Jade Selke

Multivariate Analysis of Salaries in Tech Over Time

For years, business have frowned upon employees discussing salaries, even though doing so is not illegal. Transparency in the industry is important to ensure that you are not only making enough money to thrive in the area you live, but to also ensure that employees are compensated fairly for like work. Within this project I intend to analyze a sample set of survey data to find the major factors of salary over time, whether gender still plays a major role in salary determination in technology fields, and whether it is better to stay with a company over a long period of time or to find a new job which is commensurate with skill level to ensure a better raise.

**Reasoning**

Jack Kelly of Forbes reports that new studies show that the average worker can expect a 5-10% wage increase upon switching jobs, instead of staying at a present job, while keeping job security, will only net an average 2-4% raise year over year (2019). Most companies don’t make initiatives to incentivize employees to stay, and so many feel the urge to change jobs every few years in the current market. This practice, also called job hopping, is a prevalent practice in the technology fields, given how easy it can be to find work. This is especially true now, as Covid has forced many of these businesses to allow their employees to work from home, allowing job seekers to find work anywhere in the country without having to relocate.

Along with that, the US government predicted that there would be a shortage of at least one million STEM based jobs over a 10-year period from 2016 to 2026. This predicted shortage has led to a market in which employers must attract new employees with better and more diverse forms of benefits packages, along with new methods to generate more workers within these industries (Iammartino, R. et. al. 2016). This predicted shortage by the US government was only in STEM fields within state department-based jobs, and it is likely that the shortage extends into the private sector as well.

Finally, generating a larger workforce is more difficult due to under-representation of women and minority genders in STEM fields. According to Kahn and Ginther, persistent stereotyping, culture, competition, risk aversion, and other factors continue to contribute to the large gap in gender representation in STEM education, which leads to a continuing gap in diversity of representation in STEM fields (2017). This gap in representation may make it easier in the long run for underrepresented genders to find work easier, as companies seek to make their worker population more diverse. While this will make finding a job more competitive for women and underrepresented genders, it will also likely lead to higher salaries over time.

**Hypothesis Testing**

For this project, the following hypothesis is being tested:

Multiple Linear Regression on Salaries over Time

Null Hypothesis: Increases in salary over time in technology fields cannot be determined or attributed to any specific independent variables.

Alternative Hypothesis: Increases in salary over time in technology fields can be determined or attributed to one or more specific independent variables.

Given the Null Hypothesis, none of the independent variables can be shown to covary with the dependent variable (total yearly compensation), can be rejected if:

* One or more independent variables will show covariance with total yearly compensation and be statistically significant.
* One or more independent variables may not show covariance or not be statistically significant when combined with other independent variables.

This outcome will show which independent variables can be shown to have a covariance with total yearly compensation, as well as which ones have the strongest covariance with it, thus helping to show what factors have the greatest impact when trying to get the highest salary possible.

**Data Set Creation and Use**

In generating results for this project, an anonymous sample set of employee salaries in technology fields taken from levels.fyi are analyzed and further categorized for use into a specific modelling set. This set contains 62,643 total samples, but only samples containing proper gender information are kept, reducing the size of the set to 42,701.

The independent variables used are:

* company – a nominal discrete string representing the name of the company a person works for
* title – a nominal discrete string representing the job position the person holds at the company
* gender – a nominal discrete string representing the gender of the person, divided into male and female
* location – a nominal discrete string representing the city and state where the company is located
* cityid – a discrete integer representing a set of locations as a category identifier
* yearsofexperience – a float value representing the total number of years the person has worked in their technology field
* yearsatcompany – a float value representing the total number of years the person has worked at this company

The dependent variable is:

* totalyearlycompensation – a discrete integer representing the combined salary, benefits, and bonuses of the person for the last year, rounded down

The company, title, gender, and location independent variables were converted into category codes for processing in data frames for modelling. In addition, only totalyearlycompensations, which included salaries greater than $0 were considered, and only genders which were reported were considered. Unfortunately, only the genders of male and female were available in this set. The variables were all evaluated, and the standard deviations, mean, mode, and variance can all be seen in Figure 1.

Table

Description automatically generated

**Figure 1**, Mean, Mode, Standard Deviation, and Variance of Variables

**Methodology**

The goal for modeling is to build a multiple linear regression to understand the impact of company, title, gender, location, cityid, yearsofexperience, and yearsatcompany have on totalyearlycompensation. The model may show some of the independent variables have more impact than others, and only those variables which are shown to be statistically significant will be kept, using a significance value p = 0.05. Building this regression may show that salary is correlated with one or more of the independent variables and may exhibit much stronger correlations to some than others.

**Simple Linear Regression Results**

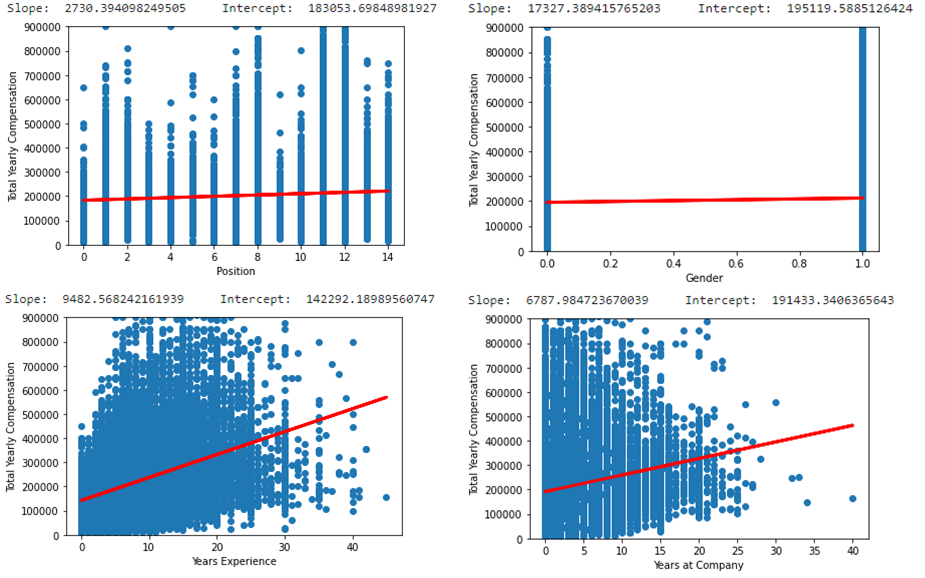
Simple Linear Regressions were run against each of the independent variables compared with totalyearlycompensation to get a better understanding of their potential individual affect. For these individual samples, company, location and cityid were shown to have the smallest effect on totalyearlycompensation, having slopes of -$13, $107, and -$2 respectively as shown in Figure 2.

Chart, scatter chart

Description automatically generated

**Figure 2**, Simple Linear Regression of Total Yearly Compensation compared with Company, Location, and CityID

The independent variable position had a greater impact, about $2,730 per year, on total yearly compensation, and sits in the middle of the slopes for the simple linear regressions, as shown in Figure 3. Yearsofexperience has the greatest effect on totalyearlycompensation, accounting for a slope of about $9,483. Interestingly, though, yearsatcompany had a slope of $6,788, which means that staying at a company may reduce salary when compared with yearsofexperience. Finally, gender had the greatest impact on totalyearlycompensation, with a slope of $17,327, which suggests that simply being male means a massively higher compensation over time.



**Figure 3**, Simple Linear Regression of Total Yearly Compensation compared with Position, Gender, Years Experience, and Years at Company

**Multiple Linear Regression Results**

After performing a multiple linear regression for the independent variables against the dependent variable, all the independent variables were found to have statistical significance because they had p-values below 0.05, as seen in Figure 4. Since all were found to be statistically significant, all independent variables were kept. An analysis of variance (ANOVA) was also conducted, which reported absolute F Values over 10 and as high as 205, shown in Figure 5. Finally, a variable inflation factor (VIF) test was run, and all independent variables were found to have factors under 2, seen in Table 1. This allowed for a coefficient equation to be built as:

totalyearlycompensation = $94,735 - $14\*company + $1,347\*title + $8,037\*gender + $104\*location - $2\*cityid + $10,203\*yearsofexperience - $2,805\*yearsatcompany

Table

Description automatically generated

**Figure 4**, Python OLS Regression Results

A picture containing table

Description automatically generated

**Figure 5**, ANOVA Model for Multivariate Analysis Results

|  |  |
| --- | --- |
| **Variable** | **Variable Inflation Factor** |
| company | 1.001698 |
| title | 1.018020 |
| gender | 1.018877 |
| location | 1.005509 |
| cityid | 1.003424 |
| yearsofexperience | 1.386071 |
| yearsatcompany | 1.378561 |

**Table 1**, Variable Inflation Factor Test Results

**Conclusion**

Based on the Multiple Linear Regression Model, the Independent Variables of gender, years of experience, and years at company have the greatest impact on the overall expected total yearly compensation of an employee in technology-based fields, while location seems to have the least impact. Every independent variable used was shown to have a p-value of near 0, well below a threshold of 0.05 which shows they all have statistical significance. The ANOVA model showed a strong chance for collinearity between independent variables, which was expected. However, using a VIF test showed that all independent variables had a VIF near 1, meaning there is actually very little chance of correlation between independent variables. Finally, because the expected slope of total yearly compensation was non-zero, even with all possible standard deviation combinations added to it, thus the null hypothesis should be rejected. In other words, increases in salary over time in technology fields can be determined or attributed to specific independent variables.

References

Iammartino, R., Bischoff, J., Willy, C., & Shapiro, P. (2016, April 19). *Emergence in the U.S. science, Technology, engineering, and Mathematics (STEM) workforce: An agent-based model of worker attrition and group size in high-density stem organizations - complex & intelligent systems*. SpringerLink. Retrieved December 11, 2021, from https://link.springer.com/article/10.1007/s40747-016-0015-7.

Kahn, S. & Ginther, D. (2017, June). *WOMEN AND STEM*. NBERG Working Paper Series. National Bureau of Economic Research. Retrieved December 10, 2021, from https://www.nber.org/system/files/working\_papers/w23525/w23525.pdf.

Kelly, J. (2019, July 26). *A New Study Concludes That It Literally Pays To Switch Jobs Right Now*. Forbes. Retrieved December 10, 2021, from https://www.forbes.com/sites/jackkelly/2019/07/26/a-new-study-concludes-that-it-literally-pays-to-switch-jobs-right-now/?sh=7ecbbcd05959.

Ogozaly, J. (2021, October). *Data Science and STEM Salaries*. kaggle. Retrieved December 1, 2021, from https://www.kaggle.com/jackogozaly/data-science-and-stem-salaries.