Study 2: Multiple Linear Regression Analysis

Introduction

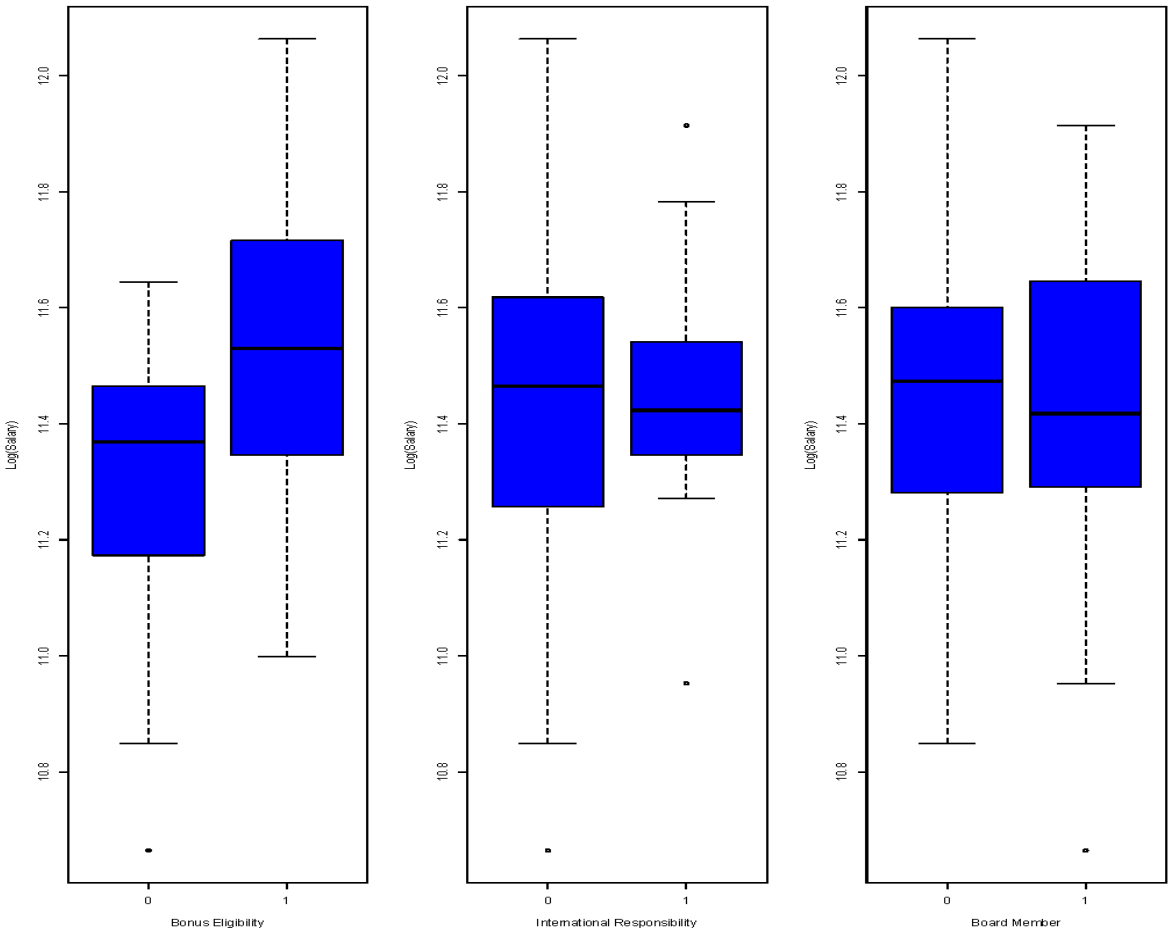
Understanding the factors influencing executive salaries is essential for consulting firms that wish to develop competitive and appropriate salaries for current and future executives. This can be explored further through an in-depth analysis, which will be done with a sample of 100 executives to predict the natural logarithm of salary, as it has better predictive power than just using salary as the response variable. The study considers ten potential predictor: experience (years), education (years), bonus eligibility indicator, number of employees supervised, corporate assets (USD millions), board member indicator, age (years), company profits in the part 12 months (USD millions), international responsibility indicator, and company's total sales in the past 12 months (USD millions). This analysis aims to determine which variables are significant in predicting salary and to create an optimized model.

Exploratory Data Analysis

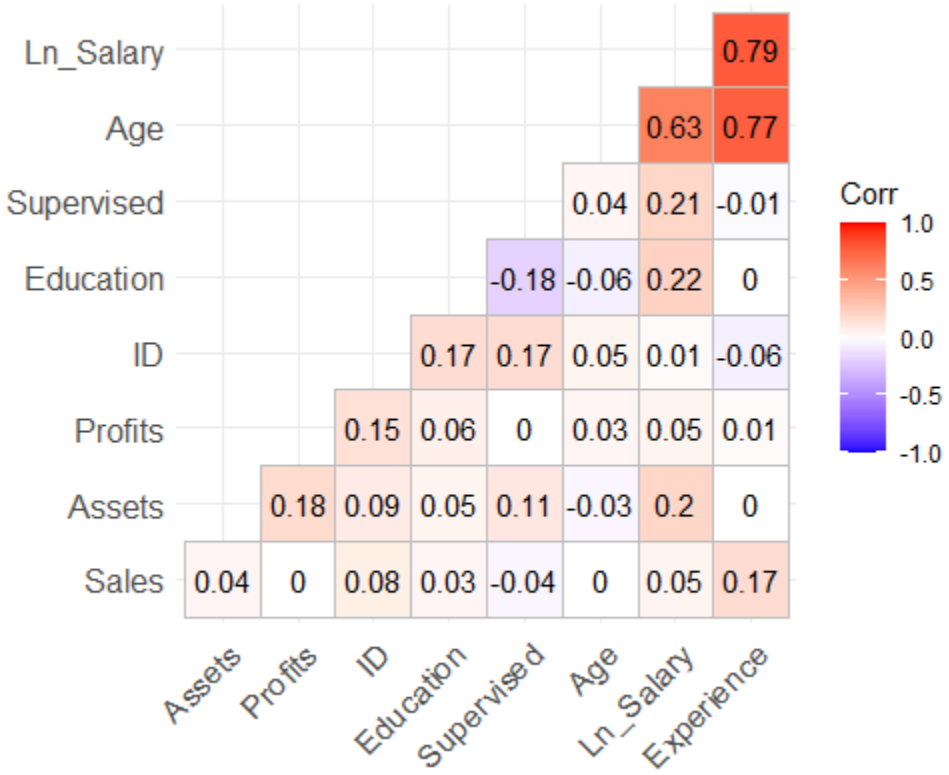
Before building our models, we analyzed the data we would be working with through exploratory data analysis. First, we identified our qualitative variables: bonus eligibility, board member indicator, and international responsibility indicator. These variables entail whether an executive was eligible for a bonus or not, if an executive was a board member or not, and if an executive had international responsibility or not. After this, we counted the occurrences of each variable. The eligibility for the bonus variable had **49** yes instances, meaning **49** executives were eligible for the bonus, while **51** weren't. The board member indicator had **49** yes instances, meaning **49** executives were board members, while **51** weren't. Lastly, the international responsibility variable had **18** yes instances, meaning **18** executives had international

responsibility, while **82** didn't. This helps us understand how much of an impact each of the variables may have on the data.

In addition, the box plots below visualize the relationship between log salary and these three qualitative variables. In the bonus eligibility vs. log salary plot, individuals eligible for bonuses have higher median log salaries than those who are not, and the interquartile range is also shifted higher. This suggests that bonus eligibility is associated with higher salaries. On the other hand, the plots showcasing international responsibility and board membership against salary demonstrate minimal differences in salary distribution between their respective groups. These initial findings indicate that bonus eligibility may be a meaningful salary predictor, while the other two variables may be less influential.



Next, we constructed a correlation matrix to understand the relationship between our variables, excluding the qualitative variables. The results showed that most of the variables have very weak or weak correlations with each other. The variables with the highest or strongest correlations were between age and salary, experience and age, and experience and salary. Age and salary have a strong positive relationship, which means that the older the executives are, the higher their salaries are. Experience and age also show a strong positive correlation, as executives have more experience as they age. Lastly, experience and salary have a strong positive correlation, which means that the more experience executives have, the higher their salary will be.



Multiple Linear Regression Model

To obtain a better understanding of what impacted our response variable, we needed to understand what role each predictor played in the overall salary of an employee. To do this, we took an algorithmic approach to ensure that the best model was fit for our data. This was done by creating a null model and a full model that used the following 3 types of stepwise regression methods: forward, backward, and bidirectional. After all algorithms completed execution, all results were the same. Each model had the same AIC and as we found the best fitted model would be one that included the following predictor variables: experience, education, bonus eligibility, number of employees supervised, corporate assets, if they're a board member, age, company profits, international responsibility, and company's total sales. After generating our linear model, the summary statistics are as follows.

```
Residuals:
      Min         1Q     Median         3Q        Max
-0.201219 -0.056016 -0.003581  0.053656  0.187251

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.9619345  0.1010567  98.578  < 2e-16 ***
Experience  0.0272762  0.0010293  26.501  < 2e-16 ***
Bonus1      0.2246932  0.0163503  13.742  < 2e-16 ***
Supervised  0.0005244  0.0000474  11.064  < 2e-16 ***
Education   0.0290921  0.0033367   8.719 9.71e-14 ***
Assets      0.0019623  0.0004972   3.947 0.000153 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07512 on 94 degrees of freedom
Multiple R-squared:  0.9206,    Adjusted R-squared:  0.9164
F-statistic: 218.1 on 5 and 94 DF,  p-value: < 2.2e-16
```

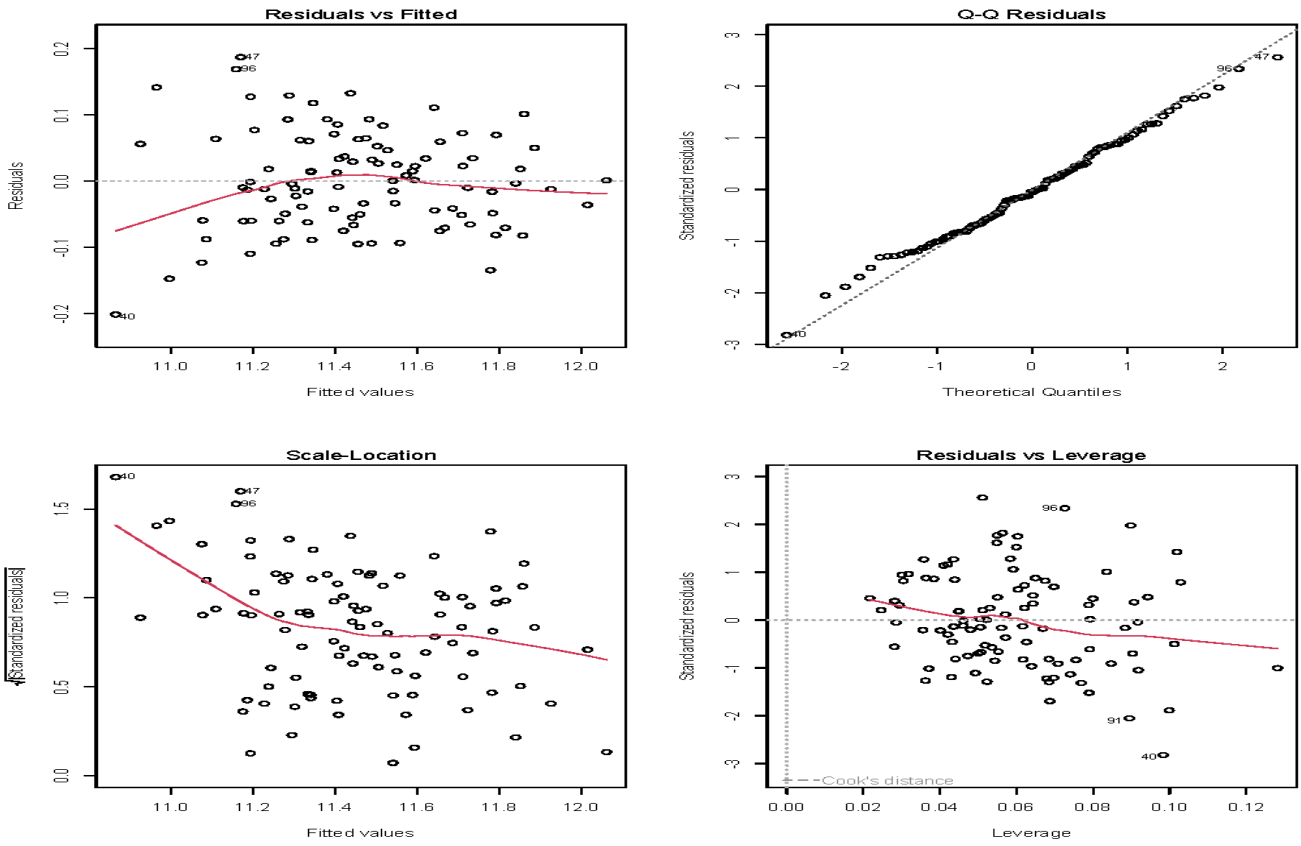
Through these findings, we concluded with the following regression equation:

$$ln(Y) = 9.9619 + 0.0273X_1 + 0.2247X_2 + 0.000524X_3 + 0.0291X_4 + 0.00196X_5$$

In this regression model, X1 represents **Experience**, measured in years. X2 is **Bonus1**, a binary variable that takes the value 1 if the employee is eligible for a bonus and 0 otherwise. X3 denotes **Supervised**, which is the number of people the employee supervises. X4 stands for

Education, which is the number of years of schooling. Lastly, X5 refers to **Assets**, representing the monetary value of assets the employee has provided to the company. Here, it is shown that the model has a **92.06**% R^2, which illustrates that the variables account for **92.06**% of the variation in log-salary, the Adjusted R^2 is also quite high at **91.64**% supporting the models account for variability. Therefore, this provides us with a strong model that gives an understanding of what may affect salary.

While the model demonstrated statistically significant coefficients and a strong explanation of the variability in the data, it is essential to verify that the key assumptions of multiple linear regression are met to ensure the validity. These assumptions include **linearity**, which requires a linear relationship between the independent variables and the dependent variable; **independence of errors**, meaning the residuals should not be correlated with each other; **homoscedasticity**, where the variance of the residuals remains constant across all levels of the independent variables; and **normality of residuals**, which assumes that the residuals are normally distributed. To check these assumptions, we used the following 4 plots depicted below.



The residuals vs. fitted plot allows us to verify the linearity and homoscedasticity assumptions. Although there is some curve at the bottom of fitted values, the residuals generally appear to be randomly distributed around the horizontal line at zero. This suggests that the relationship between the predictors and the response variable is approximately linear and that the residuals' variance stays fairly constant across fitted value levels. Some observations seem to differ marginally but do not seem to significantly affect the general trend.

The Q-Q residuals plot next evaluates residual normality. With just small tail deviations, the standardized residuals closely match the reference line. This implies that the residuals' normality assumption is fairly satisfied.

The scale-location plot looks more closely at homoscedasticity. It shows the fitted values' square root of the standardized residuals. This graph may suggest slight heteroscedasticity since it shows a small drop in the spread of residuals as fitted values rise. The trend is insufficient, though, and the general distribution of residuals seems fairly constant. Thus, the assumption of equal variance is largely acceptable.

The residuals vs. leverage graph shows significant data points that could skew the regression outcomes. There are no points with both high leverage and large residuals; all observations lie within the limits of the Cook's distance lines. This implies that the regression model is stable with respect to the data, and no one observation is driving the model.

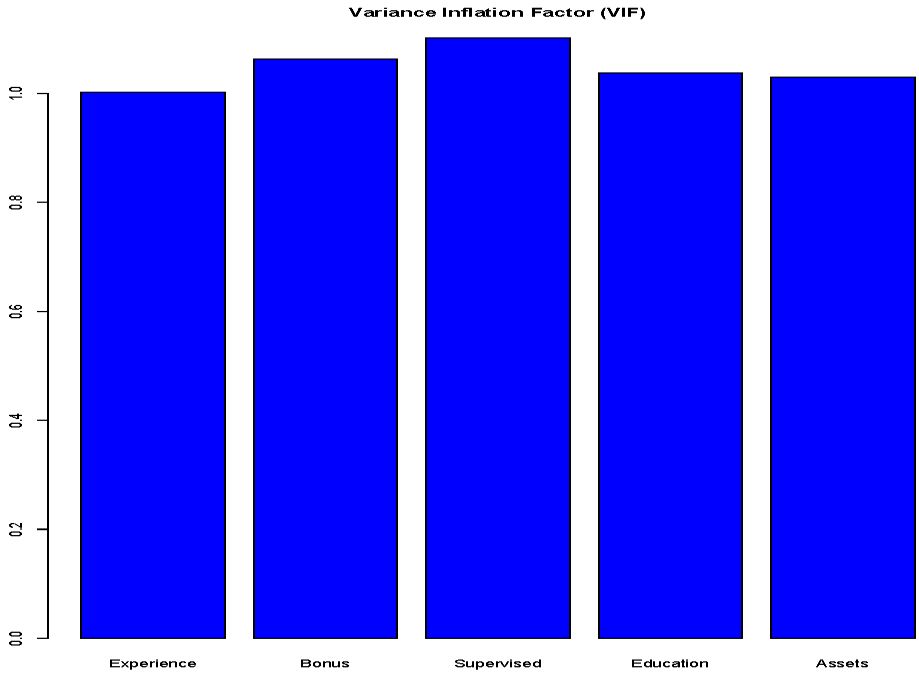
Finally, we ran a Variance Inflation Factor (VIF) study to look for multicollinearity among the predictors in our model. Two or more regression model predictors' high correlation with one another causes multicollinearity; thus, we employed VIF. We would like to avoid this since the consequences of individual predictors would overlap and so provide false relevance. A

VIF number of about one would indicate no multicollinearity; values over five would indicate strong multicollinearity.

VIF Values

Experience	Bonus	Supervised	Education	Assets
1.002071	1.063135	1.101590	1.037777	1.029408

After calculating each predictor variable, we identified that the values for these were all around 1, which indicates our predictor variables have no issues of multicollinearity. Therefore, this establishes that our model's variables are independent, which makes our model stable and the coefficients reliable. The following is a graphical representation of VIF for each predictor.



Conclusion

With all things considered, the method of choosing a model to forecast an individual's pay depending on important explanatory factors was effective. With the final model, all key assumptions of multiple linear regression linearity, independence, homoscedasticity, normality of residuals, and absence of multicollinearity were tested and met, pointing to the model's validity. Additionally, all coefficients were found to be statistically significant, indicating that each variable contributed to the prediction of salary.

The results of our study concluded that there is a strong and consistent relationship between an individual's salary and several important factors: years of experience, eligibility for a bonus, the number of individuals supervised, level of education, and the monetary value of assets contributed to the company. In conclusion, this model is incredibly useful for understanding the drivers of salary within any organization.