Ariel Reyes

BAT 3305-1

February 1, 2016

NYC Tax Fraud Case Study

**Problem**

Since taxes are based on self-reported profits, it is difficult to determine whether a business has committed tax fraud or not. The current method for catching those committing tax fraud is for ten investigators to randomly choose a company and investigate its records. With only a limited amount of information and a limited amount of personnel, finding a better solution will require some creativity. The goal is to find a method that would increase the chances of successfully finding people committing fraud without hiring more investigators and using information that is readily available.

**Dataset**

The datasets used were sets of numbers representing the taxes paid, trash bills paid, and wastewater bills paid of 51 businesses. These sets of data were used because the city has access to these records, since the city is responsible for wastewater and trash disposal. This dataset is assumed to have no outliers, that is, no company in this dataset is committing fraud. It is used for the purposes of finding what should be normal in terms of how much taxes a company should pay. Once a relationship is established using this dataset, we will test the found method to a new dataset that does include outliers. One limitation that may be considered is the size of original dataset. If given a larger amount of companies to work with, we could develop a more accurate model.

When looking at the descriptive statistics for this data, we must first note whether the distribution is considered normal. We analyzed the kurtosis and skewness of each variable. For tax, we saw that both kurtosis and skewness are -0.718 and 0.373 respectively. Both are between 1 and -1, meaning that the distribution of tax can be considered normal. We came to the same conclusion using the same method for wastewater and trash. Though some values were slightly above 1 or below -1, they were still relatively close enough to be considered normal. Since this data is normal, there is a significant correlation between taxes, wastewater, and trash and we are able to use linear regression.

**Solution**

A regression analysis was used on the sample set of data that had no tax fraud outliers. The Significance F was extremely small at 2.364E-29, supporting our assumption that taxes paid and disposal expenses are positively correlated. This regression also generated a formula which could predict the average amount of taxes a company should be paying given its cost of trash and wastewater. The formula developed is as follows: Average Predicted Tax = -16713.749+14.7366(Trash) + 14.029(Wastewater). This formula indicates that for every dollar spent in trash, a company must pay an additional $14.76 and $14.03 for every dollar spent on waste water. However, this formula calculated average tax and only considers points that are fairly close to its line to be within control. So a lower control limit must be determined in order to identify the true outliers.

The lower control limit was determined using the “Lower 95%” of both the trash and wastewater variables of the regression of the original data. Using these numbers gave us a 95% confidence level. To create a formula for the lower control limit, we took the formula for Average Predicted Tax and swapped the trash and wastewater coefficients for the coefficients listed under “Lower 95%”. This resulted in the following formula: Lowest Possible Tax = -16713.749 + 6.454085(Trash) + 6.06232\*(Wastewater). This formula indicates the lowest possible taxes a company would need to pay given how much they spend on waste disposal. The Lowest Possible Tax was calculated for each company. Outliers were determined through comparing this new predicted tax to actual taxes paid. If a company paid less than its calculated Lowest Possible Tax, it was considered to be under suspicion of tax fraud. This process determined that five companies were questionable.

**Conclusion**

This was a good way to solve this problem given the limited information, time, and resources. This solution does not require much information and is an improvement from the method that was previously used, which consisted of choosing a random company and investigating it. This method was also very simple to generate. Without investing a large amount of time, we were able to create a model to predict tax fraud fairly accurately. Other methods of solving this problem with the same given information may require more steps in the process but would likely end with the same results. If more information was available, it is certain that this would not be the best possible solution, however I believe it to be sufficient for our purposes.