**APPLIED BUSINESS ANALYTICS SAS**

JOB POSTINGS AND FORTUNE 1000 COMPANIES

By

KALYAN PRANEETH PERALA

**Introduction:**

The theme of this project is job postings in indeed and fortune 1000 companies. The datasets that are used in the project are indeed job posting for data scientist, data analyst and data engineer positions which was then combined with the dataset that corresponds to fortune 1000 companies. The indeed job posting dataset is extracted from the data gov and data world and was further cleansed for our analysis. The second dataset is extracted from the fortune 1000 US companies is extracted from various sites including fortune 1000 information websites.

As a team, we agreed to run analysis on these datasets because these datasets are quite relevant to us and this course, because each of us aspire to land in a job in a fortune 1000 company or at the least in a data analyst related position. The datasets variables such as skills required, and salary sparked more interest. We were also interested in the assets the revenue, the number of employees and locations of these fortune 1000 US companies. In compliance, with the questions that we had as a team, we conducted various analysis and analyzed the results, we have interpreted the results for these questions.

The indeed job posting dataset has columns such as job title, queried salary, number of skills and company. The dataset also has information regarding the number of reviews and the number of stars for that posting, the location of the company and the industry to which the company belongs. Each row that is each level of analysis in this dataset is about a job posting.

The second dataset that is considered is the fortune 1000 companies' dataset. The dataset corresponds to the fortune 1000 US companies, here the level of analysis is a company. The columns in this dataset includes current rank and previous rank of the company, the revenue, profits, assets, market value, number of employees, CEO's name and title.

The dataset also contains information regarding the sector and industry these companies belong to and the city, state and location details including the latitude and longitude. The two datasets that is the fortune 1000 US dataset and the indeed dataset were combined based on the company name, that is the company name is used as the primary key. The analysis and the inferences in the report are extracted from these separate datasets as well as the combination of the two datasets.

**Data Cleansing and Enhancement:**

The datasets had many empty cells, and these were cleaned. To run the analysis, we also enhanced the dataset, in the fortune 1000 US companies' dataset we added the gender of the chief executive officer. In order to enhance the data and make it simpler to run the analysis we created segregated datasets, using the job type. The job type in the dataset is a categorical variable with three categories., we segregated the three categories made it into three variables with binary categories. In order to make the analysis more efficient, the indeed dataset has a column on salary that is in various ranges. We converted these values in ranges into six different levels and use that in salary levels excel file for analysis. The following are the salary levels: <80000 is considered as level 1, 80000-99999 is level 2, 100000-119999 is level 3, 120000-139999 is level 4, 140000-159999 is level 5 and the last one level 6 is >160000.

While combining the two datasets we used the company name as the primary key to join these two tables, we used the SQL queries to combine these two tables. Redundant information such as the latitude and longitude of the company location in the data sets was eliminated. Some columns such as the name of the CEO in the fortune1000 Us companies were deleted.We combined the two datasets by adding an additional column titled “Is Fortune”. The “Is Fortune” column is a binary categorical, if the value is 1, then the company is a part of the fortune 1000, if the value is 0, then the company is not a part of the fortune 1000.

**QUESTIONS, RESULTS AND INTERPRETATIONS:**

**Question 1:** Which job type has more posting in fortune 1000 companies?

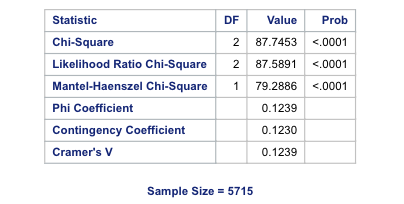
Data Set: Indeed and Fortune 1000 Combined

Test: Chi-Square Test of Independence.

Row Variable: Is fortune

Column Variables: Job Type (Data Analyst, Data Engineer, Data Scientist)

Χ calc 2 = Σrf=1Σck=1[ (ff,k – ef,k)2 / ef,k ]



**Interpretation of results:**

Setting up the hypothesis:

*H*0: Variables are independent  
 *H*1: Variables are dependent

Chi square value = 87.7453

The P value is <.0001 which is less than the level of the significance 0.05. We conclude that we reject the null hypothesis and conclude that the job postings are dependent on fortune 1000 companies.

The job postings of fortune 1000 companies depend on job type, and looking at the bar graph from the result in the appendix, we can state that data science role is most common job role for fortune 1000 and data analyst is most common role in non-fortune company.

**Question 2:** Which location has more postings of different job types?

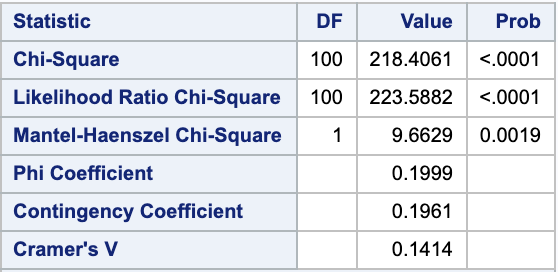
Data Set: Indeed and Fortune 1000 Combined

Test: Chi-Square Test of Independence.

Row Variable: Job Type (Data Analyst, Data Engineer, Data Scientist)

Column Variables: Location

Χ calc 2 = Σrf=1Σck=1[ (ff,k – ef,k)2 / ef,k ]



**Interpretation of results:**

Setting up the hypothesis:

*H*0: Variables are independent  
 *H*1: Variables are dependent

Chi square value = 218.4061

The P value is <.0001 which is less than the level of the significance 0.05. We conclude that we reject the null hypothesis and conclude that the variables are dependent on job type.

The job postings of different job types depend on location, and looking at the bar graph from the result in the appendix, we can state that data science, data engineer and data analyst job postings are more in CA, NY, VA, TX

**Question 3:** Is there any significant difference in salary level for each job type? Can we infer something information from that?

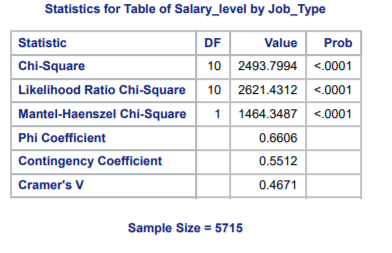
Test: Chi-Square Test of Independence

Dataset Used: Indeed

Row Variable: Job Type (Data Analyst, Data Engineer, Data Scientist)

Column Variables: Job levels (1,2,3,4,5,6) 6- being the highest

Χ calc 2 = Σrf=1Σck=1[ (ff,k – ef,k)2 / ef,k ]

‘

**Interpretation of results:**

Setting up the hypothesis:

*H*0: Variables are independent  
 *H*1: Variables are dependent

Chi square value = 2493.79

The P value is <.0001 which is less than the level of the significance 0.05. We conclude that we reject the null hypothesis and conclude that the variables are dependent.

The job postings of different salary levels depend on job type, and looking at the bar graph from the result in the appendix, we can state that data scientist posting are more common on the higher level of salary while data analyst position are more common in the lower range of salary level.

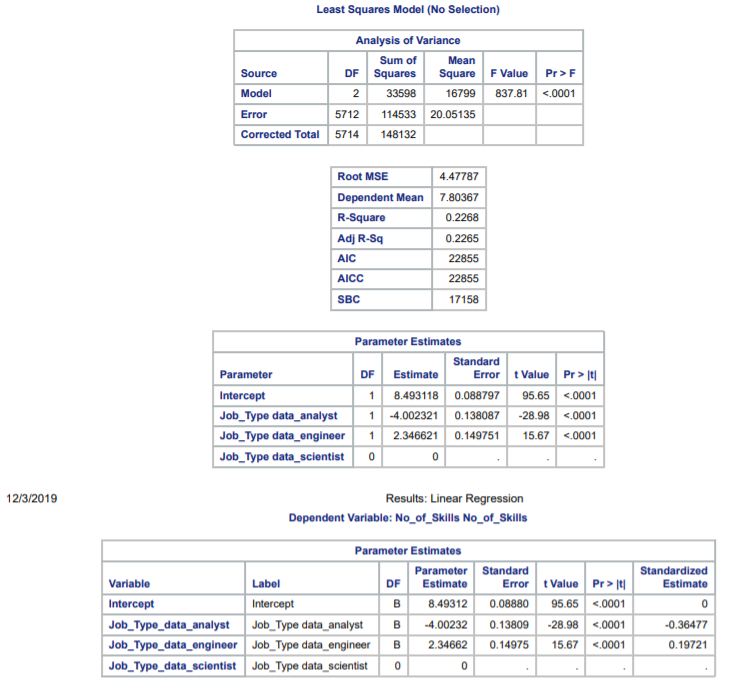
**Questions 4**: Is there any significant different in number of skills required for each job type?

Test: Linear Regression

Dataset Used: Indeed

Dependent Variable: No\_of\_Skills

Classification Variables: Job Type (Data Analyst, Data Engineer, Data Scientist)



**Interpretation of results:**

Setting up the hypothesis:

*H*0: Job type does not predict number of skills  
 *H*1: Job type predicts number of skills

F value = 837.8

The P value is <.0001 which is less than the level of the significance 0.05, we can conclude that the model is significant.

Looking at the R square value, we can say that 22.6% of variance in number of skills is predicted by the model.

From the ANNOVA table parameter estimates p-value which is also less than 0.05, we conclude that we reject the null hypothesis and conclude that job type predicts number of skills.

Also, we can say that data analyst job requires 4 (approx.) time less skills as compared to data scientist job positions, and data engineer job requires 2.3 time more skills as compared to data engineer.

**Question 5:** Does knowing SAS effect the chances of being qualified for data scientist job?

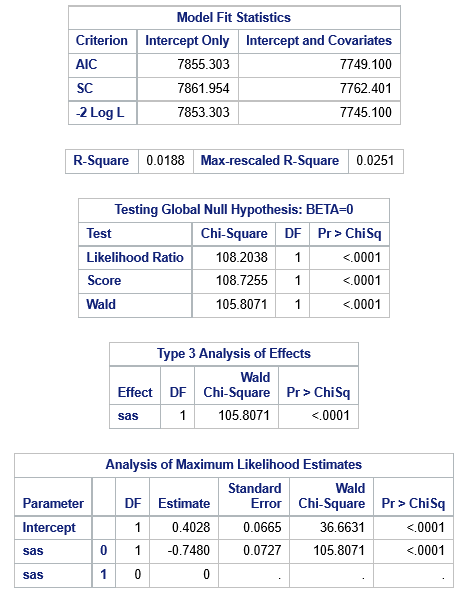
Analysis: Binary logistic regression

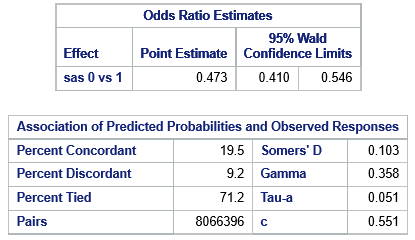
Data set used: Indeed and fortune combined

Y - Data scientist

X - SAS skill

Results:





**Interpretation of the result:**

The Wald’s Chi-Square value for the variable is 105.8071 and the p-value is <.0001

* The –2Log L values for only intercepts and intercepts with covariates is decreasing citing that the model is a good fit.
* From the p-value of Wald Chi-square we can conclude that the model is significant.
* The R square range for the model is 1.8 to 2.5%, that is the percent variance explained by the model.
* From the p-value in type 3 Analysis table, we can conclude that the skill “SAS” has a significant effect on the chances of being qualified for a data scientist job.
* From the Odds ratio estimates table, we can infer that knowing SAS increases your chances for Data scientist by 52.7%
* From the c-value, we can say 55.1% of rows in the current data are correctly predicted by the model

**Question 6:** Machine Learning influence on job posting for data scientist who already knows SAS.

Test: Moderation (binary logistic regression)

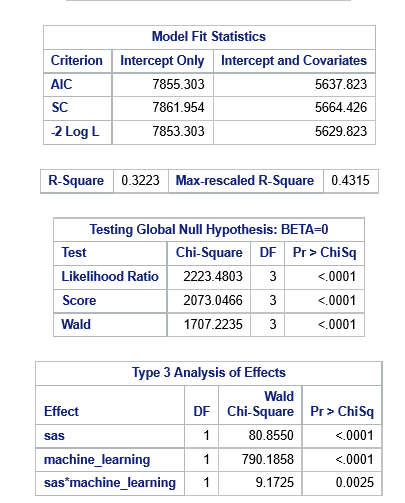
Data set used: Indeed and fortune combined

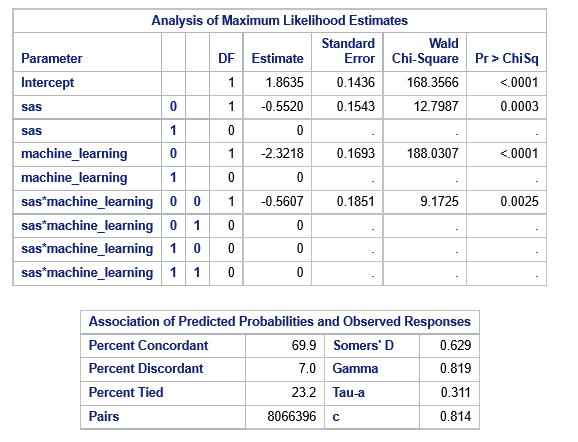
Y - Data scientist

X - SAS skill

M (moderator) - Machine Learning skill

* Results after adding Machine Learning as moderator





Y = b0 + b1 X + b2 M + b3 XM + e

Y = 1.8 – 0.5 X – 2.3M – 0.5 XM + e

**Interpretation of results:**

Hypothesis:

H0 : There is no interaction effect

H1: There is interaction effect

The p-value for the product of sas and machine\_learning is 0.0025

* The –2Log L values for only intercepts and intercepts with covariates is decreasing citing that the model is a good fit.
* From the p-value of Wald Chi-square we can conclude that the model is significant.
* The R square range for the model has significantly increased to 32.2-43.1% after the moderator is added, that means that variance explained by this model has increased after adding machine learning as moderator
* From the p-value of Type 3 analysis table, we can reject the null hypothesis and conclude that there is an interaction effect between the variables. Thus, inferring that the skill “Machine Learning” increases chances of a job posting for data scientist with existing knowledge of SAS.
* From the c-value we can say that 81.4% of rows in current data are correctly predicted by the model