Sijie Xiang, Renee Djatmiko

Determinants of individual’s Salary in car industry from Indonesia

Eon 304 / Andre Switala

April 30, 2016

**INTRODUCTION**

In this paper we are going to investigate the determinants of an individual’s salary. We are mainly interested in the relationship between an individual’s length of service in the company and how it would affect their salary. We believe that there should be a positive correlation between the two variables. Our hypothesis is that when a person works in the same company for a longer time we expect that loyalty to a company would be rewarded with a higher salary, but most importantly the worker becomes more experienced and more productive, increasing the their value in the firm’s eyes and so lead to a higher salary. If we find that there is strong evidence that the length of service affects salary positively then it tells us that current employment laws that are enforced ensure that workers are treated fairly, as the increase in their productivity is correctly rewarded. If we find that the opposite is true then there is potential for government to employ certain policies to remedy this, perhaps laws that prevent firms from decreasing a worker’s salary. However, we also know that there are many other variables that determine salary, and we are going to include some of those variables in our analysis such as the individual's age, gender, and education attainment.

Our paper is a microanalysis paper; we looked at a random sample of workers in the Automobile Industry in Indonesia. Our final model actually indicates that the Length of Service is not significant (although we have a positive correlation), we find that the Years of Education to be the more significant factor which indicates that instead of focusing on labor laws government should try to improve the education sector and increase the number of years of schooling. We also find that there is gender discrimination in Salary. Conducting a Chow Test shows that the returns of Education, Length of Service, and Age (Working Experience) is not the same for Men and Women.

**DATA**

The raw data we are using is taken from Red and White Consulting, a general consulting firm from Indonesia who conducted a survey on the Automobile industry of Indonesia in 2014. The data samples are randomly selected individual workers from the Automobile industry. We are given the data on the individual's salary, length of service, age, gender, and education attainment. The data that are useful to our analysis would be the salary, length of service, job position rank, age, gender, and education attainment. Thus the variables that we are going to use in our analysis are:

➢ Salary:The raw data on salary is the monthly salary of the individual measured in Indonesian Rupiah (Rp). It is a discrete numerical data. This is our dependent variable in our regression. When we do a Box and Whisker Plot on our data we find that there six outliers. We will run the regression with our without the outliers to see if it makes a big difference, and decide whether to drop or keep the variables in our analysis.

➢ Length of Service**:** The raw data on length of service is measured in months. It is a numerical data. This is our main dependent variable in our regression.

➢ Age**:** The raw data on age is given as a string with the number of years and months stated. For example observation one is: “34(years) and 03(months)”. To work with this data we are going to convert it to a numerical data, we are going to convert it to years. So for observation one we change it to 34+(3/12) = 34.25. Basically we convert the months into a fraction of a year and added it to the number of years so that we have a continuous numerical data to work with. Although it might be simpler to round up or down, since it is questionable that fraction of years makes a difference, keeping the fractions makes our analysis more precise.

➢ Female: The raw data on Gender is categorical. We are given the data as a string, with only “Female” or “Male” as the input. This means that we could easily convert it to a dummy variable. We are going to make female=1 and male=0.

➢ Years of Education:The raw data is a string, it a categorical data listing the type of degree you can get in Indonesia. The types of degrees are: SLTA (High School), S1 (Undergraduate), S2 (Masters), D1 (Associate Degree), D2 (Young Associate Degree), D3(Professional Medium Science Degree), D4(Undergraduate but Specialized). We know the average year it takes to finish each degree so we created a quantitative variable that measures the supposed year of education for each individual worker.

**Summary Statistics:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Ob** | **Mean** | **Std, Dev** | **Min** | **Max** |
| **Salary** | 234 | 1.02 x 107 | 5.08 x 106 | 4.54 x 10^6 | 3.68 x 107 |
| **Length of Service** | 234 | 1.33 x 102 | 86.44 | 6 | 339 |
| **Age** | 234 | 37.95 | 8.07 | 24.25 | 58.17 |
| **Female** | 44(1), 190(0) | 0.19 | 0.39 | 0 | 1 |
| **Years of Education** | 234 | 14.59 | 1.88 | 12 | 18 |

**EMPIRICAL PART**

We are mainly interested in how length of service of the individual affect their salary, however we know that there are other factors that affect salary. Our economic model tries to include all these variables that can be observed, there are other variables such as cognitive abilities that we believe influence salary but we have to omit because a lack of access to such data.

**Basic Economic Model:**

We are doing a cross-sectional regression analysis on the individuals working in the Automobile Industry. The data we are using came from a survey that obtains the individual’s information for the same period in time.

➢ **Dependent Variable:** *Salary*

➢ **Main Independent Variable:** *Length of Service*

We expect that Salary would be Positively correlated with the Length of Service. We believe that as an individual stays in the company they become more efficient in their job and thus more productive and hence will be rewarded with a higher salary. We also think that employee loyalty would also be rewarded with a higher salary.

**Control Variables:**

➢ *Age:* We think that the older you are the higher your salary is because the longer you have worked in the industry/worked in the company, and so possibly the higher rank you are. The age variable would be positively correlated to length of service. With this variable omitted during our simple regression we think that there would be a positive bias in our coefficient.

➢ *Female*: We think that there is a difference between the salary of men and women. We think that female tend to be selected for lower pay jobs and hence lower salary. We are going to control for this possible discrimination in our analysis. We also think that it correlated to our main independent variable. Because female might take maternity leave then there is a chance that the length of service for female tend to be lower than for male.

➢ *Years of Education:* We expect that Salary and Years of Education are positively associated because on average more years education give people higher return in terms of salary. We think that lower educated individuals have lower skills and thus are more easily replaceable while higher educated individuals are more valued for more specialized skills. Therefore we think that years of education and the length of service is positively correlated because the more education you get the more skilled you are and hence the less likely you are to be fired and more likely to stay in the firm. Omitting this variable in the simple regression of Salary on Length of Service would likely lead to an upward bias.

**Simple Regression:**

**1.** **Simple Regression (level-level): Salary on Length of Service**

The simple regression gives a negative correlation between the Salary and Length of Service. Our current model is a simple level-level model. The R-squared is very low and our Adjusted R-squared gives a negative value, indicating that we need more variables to explain salary. Our p-value indicates that there is no statistical significance. The coefficient says that an increase in one month of service leads to a decrease in the monthly salary of the individual by Rp.2,700. The average salary is Rp.10, 200,000 that means that there is no economic significance for our coefficient.

Looking at the scatter plot (See Appendix) of the level-level model we seem to see an evidence of heteroskedasticity in our data since we seem to see a greater dispersion at the lower end of the Length of Service. We test for heteroskedasticity using the Breusch-Pagan test and our p-value indicates that our model’s error terms do not have constant variance. Looking at the scatter plot we seem to have a linear relationship between the error term’s variance and our explanatory variable, it seems that as the Length of Service increase the variance decreases. What is saying for our model is that the longer you are in the company, the less bargaining power you have in determining your salary. Even though it seems that we can identify the functional form for the variance of error term with respect to x (negative linear), it is difficult to adjust our regression to account for this. We will simply run the robust regression to adjust for the heteroskedasticity if we use the level-level model.

**2.** **Simple Regression (level-level): Salary on Length of Service (without outliers)**

We have stated how we have found six possible outliers in our data, we found that regressing our model with these outliers significantly change our interpretation of our model. Instead of getting a negative coefficient for our model we now have a positive coefficient, which makes more economic sense, and it also fits our hypothesis more closely. Thus it makes more sense to do subsequent regression without those outliers. However, we found that when we regress the sample with and without the outliers in a multi-regression our coefficient for Length of Service drops back to a negative value. This indicates that the outliers were not significant enough to drop.

The simple regression does not truly capture the effect of Length of Service on Salary, as it does not control for the other variables that could affect Salary. The variables we control for in our multi-regression is Age and Years of Education, we want to see the unique effect of Length of Service on Salary, so we run three regressions to try to identify it. First we regress Salary on Age and Years of Education to get our first residual. The first residual captures everything explains Salary that is not explained by Age and Years of Education. This should include Length of Service if it actually matters. For our second regression we regress Length of Service on Age and Years of Education to get our second residual. The second residual captures everything that explains Length of Service that is not explained by Age and Years of Education. For our third regression we regress our first residual on our second residual, and the R-squared value we obtain is 0.0885, which means that the unique effect of Length of Service on Salary is not that significant. This means that although there is no perfect collinearity, the variables we chose are highly correlated with each other, which explains how different our interpretation of our result became when we switch to multi-regression from the simple regression.

**Trying for Different Functional Forms:**

In the previous level-level model, we find nothing significant to explain the potential relationship between explanatory variable Month of Service and dependent variable Salary, and therefore we try other potential models such as log-level, level-log, and log-log in order to further explore possible explanations.

**3.** **Simple Regression (log-log): Log (Salary) on Log (Length of Service)**

This log-log simple regression model gives us a positive correlation between Log (Salary) and Log (Length of Service). The coefficient is 0.0115, meaning for every 1% increase in one month of service leads to approximately 0.0115% increase in individual’s salary. However, as P value is 0.735, which is much bigger than 0.05, explanatory variable Log (Length of Service) is statistically insignificant. R-squared is closed to 0; adjusted R-Squared is negative, indicating that the correlation is not strong and hence we should either add more explanatory variables or try a different functional form for our model. We also tested this model for heteroskedasticity and found that our p-value indicates that there is no significant evidence for heteroskedasticity, so we do not need to run it robust if we use this functional form.

**4.** **Simple Regression (log-level): Log (Salary) on Length of Service**

This log-level simple regression model gives us a negative correlation between Salary and Log (Length of Service). P value is bigger than 1%, 5%, 10% indicating explanatory variable Length of Service is statistically insignificant. As the variable is not statistically significant, it is meaningless to explore its economic significance. It is worth noting that compared to level-level model, however, level-log model has a different explanation in terms of coefficient. In this model, one more month of service leads to 0.00844% decrease in Salary. R-squared is low; adjusted R-Squared is negative, meaning we need to either test more variables or transform to another regression models. Using the Breusch-Pagan test we find that this model seems to exhibit homoscedasticity, and so we do not need to run it robust if we use this model.

**5.** **Simple Regression (level-log): Salary on Log (Length of Service)**

The level-log model gives a negative adjusted R-squared for and a low R-squared value. The p-value indicates no significance. We can conclude that this model is not the best functional form for our regression and that the other forms are better suited (given that it has the lowest R-squared value from the other functional forms model we tested). The coefficient of -656,000 indicates that one percent increase in length of service leads to a Rp. 656,000 decrease in Salary. Testing for heteroskedasticity we find significant evidence to say that this model’s error term does not have constant variance, and if use this model we should run it robust.

**Multi-Regression:**

We can either choose to use Length of Service as a level variable or as log variable. When we compare our level-level to our level-log, we find that our level-level model has a slightly better adjusted R-squared. When we compare log-level to log-log we find that log-log has a better adjusted R-squared. We cannot use the adjusted R-squared value to compare whether level-level or log-log is the best model since they are essentially different dependent variables so we need to use economic intuition to decide which model to use. We pick the level-level model because the interpretation is easier and makes more sense; a one more increase leads to a change in salary in numerical value. If we use log-log our interpretation will be in terms of percentage which is more difficult to interpret since we would say that 1% increase in month lead to a change salary in percentage terms, which is harder to understand.

**6.** **Multi-Regression: Salary on Length of Service, Age**

We believe that omitting Age variable would lead to an upward bias for the coefficient in the simple regression and so we added Age for multi-regression to eliminate that bias. The multi-regression has confirm our earlier belief that omitting the Age variable will give us a positive bias in our coefficient for our simple regression.mn Adding Age variable into our regression also gives us a better R-squared and Adjusted R-squared then in our simple level-level model regression. Now we actually have a p-value that indicates that our two explanatory variables are significant at the 1%. The coefficient for Length of Service has fallen (become more negative) and it’s 95% confidence interval is now completely in the negative range.

**7.** **Multi-Regression: Salary on Length of Service, Age, Years of Education**

Now we add another explanatory variable to our multi-regression. We added Years of Education, and our regressions shows that it is also significant; the p-value indicating it is significant at the 1%. Our adjusted R-squared has increased showing us that our new model is a better model in explaining the variation of Salary. That is all three of our explanatory variables are significant.

**Testing for Functional Form Misspecification:**

**Testing Against Non-nested Model**

Here we are trying to decide whether an independent variable should appear in level or logarithmic form for which cannot be simply tested by standard F test so that we need to employ Davidson- Mackinnon test to tackle this problem. The Davidson- Mackinnon test is obtained from the t statistics on in the auxiliary equations (1) and (2). The fitted value in equation (1) is insignificant and that of equation (2) is significant, which means we should favor level form model over log form model. (See Appendix)

**Reset as a General Test for Functional Form Misspecification**

We prefer level form model to log form model in light of Davidson-Mackinnon test. Now, we are trying to see if we need to add quadratic terms (or higher polynomials) in our original model by implementing RESET test. Although we cannot decide how many functions of fitted value to include in an expanded regression, squared, cubed and quad terms would mostly do. The P-value for F test is approximately zero, which is small enough for us to reject the null hypothesis in favor of alternative hypothesis, which means that the original model has omitted variables. However, RESET is not perfect because it does not detect heteroskedasticity so we also need to ensure that the higher polynomials added to the original model will not make it suffer heteroskedasticity.

**F- test for the higher polynomials**

We have run multiple regression adding different combinations of the polynomials variables into our model, and then we use the F-test to test for their joint significance. Our results shows that our model is not missing any higher polynomials and so suggest that we are in fact missing interaction terms.

**8. Adding Interaction terms into the model**

We tried to add the three possible interaction terms (Length of Service \* Age, Length of Service \* Education, Age \* Education), but our result shows that only Length of Service \* Age is significant at the 5%. However when we add the interaction term into our model we find that our Length of Service variable becomes insignificant which means Length of Service alone is not an important factor in determining Salary. The coefficient for the interaction term is -983, which means that when we take the first derivative of our regression in terms of Length of Service, Age puts a downward pressure on Salary. It means for the same Length of Service, one increase in Age leads to a decrease in returns of Length of Service by Rp.983 This means that the returns to Length of Service decreases as your Age increases. If we consider Age as the proxy for one’s working experience in the industry then we say that the older you are the more experience in the industry you possess so the more Salary you get i.e. it means that the starting salary for a person with more experience is higher. If we take the first derivative in terms of Age then we see that the returns of Age decreases as the Length of Service increases, this basically says that an employee who works in the company longer is rewarded for their service with a bigger salary. This is aligned with our hypothesis that the Length of Service increases with Salary, however according to our interaction term it follows marginal diminishing returns.

**Investigating Dummy Variable:**

**9.** **Multi-Regression: Salary, Length of Service, Age, Years of Education, Years of Education Squared, Female**

When we add the dummy variable Female into our regression we are investigating whether there is potential salary discrimination between men and women. The coefficient on the Female variable measures the average difference in Salary per month between men and women who have the same Length of Service, Age, and Years of education. The coefficient from our regression then says that women receive, on average Rp.409, 000 less than men per month. Note that we have controlled for the other variables so the salary differential cannot be explained by average length of service, age, and years of education, so the difference can only be explained by gender or the other factors associated with gender than we have not controlled for. But our dummy variable is not even significant so we cannot say that there is a difference between men and women.

**Chow Test: Test for differences in Salary between Women and Men**

In the earlier regression we find that there is no difference in intercepts for men and women, implying that there is no difference between the two groups, so we are going to re-verify our findings using the Chow Test. The Chow test allows us to test whether Women and Men’s Salary follow the same regression function (slope and intercept). We now try to see if there is a significance difference between the slope for Women and Men. Our null hypothesis is that there is no difference between the slope and the intercept. A same slope means that for each additional month of service, one-year increase in age, and one year of education; the change in salary for women and men increase the same way. While the intercept can be interpreted as the starting salary of an individual who has not worked in the company, is of zero age, and has no education. When we conduct the Chow Test we actually find that we reject the null hypothesis at the 5% significance level, which suggest that there is in fact salary discrimination between Men and Women, the difference between the two seems to be the slope. (See Appendix for Chow Test Calculation)

**LIMITATIONS**

There are many variables that we know affects Salary such as Natural Ability, Work Experience in the Industry, and Family Connections, etc.… We have omitted some of these variables because some of these are unobservable. Things like Natural Ability would be positively correlated to Salary because more talented individuals are most likely better workers, this would lead to an upward bias for most of our variables. We can use a proxy variable like IQ to measure Natural Ability if we conduct the research again, it will eliminate some of the bias. In our regression we have used the variable Age as a proxy for work experience, and although it seems like a good proxy variable, we could have actually asked the survey their actual Work Experience in the industry. But there might be an incentive for workers to overreport their working experience years. Also our variable Years of Education is in fact not the actual years of education each subject has but the expected years of education based on their degree, there is measurement error involved here because some people might finish earlier or later, so it is possible to ask survey their actual years of education instead. However, like the Work Experience there is incentive for subjects to overreport or underreport. We think that Family Connections also affect salary, as nepotism would mean that workers who are well connected would most likely get promoted to better jobs and earn higher salaries. For Family connections perhaps we can ask subjects whether they have a share in the company or a family member working in the company or in the industry, and then create a dummy variable that measures this. However, looking back at our data and our research perhaps we can try a panel data instead. Looking at the Change of Salary over time would also capture the effect of the unobservable variable like Natural Ability, assuming it does not change then its effect will be taken out when we do the difference-difference.

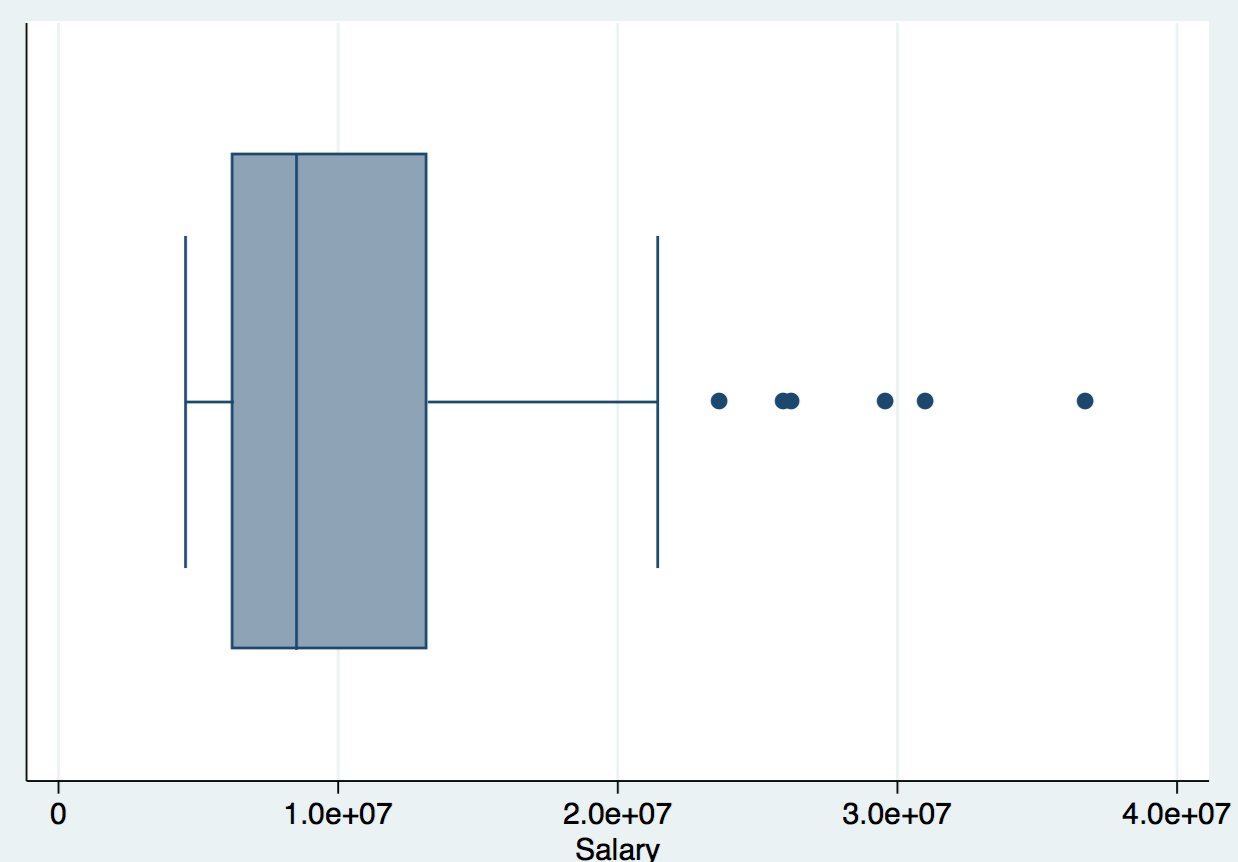
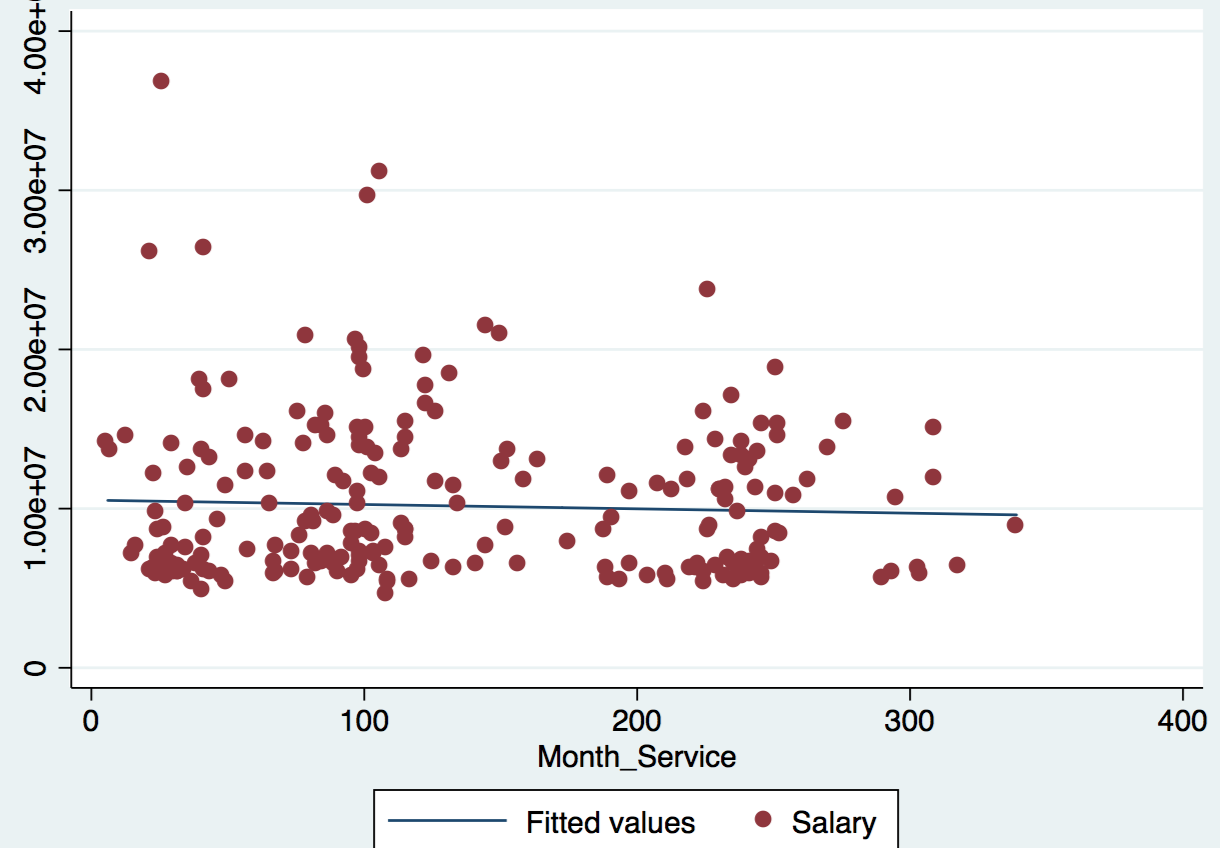
**CONCLUSION**

**10. Final Model:**

We have settled for this model as our final, we test for heteroskedasticity and find that our p-value is almost zero which means the model is heteroskedastic so we need to run our model robust. The interpretation for our model is that Age, Years of Education, and interaction term Length of Service \* Age is statistically significant at 5% whereas Female dummy and Length of Service becomes insignificant at 5%. However, Length of Service, Age, and Years of Education all have positive coefficients, indicating they these variables are positively correlated with Salary so an increase of these independent variables will lead to an increase of dependable variable. More specifically, Ceteris Paribus, one more year of education on average leads to an Rp.708, 000 increase in monthly salary; one more year of age, potentially more working experience would be rewarded with an Rp.552, 000 boost in salary. Our model supports our initial hypothesis.

**Table of Regression Result:**

**APPENDIX**

**1. Scatter Plot and The Box & Whisker Plot**

The box and whisker plot shows six extreme data points for Salary.

**2. Unique effect of Length of Service on Salary**

Regression 1:

Regression 2:

Regression 3: R-Square: 0.0885 Adjusted R-Squared: 0.0846

**3. Davidson- Mackinnon**



**4. Chow Test**

According to the data, SSR1 = 4.9367 \* 10 ^ 14 (When Female = 1)

SSR2 = 3.0431 \* 10 ^ 15 (When Female = 0)

SSRr = 3.7181 \* 10 ^ 15

F = 2.297, C = 3.02 \*\*\* / C = 2.21 \*\* / C = 1.85 \*