Statistical Exploration of the Data for the Georgia Mountain Food Bank

Ping Ye*, Jordan Boeckman
University of North Georgia, USA
Email: ping.ye@ung.edu

Abstract

Hunger relief is one of the major needs during humanitarian emergencies. Georgia Mountain Food Bank's (GMFB) is to address hunger, health and quality of life by serving those in need throughout North Georgia. GMFB is invited to participate in this project by providing data on the "People in Need" program. GMFB provides the demographic and food picking up information for FY17 & FY18. The data are completely anonymous without any ethical concerns. GMFB would like to obtain a better understanding of the data through mathematics research and analytic study with this project.

Keywords: service station, hypothesis testing, regression, decision tree, analytic study

1. Introduction

Georgia Mountain Food Bank (GMFB) is a non-profit organization which mission is to address hunger, health and quality of life by serving those in need throughout North Georgia. GMFB provides a vital link between sources of food supplies and hardworking community-based partner agencies who help get the food into the hands of families and individuals who need it. Whether it is distributing food or serving the community through outreach programs, initiatives and resources.

This project works as a senior project for the undergraduate research of the student author. The goal of the project is to find out if Hall County and other counties need new service stations across North Georgia. The studies are based on data taken from 2017 and 2018. The data includes the following variables: date, time, number of people in the households, number of seniors and children, street number, street name, city, zip code, county, phone number, pounds, second visits, and benefit packs. There were 367 entries for the year 2017 and 409 entries for the year 2018. The statistical methods of data cleaning, box plots, hypothesis testing, multiple linear regression, logistic regression and decision tree are applied to explore the data. These different tools and techniques are used to visualize the data, test its significance, identify the significant factors, and help GMFB to make decisions.

The first steps were to clean and organize the initial data. To deal with the missing data, zeros were added for the "Senior" and "Children" columns where there were blanks before. If any "pounds" data was missing, it was replaced by the average in that respective column to fill in the missing value. A new column titled "Adults" = "Household" – "Seniors" – "Children" was created. The next step was to organize the data into four tables on Excel to further organize and categorize the data. After cleaning data, four data sets were generated. Those are the data set for the first half of the year 2017, the data set for the second half of the year 2017, the data set for the whole year of 2017, and the data set for the year 2018.

2. Experimental Section

2.1 Comparison of Mean Pounds by Boxplots

By using boxplots to compare the means of the "Pounds" columns, the categories in which the groups coincide or differ can be visualized. The boxplots of the hall county versus other counties in terms of the "Pounds" variable for the year 2017 and the year 2018 are shown in the following Figure1 and Figure2. Based on the boxplots, it shows that the mean pound between hall county and other counties is pretty close for both FY2017 and FY2018. To look into more details of the data, the boxplots of the hall county versus other counties in terms of the "Pounds" variable for the first half and second half of 2017 are shown in the following Figure3 and Figure4. The figures indicate the mean pounds are similar during the first half-year of 2017 but look somewhat different during the second half-year of 2017.

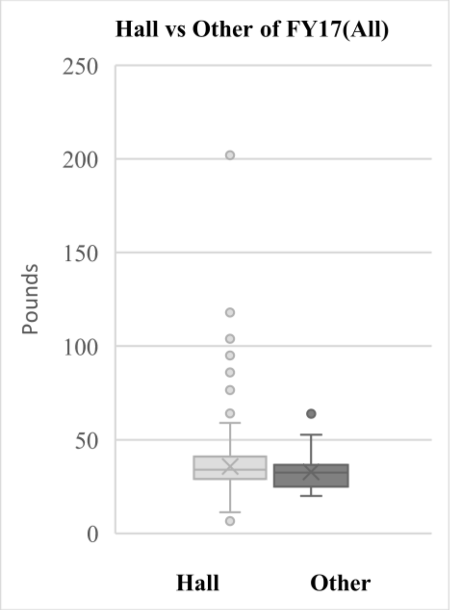


Figure1: Boxplot of 2017

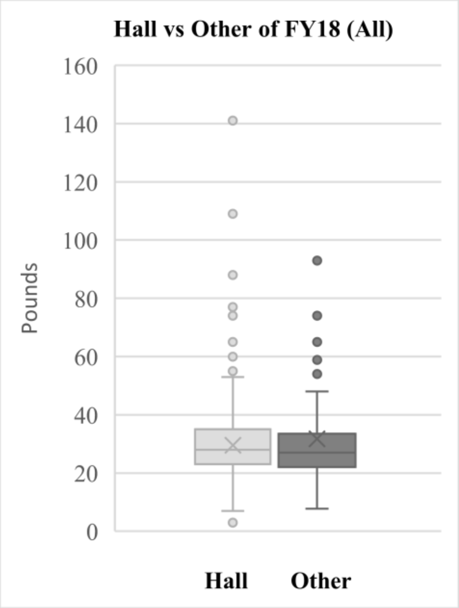
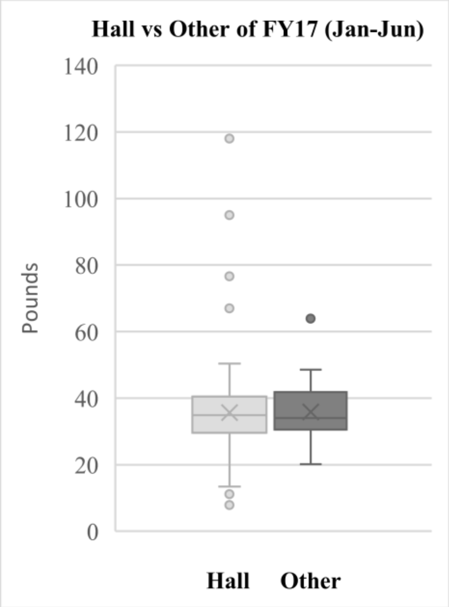


Figure2: Boxplot of 2018



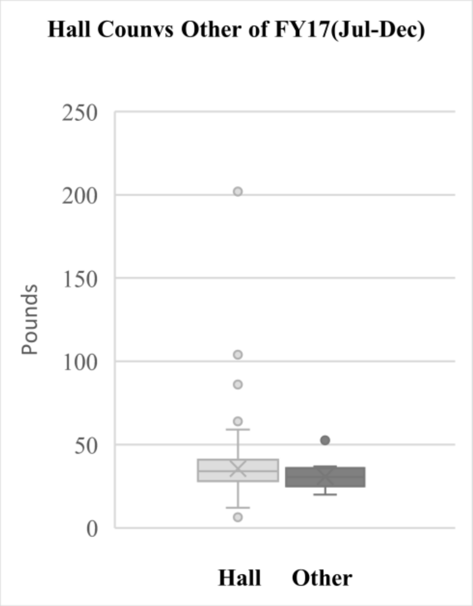


Figure3: Boxplot of the first half of 2017 Figure4: Boxplot of the second half of 2017

To take a more in-depth look into the result, the hypothesis testing of the difference of two means method is used for the above four cases, respectively. The hypothesis testing goal is to test the significant difference in the mean weight of foods people picked up at GMFB between hall county and all other counties to determine whether new service locations are needed or not. If H_0 is rejected, then new service locations are needed. To do the hypothesis test, both the software Minitab and R are used. Next, the independent t-test is applied for each of the four tables with hall county versus other counties. The results are showed in the following Table1. Here μ_1 represents the mean pound for hall county, μ_2 represents the mean pound for all other counties.

Time Period	Ho	Ha	Paired T Hypothesis Test Output		Statistical Conclusion
			T-Stat	P-value	
FY17	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	1.79	0.077	Not statistically significant
FY18	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	-0.87	0.386	Not statistically significant
FY17 Jan - Jun	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	-0.05	0.959	Not statistically significant
FY17 Jul - Dec	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	2.68	0.009	Statistically significant

Table1: Hypothesis testing results of the mean food weight between hall county and other counties.

Based on the P-values of the hypothesis testing for the difference of means, all the results are consistent with the boxplot results. The hypothesis tests help to get an even more in-depth look at the data.

These tests give more details than the boxplots and help to determine the statistically significant data. During both FY17 and FY18, the results are not statistically significant even the pattern was slightly changed during the second half of 2017. It may be due to the seasonal demands. Overall, the food bank has reasonable service locations, and they do not need to set up new locations in other counties.

2.2 Identifying Significant Factors

Other interesting questions are raised to draw GMFB's attention such as "How to identify the family who needs help?", "How to prepare the food amount according to different family types?", "Any information indicates the second visit?", "In general how much food should we prepare for the first visit?" By developing appropriate models, we can narrow the full range of data to a few critical predictor variables. Furthermore, we can probe the answers for the above questions in which GMFB can focus its efforts.

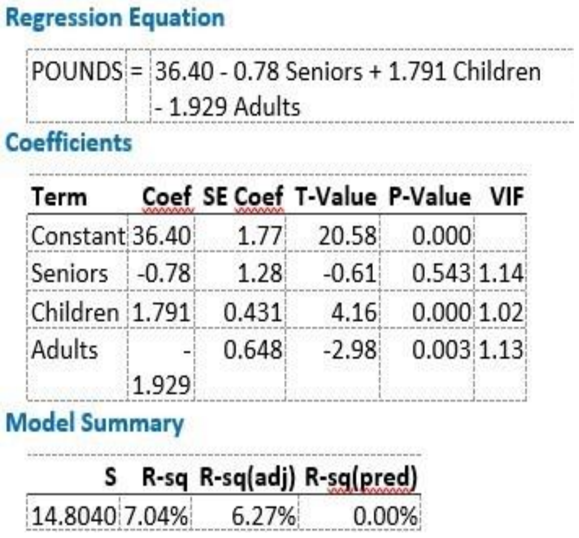
2.2.1 Linear Regression Model

To find the connection between the number of family members and the picking up food weight, a multiple linear regress model is built with the following form:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

Where y is the food weight, x_1, x_2, x_3 represent the number of seniors, the number of children, and the number of adults in a family, respectively.

After applying the model to both the FY17 and FY18 datasets, we get the results as shown in Figure5 and Figure6.



Regression Equation

POUNDS	=	36.40 - 0.78 Seniors + 1.791 Children - 1.929 Adults

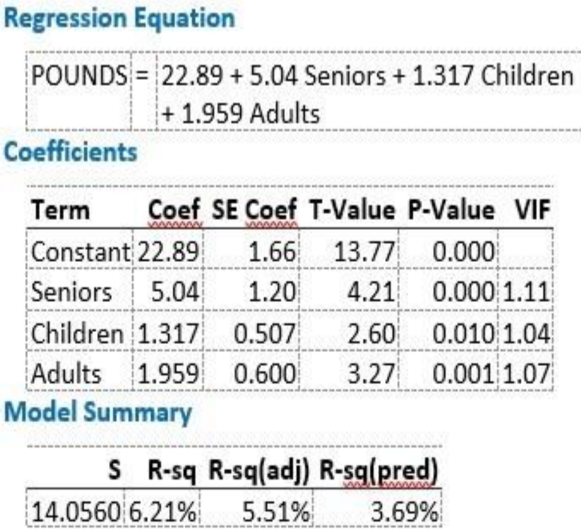
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	36.40	1.77	20.58	0.000	
Seniors	-0.78	1.28	-0.61	0.543	1.14
Children	1.791	0.431	4.16	0.000	1.02
Adults	-1.929	0.648	-2.98	0.003	1.13

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
14.8040	7.04%	6.27%	0.00%

Figure5: Linear regression model of FY17



Regression Equation

POUNDS	=	22.89 + 5.04 Seniors + 1.317 Children + 1.959 Adults

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	22.89	1.66	13.77	0.000	
Seniors	5.04	1.20	4.21	0.000	1.11
Children	1.317	0.507	2.60	0.010	1.04
Adults	1.959	0.600	3.27	0.001	1.07

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
14.0560	6.21%	5.51%	3.69%

Figure6: Linear regression model of FY18

In above Figure5 the results of the regression analysis of FY17 show that the P-value for "Seniors" is 0.543, the P-value for "Children" is 0.000, and the P-value for "Adults" is 0.003. This means that "Children" and "Adults" are statistically significant to "POUNDS" because both of their P-values are less than 0.05. The adjusted R squared value for this model is 6.27%. Therefore, we are unable to use this model to predict pounds because the R-squared value is not plausible. The linear regression model of FY18 has the same problem since its adjusted R-Squared value is only 5.51% (Figure6). After checking the Fits and Diagnostics for Unusual Observations and the Residual Plots (Figure7 and Figure8), the lack of fit is because the data has too many unusual observations causing the data not to follow the normal distribution. We can solve the problem by removing the unusual observations, but it needs more data to improve the model's accuracy. Other models may be selected to research the relationship between the response variable "POUNDS" and the predicted variables for future studies, such as polynomial regression or nonlinear regression.

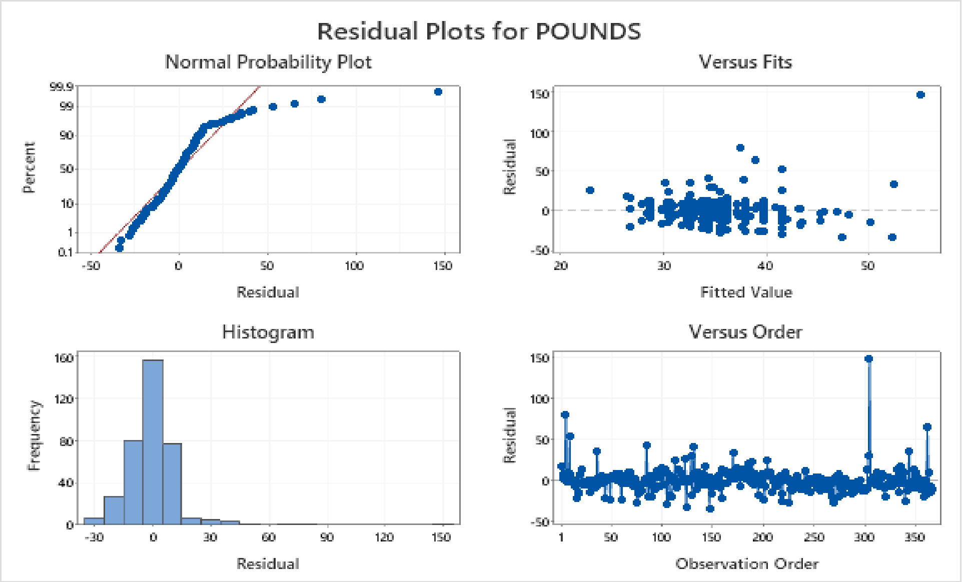


Figure7: Residual plots of FY17

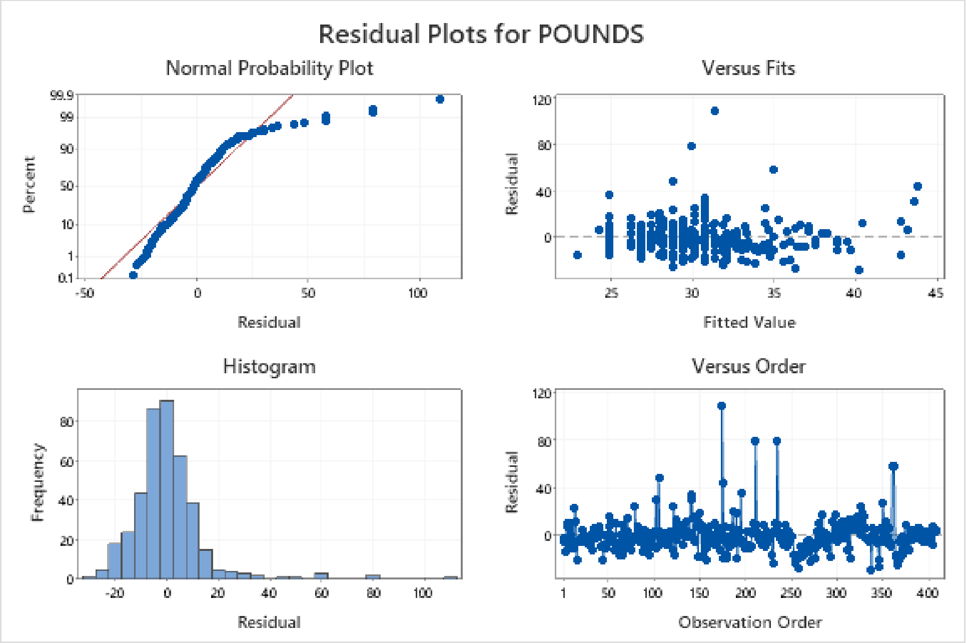
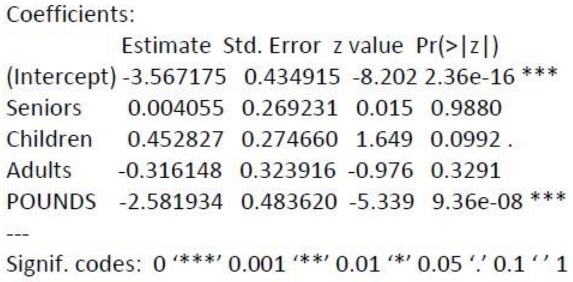


Figure8: Residual plots of FY18

2.2.2 Logistic Regression Model

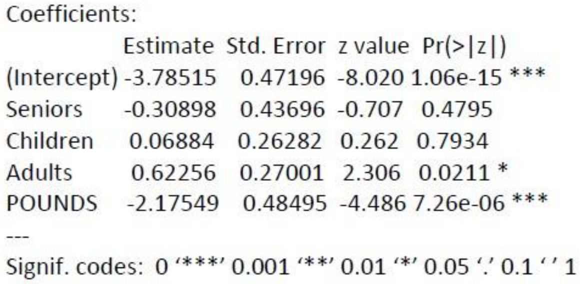
To answer the question about the second visit of GMFB, we set the response variable as the binary of 1 or 0, which means the family will have the second visit or not. The ideal model when predicting a dichotomous categorical variable would be logistic regression. Here, the logistic regression model is used to reduce the number of independent variables by removing variables that are not impactful in resulting in the second visit to GMFB.

For all models, the same training set and testing set are used. The training set and testing set are split in the way that the training set contains 70% of the dataset while the testing set contains the remaining 30%.



```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -3.567175  0.434915  -8.202 2.36e-16 ***
Seniors      0.004055  0.269231   0.015  0.9880
Children     0.452827  0.274660   1.649  0.0992 .
Adults      -0.316148  0.323916  -0.976  0.3291
POUNDS      -2.581934  0.483620  -5.339  9.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure9: Logistic regression model of FY17



```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -3.78515   0.47196  -8.020 1.06e-15 ***
Seniors     -0.30898   0.43696  -0.707  0.4795
Children     0.06884   0.26282   0.262  0.7934
Adults       0.62256   0.27001   2.306  0.0211 *
POUNDS      -2.17549   0.48495  -4.486 7.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

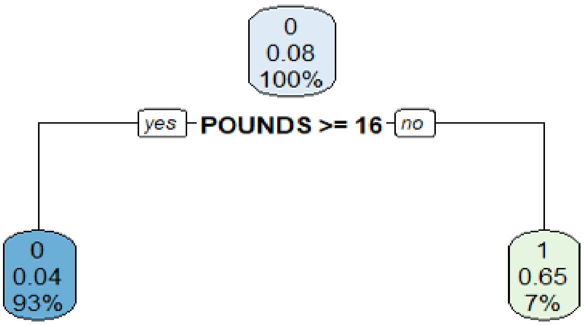
Figure10: Logistic regression model of FY18

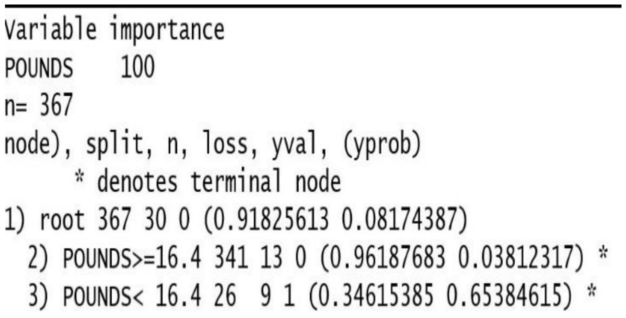
After applying the logistic regression with significance level 0.05, “POUNDS” is statistically significant to the “2nd Visit”, which indicates the food weight people pick up during their first visit will affect the potential 2nd visit. Although the types of people in the household, including seniors, children, and adults are not statistically significant with effect to the 2nd visit, we can see their p-values are in the order of children <adults<seniors. This indicates the influence to affect the 2nd visit will be children>adults>seniors. To avoid overfitting and to calculate the model accuracy, FY17 data has been split as 70% training and 30% testing. The model accuracy is 93.7% based on the confusion matrix. It indicates that we can use this logistic model to predict the 2nd visit based on “POUNDS”.

Similarly for the FY18 data, “POUNDS” and “Adults” are statistically significant to the “2nd Visit”, which indicates the amount of food weight people pick up during their first visit will affect the potential 2nd visit, and the number of adults in the household will also affect the potential 2nd visit. Although the types of people in the household except “Adults” are not statistically significant to affect the 2nd visit, we can see their p-values are in the order of adults<seniors<children. This indicates the influence to affect the 2nd visit will be children<seniors<adults. To avoid overfitting and to calculate the model accuracy, FY18 data has been split as 70% training and 30% testing. The model accuracy is 94.9% based on the confusion matrix. It indicates that we can use this logistic model to predict the 2nd visit based on “POUNDS” and “Adults”.

2.2.3 Decision Tree Model

Finally, a decision tree model is used to visualize the prediction of the second visit. Meanwhile, it is the second classification model applying to our data. We can compare its result with the previous logistic regression model.

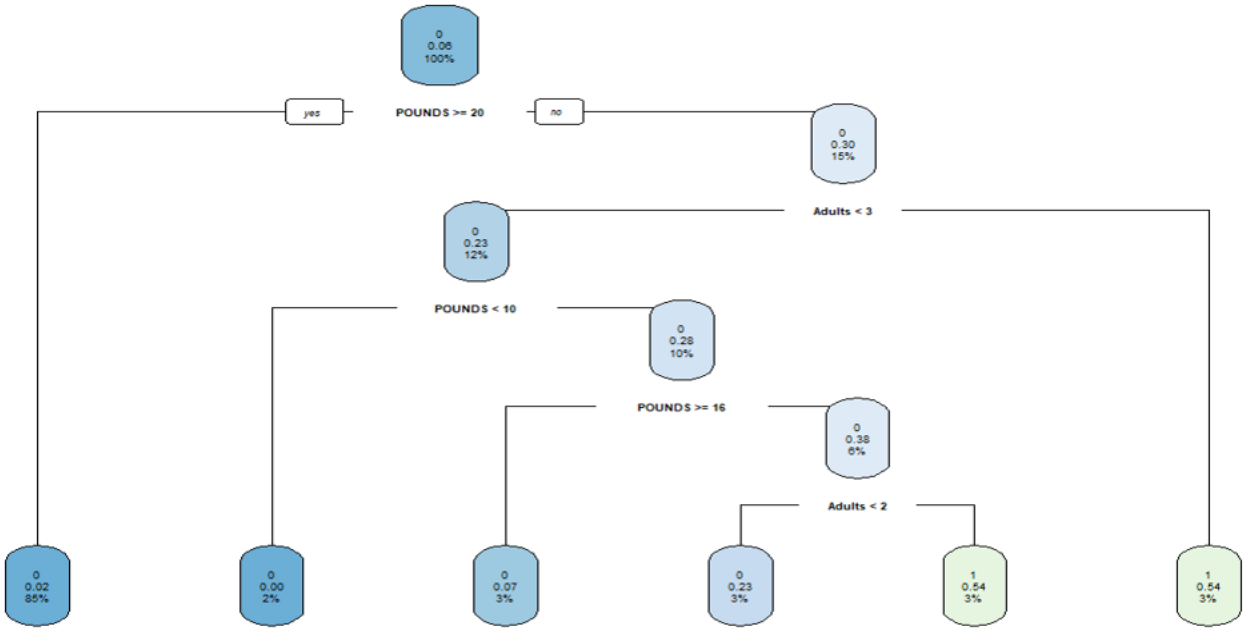


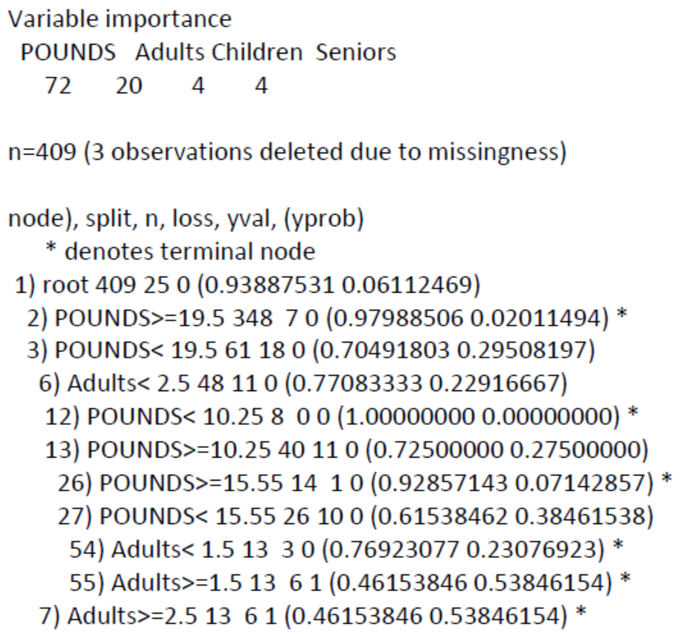


```
Variable importance
POUNDS    100
n= 367
node), split, n, loss, yval, (yprob)
      * denotes terminal node
1) root 367 30 0 (0.91825613 0.08174387)
  2) POUNDS>=16.4 341 13 0 (0.96187683 0.03812317) *
  3) POUNDS< 16.4 26  9 1 (0.34615385 0.65384615) *
```

Figure11: Decision tree model of FY17

From the above Figure11, again, the decision tree model results of FY17 show that "POUNDS" is statistically significant to the "2nd Visit", which is consistent with the logistic regression results. The model predicts that if the household picks up food at least 16.4 pounds during the 1st visit, then the probability that they will come to have the 2nd visit is 0.04; if the household picks up food less than 16.4 pounds during the 1st visit, then the probability that they will come to have the 2nd visit is 0.65. This provides the valuable input of the number "16.4 pounds" for the question "In general how much food should we prepare for the first visit?"





```
Variable importance
 POUNDS   Adults Children  Seniors
    72      20       4       4

n=409 (3 observations deleted due to missingness)

node), split, n, loss, yval, (yprob)
     * denotes terminal node
 1) root 409 25 0 (0.93887531 0.06112469)
   2) POUNDS>=19.5 348  7 0 (0.97988506 0.02011494) *
   3) POUNDS< 19.5 61 18 0 (0.70491803 0.29508197)
     6) Adults< 2.5 48 11 0 (0.77083333 0.22916667)
      12) POUNDS< 10.25 8  0 0 (1.00000000 0.00000000) *
      13) POUNDS>=10.25 40 11 0 (0.72500000 0.27500000)
        26) POUNDS>=15.55 14  1 0 (0.92857143 0.07142857) *
        27) POUNDS< 15.55 26 10 0 (0.61538462 0.38461538)
          54) Adults< 1.5 13  3 0 (0.76923077 0.23076923) *
          55) Adults>=1.5 13  6 1 (0.46153846 0.53846154) *
     7) Adults>=2.5 13  6 1 (0.46153846 0.53846154) *
```

Figure12: Decision tree model of FY18

For the FY18 data, the decision tree model results show that "POUNDS" is the most important variable to affect the "2nd Visit" and the "Adults" variable is the next influence one. This is consistent with the logistic regression results. If the household picks up food of at least 19.5 pounds during the 1st visit, then the probability that they will come to have the 2nd visit is 0.02. If the household picks up food less than 19.5 pounds during the 1st visit, then "Adults" is the second most important variable to affect the 2nd visit. If the number of adults is more than 2.5, then the probability that they will come to have the 2nd visit is 0.54. If the pounds for the 1st visit are less than 19.5 and the number of adults is less than 2.5, then we need to consider the pounds again for the next step.

3. Discussions and Future Study

Through different statistical methods and analytic studes, this paper shows how the Mountain Food Bank's data can be used to answer the following questions:

1. If new service stations are needed in the north Georgia area?

Both the boxplots and hypothesis tests show that the current service locations are quite reasonable in the area for both 2017 and 2018. So new service stations are not necessarily needed.

2. How to identify the family who needs help?

The logistic regression reveals that the second visit is affected by the number of particular family members. For example, the number of kids is significant to the second visit in 2017 and the number of adults is significant to the second in 2018.

3. How to prepare the food amount according to different family types?

Unfortunately, the multiple linear regression does not work well due to the lack of fit caused by too many unusual observations. The problem could be fixed by increasing the sample size or by trying other nonlinear regressions.

4. Does any information indicate the second visit?

The number of household members and the food weight picked up during the first visit can be used to predict the probability of the return. The suggestion for GMFB to prepare for the second visit is to record the information of the household during their first visit.

5. In general how many foods should we prepare for the first visit?

Both the logistic regression and the decision tree show that if the household does not pick up enough foods during the first visit, then most likely it will lead to the second visit. The decision tree model reveals the magic number is 16.4 pounds in 2017 and 19.5 pounds in 2018.

For future studies, the authors will build other statistical models to explore the data. The authors will also work with GMFB to analyze the FY19, FY20, and FY21 data. Especially for the FY20 and FY21 data to find out the influence of Covid-19.

References

[1] Georgia Mountain Food Bank Website, *https://www.gamountainfoodbank.org/*

[2] Statistics for Engineers and Scientists *by Navidi, McGraw Hill, 2020*

[3] An Introduction to Statistical Learning, with Applications in R *by James, Witten, Hastie and Tibshirani, Springer, 2013.*

[4] An outcome evaluation of a food bank program *by Cotugna, Vickery and Glick, Journal of the Academy of Nutrition and Dietetics, VOL 94, ISSUE 8, P888-890, 1994.*

[5] Data Science For Business-What You Need to Know About Data Mining and Data-Analytic Thinking. *Foster Provost and Tom Fawcett.*

[6] Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies. *John D Kelleher, Brian Mac Namee, and Aoife D'arcy.*

[7] Probability and Statistics. *MIchael J Evans and Je_rey S Rosenthal.*

[8] Statistical Design and Analysis of Experiments. *Wiley, NY. 1989.*