12/2/2018

Andrew Marshall

Minitab Project - 4

INFS 608 Applied Statistics

## Overview

We have been tasked with investigating certain consumer characteristics to help predict how much credit card users will charge. The characteristics will be using for this investigation are household size and annual credit card charges for a sample size of 50 consumers. The result of this investigation should lead to the creation of a prediction model for potential future investigations.

## Initial Findings



Figure

With the consumer data we were provided on the 50 consumers we were able to determine our initial statistical results for Income, Size and Amount as shown in Figure 1. The mean household income is $43,480.00, mean household size is 3.42 people and the mean credit card amount spent is $3,964.00. We also noticed that the coefficient of variance for household size is 50.85 and is the highest of the 3. Based on this we can determine that the household sizes are significantly dispersed around the mean.

## Initial Findings, Part 2



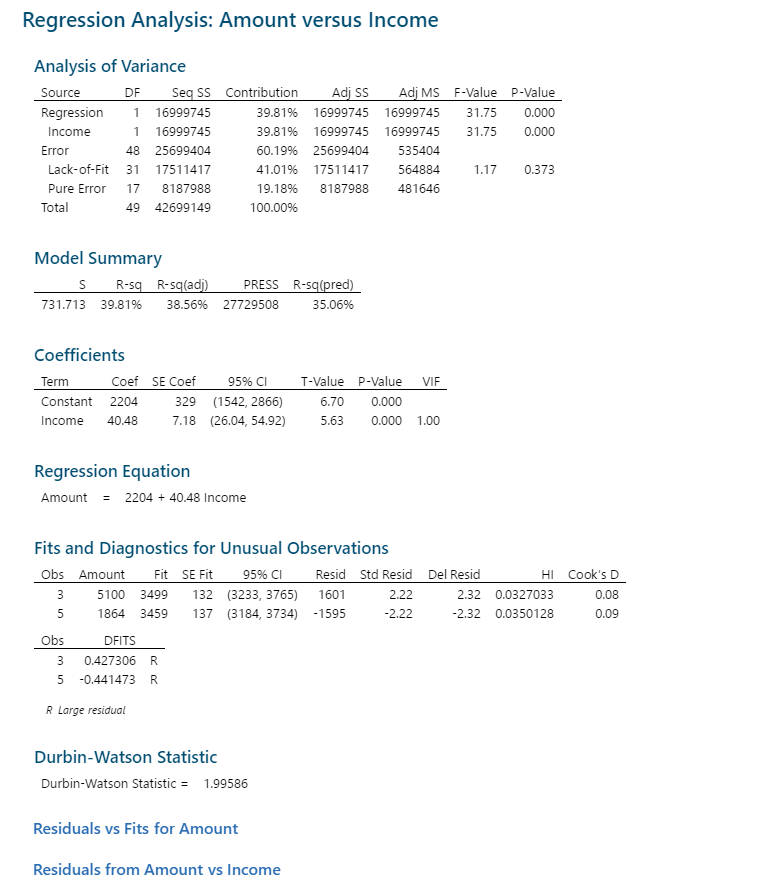
Figure



Figure

As Figure 2 shows, the distribution for the Amount data can be considered normal and therefore useful for doing a valid T-test. However, as shown in Figure 3, that is not the case for the Income data since most of the data falls outside the normal curve and in the tails. This means its distribution cannot be considered normal, so no valid T-test would be possible.

## Regression Models and Analysis



Figure



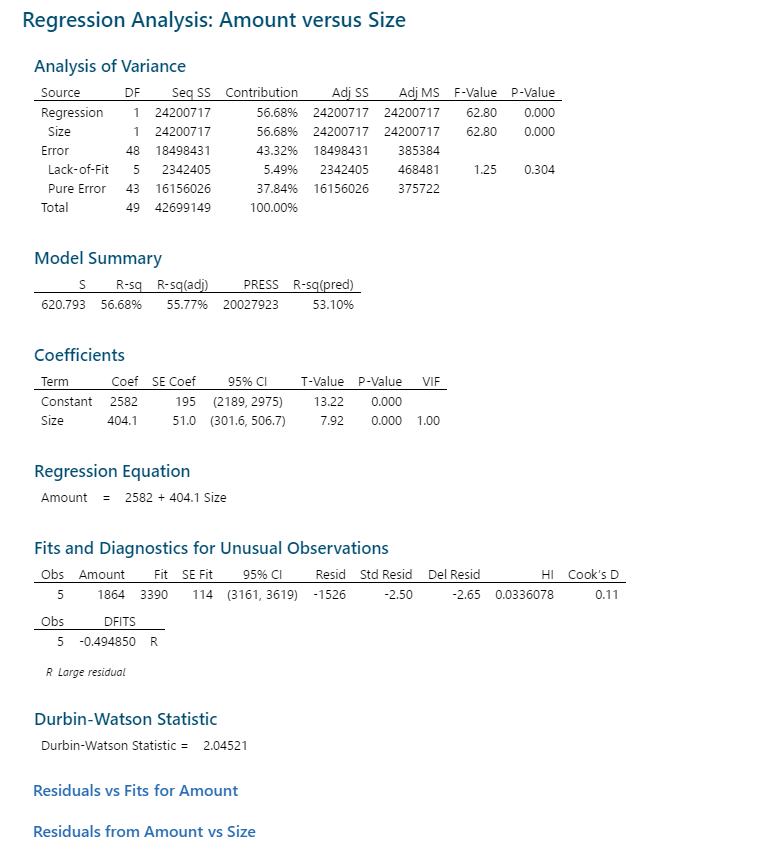
Figure



Figure

## Analysis Findings: Amount vs. Income

Given that the R-squared value is 39.81% and the R-squared adjusted is 38.56%, both of which are less than the ideal 90%, we can determine that there is a “loose fit”. This model implies there is little correlation between the Amount spent with credit cards to Household Income. Building additional models should assist in getting as close to 90% as possible. Also, we have a T-Test statistic of 5.63 which makes it significantly different from zero. Therefore, we must reject the null hypothesis that the coefficient is equal to zero considering the p-value is 0.000, which is less than 0.05.



Figure



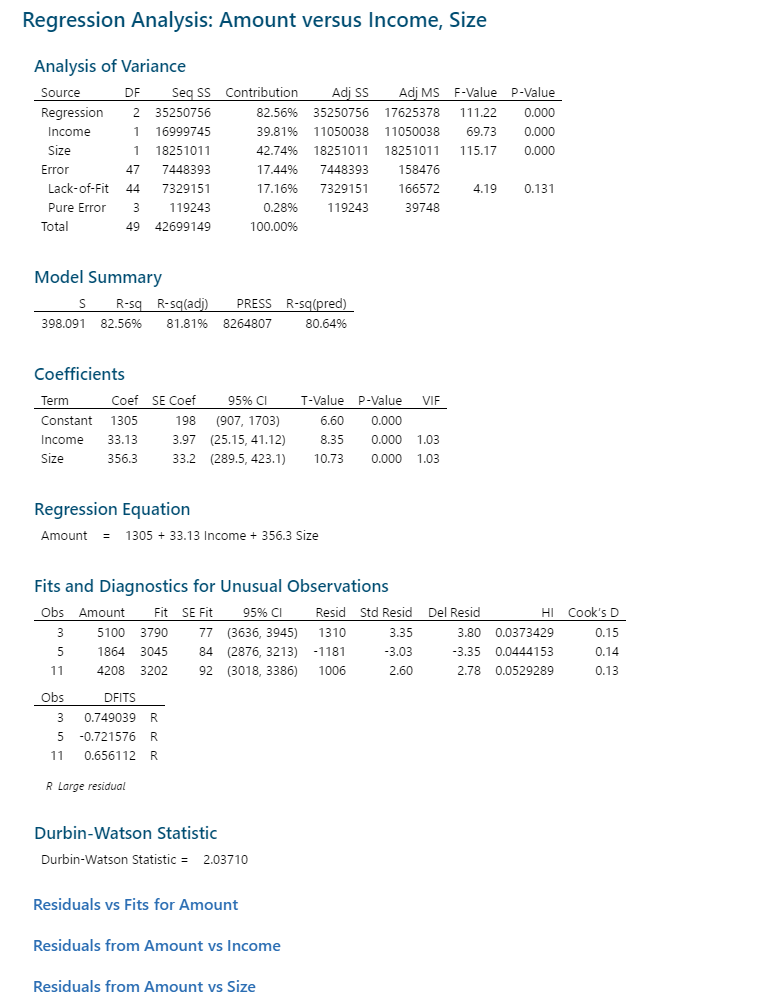
Figure



Figure

## Analysis Findings: Amount vs. Size

Given that the R-squared value is 56.68% and the R-squared adjusted value is 55.77%, which is less than though closer to, the ideal 90%. The model projects that for every 1 household member present, there is $404.01 increase in credit card spending. This implies more of a correlation between the Amount spent with credit cards to Household Size. Also, we have a T-Test statistic of 7.92 which makes it significantly different from zero. Therefore, we must again reject the null hypothesis that the coefficient is equal to zero considering the p-value is 0.000, which is less than 0.05.



## Analysis Findings: Amount vs. Income, Size

Given that the R-squared value is 82.56% and the R-squared adjusted value is 81.81, are much closer the ideal 90%. This model provides the best correlation of the 3 models. Since this is a multiple regression model, you will notice that there has been a change in the coefficients. Therefore, we can now see that for every 1 household member there is a $365.30 projected increase in credit card spending; we can also see now that for every $1 increase in Income, there is a projected $33,000 increase in credit card spending.

## Credit Card Spending Forecast



Figure



Figure

Based on this data, we can conclude that the household size of 1 offers the least reliable. This is because the max income for this household size in our dataset is $44,000 as shown in Figure 11, which would make the forecast data an outlier.

## Final Recommendations

Based on our investigation we have determined that the multiple regression model provides a more accurate results for our needs. The only issue of note is the lack of a normal distribution for the household income data.

However, it is possible that adding additional variables could assist with mitigating that issue. One such variable could be the household credit score. This would give us a better indicator of the household’s financial health and the marketing department could act according based on those credit scores. Another useful variable would be the household location data, such as the zip code. This variable will provide specific household purchasing history based on location, which would allow the marketing department to more accurately products specific to those locations.

In conclusion, we recommend using the multiple regression model and adding the following additional variables:

* Household credit score
* Household location data (i.e. zip code)