fewer than 400 female data scientists and more than 1200 male data scientists in our dataset.

From the figures shown, we can see that our continuous predictors including years of experience and years at the company have similar distribution as our response variable, which justifies the use of linear regression in the analysis.



c. Variable Selection:

To build a model, the first thing to start with is selecting reasonable predictors and responses that could provide an answer to the research question. Firstly, a linear model with all possible and reasonable variables related to the research question was fitted. Based on this model, we calculated the p-values for each predictors, and removed the predictors randomly until all the predictors have been checked. ANOVA partial F-Test and T-Test were used to evaluate if those predictors can be removed.

- if the p-value of ANOVA partial F-Test is smaller than 0.05, use T-Test to make further decision, larger than 0.05, the predictors should be removed
- if the p-value of T-Test is smaller than 0.05, the predictors cannot be removed; larger than 0.05, the predictors should be removed

All the predictors that cannot be removed are the appropriate variables to build our model.

d. Analyze the data

My research question can be answered using a **linear regression** model because we expect that there exists a linear relationship between our predictors and responses. And each row of our data represents an independent individual, which indicates that each row in our dataset should be independent to one another.

I will be using total yearly compensation (Y) as my response and use the predictors years of experience, years at the company, gender, Bachelors_Degree, Masters_Degree, Doctorate_Degree because we are interested in the annual salary level which is total yearly compensation, and we would like to see how years of experience, years at company and gender affect the total yearly compensation.

Some anticipated issues might be that the outliers have some impact when estimating the coefficients of our model and need to be removed. Another issue is that two of our predictor variables, years of experience and years at the company might be correlated to each other in a way that years of experience = years at the company + years at other companies. In fact, there are 543 individuals whose years of experience are exactly equal to their years at the company. The correlation between our predictor variables will make our prediction less accurate.

Then we would like to build some machine learning model to make prediction.

We would like to fit our training data set to a regression tree. First, we split the whole dataset into training and testing data (70/30). Since we are interested in the yearly salary, we set the yearly salary as the reponse and all other variables as predictors. Then we pruned the tree based on the optimal complexity parameter.

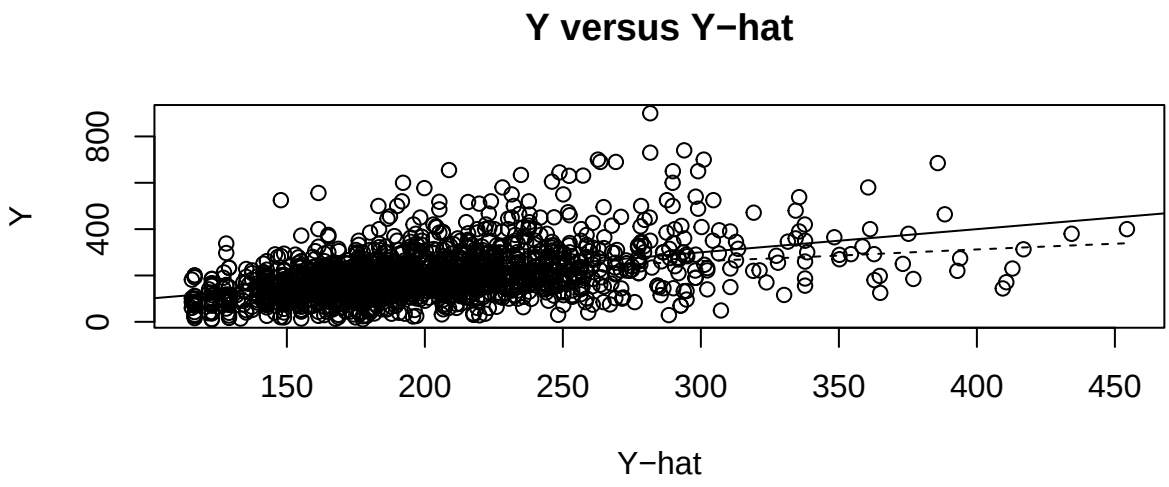
Other than the regression tree, we also fit our model to bagging model, random forest model and boosting with 1,000 trees for a range of values of the shrinkage parameter λ.

C. Result

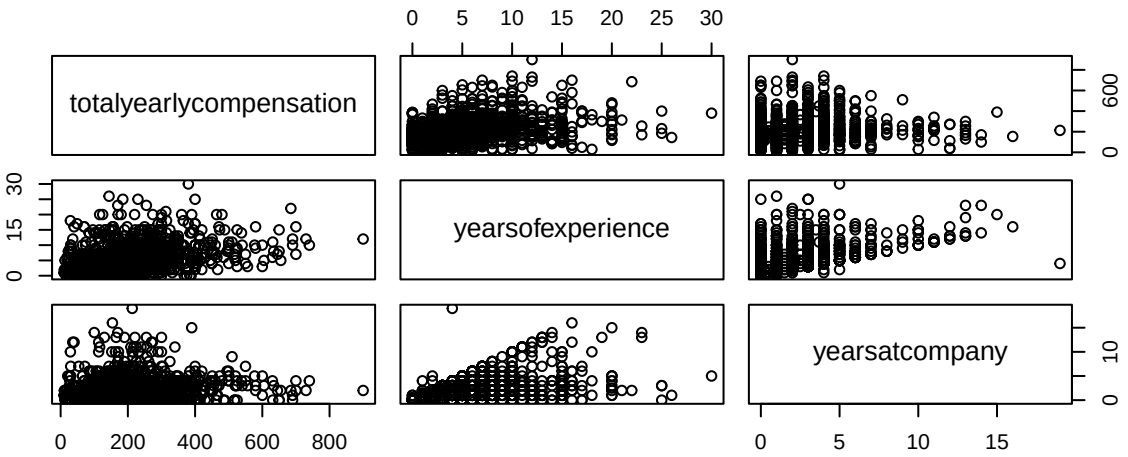
1. Process of obtaining the final model

After removing the variables with insignificant p-values (except Genders since we are interested in gender), the **ANOVA** partial **F-Test** is in-significant. Thus, all these predictors can be removed. For the remaining variables, w checked one by one, and for removing every one of them, the values of AN OV A partial F-Test is significant, which indicate that all the remaining cannot be removed. Then the model is only having predictors *yearatcompany*, *yearofexperience*, *gender, Bachelors degree, Masters degree, and Doctorate degree, and response totalyearlycompensation.*

2. Goodness of the result



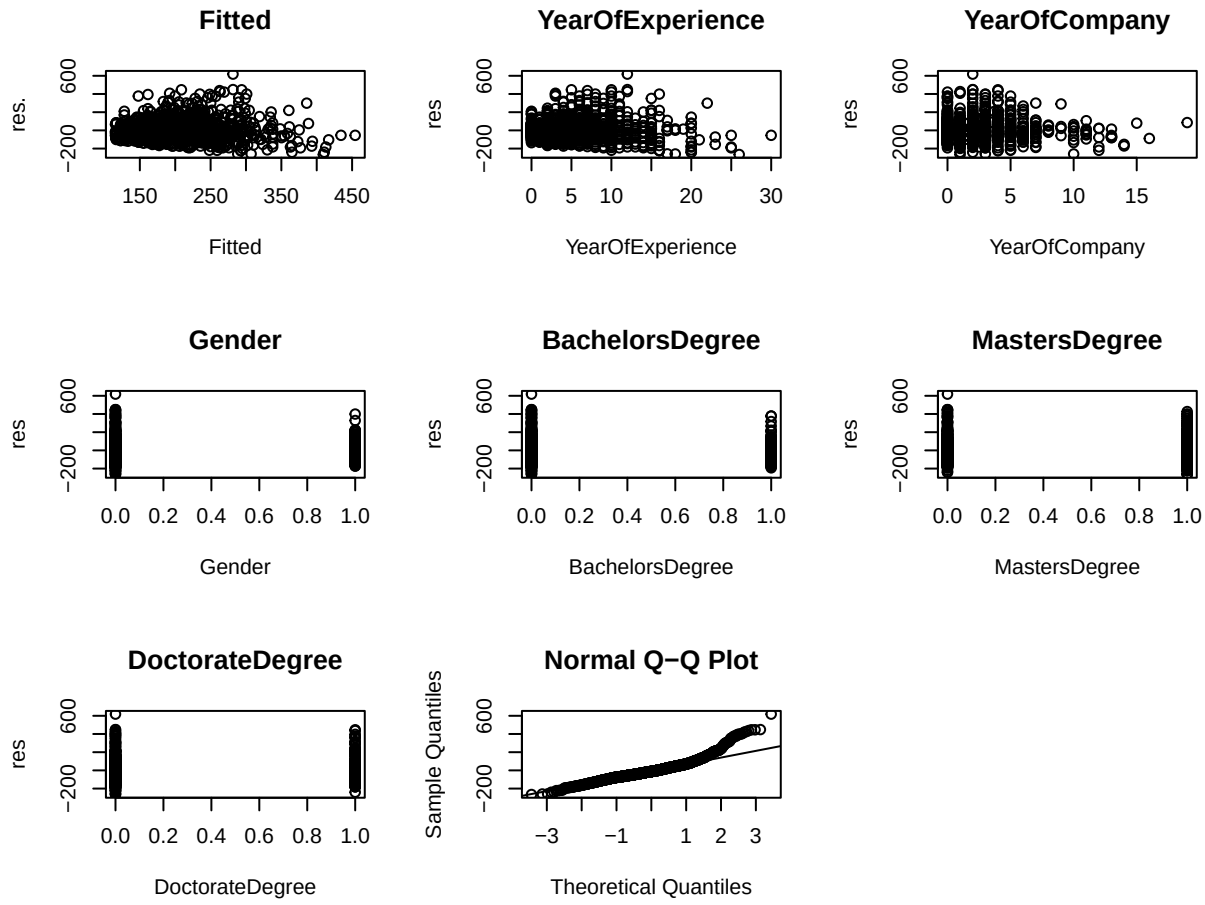
The plot above showing the relationship between the response Y and the fitted values of Y. By observing the plot of response against the fitted values, we see that there is a non-random scatter around the identity function, indicating that there is not a simple function of a linear combination of predictors. And most of the points in the graph are having a positive values. Thus, conditional means response is not a perfect single function of a linear combination of the predictors.



Based on this plot, we noticed that there is some patterns between each pairs of variables. Most of the plots have points concentrating at the lower left and moving upward to the upper-right corner. Therefore, most of the associations between each possible pairs of the association is positive. Only the scatter plot for total yearly compensation and years at company seems to have an unclear association. Above this, we can conclude that conditional mean of each predictor has a linear function with another predictor.

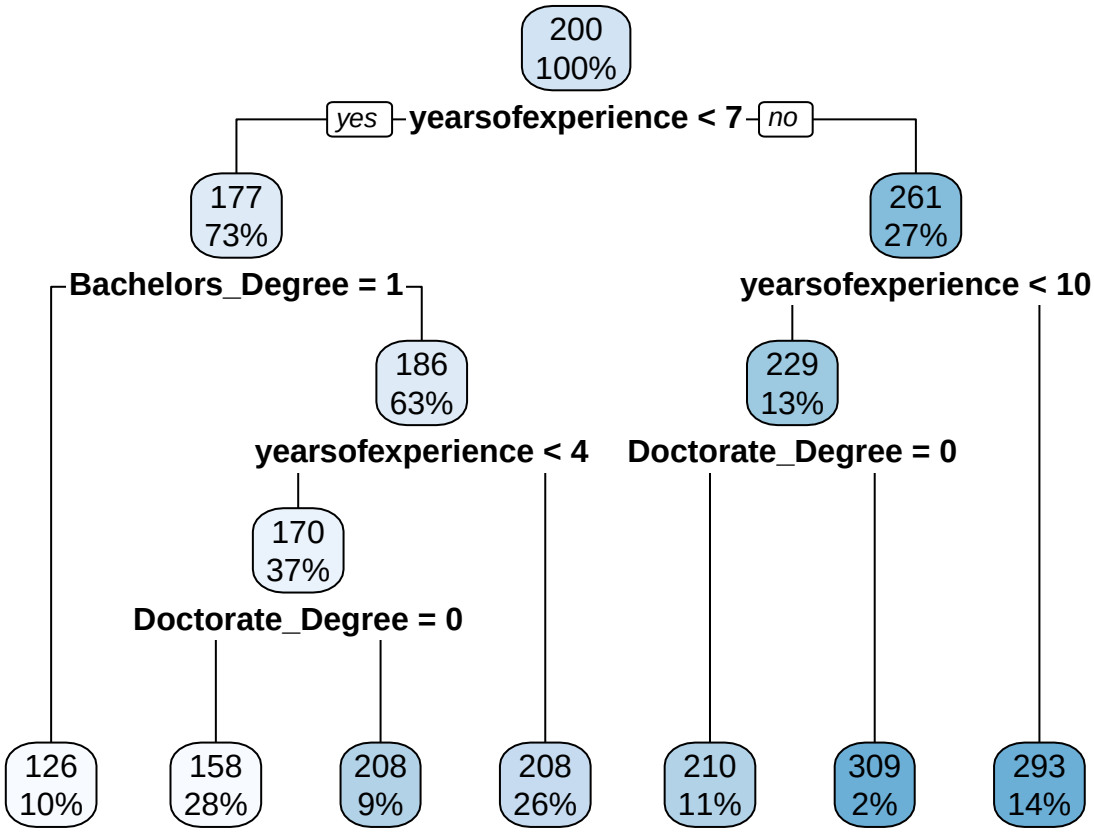
From the plot below, We recognized that the scatters does not spread randomly and equally. But we noticed that all the points are more equally distributed at both side of 0,. This implies that **assumptions 1** is likely satisfied, which is the population errors have mean zero. However, assumption 2, 3 still do not hold perfectly, which are the population responses (equivalently errors) are somhow correlated with each other, and the population errors (equivalently responses) do not have constant spread/variance around conditional mean.

By the normal QQ-plot, there is a straight diagonal string of points along an identity function, with some deviation on both ends. This shows that there is a **normality**, the populations errors are normally distributed with the mean of assumption 1.




```
##                              term   Estimate Std. Error    t value      Pr(>|t|)
## 1                     (Intercept) 165.033014  5.2529025 31.4174908 1.923442e-171
## 2 datascience$yearsofexperience  10.415917  0.6497583 16.0304484  5.460427e-54
## 3    datascience$yearsatcompany  -4.183001  1.1953335 -3.4994428  4.781718e-04
## 4  datascience$Bachelors_Degree -48.450760  7.6040814 -6.3716783  2.397603e-10
## 5    datascience$Masters_Degree -22.281666  5.5454344 -4.0180200  6.123390e-05
## 6  datascience$Doctorate_Degree  41.485401  6.7276924  6.1663642  8.699036e-10
## 7           datascience$gender1  -1.003891  5.6693093 -0.1770746  8.594707e-01
```

This is the summary table of our model. Based on the p-values in the table, we noticed that the p-values for **year of experience** and **year at company** are extremely small, which indicates that these two are two main factors of the salary level.



Here is the tree that we built to make the prediction. The max depth after pruning is 5. And the factors that appear in the tree are year of experience, year at company, Doctorate degree.

Table 1: MSE of all models

models	MSE
Pruned Regression Tree	3797.952
Bagging	8821.297
Random Forest	8199.548
Boosting	8474.451

Based on the table above, the regression tree model have the smallest MSE.

D. Conclusion and Summary

The most significant factors that would affect salary are years working at company, years of experience, Gender.

The relationships between response and each predictors are as follow:

- years working at company: negative

- years of experience: positive
- genders: positive

According to the above relationships, we are not surprising to see that if you are more experienced, you will earn more.

However, there are some interesting facts that for people works longer in a company, they will earn less. This could be caused by some extremely outliers in the dataset. Usually, employees will at least stay at the same salary level through out the duration of working at the companies. Another one is that people having a bachelor's or master's degree earn less than those does not have. All the information of degree are recorded by 0 and 1 without NA. Thus, 0 may be the default values for the column, which may represents not having that degree or unknown. This may influence the result and cause the interesting facts.

Another funny fact is that female earns more than males. Because from many facts of gender inequality in reality, females earn less than males at the same positions, with same degree. I think this may show an improvement in the gender equality nowadays. Another reason may be the percentage of female in this dataset is much smaller than male, thus, the data of female in this dataset is less representative. However, I would like to believe that we have a more open, less descrimination working environment for female than in the past.

For the machine learning model, I have tried all kinds of models that we learned, only the regression tree models give us the least mse, which is 3797.952. However, the mse is still significantly large.

Overall, only according to the result, experience is the most important factors to earn more. And generally, females earns more than males when working as a datascientists.

F. Reference

[1] Shambrook, Jennifer, et al. *"Research administrator salary: association with education, experience, credentials, and gender."* Journal of Research Administration, vol. 42, no. 2, fall 2011, pp. 87+. Gale In Context: Canada,

https://go.gale.com/ps/i.dop=CICu=utorontomainid=GALEA276807849v=2.1it=rsid=bookmark-CICasid=809af152.

[2] "Data Science and STEM Salaries", Kaggle.com, 2021. [Online].

Available: *https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries.*

[3] License: According to the terms and conditions of the website, people may use the data for personal, non-commercial purposes (mentioned in the last sentence in the "License" section). This is the link to the page of terms and conditions of level.fyi:

https://www.levels.fyi/about/terms.html