**Task 1 - Data Science for Engineers**

To solve the task I decided to develop a multiple regression analysis. This method is suits the problem greatly because we want to predict the value of a variable based on the value of two or more variables, (Research Ltd, 2018) precisely, we want to predict the value of the tip based on variables as the total bill, sex, smoker status, day, time and size.

Now, in order to obtain the model for the prediction I followed a several steps process which will be briefly described below, but feel free to check the Excel and Minitab files for more detail.

1. **Collect the data.** Which in this case was handled in as a [csv file](http://bit.ly/tips_data).
2. **Clean the data.** In order for tools such as Excel and Minitab to work, data must be cleaned. In this case, all I had to do was replace categorical variables with numbers.
3. **Examine relationships.** When developing a multiple regression analysis you have to determine which independent variables are significant for the dependent variable and which not to avoid overfitting, as well as which independent are not related to other independent variables to avoid collinearity. To do this I used Pearson's formula for continuous and 1/0 categorical variables (Priority Digital, n.d), for categorical variables with more than 2 categories I used the regression analysis results from both tools..

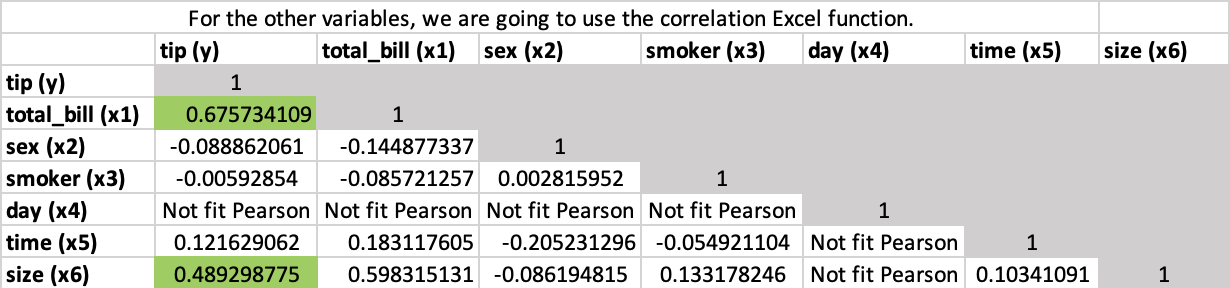


Table 1. Pearson correlation table

1. **Run the regression analysis with non-redundant and significant variables.** For educational purposes I decided to do first the analysis in Excel and develop the formula myself. From the above table I concluded that it was safe to use variables such as total bill, day, time or size since they held a relationship with the tip, unlike variables such as sex or smoker status, that apparently held a negative relationship with the tip. Nevertheless I decided to double check with Excel and develop two multiple regression analysis, one with all the variables and the second discarding unuseful variables.

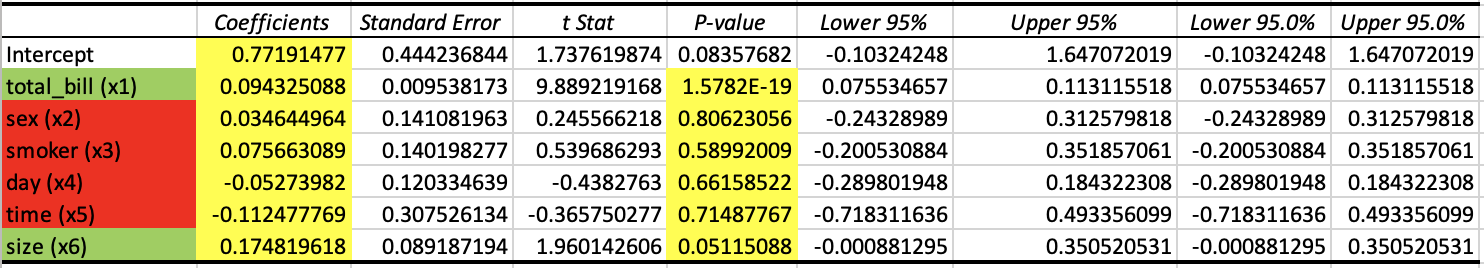
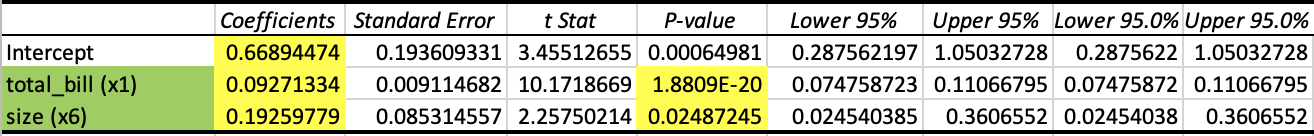
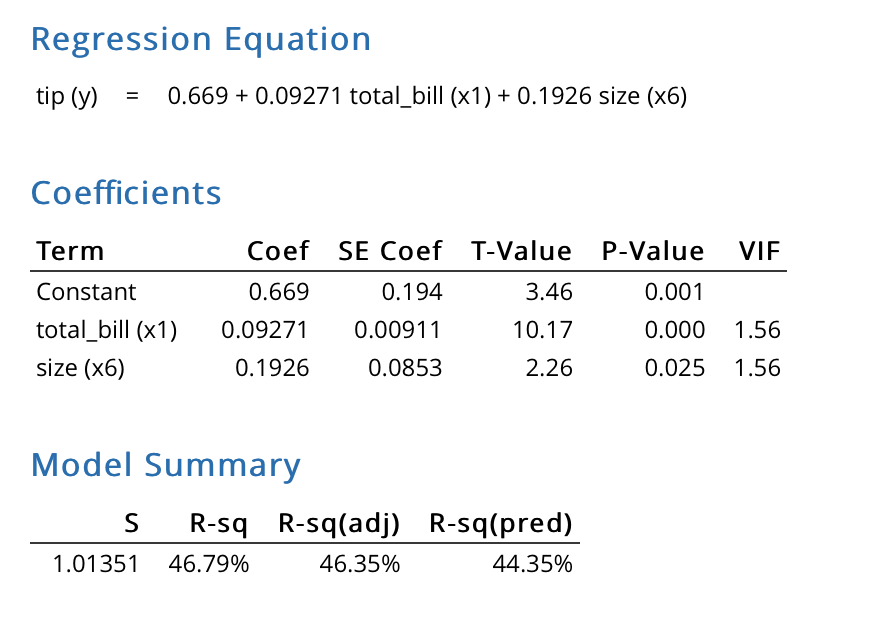


Table 2. First regression analysis with Excel

With the information from the above table I discarded sex, smoker status, day and time as potential variables for the regression model because they had a high P value (> .05), which meant that they were not significant thus they did not contribute to the model. Next, I developed another model with the significant variables, total bill and size.

Table 3. Second regression analysis with Excel

In this new analysis, both individual P-values were significant, which meant that they would contribute to the model. Finally, I also developed the same model in Minitab using the Fit Regression Model feature, which develops the regression equation too.

Table 4. Regression analysis with Minitab

It is to note that with the table above I concluded it would be ok to use both variables in the same equation, since they had a low Variation Inflation Factor (VIF). According to Minitab’s blog, it is problematic to use variables with a VIF from 5 - 10, and unacceptable to use them if they hold a VIF > 10 due to the high collinearity. (Heckman, 2015)

1. **Use the best fitting model to run predictions.** As seen above I developed multiple models and then decided to go with the best. The chosen model, developed with Minitab, derives the following equation:

which is in the form .

This model was quite acceptable considering the following factors.

* **Highest R-sq:** From all potential models, the chosen one had the highest R-squared, which meant that it was the model that had the better prediction according to the chosen variables. With this model roughly 47% of the variability could be explained by the set of independent variables.
* **Low drop off between R-sq adjusted and R-sq predicted:** A large drop off between these scores indicate overfitting.
* **Acceptable P-Value:** The overall regression had a low P-Value, which means that the developed model was significant enough, or in other words, a good model. Also, the individual P-Values for both total bill and size were acceptable.

To confirm that this model performs good predictions, let us perform a few substitutions on our model and compare them with entries from our data set. Let us recall our model too:

// Here we are adding the standard error

The expected result was 5, which is inside our range. So it was a **good prediction.**

The expected result was 2.64, which is inside our range, so it was a **good prediction.** Without the error range, we also had a really close prediction, since 2.865 is very close to 2.64.

The expected result was 3.5, which is inside our range, so it was a **good prediction.** Without the error range, we also had a really close prediction, since 3.417 is very close to 3.5.

With the above information we can conclude with evidence that our model makes predictions that are **truth**. Our equation can be read as the following:

* If size is held constant, then tip is expected to increase by .092 dollars for each additional dollar spent on totalBill. This **makes sense** both **statistically and logically,** considering that a tip should be around 10%, at least in Mexico.
* On the other hand, if the totalBill is held constant, then tip is expected to increase by .19 dollars for each additional person on the table.

Please see attached Excel and Minitab files for further detail on the procedures.

**References**

1. L. Research Ltd, “Multiple Regression Analysis using SPSS Statistics,” *Laerd Statistics,* 2018. [Online]. Available: <https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php.>[Accessed: 24-Jan-2021].
2. Priority Digital, “The Excel PEARSON Function,” *Excel PEARSON Function,* no date. [Online]. Available: <https://www.excelfunctions.net/excel-pearson-function.html.> [Accessed: 24-Jan-2021].
3. E. Heckman, What in the World Is a VIF?, 20-Jul-2015. [Online]. Available: <https://blog.minitab.com/blog/starting-out-with-statistical-software/what-in-the-world-is-a-vif#:~:text=This%20post%20will%20give%20you,VIF%2C%20or%20Variance%20Inflation%20Factor.&amp;text=The%20VIF%20measures%20how%20much,if%20your%20predictors%20are%20correlated.> [Accessed: 25-Jan-2021].