Anthony Davis & Tyson Jeffrey

STAT311 Regression Analysis

Dr. Iresha Premarathna

12/02/2024

Predicting Tip Amount Received by Waitresses/Waiters

Introduction:

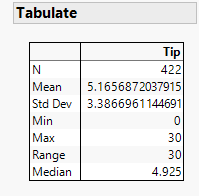
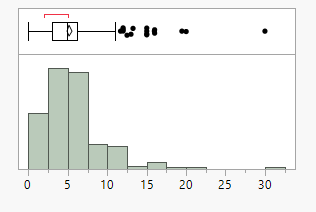
This case study was done to find a relationship between the tip amount received by waitress with several factors related to the customers at a restaurant. To do so, linear regression was used to analyze the relationship between eight qualitative variables (Day-day of week; Meal-lunch, dinner, or late night; Payment-credit, cash, credit with cash tip; Age-young adult, middle aged, senior citizen; GiftCard-was one used; Comps-were there complimentary items; Alcohol-was it purchased; Bday-was there a free birthday item) and two quantitative variables (Bill-total size of the bill in $, and Tip Percentage-in decimal form) with our response variable (Tip-tip amount in $). Our intent was to find as accurate of a prediction as possible for tip amount a waitress/waiter may receive, which would be useful for those working in restaurants.

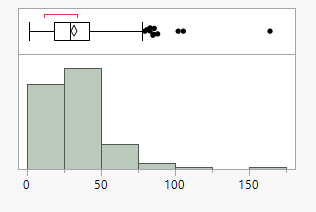
Research Questions:

1. Can we find a statistically significant relationship between the given information and the mean tip amount?
2. Do any of our variables show significant interaction that would improve our model without compromising simplicity?
3. Are there any outliers in data or issues regarding assumed normal distribution with the data we have?

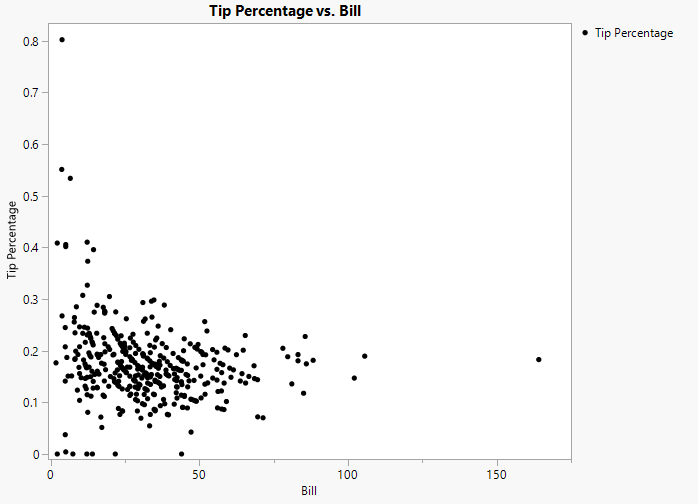
Summary Statistics and data:

Our dataset (Waitress Tips) has 422 points of data. Below we have collected some simple statistics about our response variable of tip amount. We can see that our response variable does have a skewed distribution (top histogram/box plot), but as the points mostly coincide with the skewed distribution of bill amount (bottom histogram/box plot), which will be used to help predict tip amount, it should be okay for the sake of prediction.

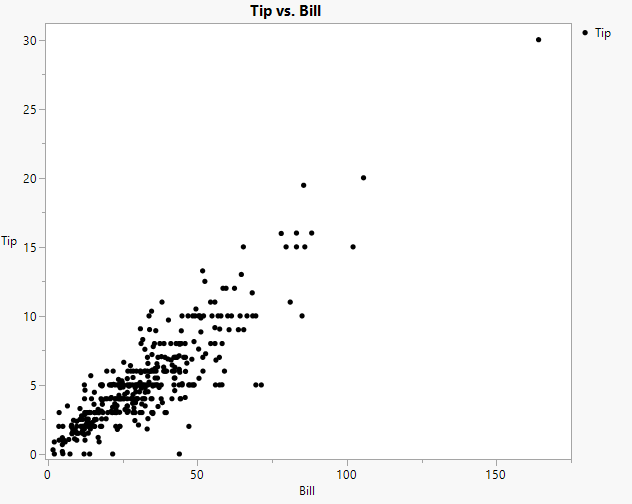
 



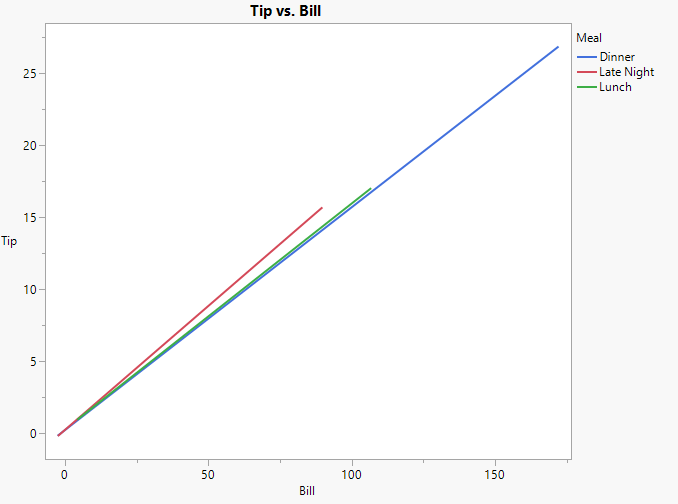
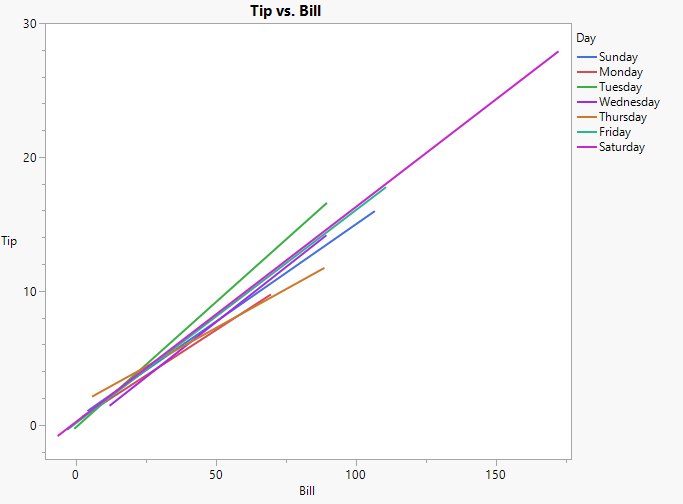
For our models, since we wanted to determine the tips waitresses/waiters can expect, we needed to decide between using tip amount or tip percentage as our response variable. We decided to not utilize the tip percentage, as it appears to have consistent values for most points of data when compared with bill amount. Except for some very high tip amounts amongst lower bill amounts, and a handful of very small or $0 tips, most of the tip percentages appear to range around 10%-30%, while averaging around 20%:

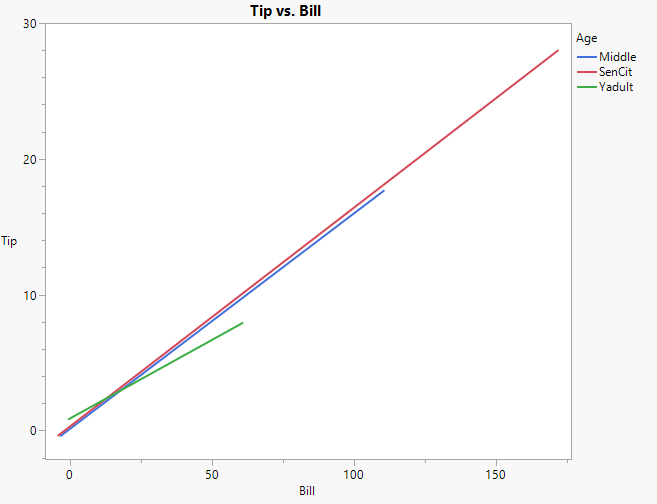
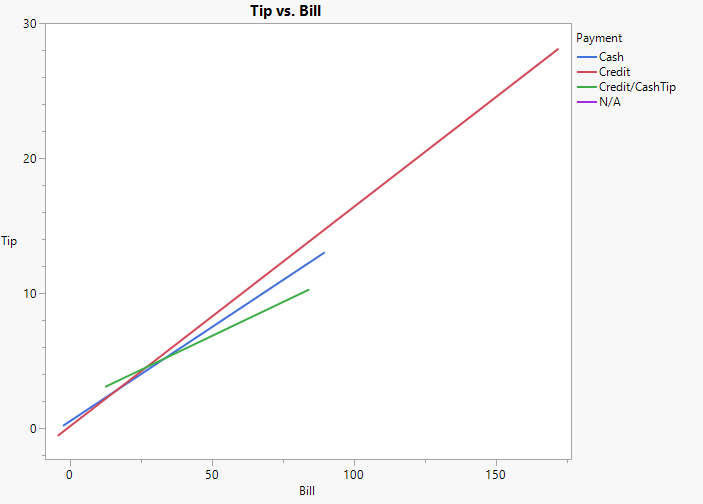


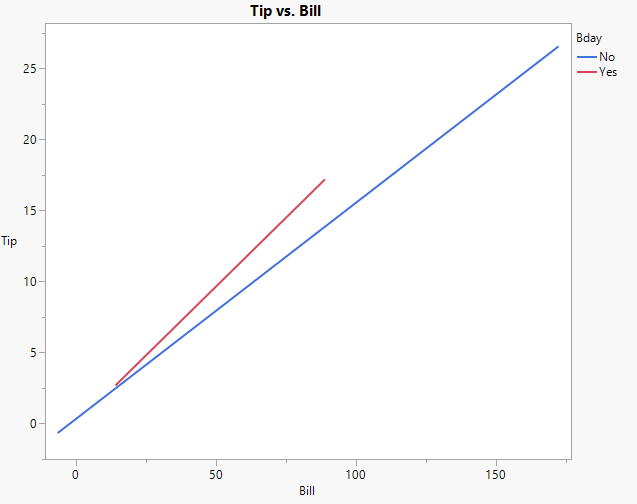
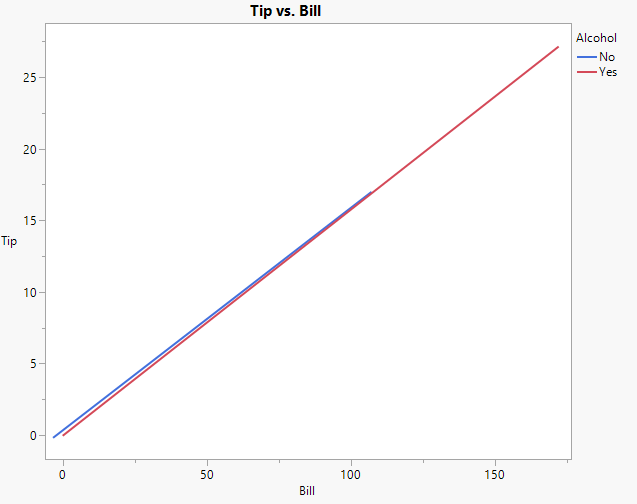
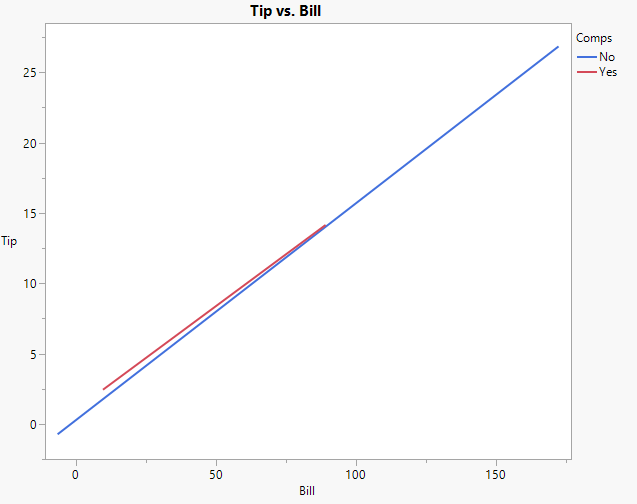
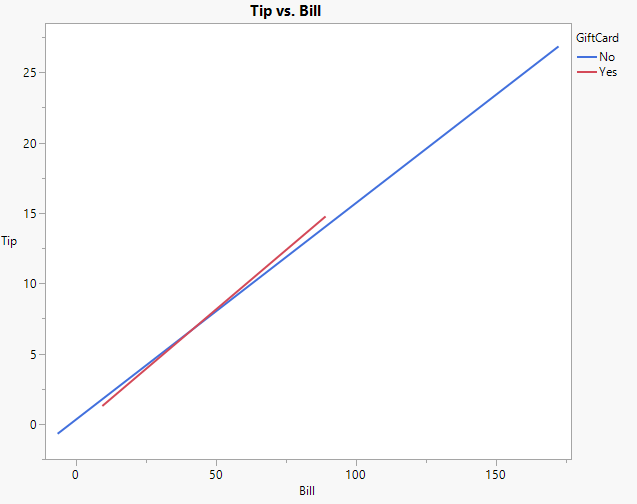
On the other hand, tip amounts appear to have a very linear relationship with bill amount, and for that reason linear regression has been used when relating bill and tip amount, and why a second-order term would appear unnecessary. Tip percentage will be added to the models to improve accuracy:



There does appear to be some interaction between bill and some of the qualitative variables, as shown in the graph below, so interaction terms will be added in later models. There are clear interactions with bill versus tip with the variables Day, Payment, Age, and Bday. Other variables will be given interactions an compared to a model with only the clear interactions:







Hypothesized Models:

For all our models, row 239 was removed prior to running the models. While running the interaction models, we found that there existed interdependency amongst multiple terms. We discovered that this row of data was creating a new level for payment method, non-applicable. For that reason, we removed it, fixing the issue of interdependency. So, for all these models, 421 out of the 422 data points were utilized.

Model 1:

For our first model we will compare tip to all the variables we will be using with no interaction terms. This is a simple model that assumes that there are no significant interactions between any of the variables. (Bases for qualitative variables [ through set to 0] are Saturday, lunch, Credit payment w/ cash tip, young adult, gift card used, complimentary item given, alcohol purchased, birthday item given).

Tip amount

{1 if Sunday, 0 otherwise} {1 if Monday, 0 otherwise}

{1 if Tuesday, 0 otherwise} {1 if Wednesday, 0 otherwise}

{1 if Thursday, 0 otherwise} {1 if Friday, 0 otherwise}

{1 if Dinner, 0 otherwise} {1 if Late Night, 0 otherwise}

{1 if Cash payment, 0 otherwise} {1 if Credit payment, 0 otherwise}

{1 if Credit payment with Cash tip, 0 otherwise}

{1 if Middle aged, 0 otherwise} {1 of Senior Citizen, 0 otherwise}

{1 if no gift card used, 0 otherwise}

{1 if no complimentary item given, 0 otherwise}

{1 if no alcohol purchased, 0 otherwise}

{1 if no free birthday item given, 0 otherwise}

Bill amount in $ Tip Percentage

Model 2:

In model 2 we used stepwise regression to see if we could find an even simpler model that would not compromise the accuracy of our predictions. We found that while using mixed stepwise regression with p values to leave and enter both being 0.05, we get a model that only uses bill, tip percentage, and Bday.

A screenshot of a computer

Description automatically generated

{1 if no free birthday items given, 0 otherwise}

Bill amount in $ Tip Percentage

Model 3:

In model 3 we added the interaction terms that appeared significant in the graphs shown in the summary statistics and data section to model 2. The terms without interaction that were not used in model 2 were also added.

{1 if Sunday, 0 otherwise} {1 if Monday, 0 otherwise}

{1 if Tuesday, 0 otherwise} {1 if Wednesday, 0 otherwise}

{1 if Thursday, 0 otherwise} {1 if Friday, 0 otherwise}

{1 if Cash payment, 0 otherwise} {1 if Credit payment, 0 otherwise}

{1 if Credit payment with Cash tip, 0 otherwise}

{1 if Middle aged, 0 otherwise} {1 of Senior Citizen, 0 otherwise}

{1 if no birthday, 0 otherwise}

Bill amount Tip percentage

Model 4:

Model 4 also utilizes interaction terms but includes all of them along with all the variables in model 1. This way we can compare if added the extra terms benefits the model overall.

(same as model 1 variables)

(interaction terms between qualitative variables and bill amount)

Model reruns:

We will take the best models and try to improve it by removing any outliers and making modifications to improve model significance by looking at residuals once the best models are found.

Results:

Model 1:

**A screenshot of a computer

Description automatically generated**A screenshot of a graph

Description automatically generated

A screenshot of a graph

Description automatically generated

Model 1 does immediately give us a statistically significant model with an F value of 174.25 and p-value < 0.0001. It also appears to be a reasonable model overall with an adjusted R Square of 0.881554 and rMSE of 1.165, but we will see that later models may improve on this. One downside of this model is that some of the levels for the quantitative variables are not significant.

Model 2:

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

A screenshot of a computer error

Description automatically generated

As can be seen in the reduced version of model 1, model 2 is also a significant model, with a very large F value of 1031.25 and p-value < 0.0001. The adjusted R Square and rMSE are slightly worse than in model 1, but by very minimal amounts (adjusted R Square by about 0.0012 and rMSE by about 0.005). On the other hand, we can see that all variables are significant to the model.

If we compare models 1 and 2 with a nested F test we get the following results:

Sum of Squares 25.862190637

Numerator DF 15

F Ratio 1.2696775957

Prob > F 0.2180198298

In other words, despite removing fifteen levels from 7 variables, with an F value of 1.27 and p value of 0.218, testing the models suggests that the removed terms cannot be shown to contribute significantly to the prediction of tip amount. This implies that if a simple model is desired, that is one without interaction, model 2 would be sufficient when compared with model 1.

Model 3:

A screenshot of a computer

Description automatically generated A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

Although we can see that model 3 does improve the adjusted R square and rMSE values of the previous models, it is again by a very minor amount. In fact, when compared to the values in the simplest model, model 2, the adjusted R Square only improves by around 0.01 and rMSE by 0.048. The model is still significant at predicting tip amount with F value of 142.48 and p value < 0.0001. Many of the levels from the variables do show to be insignificant to the model overall.

Model 4:

A screenshot of a computer

Description automatically generatedA screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

Once again we can see that although model 4 remains significant overall with and F value of 43.27 and p value < 0.0001, the adjusted R Square and rMSE are only slightly improved. In fact, from model 2 it manages to only improve by an adjusted R Square of about 0.012 and rMSE of 0.06.

When using a nested f test between model 4 and 3, we get the following result:

Sum of Squares 21.916950009

Numerator DF 10

F Ratio 1.7742438053

Prob > F 0.0635279309

If we were to use a significance level of 0.05, we would see that in this case (with an F value of 1.774 and p-value of 0.0635), we manage to fail to show that the added terms in model 4 contribute significantly to the overall prediction. For this reason, model 3 will be used in favor of simplicity without sacrificing much for significance.

Removing outliers:

Based on the results, if high accuracy was desired, the slight improvements found by using interaction terms may be beneficial, therefore making model 3 the best model whilst maintaining some simplicity. But for the sake of predicting tip amount, a simpler model that hardly sacrifices accuracy may be better, which would imply that model 2 is the best model. With that being said, we will try to improve the results found in model 3 and in model 2.

When using multiple confidence intervals for externally studentized residuals, we can see four obvious outliers in rows 37, 39, 49, and 287 for both models (highlighted in both models). We will remove these outliers and rerun model 2 and model 3 to see if we made improvements.

Model 2 externally studentized residuals:

A graph of a number of dots

Description automatically generated with medium confidence

Model 3 externally studentized residuals:

A graph of numbers and lines

Description automatically generated with medium confidence

Model 2 minus outliers:

A screenshot of a computer

Description automatically generated A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

Removing the four outliers caused adjusted R square of model 2 to improve by 0.03 and rMSE by 0.16, which is a fair improvement that will help to improve accuracy. Furthermore, the terms have all become more significant, with only one level of payment not being significant to the model.

Model 3 minus outliers:

A screenshot of a computer

Description automatically generated A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

We can see that model 3 has also improved. The adjusted R square has improved by about 0.033 and rMSE by about 0.183. Once again, removing just four rows of data has a noticeable improvement on model accuracy.

Residuals vs predicted (after outliers removed):

Model 2:

A screen shot of a graph

Description automatically generated

Model 3:

A screen shot of a graph

Description automatically generated

It appears that the models do have a clear trend. However, upon attempting to fix this by adding second-order terms, transforming y (by Poisson, binomial, or multiplicative transformation), standardizing values, and removing even more outliers, we either do not fix the problem or make the model significantly worse to the point of model insignificance or extremely low R Square/high rMSE values. For that reason, we will assume that the models we have declared our best models are normally distributed, even though we can see that they are not.

Conclusion:

Overall, we have managed to find multiple models that are fairly accurate at predicting tip amount for waitresses/waiters. In an application where higher accuracy is wanted, after removing five points of data (four outliers and 1 problematic data point) model 3 offers an accurate if more complex prediction. On the other hand, if such high accuracy is not needed, model 2 offers an extremely simple model and prediction that does not sacrifice much of the accuracy from model 3, making it remain an accurate model and a good fit for predicting tip amount.