Report of MA678 Midterm Project

Tao He

12/01/2021

Abstract

It is commonly recognized that the educational degree plays a significant role in getting a job and how a degree affects the salary has sparked heated discussion. While salary is not the first consideration for most people who choose to pursue a higher degree, we still think it is an interesting question. Therefore, I used salary data for employees who graduated from data science and STEM program and works on 12 companies, including Apple, Amazon, and Microsoft. Then, based on the multilevel regression analysis, I explored that having a higher degree had a positive impact on salary. This report consists of 5 main sections: Introduction, Method, Results, and Discussion.

Introduction

Salaries often depend on the value of each employee and their value to the company as well. However, each employee has unique characteristics but also has similar backgrounds to other employees beyond various education levels, such as work experience and stock gifted. And certain specific qualities will allow the employee to contribute more to the company and stand out among the candidates, for example, a candidate with a lot of work experience is more likely to be hired by a company since not only he can get started in the shortest time, but also does not need to spend extra time and money to train him.

Therefore, I used various educational levels to classify each employee in different companies and also consider work experience and being gifted stock to determine whether pursuing a Ph.D. is worthwhile. Before that, I would clean the data and select some needed variables.

Method

Data Cleaning and Processing

The original data set is published on Kaggle: Data Science and STEM Salaries, which is a training file for this data set that has over 60,000 application salary records and 29 variables.

There is a lot of information in this data set, including characters and binary variables which indicate whether a worker has tho those characteristics, e.g., he/she is an Asian with a Ph.D. working in Apple. Since we do not use all the columns in the dataset, I chose the following variables: ***Total Yearly Compensation***, ***Education***, ***Company***, ***Years of Experience***, ***Stock Grant Value***.

- **Education**: High school, Bachelor's Degree, Some College, Master's Degree, PhD

All those columns are 0 and 1 variables. For example, when "High school" equals 0, the worker has only a high school degree. Also, some college means someone started university but not finished yet.

- **Company**: Amazon, Apple, Capital One, Cisco, Facebook, Google, IBM, Intel, Microsoft, Oracle, Sales force, VMware

After getting my new data set, I removed the useless space characters for some columns with the "dictionary" type, otherwise, the subsequent filtering section would not be able to filter all the eligible rows and maximize the use of the original data set. Then, I selected the twelve companies with the largest amount of data to see how education level affects workers' annual income. Finally, I got the cleaned data with 4657 observations.

Exploratory Data Analysis

For `Total Yearly Compensation` and `Stock Grant Value` have a large range and also if we see the density plots of them, there will be a long tale. Therefore, in order to make the plot more easy to read, I take log of these variables and draw some distribution plot and scatter plots to see if there is correlation between some variables with total yearly compensation, since my question is how education levels affects the salary among different companies.

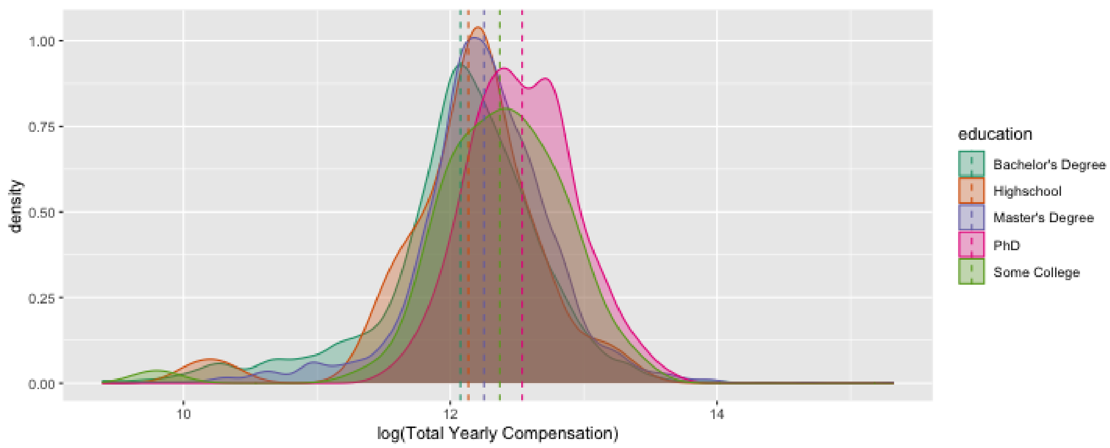


Figure 1: Distribution of annual income among different education levels

Figure 1 shows that in most cases, the higher the degree, the higher the salary. In detail, the salary of Ph.D. employees is higher than those of other degrees. Moreover, one interesting thing is that bachelor's degree employees are paid more than those who work graduate from high school. This can happen since there is a gap between college and university and their efforts on the study are not the same as well, therefore, after spending a lot of money on college, some graduates with bachelor's degrees are not as good as high school students.

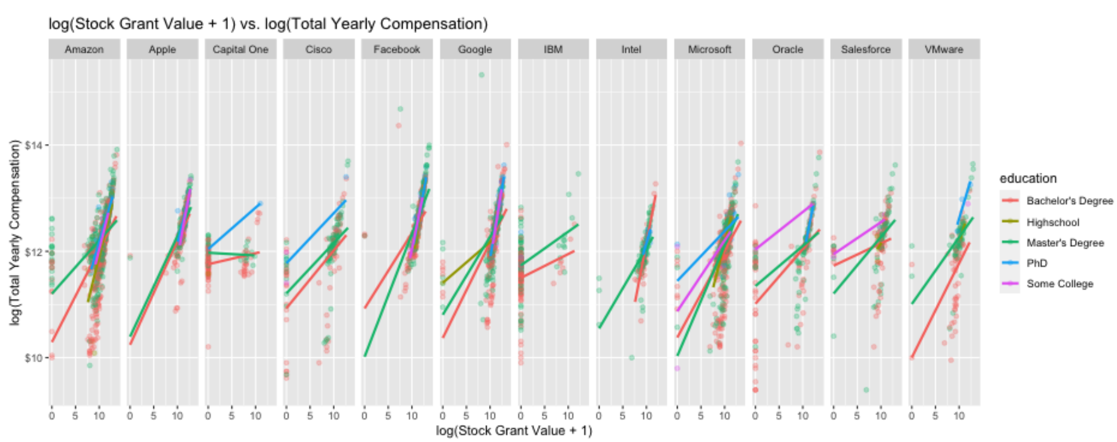


Figure 2: Data was separate into groups with different company. Different colors represent individuals are in different education level.

Figure 2 shows the relationship between the value of stock grants owned by individuals and total annual compensation. In the majority of companies, there is a positive relationship. However, the effect varies in different companies with different intercepts and slopes. In detail, with the same stock grant value, compared with other education levels, Ph.D. workers have a higher annual salary, especially in some companies, like `Apple`, `Google`, and `Oracle`, which are all Technology Companies.

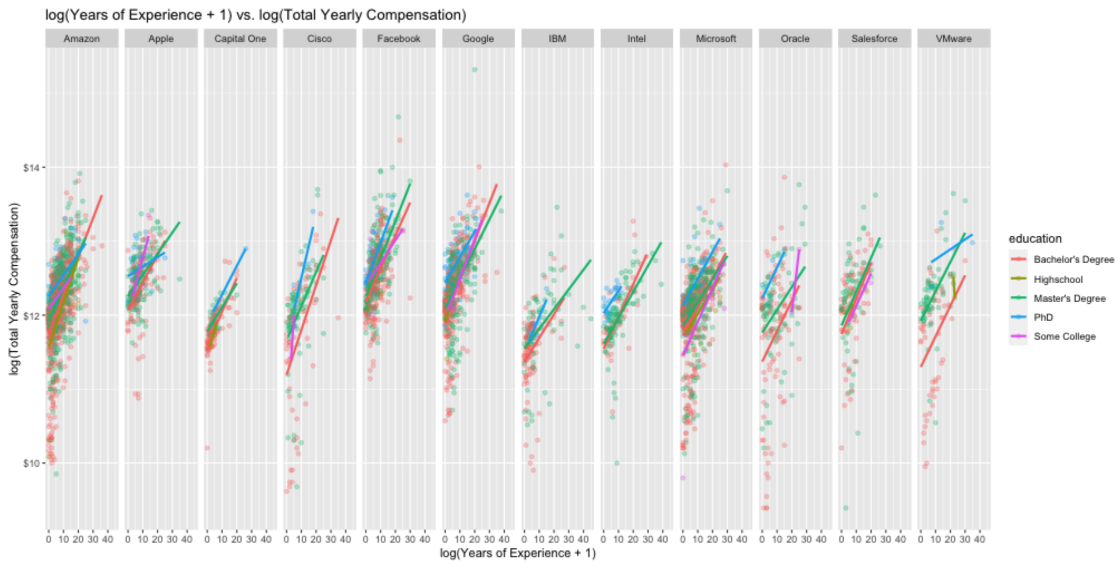


Figure 3: correlation between years of experience and total yearly compensation.

Figure 3 indicates that there is also a positive correlation between work experience and annual income in different company groups. However, the magnitude and intercept of the effect vary widely across education levels. This makes sense since some people drop out of college without a degree and start their own business or are tapped by companies for special talent. Therefore, they are highly capable and valuable to the company, especially after gaining work experience.

Moreover, we noticed that in almost all companies, some college workers' salaries increase rapidly with work experience growing. Additionally, in VMware, when a worker graduated from high school has more work of experience, he/she will have less yearly income, which is quiet strange. therefore, I decided to go in detail to figure out how the education level affect the yearly income in various company.

Model Fitting

Since the annual income of workers varies in different companies, especially those with different levels of education, then I decided to use a multilevel model to fit the "Total Yearly Compensation". As for variable selection, in addition to the binary variable of education level, I also included "Stock Grant Value" and "Years of Experience", which directly affect the annual salary as we mentioned before. Furthermore, as these two continuous variables are more or less skewed and have heavy tails, I used `log(variable + 1)` to create new variables. Their distribution plots can be found in the Appendix of this report. Since it is clear from the EDA that different levels of education and annual salary are correlated across companies, I use different slopes and intercepts in the multilevel model. Below is the function:

$$log(TotalYearlyCompensation) = -0.02 + 0.09 \cdot log(StockGrantValue + 1) + 0.23 \cdot log(YearOfExperience + 1) +$$

$$0.13 \cdot MastersDegree + 0.18 \cdot BachelorsDegree + 0.12 \cdot DoctorDegree +$$

$$0.12 \cdot HighSchool + 0.03 \cdot SomeCollege + effect_{company}$$

And to see the fixed effects below, all variables are significant at alpha = 0.05 level.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	0.11	0.43	<0.00	25.81	0.992
masters_degree	-0.12	0.42	<0.00	-0.27	0.993
bachelors_degree	-0.22	0.42	<0.00	-0.53	0.991
doctor_degree	0.09	0.42	<0.00	0.22	0.995
high_school	-0.22	0.43	<0.00	-0.53	0.987
some_College	0.03	0.43	<0.00	-0.08	0.997
log_years_of_experience	0.23	<0.00	4621.00	30.07	<2e-16 ***
log_stock_grant_value	0.09	<0.00	3850.00	33.18	<2e-16 ***

Result

Model Coefficients

Just take some example here, for company Apple, we can conclude this formula:

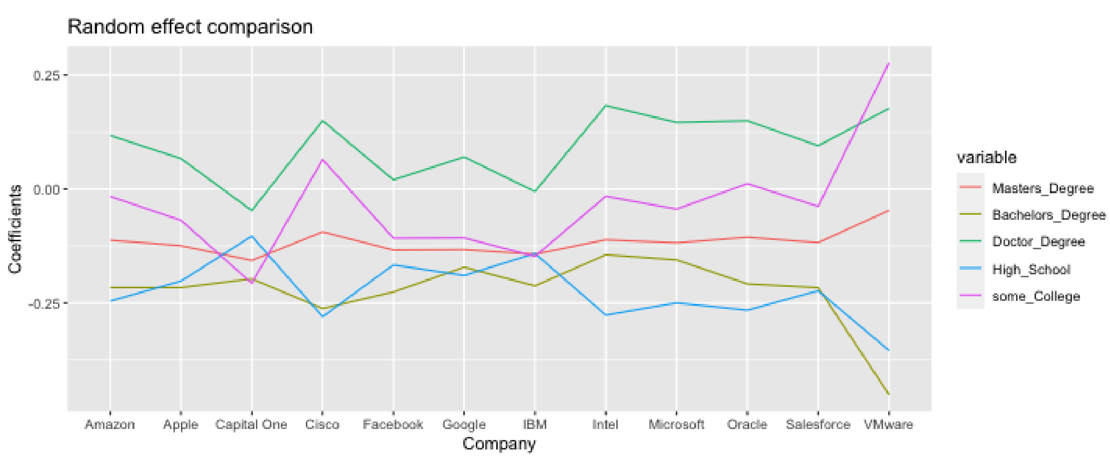
$$log(TotalYearlyCompensation) = 10.96255 + 0.23 \cdot log(StockGrantValue+1) + 0.09 \cdot log(YearOfExperience+1) -$$

$$0.11 \cdot MastersDegree - 0.22 \cdot BachelorsDegree + 0.12 \cdot DoctorDegree - 0.25 \cdot HighSchool - 0.02 \cdot SomeCollege$$

- explain why PhD's intercept is the highest *

And all the parameters of three predictors are all bigger than 0, which means they all have positive impact on number of comments. For each 1% difference in stock grant value, the predicted difference in total compensation is 0.53%. And the same for base salary.

	(Intercept)	Masters_Degree	Bachelors_Degree	Doctor_Degree	High_School	some_College
Amazon	11.07	-0.13	-0.36	0.28	-0.37	0.01
Apple	11.07	0.75	0.18	2.31	-0.19	0.07
Google	11.06	>0.00	-0.16	0.38	-0.16	0.05
IBM	11.15	0.14	-0.06	0.47	0.03	0.14
Microsoft	11.00	-0.18	-0.34	0.27	-0.41	-0.03



Model Validation

- plot random effect *

Discussion

limitation: test, correlation between binary predictors and continuous predictors

Next steps And since we will be graduating from the MSSP program soon, is a PhD worth it?

Citation

Hadley Wickham (2017), tidyverse: Easily Install and Load the 'Tidyverse', R package version 1.2.1.: https://CRAN.R-project.org/package=tidyverse

Jack Ogozaly, Accessed October 2021, Kaggle: Data Science and STEM Salaries, https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries

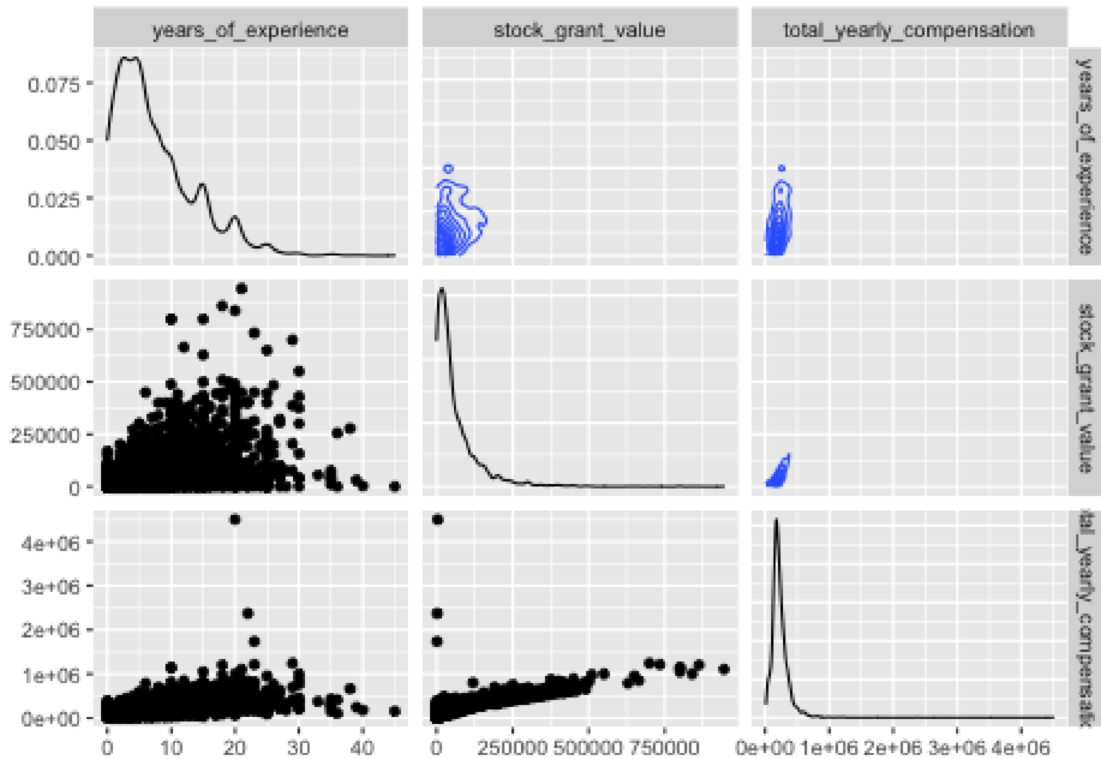
Marco Murtinu, Marta-Bar and Zarasim (November 14, 2021), Is a PhD worth it (Python)? https://www.kaggle.com/marcomurtinu/is-a-phd-worth-it

Rune Haubo Bojesen Christensen, lmerTest: Tests in Linear Mixed Effects Models, R package version 3.1.3.: https://CRAN.R-project.org/package=lmerTest

Plot random effects from lmer (lme4 package) using qqmath or dotplot: How to make it look fancy?,https://stackoverflow.com/questions/13847936/plot-random-effects-from-lmer-lme4-package-using-qqmath-or-dotplot-how-to-mak

Appendix

Check distribution and correlation



More EDA

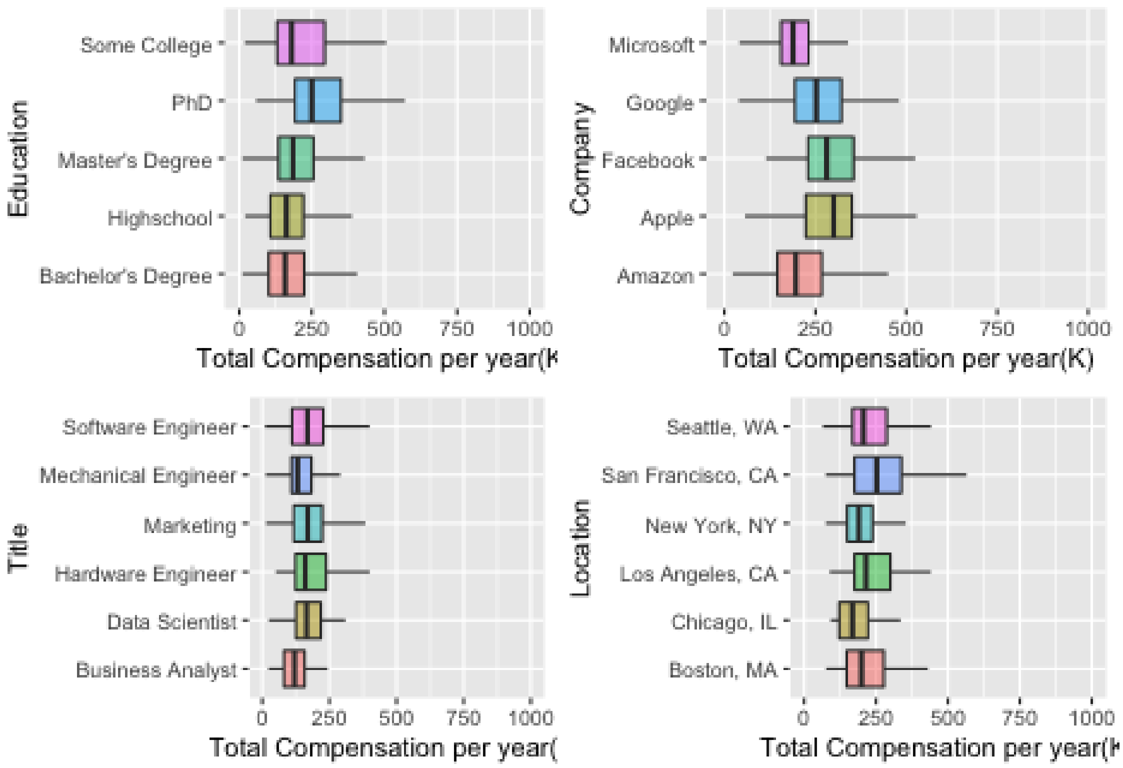


Figure 4: Various distribution of total yearly compensation in thousands by different measurement

Model check

Model coefficients

```
FALSE $company
```

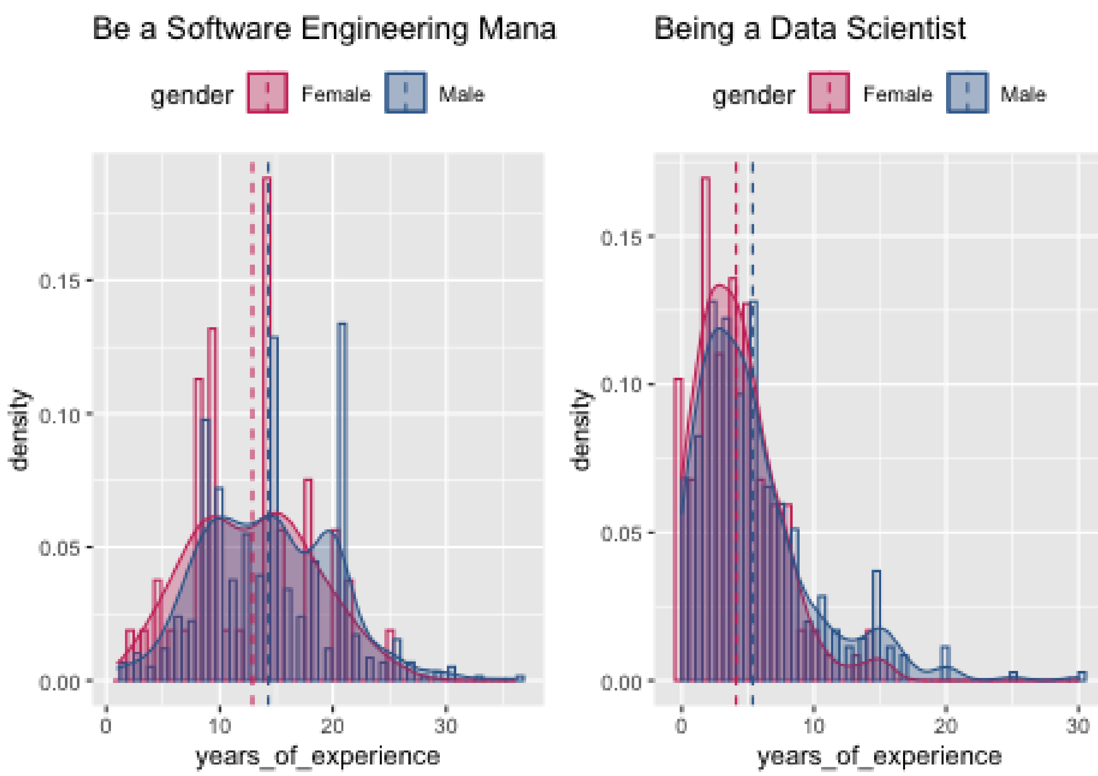


Figure 5: How many years of working experience we need to be a software engineering manager and a data scientist? It takes less time for a woman than a man. Moreover, if you would like to become a software engineering manager, you need about 13 to 15 years of work experience and if you want to become a data Scientist, you need about 4 to 5 years of work experience.

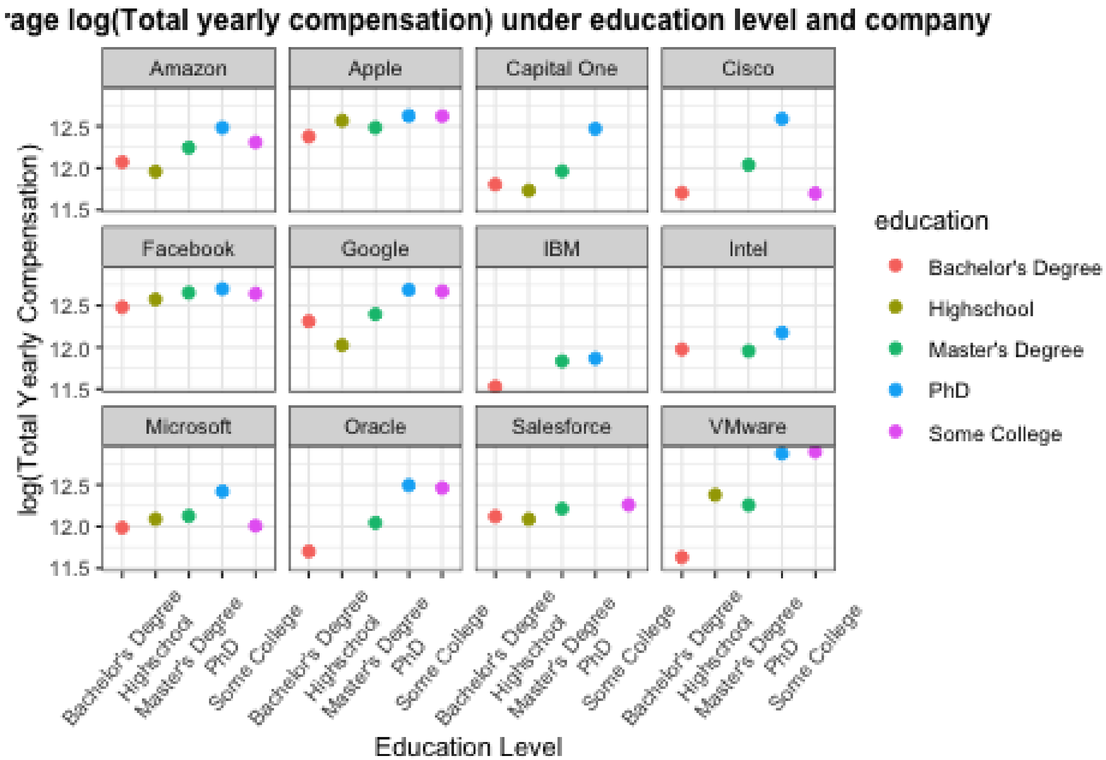


Figure 6: Average log(Total yearly compensation) under education level and company. In almost all companies, PhD workers are paid the highest salary.

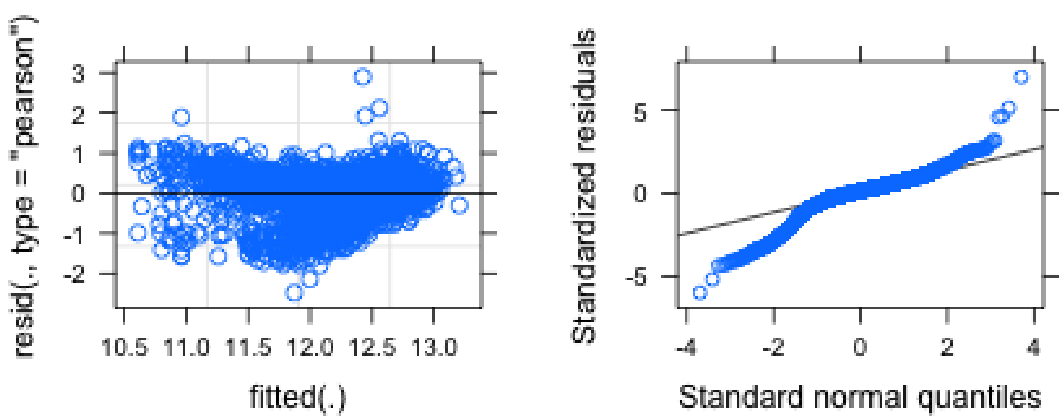


Figure 7: Residual plot and Q-Q plot.

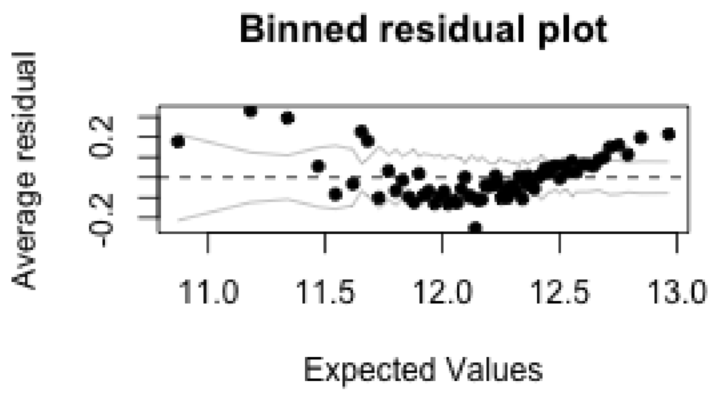


Figure 8: Binned residual plot

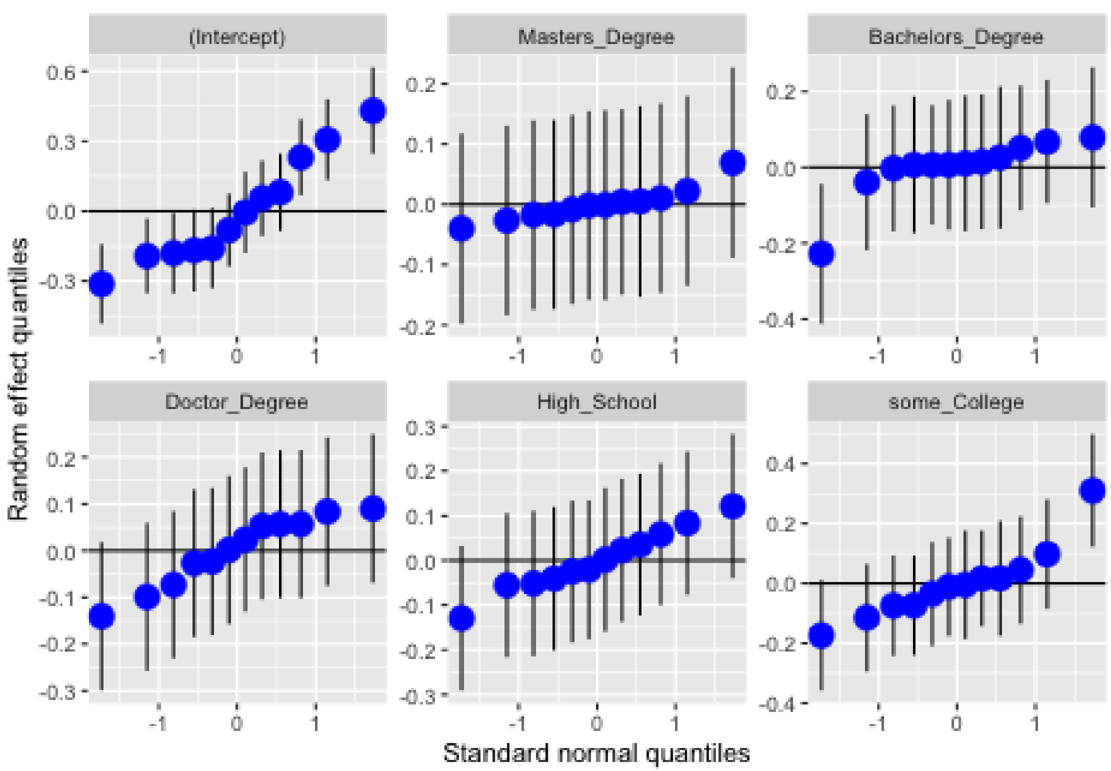


Figure 9: Random effect coefficients. Each point in various education level and intercept represents various group levels, which are companies

Full Results

Random effects of model

```
$company
            (Intercept) Masters_Degree Bachelors_Degree Doctor_Degree
Amazon      -0.08219157    0.004171669      0.006900100   0.023925516
Apple        0.08026156   -0.008080547      0.006925094  -0.026822705
Capital One  0.43077703   -0.040061802      0.025884466  -0.140761535
Cisco       -0.15955965    0.022274435     -0.039059354   0.056398020
Facebook     0.23048026   -0.017399468     -0.002868700  -0.073353453
Google       0.05548276   -0.016666940      0.051504211  -0.023386024
IBM          0.30633800   -0.026551898      0.010602947  -0.099040639
Intel       -0.31265777    0.005044646      0.078860419   0.089590660
Microsoft   -0.19266392   -0.001646319      0.067681238   0.052646846
Oracle      -0.18109050    0.010648512      0.014835000   0.056207749
Salesforce  -0.00584261   -0.001063962      0.006750927   0.001337921
VMware      -0.16933358    0.069331676     -0.228016349   0.083257644
             High_School some_College
Amazon      -0.021045938  0.016970177
Apple        0.022859605 -0.035320778
Capital One  0.121417897 -0.173382769
Cisco       -0.055215569  0.097967426
Facebook     0.058405171 -0.074581598
Google       0.035184038 -0.073918797
IBM          0.083520823 -0.114506427
Intel       -0.051667722  0.017255171
Microsoft   -0.024870944 -0.010807819
Oracle      -0.040874909  0.045013573
Salesforce   0.001563645 -0.004826691
VMware      -0.129276097  0.310138532

with conditional variances for "company"
```

Fixed effects of model

```
            (Intercept)          Masters_Degree        Bachelors_Degree
            11.03098198             -0.11635990             -0.22303899
          Doctor_Degree             High_School            some_College
             0.09361419             -0.22483142             -0.03323250
log_years_of_experience   log_stock_grant_value
             0.23154599              0.09327804
```

Coefficients of model

```
$company
            (Intercept) Masters_Degree Bachelors_Degree Doctor_Degree
Amazon         10.94879    -0.11218823       -0.2161389   0.117539709
Apple          11.11124    -0.12444045       -0.2161139   0.066791488
Capital One    11.46176    -0.15642171       -0.1971545  -0.047147342
Cisco          10.87142    -0.09408547       -0.2620983   0.150012213
Facebook       11.26146    -0.13375937       -0.2259077   0.020260740
Google         11.08646    -0.13302684       -0.1715348   0.070228169
IBM            11.33732    -0.14291180       -0.2124360  -0.005426447
Intel          10.71832    -0.11131526       -0.1441786   0.183204852
Microsoft      10.83832    -0.11800622       -0.1553578   0.146261039
Oracle         10.84989    -0.10571139       -0.2082040   0.149821941
```

```
Salesforce     11.02514   -0.11742386       -0.2162881   0.094952114
VMware         10.86165   -0.04702823       -0.4510553   0.176871837
            High_School some_College log_years_of_experience
Amazon       -0.2458774  -0.01626232                0.231546
Apple        -0.2019718  -0.06855328                0.231546
Capital One  -0.1034135  -0.20661527                0.231546
Cisco        -0.2800470   0.06473493                0.231546
Facebook     -0.1664263  -0.10781410                0.231546
Google       -0.1896474  -0.10715129                0.231546
IBM          -0.1413106  -0.14773892                0.231546
Intel        -0.2764991  -0.01597733                0.231546
Microsoft    -0.2497024  -0.04404032                0.231546
Oracle       -0.2657063   0.01178108                0.231546
Salesforce   -0.2232678  -0.03805919                0.231546
VMware       -0.3541075   0.27690603                0.231546
            log_stock_grant_value
Amazon                 0.09327804
Apple                  0.09327804
Capital One            0.09327804
Cisco                  0.09327804
Facebook               0.09327804
Google                 0.09327804
IBM                    0.09327804
Intel                  0.09327804
Microsoft              0.09327804
Oracle                 0.09327804
Salesforce             0.09327804
VMware                 0.09327804

attr(,"class")
[1] "coef.mer"
```