A Statistical Analysis of Factors Contributing to Obesity

Avery McKeaney and Grace Flitsch

2023-04-21

Abstract

This research paper analyzes data on the obesity levels of individuals from Mexico, Peru and Columbia and how they are influenced by ones eating, drinking, transportation, and smoking habits. Our goal for this project is to determine which attributes have a strong correlation with obesity (high BMI), in other words which attributes are statistically significant predictors in the data. We created graphs to visualize patterns in the data and then proceeded to run hypothesis tests on the different attributes that have an affect on an individuals BMI. We find that age, number of main meals, consumption of water, physical activity frequency, family history with obesity, consumption of food between meals, and mode of transportation all are statistically significant predictors for BMI.

Introduction

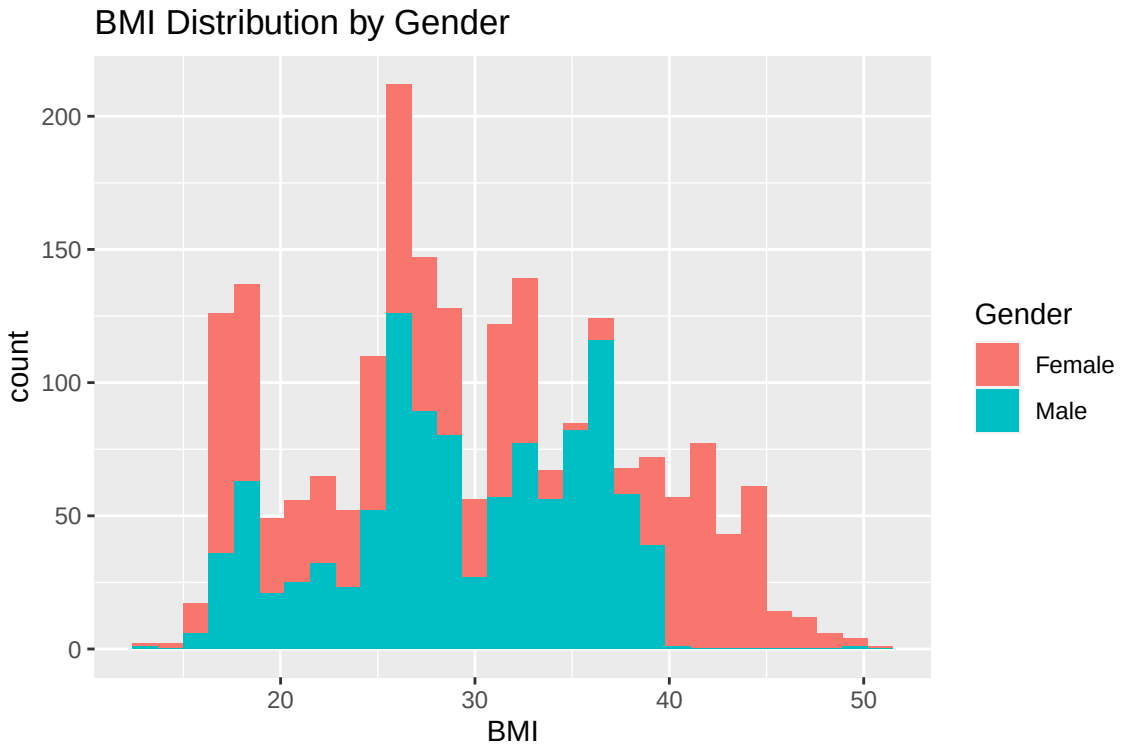
The data set we are analyzing for our project is an estimation of obesity levels from individuals in Peru, Columbia and Mexico. 77% of the data was generated synthetically, while 23% was collected from individuals in the given regions. The synthetically generated data was created using Weka software and the SMOTE (Synthetic Minority Oversampling Technique) filter but the creators of the dataset. This means that this portion of the data is not from real subjects, but it has been generated in a way that helps avoid overfitting. This data was collected and researched by Fabio Mendoza Palechor and Alexis de la Hoz Manotas which can be found in their article, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico"(https://doi.org/10.1016/j.dib.2019.104344 or https://www.kaggle.com/code/mpwolke/obesity-levels-life-style/notebook). The data contains 17 attributes (or variables) and 2111 observations. The attributes are listed in the table below. We are analyzing the affect that these various attributes have on an individual's obesity classification and body mass index (BMI). BMI is not known to be a particularly reliable indicator for over overall health and body fat percentage, but it is the only statistic available to us in our data. A BMI of over 25 is considered overweight, with the obesity classes listed in this project all falling above that threshold. Obesity is a worldwide health crisis, we use statistical analysis to decipher which lifestyle habits may impact a subjects' BMI. This analysis may be helpful in recommending changes one can make to improve overall health. Through regression and hypothesis testing, we report both unexpected and expected correlations in the data, along with determining which variables are statistically significant predictors.

Variable Name	Definition
FAVC	Frequent consumption of high caloric food
FCVC	Frequency of consumption of vegetables
NCP	Number of main meals
CAEC	Consumption of food between meals
CH2O	Consumption of water daily
CALC	Consumption of alcohol
SCC	Calories consumption monitoring
FAF	Physical activity frequency
TUE	Time using technology devices
MTRANS	Mode of transportation

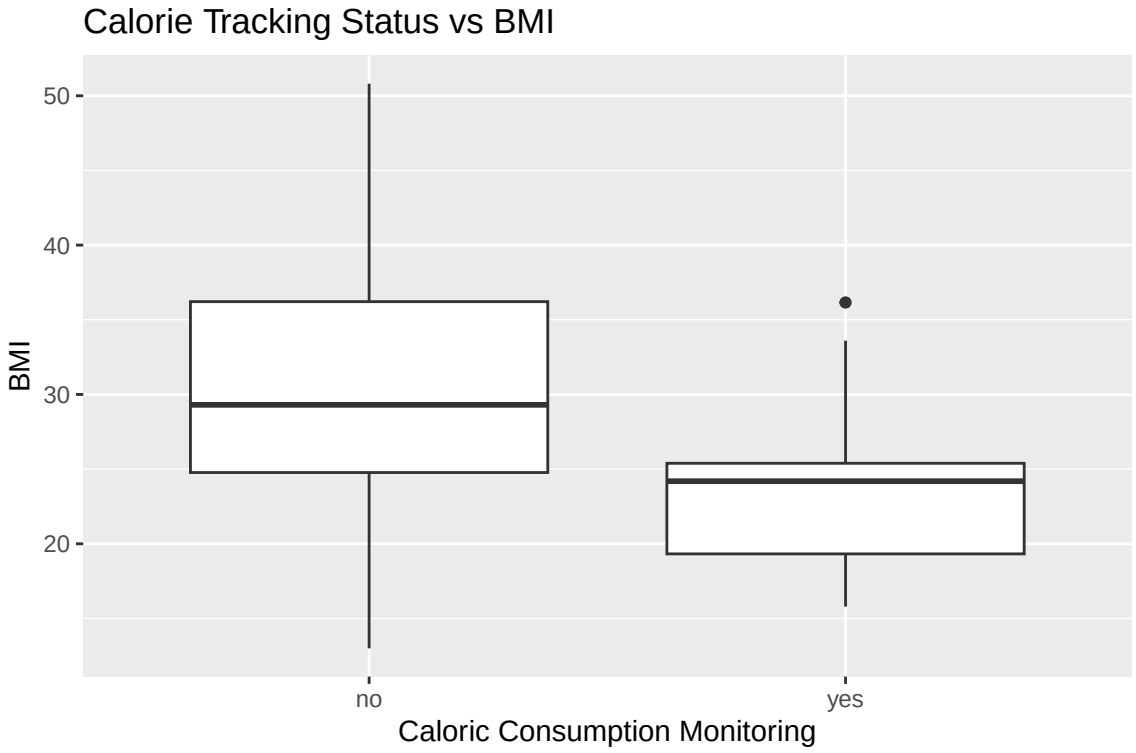
Exploratory Data Analysis

Numerical statistics: There are 2111 observations of 18 variables in the dataset. We created a BMI variable for use in EDA, which is going to be directly correlated with the categorical obesity index outcome variable, so we will not be using that in any predictions. There are no missing values in the dataset. In the data, 49% of the participants are women, and 51% are men. 83% of our data is from people aged 30 or below, which we need to note when making any conclusions. This means that the data is not representative of people of all ages. The distribution of samples across the obesity index outcome variable is approximately even, which is a positive attribute of the dataset. We plotted all of the variables against our outcome variables, with notable plots and comments shown below.

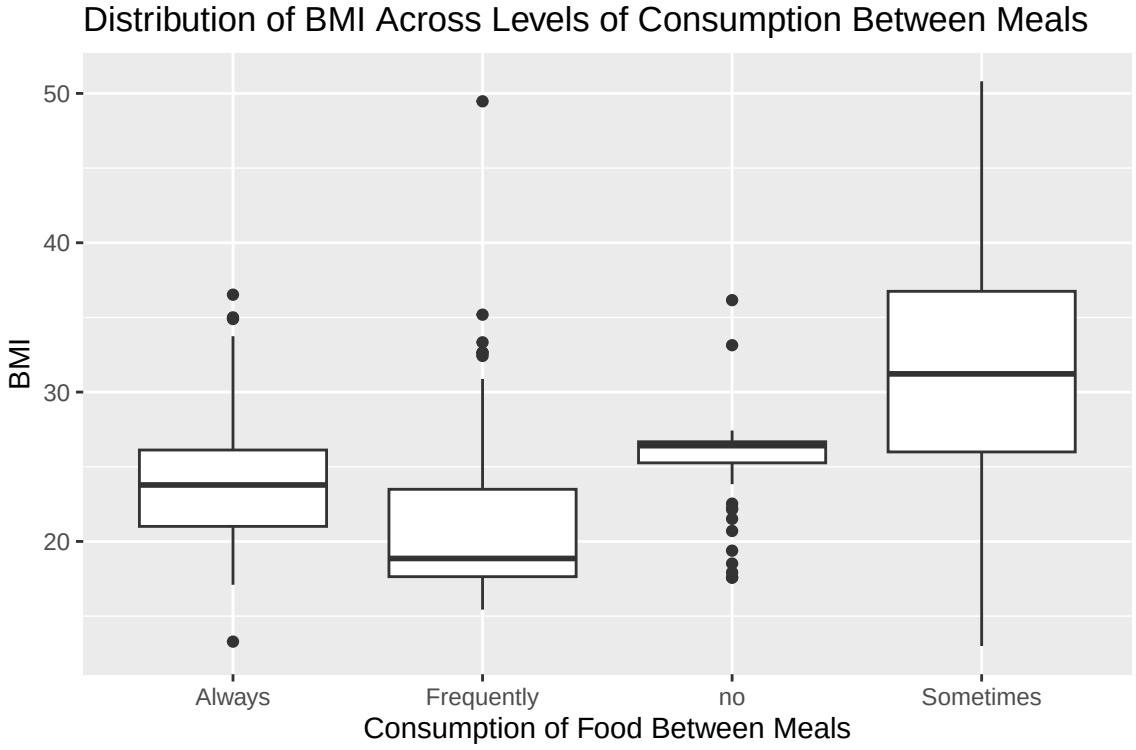
From this plot, we see that as the samples' weight increased, so did the probability that they had a history of family obesity. This was something that we expected to see, but this plot led us to wonder how much that family history affected the expected BMI, which we explored later on.



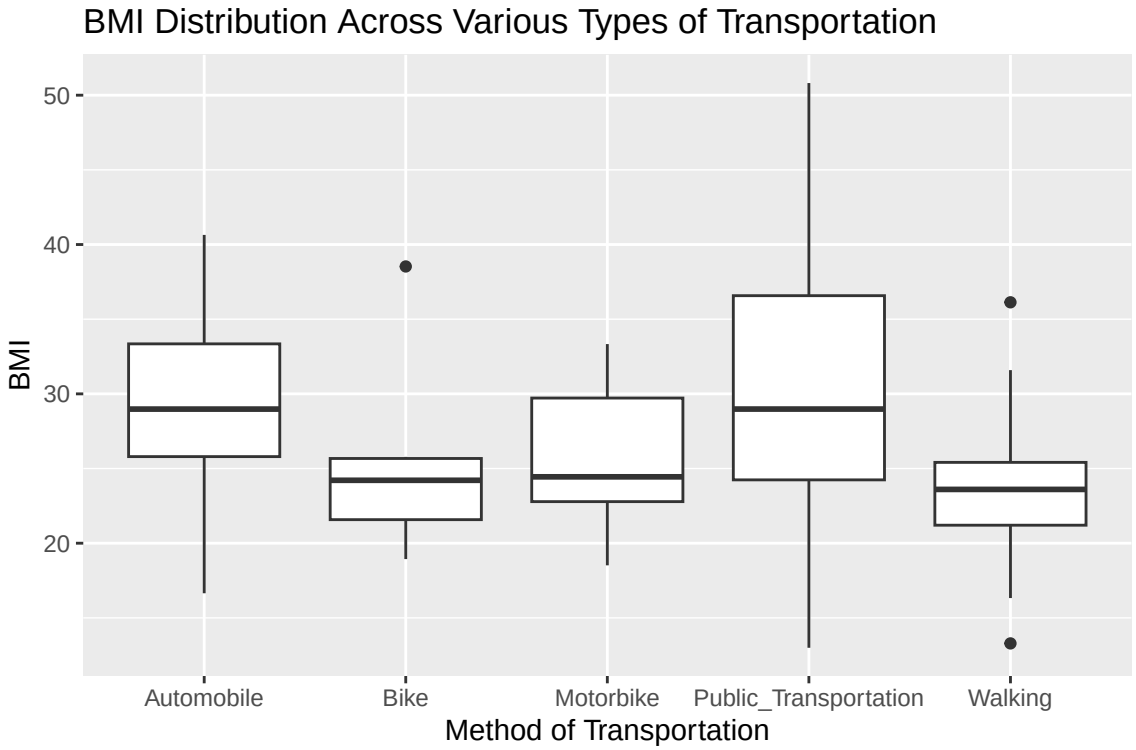
This plot displays a histogram of the BMIs of the people surveyed, colored by gender. We were surprised to see that women seemed to dominate the higher BMIs. Both groups could roughly be described as normally distributed, we do note that women also had low outliers, therefore are less normally distributed.



This plot suggests that subjects who are not monitoring their caloric intake are more likely to be overweight. This didn't immediately make sense to us. We then considered that some individuals track calories at an unhealthy level to maintain extremely low weights, which might skew the median BMI of those who track calories to be lower. We again are curious how different the BMI is predicted to be for an individual who is tracking caloric intake.



These results are interesting because it is not what we expected. The survey results may be skewed depending on the subjects' definition of "sometimes". We think that most people would probably say sometimes as an answer when they are just not sure. This would be an example of response bias, which would mean we may not want to use this predictor. We will still perform a difference of means hypothesis test, since the spread in the responses for "sometimes" is so large.



This was one of the most interesting plots to us. From this, it looks like the median BMI may be the highest for those who take public transportation, and lowest for those who walk. Method of Transportation is one of the only variables we have that may be a predictor of socioeconomic status. There are other factors at play that may have to do with the BMIs being higher for those who take public transit. This difference in BMI does not indicate that more steps from walking make you lose weight, but more likely that those who are wealthy enough to afford a car or to live close enough to work to walk can also afford healthier food, which tends to be more expensive and time consuming to make. This plot motivated us to perform a hypothesis test to check whether the means are different. Looking at the plot, we know we need to check the assumptions, since the variances look very different.

Through our exploratory data analysis, we note some trends that we expected in the data, and others we did not. There are not any significant outliers that we need to deal with. From this section, we are motivated to first examine a multiple regression with all of the variables as predictors. From there we would like to conduct multiple hypothesis tests and create some confidence intervals for how much certain variables contribute to the BMI predictions.

```
##
## Call:
## lm(formula = BMI ~ Age + FCVC + NCP + CH2O + FAF + TUE + family_history_with_overweight +
##     CAEC + Gender + SMOKE + SCC + CALC + MTRANS, data = obesity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -19.0880  -4.0426   0.2419   3.7172  23.7194
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        5.51870    5.98135   0.923  0.35630
## Age                                0.30678    0.02721  11.273  < 2e-16 ***
## FCVC                               3.37933    0.25175  13.424  < 2e-16 ***
## NCP                                0.39541    0.16792   2.355  0.01863 *
## CH2O                               0.63978    0.21865   2.926  0.00347 **
## FAF                               -0.85211    0.15919  -5.353 9.61e-08 ***
## TUE                               -0.40073    0.22231  -1.803  0.07161 .
## family_history_with_overweightyes  7.08862    0.36634  19.350  < 2e-16 ***
## CAECFrequently                    -3.84984    0.89263  -4.313 1.69e-05 ***
## CAECno                             2.39994    1.17183   2.048  0.04068 *
## CAECSometimes                      3.28285    0.82748   3.967 7.52e-05 ***
## GenderMale                        -0.44384    0.27478  -1.615  0.10640
## SMOKEyes                          -0.51638    0.89879  -0.575  0.56567
## SCCyes                            -2.54475    0.63122  -4.031 5.74e-05 ***
## CALCFrequently                    -4.48772    5.84076  -0.768  0.44237
## CALCno                            -5.62601    5.80541  -0.969  0.33261
## CALCSometimes                     -3.05930    5.81035  -0.527  0.59858
## MTRANSPublic_Transportation        4.47781    0.39556  11.320  < 2e-16 ***
## MTRANSWalking                      0.95535    0.85940   1.112  0.26642
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.723 on 2074 degrees of freedom
## Multiple R-squared:  0.4954, Adjusted R-squared:  0.491
## F-statistic: 113.1 on 18 and 2074 DF,  p-value: < 2.2e-16
```

The outcome of our multiple regression model, shown above, suggested that several variables are contributing to the BMI prediction significantly. The R-squared value of .494 is not great, but since we are more focused on variable selection than predictive power, we still used these results. The best predictors for BMI are FCVC, Age, NCP, CH2O, FAF, family_history_with_overweight, CAEC, responding "yes" for SCC, taking a motorbike, and taking public transportation.

```
model2 <- lm(BMI ~ Age + FCVC + NCP + CH2O + FAF + family_history_with_overweight + CAEC + SCC + MTRANS
summary(model2)
```

```
##
## Call:
## lm(formula = BMI ~ Age + FCVC + NCP + CH2O + FAF + family_history_with_overweight +
##     CAEC + SCC + MTRANS, data = obesity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.5976  -4.0848   0.2989   3.7994  24.7274
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.89173    1.42615  -0.625 0.531863
## Age                                0.32815    0.02662  12.329  < 2e-16 ***
## FCVC                               3.71867    0.24457  15.205  < 2e-16 ***
```

```
## NCP                                 0.58689    0.16903   3.472 0.000527 ***
## CH2O                                0.76455    0.22105   3.459 0.000554 ***
## FAF                                -1.11151    0.15779  -7.044 2.53e-12 ***
## family_history_with_overweightyes   6.69145    0.36972  18.099  < 2e-16 ***
## CAECFrequently                     -4.01568    0.90725  -4.426 1.01e-05 ***
## CAECno                              3.17491    1.18433   2.681 0.007403 **
## CAECSometimes                       3.57765    0.83938   4.262 2.11e-05 ***
## SCCyes                             -2.55665    0.63901  -4.001 6.53e-05 ***
## MTRANSPublic_Transportation         4.83151    0.39886  12.113  < 2e-16 ***
## MTRANSWalking                       0.96527    0.86883   1.111 0.266696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.838 on 2080 degrees of freedom
## Multiple R-squared:  0.4734, Adjusted R-squared:  0.4704
## F-statistic: 155.8 on 12 and 2080 DF,  p-value: < 2.2e-16
```

We create another multiple regression, with the predictors with large p-values in the initial model removed. Our standard error decreased from 5.9 to 1.4 upon changing the model. The same predictors are still relevant, but some of their p-values decreased, specifically CAECno.

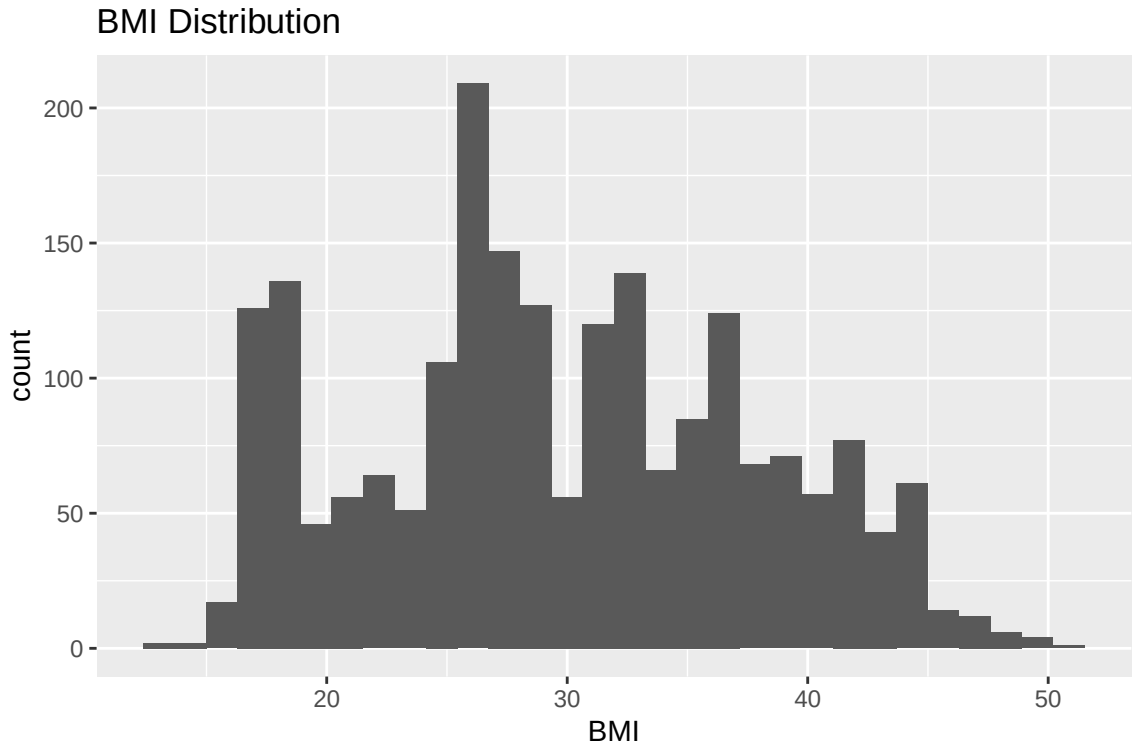
For all of these variables, we have sufficient evidence to suggest that the beta values are different than zero. These values are estimated to be anywhere between -3 and 7. This means that some of these values negatively impact the BMI while others suggest a higher BMI. To break these findings down further, we will conduct Tukey multiple comparison of means.

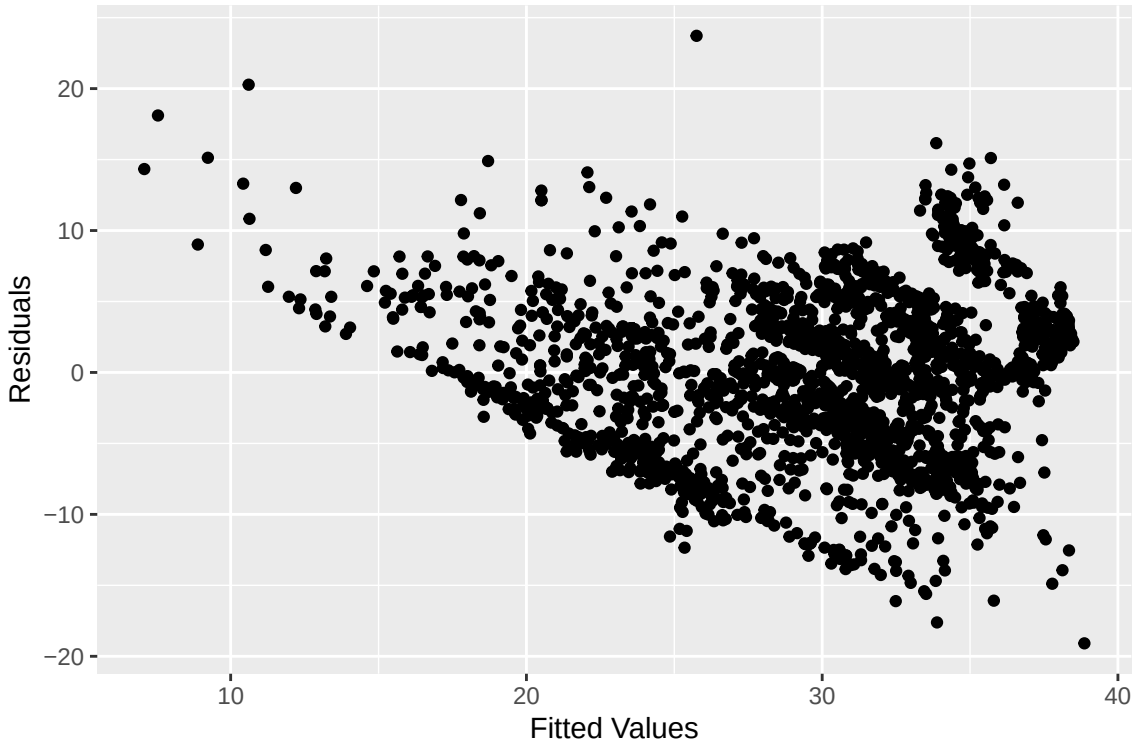
Statistical Analysis

Before modeling, we check the assumptions necessary for hypothesis testing. We meet the large-sample requirement with an n of 2111, and all of the sub categories have greater than 5 observations. The distribution of BMI (below) follows a nearly normal shape, thus we assume that the population as a whole is normally distributed. Each real sample is independent because each individual survey response is required to be from an individual web session. We can assume that synthetically generated data is independent.

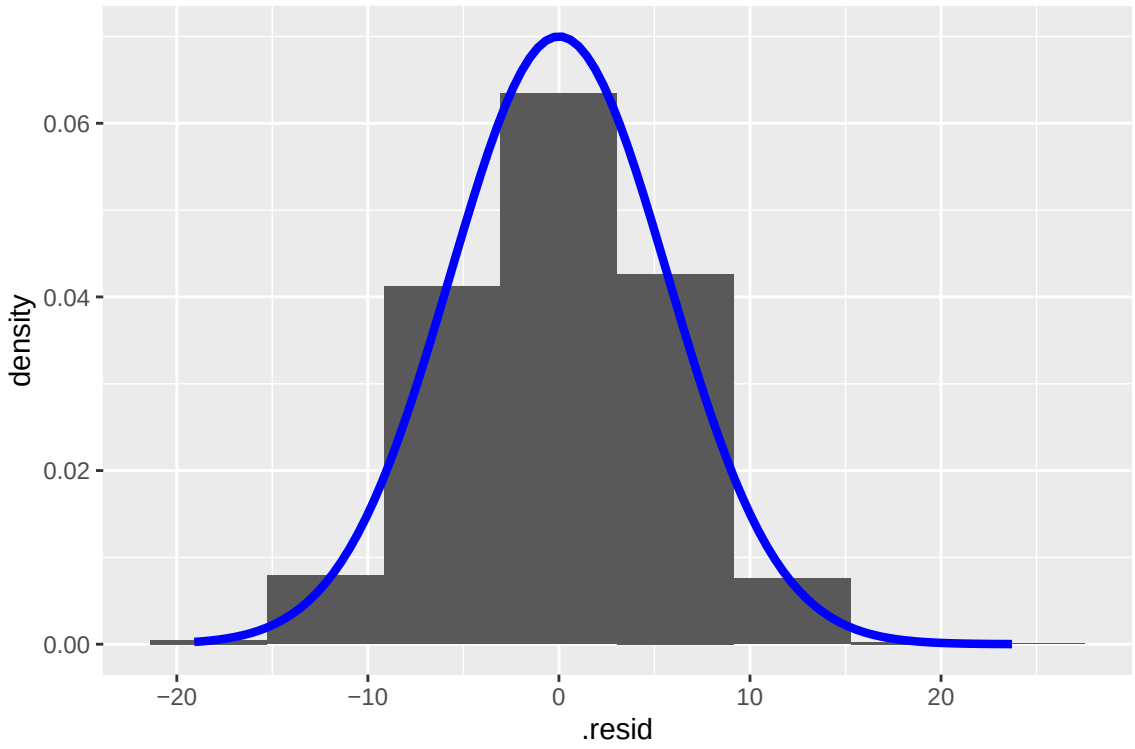
```
obesity %>% filter(MTRANS == "Bike")
```

```
##  [1] Gender                          Age
##  [3] Height                          Weight
##  [5] family_history_with_overweight FAVC
##  [7] FCVC                            NCP
##  [9] CAEC                            SMOKE
## [11] CH2O                            SCC
## [13] FAF                             TUE
## [15] CALC                            MTRANS
## [17] NObeyesdad                      BMI
## <0 rows> (or 0-length row.names)
```





As shown in the residuals plot above, there is a visible pattern in the plot, indicating that the variances change as BMI values increase, and we begin to systematically over predict near the large fitted values. This displays the limitations in our data and signifies that a linear regression model is not a good fit for this data. Instead, a polynomial model might be a better fit for this data. Next, we graphed the residuals to determine if they are normally distributed.



Interestingly, the nearly normal condition of the residuals is met. The distribution is almost perfectly normal, so even though the residual plot had a pattern, we are going ahead with the interpretation of the multiple regression model.

Moving forward with modeling, we first look at the relationship between BMI and family history, since this was the first thing we visualized in EDA.

```
##
##  Welch Two Sample t-test
##
## data:  obesity$BMI by obesity$family_history_with_overweight
## t = -35.756, df = 980.96, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -10.653591  -9.545026
## sample estimates:
##  mean in group no mean in group yes
##          21.46065          31.55996
```

With a small p-value of <2e-16, we can conclude that family history of obesity is a statistically significant predictor and is associated with the obesity level of an individual. We reject the null hypothesis and conclude that there is a difference in mean BMI between the "yes" and "no" family history groups. We are actually 95% confident that an individual with no family history of obesity has a BMI that is predicted to be between 9.4784 and 10.57888 points below one who has family history.

Next, we run a Tukey hypothesis test to test comparisons of means of modes of transportation. We remove the Bike and Motorbike variables since the large sample assumptions were not met by their sample sizes of 8 and 11. All other conditions are met as the sample sizes for each transportation group are greater than 30.

```
##   Tukey multiple comparisons of means
```

```
##     95% family-wise confidence level
##
## Fit: aov(formula = BMI ~ MTRANS, data = obesity)
##
## $MTRANS
##                                         diff        lwr       upr     p adj
## Public_Transportation-Automobile  0.9214343 -0.0692975  1.912166 0.0746191
## Walking-Automobile               -5.5261104 -8.1670215 -2.885199 0.0000030
## Walking-Public_Transportation    -6.4475447 -8.9839363 -3.911153 0.0000000
```

Using these results, we can conclude that there is a difference between the means of different forms of transportation. For example, Walking and Public Transportation have a small p-value of 0, suggesting that we have enough evidence to reject the null-hypothesis that the means are equivalent and can conclude that walking affects the obesity level on an individual differently than Public-Transportation does. In other words, walkers have a lower average BMI than those who take public transit.

Next, we run a T-test to see if there is a difference in means between smokers and non-smokers.

```
##
##  Welch Two Sample t-test
##
## data:  obesity$BMI by obesity$SMOKE
## t = 0.11658, df = 44.605, p-value = 0.9077
## alternative hypothesis: true difference in means between group no and group yes is not equal to 0
## 95 percent confidence interval:
##  -1.959741  2.200482
## sample estimates:
##  mean in group no mean in group yes
##          29.73848          29.61811
```

We do not have enough evidence to reject that the BMI means are the same (p-value=0.97) for smokers and non-smokers. Therefore, smoking is not a statistically significant predictor when determining the obesity level of an individual. We only have 44 observations where the subject is a smoker, which is extremely small, leading to a class imbalance. This means we cannot confidently say that smoking is not a predictor, our data just doesn't have enough data points to draw any conclusions about smoking.

The next T-test we run is on a person's gender and BMI. The sample size of men and women are large and nearly equivalent in the dataset. It is important to note that the Welch's test scales the degrees of freedom to account for any differences in variation.

```
##
##  Welch Two Sample t-test
##
## data:  obesity$BMI by obesity$Gender
## t = 2.2465, df = 1821.4, p-value = 0.02479
## alternative hypothesis: true difference in means between group Female and group Male is not equal to
## 95 percent confidence interval:
##  0.1001272 1.4770989
## sample estimates:
## mean in group Female   mean in group Male
##             30.13239             29.34377
```

After running this hypothesis test, we can conclude that we have enough evidence (p-value=0.01527) to reject mean equivalence for the BMI of women and men in our sample. Also, we are 95% confident that womens' average BMI is 0.1633 to 1.53 higher than mens' BMI.

From our earlier EDA, we were interested to see if the consumption of food between meals did make an impact on the prediction of BMI. Rather than a T-test, we are running a Tukey test to determine if there is a difference in BMI means between people who sometimes eat between meals, people who always do, people who frequently do, and people who never do. It is important to recognize that the number of individuals in each category tested in this Tukey model is greater than 30.

```
model <- aov(BMI~CAEC, data=obesity)
TukeyHSD(model, conf.level=.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BMI ~ CAEC, data = obesity)
##
## $CAEC
##                             diff       lwr        upr     p adj
## Frequently-Always     -3.631230 -6.504440 -0.7580205 0.0064492
## no-Always              0.985726 -2.700316  4.6717681 0.9018746
## Sometimes-Always       6.780750  4.136704  9.4247970 0.0000000
## no-Frequently          4.616957  1.743747  7.4901665 0.0002191
## Sometimes-Frequently 10.411981  9.123800 11.7001624 0.0000000
## Sometimes-no           5.795024  3.150978  8.4390710 0.0000001
```

From our Tukey test, we see that there is a difference in BMI means. This shows that responding "sometimes" gives the model information to predict a higher BMI, even higher than "always". This may mean that people with high BMI's were more likely to respond "sometimes" than "always" which is interesting and shows response bias in the data.

Finally, we test if the means of BMI for the various levels of consumption of alcohol differ. The sample sizes for each alcohol consumption level are all large enough to run a Tukey test (>30).

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BMI ~ CALC, data = obesity)
##
## $CALC
##                              diff        lwr       upr     p adj
## Frequently-Always     4.48980141 -15.701238 24.680841 0.9405133
## no-Always             4.58111486 -15.483110 24.645339 0.9360199
## Sometimes-Always      8.59676489 -11.458785 28.652315 0.6882817
## no-Frequently         0.09131344  -2.434340  2.616967 0.9997123
## Sometimes-Frequently 4.10696348   1.651175  6.562752 0.0001052
## Sometimes-no          4.01565004   3.053394  4.977907 0.0000000
```

With a p-value of 0.0001268 and 0, respectively, we have enough evidence to reject the null hypothesis, suggesting a difference in BMI means between an individual who frequently, sometimes, and never consumes alcohol.

Conclusion

In our research, we have built a multiple regression model that has an acceptable residual plot. This model shows us that there are 12 significant predictors for BMI in the given variables. These predictors are

age, FCVC, NCP, CH2O, FAF, family_history_with_overweight, CAEC, SCC, and MTRANS. We show through hypotheses and analysis of variance testing which variables were predictive, and gave 95% confidence intervals for the estimated effect on BMI. As shown in the results we analyzed, there are several limitations we encounter in our data. For example, there is both response and non-response bias in the sample as individuals had the choice to respond and those responding did not necessarily respond accurately. That being said, the sample wasn't a good and accurate representation of the population in Columbia, Peru and Mexico. Additionally, many of the attributes collected are difficult to measure and were not specific enough to come to an accurate conclusion. Also, the number of observations that fall in each "level" for specific variables are not all equivalent, making it difficult to accurately analyze the data. Finally, the sample collected was from individuals mainly under the age of 45. This has an affect on the results of our data as the sample is not an accurate representation of the population. We are unaware of the effect that these attributes may have on individuals who are "old". After concluding our research on obesity in specific regions, we are left with several questions: Are the results consistent with obesity in other countries? Are there additional attributes that were not analyzed that contribute significantly to one's health and body weight? How would our results change if we collected more data from "seniors"?