Final Project Report
August 17th, 2019

Key Factors That Foreshadow Employee Contribution @ IBM

MSBA 6120
Intro to Stats for Data Scientists

Brady Engelke
Sherlock Zhang
Penny Fan
Jashyant Sikhakolli

Content

Introduction

Business Problem

Many employers today struggle when conducting a candidate search because skilled talent is in short supply. Recruiters are usually bombarded with hundreds, even thousands of resumes. They spend a lot of time filtering for candidates who are not only a perfect fit for the firm's culture but also well-equipped to make an appreciable contribution to the firm. However, in some rare cases, an organization will find two incredible candidates that are equally impressive, but they might have the budget allocated only for one of them.

The hiring manager is likely to be apprehensive by the thought of hiring the wrong candidate. The longer a hiring manager takes to make the decision, the greater the risk of either candidate turning to a competing organization. So, what should the hiring manager do to make the best decision with the long-term growth of the company in mind? Can the team leverage multivariate regression to uncover which aspects of an applicant's background relate to an impactful employee contribution to the firm? If so, the hiring manager could utilize these insights to supplement the standard behavioral evaluation of motivational fit and cultural fit to make a more informed decision between these two promising candidates.

Data Set Information

In order to answer the question above, we obtained a dataset gathered by IBM. It contains 31 variables that encompass a wide range of employee attributes, all relating to an employee id. These attributes include marital status, work-life balance, environment satisfaction, business travel, income, stock options, performance evaluations, years with the current manager, and job involvement, etc. However, only 8 of these variables were deemed relevant by the team to accomplish the analysis described above. Those 8 variables are listed below:

Table 1: Original Variables of Interest

Variable	**Description**
Gender	Male or Female
Age	Calculated based on an employee's Date of Birth
YearsAtCompany	Number of years the employee has worked at IBM
EducationField	Field majored in such as Marketing, HR, Medical

Education Level	Level of Degree obtained prior to joining IBM on a scale of 1, 2, 3, 4 (Ph.D., Master's, Bachelor's, High School Diploma)
TotalWorkingYears	Total years of work experience
NumCompaniesWorked	Number of companies worked for including IBM
PercentSalaryHike	Total salary increase earned by an employee during their time with IBM
Performance Rating	Good or Bad

As the team is trying to engineer a tool that will help HR identify the candidate who will be the best fit at IBM, it was necessary to manipulate some of the variables above so that an employee's pre-hire attributes could be related to their resultant salary hike over their tenure at the company. Below is a table of the new variables in which the team engineered and some of the original variables. Additionally, included is a description of the manipulation required to create the new variables, and the utility each of the selected variables served in the analysis.

Table 2: Variables Used in Analysis

Variables	**Description**	**Type**
(New) Average%Yearly Salary Hike	Manipulation: (PercentSalaryHike / (Years at Company)) Utility: This measure was chosen over PercentSalaryHike because it defines the average rate at which an employee provides value. By normalizing PercentSalaryHike, we are accounting for employees that have been with IBM for a long time who are likely to have a higher cumulative PercentSalaryHike than recent hires.	Interval, Response
(New)Age When Hired	Manipulation: (Age – Years at Company) Utility: What is the age when hired that is most indicative of a valuable contribution to the company?	Interval, Predictor
(New) Previous Years of Experience	Manipulation: (Total Years of Experience - Years at Company) Utility: Are employees with more experience better prepared to provide impact?	Interval, Predictor

(New) Number of Previous Companies	Manipulation: (NumCompaniesWorked - 1(IBM)) Utility: Does an employee gather more dynamic experience by working for a large or small number of employers?	Interval, Predictor
Gender	Utility: Is there any disparity in average % yearly salary hike due to gender?	Nominal, Predictor
Education Field	Utility: Are there fields that have historically provided more productive employees for IBM?	Nominal, Predictor
Education Level	Utility: Does the average % yearly salary hike vary with the level of degree obtained?	Ordinal, Predictor
Performance Rating	Utility: Can we verify whether IBM awards salary hikes reasonably based on the performance rating?	Ordinal, NA

It is important to note that the *Average % Yearly Salary Hike* is a credible indicator as to how much value an employee has contributed to the company, only if IBM awards salary hikes judiciously. If there is any bias in the way IBM awards salary hikes, it may be detrimental to the usefulness of the tool for predicting whether an applicant will provide a valuable contribution or not.

Assumptions

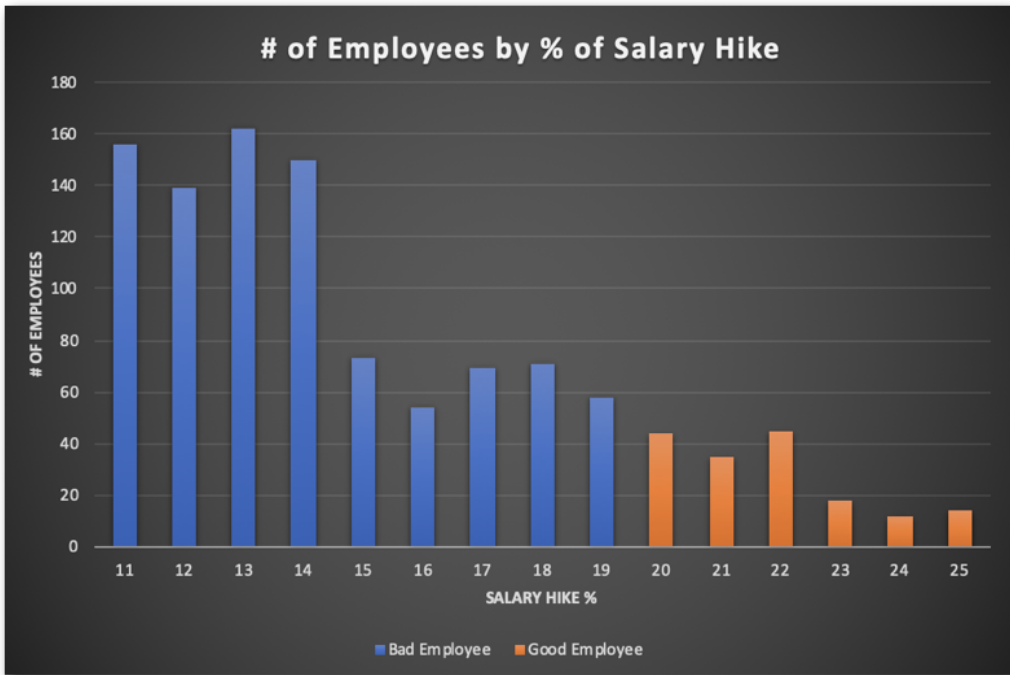
1) The data set was collected via a random sampling process
2) The data collection process is stable over time
3) The differences between the observations of our response variable and the best-fit-line provided by the regression model will be normally distributed

Analysis

Pre-Modeling

To assess the credibility of using *Average % Yearly Salary Hike* as a measure of employee contribution to the firm, the team tabulated the groups of employees who were rated "good" or "bad" by their supervisors by the level of total salary hike each employee received while at the company. The team found a discernible threshold at a total salary hike of 20% which can be seen below in Figure 1.1. The "good" employees all earned a salary hike above this threshold and "bad" employees earned total salary hikes below it. Additionally, a majority of employees are receiving salary hikes at the lower end of the range (11% - 14%) while only a few high-performing employees are earning total salary hikes in the 23% - 25% range. These results provide reasonable evidence that IBM does not allocate salary hikes haphazardly and can be used as a metric to assess the employee contribution to the organization.

Figure 1.1



The next step in the team's analysis was to map what predictor variables may be related to each other. This is pertinent due to the fact that if two predictors are correlated, the magnitude of the relationship between these predictor variables and the *Average % Yearly Salary Hike* is not interpretable within a regression model. Therefore, it is a reasonable endeavor to limit the occurrence of this phenomena within the modeling process. After creating a correlation matrix (See Figure 1.3 in Appendix), we discovered that *Previous Years of Experience* and the *Age When Hired* have the strongest linear relationship - which is intuitive. Furthermore, the *Number of Previous Companies* is moderately correlated with *Age When Hired* and *Previous Years of Experience* amongst employees at IBM, both intuitive as well. The reason that these correlations are notable is that it is likely these predictors are capturing similar aspects of the variation

within the *Average % Yearly Salary Hike* due to the fact they are collinear in nature. This information will be valuable when identifying which predictors to remove from the model in the next section.

Modeling

The objective of our modeling process was to refine the number of predictors we used to a minimal number, simplify the data gathering process for the model in practice, all the while ensuring the model will be the best tool for the HR department to use when trying to predict what the *Average % Yearly Salary Hike* will be for an applicant. The team decided to utilize a top-down approach for the modeling and to chisel away the insignificant predictors that did not capture any of the variation in the response variable within each successive model. Thus, we included all of the predictor variables into the first model to get an initial sense of how all of the predictors relate to an employee's *Average % Yearly Salary Hike*. The least significant relationship from the initial model indicated that there is a 65% chance we would have gotten these results if in fact there is no relationship between *Average % Yearly Salary Hike* and *Education Field* of the applicant, holding the other predictor variables fixed (See Figure 2.1 in Appendix). This result was counter-intuitive. We hypothesized that *Education Field* would be the richest predictor amongst our predictor set and would be a critical aspect of our final recommendation to the HR department. This was not the case based upon this finding and the team chose to remove *Education Field* from the subsequent model.

The second model depicted that both *Previous Years of Experience* and the *Number of Previous Employers* the applicant have had were insignificantly related to the response variable of interest. Statistically, there is a greater than 50% chance we would have gotten the results we obtained for *Previous Years of Experience* and *Number of Previous Employers*, holding the other predictor variables fixed, if in fact there is no relationship between these predictors and the *average % yearly salary hike* (see Figure 2.2 in Appendix). Thus, we deemed it reasonable to remove these predictor variables from our following model formulation.

All that was left as predictor variables for the third model were *Gender* of the applicant, *Education Level*, and *Age When Hired*. The results of the third model showed two substantial relationships and one relationship that could be deemed significant depending upon the HR department's acceptable level of error (see Figure 2.3 in Appendix). Below are the relationships identified in descending order of significance.

1. It is extremely unlikely we would have gotten the sample result we obtained, holding the other predictor variables fixed if in fact there is no relationship between *Age When Hired* and the *Average % Yearly Salary Hike* that employee ends up earning.
2. There is a 0.001% chance we would have gotten the sample result, holding the other predictor variables fixed, if in fact there is no relationship between the *Education Level* the applicant has obtained and the *Average % Yearly Salary Hike* that the employee eventually earns.
3. There is a 27% chance we would have gotten the sample result, holding *Age When Hired* fixed, if in fact there is no relationship between the Gender of the applicant and the Average % Yearly Salary Hike that the employee actually earns.

Since the acceptable level of error is unknown at this point, we developed a fourth and final model, removing *Gender* as a predictor from the model (see Figure 2.4 in Appendix). The resulting significance of *Education Level* and *Age When Hired* were about the same as the third model. Equipped with these findings

from each of the four models, the team was ready to weigh the pros and cons of each model and decide upon the most logical model for the HR department to utilize in their hiring efforts.

Post Modeling

Below is a table that contains several key elements of the four models that assisted the team in the model selection phase:

Table 3: Metrics for Model Selection

	Number of predictors	S	R^2	Highest p-value
Model 1	6	3.127	0.071	0.944
Model 2	5	3.124	0.068	0.613
Model 3	**3**	**3.122**	**0.068**	**0.270**
Model 4	2	3.123	0.067	0.036

In table 2, the number of predictors is a measure of how easy it will be to gather the necessary information to implement this model in practice. S represents the standard deviation amongst the observations of our response variable from the best-fit-line of the regression (aka. residuals) that is still unexplained by each model. The R^2 captures the proportion of variability explained in the model. The highest p-value summarizes the likelihood at which that the team would have observed the sample results if in fact there was no relationship between the least significant predictor and the response variable.

On observing the modeling results, it is clear that the R^2 and S aren't changing substantially as predictors are removed from the model. However, it is informative that S increases as predictors are added to each model, except when going from model 4 to model 3. This was a key driver as to why the team chose model 3, the addition of *Gender* to model 3 reduced the S value but increased the R^2 value. The last thing an analyst wants is to add predictors to a model and for the standard deviation of the residuals to increase. This undermines a foundational objective of any regression model of minimizing the standard deviation that is still unexplained within the response variable. This insight, amongst considering model 3's simplicity consisting of 3 predictor variables, competitive R^2 in comparison with the other models of 0.068 and having the highest predictor p-value of 0.270 solidified are the decision to select model 3 as our deliverable for the HR department.

Below are the guiding insights the team concluded upon after analyzing the selected model's results in descending order of importance:

1) Higher Education Level did not always align with a higher *Average % Yearly Salary Hike*. Intuitively, Ph.D. had the highest Average % Yearly Salary Hike while a high school diploma provided the lowest Average % Yearly Salary Hike. The fact that is interesting is those who obtained a Bachelor's degree had a higher Average % Yearly Salary Hike than those who held a Master's degree, holding *Age When Hired* fixed.

2) In terms of *Age When Hired*, we found its relationship with *Average % Yearly Salary Hike* was in line with intuition i.e. *Average % Yearly Salary Hike* goes up when *Age When Hired* goes up.
3) In terms of *Gender*, if an applicant is a male, on average he will receive an increase of 0.21% in *Average % Yearly Salary Hike* compared to females, holding *Age When Hired* fixed.

However, an R^2 of 0.068 is far from satisfactory and upon further analysis, the team found that assumption 3 did not hold for the selected model (see Figure 3.1 & 3.2 in Appendix). Typically, when interpreting the value of R^2, the closer it is to 1.0 (which is the statistical roof for R^2), the better [1]. Normally, 0.4 is the cutoff in deciding whether a reasonable model has been developed or not. Since the R^2 value of the selected model is 0.068 and the predictive ability of the selected model is not consistent across the full range of the response variable, the team does not recommend the HR department to use this model for hiring decisions.

Conclusion

From the analysis, it was found that well-rated employees indeed earned better salary hikes at IBM. As rating depends on employee contribution to the organization, our assumption to consider *"Yearly Salary Hike"* as a response variable was validated. We observed that as the *"Age when hired"* increased, *"Average % Yearly Salary Hike"* increased as well. This corroborated our expectation that older people who take up senior positions in a firm are more likely to provide a greater rate of contribution to the firm's growth. Also, contrary to our belief of the level of education being a significant parameter behind an individual's contribution to the firm seems to be untrue for one case. We identified that employees who obtained a Bachelor's degree earned a higher rate of salary increase compared to employees with a Master's degree.

Using multivariate regression to determine a candidate's contribution to the firm based on existing employee data seems to be infeasible given the information at the team's disposal. In order to provide a credible tool for the HR department, the team will need to look into additional data sets to see if there are predictors available that can describe more of the variation within the observations of *Average % Yearly Salary Hike* that also will produce a model that satisfies the requirements of assumption 3.

Appendix

Data source:

https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

Figures:

Figure 1.2

	PercentSalaryHike														
PerformanceRating	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
3	156	139	162	150	73	54	69	71	58	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	44	35	45	18	12	14

Evidence for using *PercentSalaryhike* as an indicator of employee contribution to the company.

Figure 1.3

	yrly_salary_hike	firstday_age	prev_yrs_exp	num_previous_companies	education_level
yrly_salary_hike	1.00000000	0.2309108	0.16867039	0.0628747	-0.06889176
firstday_age	0.23091080	1.0000000	0.66692841	0.3571078	0.15196472
prev_yrs_exp	0.16867039	0.6669284	1.00000000	0.4052565	0.08753347
num_previous_companies	0.06287470	0.3571078	0.40525648	1.0000000	0.12372829
education_level	-0.06889176	0.1519647	0.08753347	0.1237283	1.00000000

Assessment of collinearity amongst predictors

- *prev_yrs_exp* & *firstday_age* = 0.667
- *prev_yrs_exp* & *num_previous_companies* = 0.405
- *num_previous_companies* & *firstday_age* = 0.357

Figure 2.1 - Model 1

```
Call:
lm(formula = yrly_salary_hike ~ Gender + firstday_age + EducationField +
    education_level + prev_yrs_exp + num_previous_companies)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1599 -1.7760 -0.9803  0.7545 18.0805

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.41596    0.85230   1.661 0.096932 .
GenderMale                     0.20459    0.19288   1.061 0.289045
firstday_age                   0.08593    0.01416   6.066 1.81e-09 ***
EducationFieldLife Sciences    0.15513    0.73458   0.211 0.832787
EducationFieldMarketing       -0.28149    0.77172  -0.365 0.715364
EducationFieldMedical          0.05248    0.73986   0.071 0.943463
EducationFieldOther            0.37330    0.82127   0.455 0.649532
EducationFieldTechnical Degree -0.14146    0.79081  -0.179 0.858065
education_level2              -0.76374    0.35745  -2.137 0.032850 *
education_level3              -0.66531    0.31967  -2.081 0.037646 *
education_level4              -1.24103    0.33439  -3.711 0.000217 ***
prev_yrs_exp                   0.01236    0.02142   0.577 0.563977
num_previous_companies        -0.02249    0.04399  -0.511 0.609160
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.127 on 1087 degrees of freedom
Multiple R-squared:  0.07086,   Adjusted R-squared:  0.0606
F-statistic: 6.908 on 12 and 1087 DF,  p-value: 3.698e-12
```

Figure 2.2 - Model 2

```
Call:
lm(formula = yrly_salary_hike ~ Gender + firstday_age + education_level +
    prev_yrs_exp + num_previous_companies)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0814 -1.7789 -0.9645  0.7166 18.0005

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.48782    0.45457   3.273 0.001097 **
GenderMale             0.20747    0.19237   1.078 0.281061
firstday_age           0.08501    0.01412   6.022 2.35e-09 ***
education_level2      -0.73040    0.35603  -2.052 0.040454 *
education_level3      -0.66476    0.31878  -2.085 0.037273 *
education_level4      -1.24206    0.33244  -3.736 0.000197 ***
prev_yrs_exp           0.01371    0.02137   0.642 0.521313
num_previous_companies -0.02232    0.04383  -0.509 0.610727
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.124 on 1092 degrees of freedom
Multiple R-squared:  0.06834,   Adjusted R-squared:  0.06237
F-statistic: 11.44 on 7 and 1092 DF,  p-value: 4.494e-14
```

The correlation between *prev_years_exp* and *firstday_age* is around +0.67 and the correlation between *prev_years_exp* and *num_previous_companies* is +0.41, indicating moderate collinearity amongst these predictor pairs. In order to address this, we first removed the *prev_years_exp* due to its +0.67 correlation with *firstday_age*. We then tested the updated model and found that there was no substantial change in the results, motivating the team to remove *previous years of experience* from the following model as well.

Figure 2.3 - Model 3

```
Call:
lm(formula = yrly_salary_hike ~ Gender + firstday_age + education_level)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0997 -1.7813 -0.9571  0.7039 18.0196

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.38673    0.41306   3.357 0.000814 ***
GenderMale        0.21087    0.19207   1.098 0.272497
firstday_age      0.08918    0.01048   8.509  < 2e-16 ***
education_level2 -0.73958    0.35554  -2.080 0.037744 *
education_level3 -0.66825    0.31836  -2.099 0.036041 *
education_level4 -1.26038    0.33102  -3.808 0.000148 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.122 on 1094 degrees of freedom
Multiple R-squared:  0.06788,   Adjusted R-squared:  0.06362
F-statistic: 15.93 on 5 and 1094 DF,  p-value: 3.548e-15
```

Figure 2.4 - Model 4

```
Call:
lm(formula = yrly_salary_hike ~ firstday_age + education_level)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2178 -1.7978 -0.9600  0.6696 18.1100

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.53070    0.39173   3.908  9.9e-05 ***
firstday_age      0.08886    0.01048   8.481  < 2e-16 ***
education_level2 -0.74799    0.35549  -2.104 0.035597 *
education_level3 -0.68438    0.31805  -2.152 0.031631 *
education_level4 -1.26467    0.33103  -3.820 0.000141 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.123 on 1095 degrees of freedom
Multiple R-squared:  0.06685,   Adjusted R-squared:  0.06344
F-statistic: 19.61 on 4 and 1095 DF,  p-value: 1.328e-15
```

Figure 2.5 - Model 5

```
Call:
lm(formula = yrly_salary_hike ~ firstday_age + firstday_age_sqd +
    education_level)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9713 -1.8001 -0.9489  0.6747 18.1226

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.327413   1.075181   0.305   0.7608
firstday_age      0.169437   0.067868   2.497   0.0127 *
firstday_age_sqd -0.001205   0.001003  -1.202   0.2297
education_level2 -0.782042   0.356548  -2.193   0.0285 *
education_level3 -0.709520   0.318671  -2.227   0.0262 *
education_level4 -1.305900   0.332734  -3.925 9.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.122 on 1094 degrees of freedom
Multiple R-squared:  0.06808,   Adjusted R-squared:  0.06382
F-statistic: 15.98 on 5 and 1094 DF,  p-value: 3.163e-15
```

Model 5 served as a test to see whether there was a curvilinear relationship between *firstday_age* and *yrly_salary_hike*. There is a 22% chance we would have seen these results if in fact there is no curvilinear relationship between *firstday_age* and *yrly_salary_hike.* Depending upon management’s acceptable level of error, the *firstday_age_sqd* may be a relevant term to include in the model.

Figure 3.1 - Standardized Residual Plot for Model 3

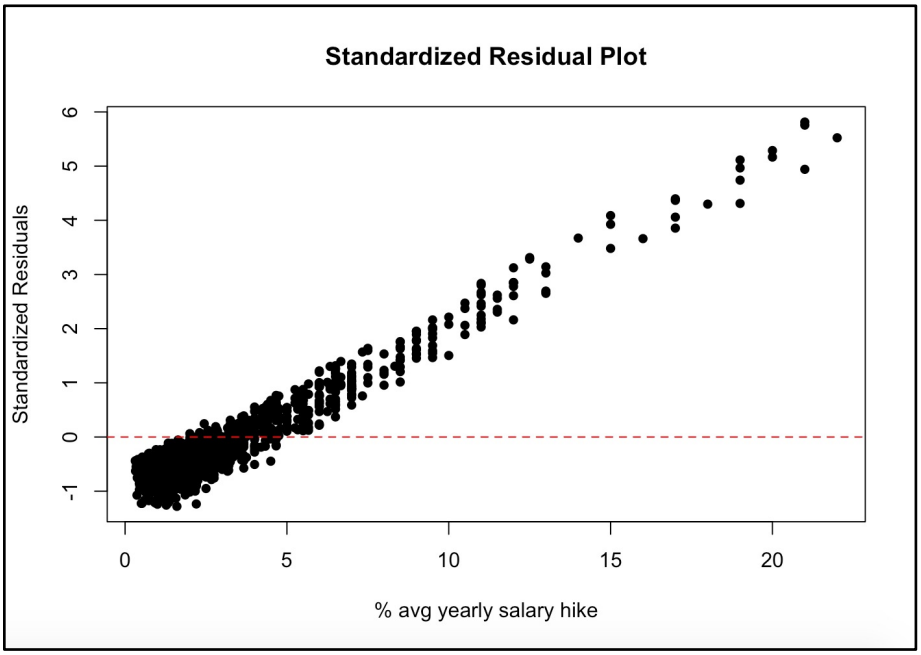
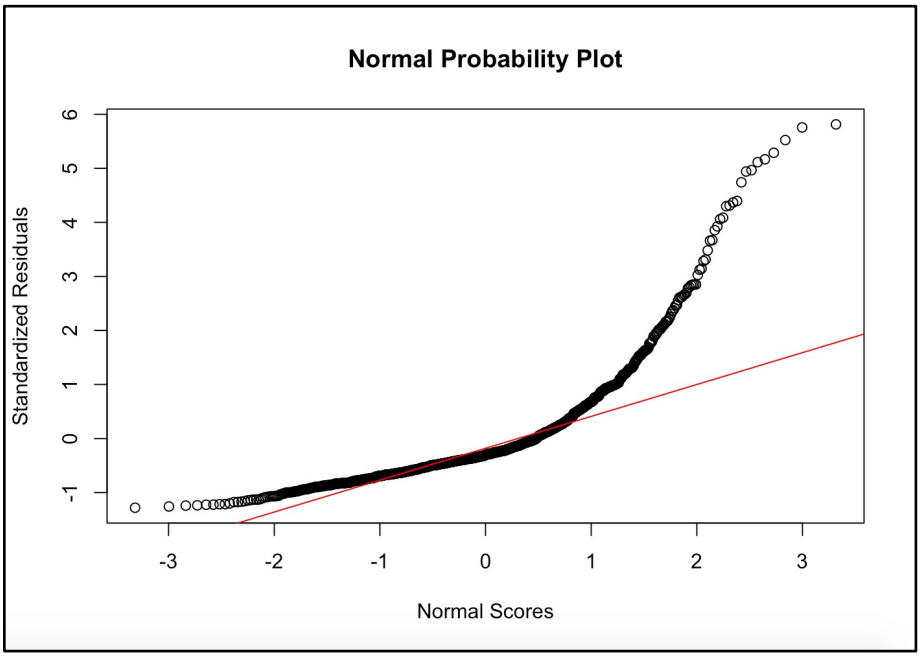


Figure 3.2 - QQ-Plot for Model 3



References

[1] Robert Nau, “What’s a good value for R-squared?”, *Duke University,* [Online], Available: https://people.duke.edu/~rnau/rsquared.htm. [Accessed August 17, 2019]