Steven Francis

Categorical Data Analysis

Take-Home Quiz

Spring 2018

Objective

The objective of this analysis to the help the company (Company X) understand which employees are the most satisfied. Since Company X is a national business compromised of people of different gender, age, and race, the data generated from an employee survey was broken into different subgroups:

|  |  |  |  |
| --- | --- | --- | --- |
| **REGION** | **GENDER** | **AGE** | **RACE** |
| 1 = Northeast | Male | Younger than 35 | White |
| 2 = Mid-Atlantic | Female | 35 – 44 | Other |
| 3 = Southern |  | Older than 44 |  |
| 4 = Midwest |  |  |  |
| 5 = Northwest |  |  |  |
| 6 = Southwest |  |  |  |
| 7 = Pacific |  |  |  |

Even though having a 100% satisfaction rate would be ideal, it is unrealistic. However there is nothing wrong with wanting to improve satisfaction rates. In order to implement effective changes, one must examine the data to understand which groups are currently the most satisfied and which groups satisfaction rates could use a boost.

Why would a company want to conduct this type of analysis?

Company X is in the business of making money! Therefore, the reason this type of study would be beneficial for Company X is because satisfied employees tend to be productive employees. Since it has already been established that Company X does not have a 100% satisfaction rate (although it was previously established as unrealistic, Company X wouldn’t waste money on a study like this one if all seemed to be well), Company X should conduct research on different morale building tasks and exercises (such as a company Happy Hour or a company picnic) to increase satisfaction. In order to infer which groups would benefit most from morale building actions, one should first look at the raw data to gain a sense of what is going on within Company X.

Breakdown of Raw Data

The following graphs are the satisfaction rates (as of today, with no morale building actions implemented) represented in from the data by region, gender, age and race:

This graph is based off the simple calculation of taking the sum of all satisfied employees from each region (regardless of gender, age, and race) and divided by the total number of employees from each region. It seems as if the Mid-Atlantic region has the most satisfied employees, and employees from the Northeast region are the least satisfied.

These two graphs have also been generated by very simple calculations. The graph on the left has been calculated by taking the sum of **all satisfied employees** **within each age group** and dividing those numbers by the **total number of employees within each age group**. As one is able to see, the most satisfied age group is the over 44 age group and the younger than 35 age group is the least satisfied. Likewise, the graph on the right has been generated by taking the sum of **all satisfied employees that fall within the white and other race categories** and divided those numbers by the **total number of employees that respectively fall within the aforementioned race categories.** It appears that white employees are more satisfied than the employees that fall under the “other” race group.

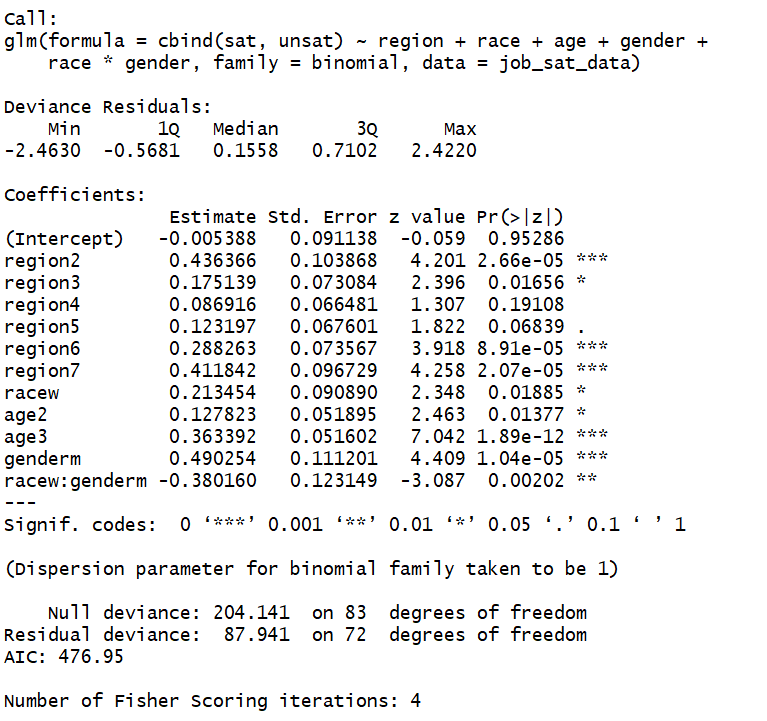
This graph has also been generated by using very simple calculations. It was calculated by taking the number of **all satisfied females** and **males** and dividing those numbers by the **total number of females** and **males** respectively. It is apparent that male employees are more satisfied than female employees.

Statistical Modeling

In order to gain more inference from the data, building a statistical regression model can be very influential. Since the point of the survey is to figure out which group of employees are could be more satisfied, a regression model can give further insight because it is built from estimations generated from the data. Because there are many different relationships one could investigate through regression modeling, the goal is to choose one that provides good enough estimations to measure how the variables included in the model are numerically related by one another. The manner in which the model was built was by using a variation of the backwards elimination stepwise regression analysis which begins by looking at every possible combination of variables and eliminating variable interactions that are deemed to be statistically insignificant in order to create a statistically significant regression model. [[1]](#footnote-1)

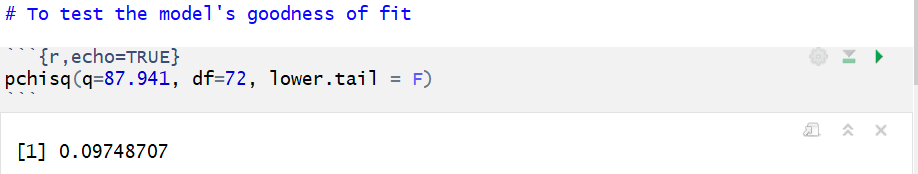
Since the purpose of this analysis is to determine whether particular groups are satisfied or not, creating a regression model with the response terms being of binomial nature would be better served by using a logistic regression model with a response variable () indicating a satisfied employee as a success and a dissatisfied employee as a failure. After completing the aforementioned elimination model, the final model included the following variables and interaction:

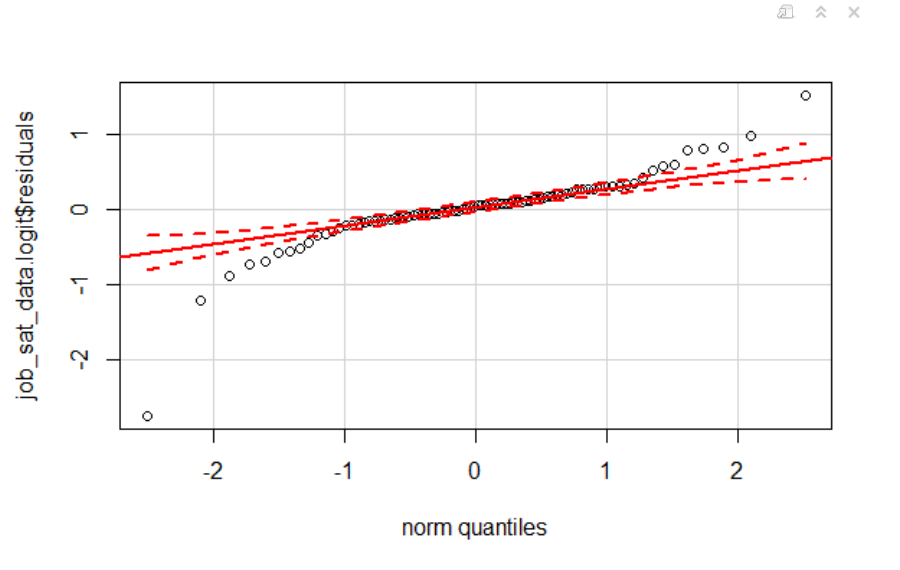
Each time a variable is added to the model, it adds to the structure of the model meaning that the information each variable possesses is incorporated into the model. Therefore the addition of each main effect should be intuitive. However, the race\*gender interaction may not be so instinctive. It means that when we hold gender constant, it provides two updated versions of the regression model for both race groups and while simultaneously holding race constant, it provides two more updated versions of the regression model therefore creating four equivalent numerical equations to represent the data. What this does is provide the model with more of “explanatory” power. The following is the output summary of the built model:



The way to read this output is by understanding that when an employee falls within a certain classification, the corresponding estimated coefficient is used to calculate a probability that the particular employee is satisfied. Another very important note is interpreting the coefficients is that since this is a logistic regression model, one must exponentiate both sides of the equation to rid the model of the logarithms.

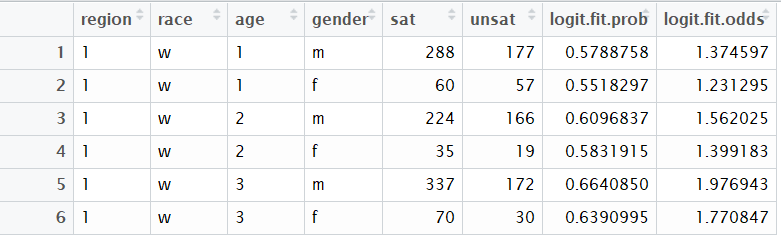
Even though, on an individual level, not all of the variables would be considered statistically significant (could be determined a number of ways but the most common is by looking at the last column), one could show that the model itself is statistically significant by obtaining a probability value of the residuals of the model as well as by showing that the residuals somewhat follow the behavior of a standard normal variable.

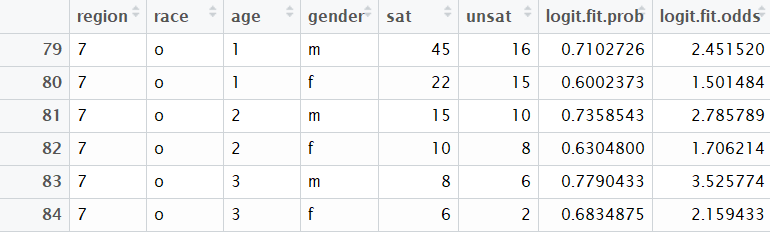


Since this probability value is above the 1 – α probability threshold, the logistic regression model is statistically significant.

Example of a QQ-Plot that tests the normality of the models residuals. Since most of the fitted data falls between the two dotted lines and along the solid line, the residuals of the model somewhat follow a standard normal distribution.

Since the model passes the goodness of fit test and its residuals behave in a somewhat standard normal behavior, one could use the fitted probabilities from the model to calculate the fitted log odds of each variable. A snapshot of a table containing the log fitted probabilities and the fitted log odds is below.





Having a table such as this one allows one to make statistical inferences (basically a deeper analysis of what was explained by the raw data alone), based off of the logistic regression model, pertaining to region, gender, age, and race. For example, the logistic odds calculation allows one to see that according to the logistic regression model, the employee classification with the lowest satisfaction log odds are from non-white females that are younger than 35 that work in region 1 (Northeast). The corresponding value is 0.9946267 indicates that the log odds of satisfaction for females from this group are 0.99 times that of dissatisfied females of the same aforementioned group. Conversely, the highest satisfaction log odds from this data set were from non-white male employees, from region 2 (Mid-Atlantic) and of the over 44 age group. [[2]](#footnote-2)

Conclusion

What the model and the data tells is that younger employees are the least satisfied within Company X. Also females are less satisfied than males. Another important indicator of satisfaction is region as employees from region 2 tend to be happier than employees from all other regions. Finally, non-white employees tend to be more satisfied than white employees. Company X should look for morale building exercises that cater to the younger employees and in particular the young females. If Company X does this, productivity would increase and profits would soon follow.

1. Statistical significance will be tested on the 95% level, α = 0.95, therefore 1 – α = 0.05 [↑](#footnote-ref-1)
2. For the table in which I used to determine the fitted log odds, please refer the back of the packet [↑](#footnote-ref-2)