Technical Report - Stat 320

Darwin Stahlback , Ani Lamichhane, Karen Vuong

12/22/2021

DETERMINING OBESITY

glm(formula = is_obese ~ gender + age + height + family_history + food_between_meals + main_meals + monitor_calories + transportation, family = binomial, data = obesityuse)

Abstract:

This project aimed to utilize data collected via a University de la Costa web survey and from the World Health Organization (WHO) on life choices in central and south American countries to predict obesity. This data was used to create a logistic regression model to infer that of the multiple predictors provided, that meal habits, precisely the number of meals one chose to eat, had the most impact on determining if a person became obese. Our team used backward elimination to remove eleven of the seventeen variables that we found were not statistically significant. The resulting model allowed us to predict the odds of becoming obese in the countries of Mexico, Columbia, and Peru in 2019.

Introduction:

Obesity is becoming a major global issue in society. Obesity is a condition for a person who gains an excessive amount of fat that takes their Body Mass Index, a calculation based on height and weight, above thirty. Being obese can lead to multiple health issues, including Diabetes Type 2, sleep apnea, high blood pressure, heart and liver disease, pregnancy complications in the mother and child, and risk of a heart attack. Although genetics can play a minor factor, obesity is mainly preventable through personal lifestyle choices. By determining which predictors provide the best chance of reducing obesity, a person can increase, decrease, or eliminate the predictor's role in not being obese in their own lives.

Data:

The data set is a collection of observations to estimate obesity levels in individuals between the ages of 16 and 64 from Mexico, Peru, and Colombia, based on their eating habits and physical condition. The information was gathered by the University de la Costa via an online web survey over 30 days through the tool Weka covering seventeen questions related to their personal and life choices. The resulting dataset contains 2111 records that the University cleaned up of incomplete and invalid values. The response variable is whether or not a person is obese and was generated by calculating BMI (Weight in Kilograms / (Height in Meters x Height in Meters) then applying if BMI is equal to or greater than 30 variable is True/1. The significant predictor variables taken from the dataset based on this reports were two numerical predictors (Age, and Height) and five categorical: gender (binary), family history of obesity (binary), transportation used (categorical), if they eat a main meal (categorical) and if one monitored their caloric intake (binary).

Although this data set was thorough, data wangling was still needed before running a model. Some of the variable names were either too long or required to be renamed to make better sense of what was being analyzed. The response variable of obesity was also in different obesity level categories which posed two problems. The first issues being that the response was not binary. to make it so calculations for BMI were done. After finding BMI to get the response binary, since the BMI threshold for obesity is 30 a new variable "is_obese" was made with 1 if the participant had a BMI of 30 or greater and 0, if participant's BMI was lower.

```
obesity <- obesity %>%  clean_names() #all column names to lower case
obesity$caec <- tolower(obesity$caec) #all column values to lower case
obesity$mtrans <- tolower(obesity$mtrans)#all column values to lower case
obesity$n_obeyesdad  <- tolower(obesity$n_obeyesdad)#all column values to lower case

#this takes original n_obeyesdad categories back to numeric
obesity <- obesity %>%
  mutate(massbodyindex = obesity$weight/(obesity$height*obesity$height))
# creates binary if obese 0 = no / 1 = yes
obesity$is_obese = ifelse(obesity$massbodyindex<30,0,1)
is.num <- sapply(obesity, is.numeric)
# format all numeric column to three places
obesity[is.num] <- lapply(obesity[is.num], round, 3)
```

```
#renaming variables
colnames(obesity)[colnames(obesity) == "family_history_with_overweight"] = "family_history"
colnames(obesity)[colnames(obesity) == "favc"] = "food_between_meals"
colnames(obesity)[colnames(obesity) == "caec"] = "main_meals"
colnames(obesity)[colnames(obesity) == "scc"] = "monitor_calories"
colnames(obesity)[colnames(obesity) == "calc"] = "alcohol"
colnames(obesity)[colnames(obesity) == "mtrans"] = "transportation"
```

```
obesityuse = obesity%>%
  select(gender, age, height, weight, family_history, food_between_meals,
         main_meals, smoke, monitor_calories, alcohol, transportation, is_obese)
```

Variable sectional was done through a process of backwards elimination by hand leading to the final model. There were also a few obstacles while coming to the final model. First, not all the polytomous variables are all significant to the model. Such as how "main_mealsfrequently" is significant to the model but "main_mealsno" is not. To come to our final model, we compared it to others without "main_meal" and "transportation" which both have non-significant variables. In our comparison, we did a nested likelihood test of the models and compared AIC values. Another issues while finding the model was the predictor weight, which caused the algorithm to not converge when running the logistic regression, this was do to weight being tied so closely to obesity itself. If someone is 400 pounds they will be obese.

```
finalmodel = glm(formula = is_obese ~ gender + age + height + family_history +
    food_between_meals + main_meals + monitor_calories + transportation,
    family = binomial, data = obesityuse)
summary(finalmodel)
```

```
##
## Call:
## glm(formula = is_obese ~ gender + age + height + family_history +
##     food_between_meals + main_meals + monitor_calories + transportation,
##     family = binomial, data = obesityuse)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3108  -0.6408  -0.0454   0.8081   3.9639
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                          -14.33923    1.58161  -9.066  < 2e-16 ***
## genderMale                            -0.56890    0.14303  -3.977 6.97e-05 ***
## age                                    0.11043    0.01199   9.208  < 2e-16 ***
## height                                 2.63341    0.83732   3.145 0.001661 **
## family_historyyes                      3.59157    0.37848   9.489  < 2e-16 ***
## food_between_mealsyes                  1.95822    0.26281   7.451 9.26e-14 ***
## main_mealsfrequently                  -2.02824    0.57553  -3.524 0.000425 ***
## main_mealsno                          -0.80636    0.88471  -0.911 0.362068
## main_mealssometimes                    1.33660    0.43644   3.062 0.002195 **
## monitor_caloriesyes                   -2.12266    0.63673  -3.334 0.000857 ***
## transportationbike                     0.26834    1.69853   0.158 0.874470
## transportationmotorbike                2.58546    0.95646   2.703 0.006868 **
## transportationpublic_transportation    1.53575    0.17387   8.833  < 2e-16 ***
## transportationwalking                 -0.85450    0.67043  -1.275 0.202463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2913.9  on 2110  degrees of freedom
## Residual deviance: 1958.1  on 2097  degrees of freedom
## AIC: 1986.1
##
## Number of Fisher Scoring iterations: 7
```

The model passes the conditions of independence, randomness, and linearity of the logit. The model does not violate independence because the rows do not include time or spatial units. Randomness is met because whether someone has obesity is a random factor.

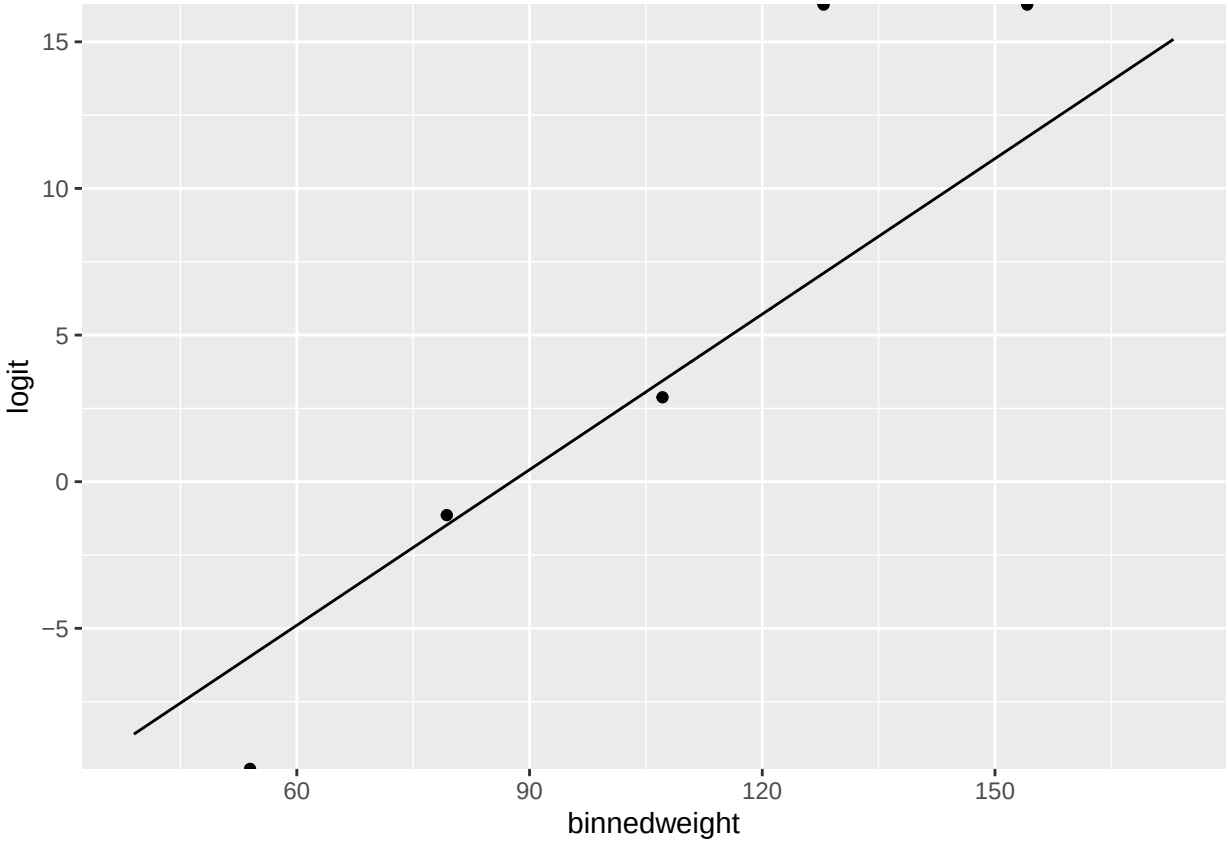
```
slogweight1 = glm(is_obese ~ weight, data = obesityuse, family = binomial)

slogweight <- augment(slogweight1, data = obesityuse)
slogweight <- slogweight %>%
  mutate(odds = exp(.fitted),
         probability = odds / (1 + odds))
```

```
obesityuse <- obesityuse %>%
  mutate(weightGroup = cut(weight, breaks = 5))
```

```
obesityuse_binned <- obesityuse %>%
  group_by(weightGroup) %>%
  summarize(binnedis_obese = mean(is_obese), binnedweight = mean(weight)) %>%
  mutate(logit = log(binnedis_obese/(1-binnedis_obese)))
```

```
ggplot(obesityuse_binned) +
  geom_point(aes(x = binnedweight, y = logit)) +
  geom_line(data = slogweight, aes(x = weight, y = .fitted))
```



At first glance linearity does not look to be met, but the points at the very top of the plot are the ones where everyone is obese, so it is 100%. As said before, if someone is 400 pounds, they will be obese, and looking at the points near the center they follow a linear trend. So, the linearity condition is met.

Results:

The final model is best used for predicting obesity in countries of Mexico, Peru, and Colombia, rather than making inference about a larger population.

According to our model and interrupted in the odds space:

```
exp(coef(finalmodel))
```

```
##                          (Intercept)                           genderMale
##                         5.923114e-07                         5.661471e-01
##                                  age                               height
##                         1.116761e+00                         1.392117e+01
##                    family_historyyes                food_between_mealsyes
##                         3.629087e+01                         7.086735e+00
##                  main_mealsfrequently                          main_mealsno
##                         1.315667e-01                         4.464821e-01
##                  main_mealssometimes                 monitor_caloriesyes
##                         3.806072e+00                         1.197130e-01
##                     transportationbike             transportationmotorbike
##                         1.307792e+00                         1.326943e+01
## transportationpublic_transportation                transportationwalking
##                         4.644814e+00                         4.254944e-01
```

Compared to female's males are associated with multiplying the odds of having obesity by a factor of 0.57773 holding all else constant.

Compared to people that do not eat food between meals people that do are associated with multiplying the odds of having obesity by a factor of 6.80202 holding all else constant.

Compared to people that do not monitor their calories People that do are associated with multiplying the odds of having obesity by a factor of 0.08072 holding all else constant.

Compared to people that always eat a main meal, people that frequently do are associated with multiplying the odds of having obesity by a factor of 0.14302 holding all else constant.

Compared to people that always eat a main meal, people that do not are associated with multiplying the odds of having obesity by a factor of 0.28873 holding all else constant.

Compared to people that always eat a main meal, people that sometimes* do are associated with multiplying the odds of having obesity by a factor of 4.89057 holding all else constant.

Compared to people that use an automobile, people ride a bike* are associated with multiplying the odds of having obesity by a factor of 0.72758 holding all else constant.

Compared to people that use an automobile, people that ride a motorbike* are associated with multiplying the odds of having obesity by a factor of 7.85224 holding all else constant.

Compared to people that use an automobile, people that use public transportation are associated with multiplying the odds of having obesity by a factor of 5.25079 holding all else constant.

Compared to people that use an automobile, people that walk* are associated with multiplying the odds of having obesity by a factor of 0.41876 holding all else constant.

*these categorical variables are not significant in our multiple regression

```
ll.null = finalmodel$null.deviance/-2
ll.proposed = finalmodel$deviance/-2

(ll.null - ll.proposed)/ll.null
```

```
## [1] 0.3280139
```

Although our model built with significant variables after calculating the McFadden Pseudo R^2 our model was only able to explain 32.80% of variability.

Conclusion:

The data shows that people that frequently or don't have a main meal are less likely to be obese because there is less food consumption compared to people who always have a main meal. This is probably due to the fact that people that are sometimes eating their main meals being more susceptible to eating snacks throughout the day leading to more food consumption overall.

Compared to people that use a car the data shows, people that use public transportation are most likely to be obese which was surprising because people that use public transportation typically have to walk to get to the transport which is exercise which should lead to fat loss. The data shows the opposite this is due to the fact that people that are using frequently public transportation come from a lower income backgrounds and being in poverty causes people to consume less nutritional food.

The data also found that women are more likely to be obese than men this is probably because women in Central and South America take on the traditional role of a caretaker in their families. Which causes them to be more likely to be a single parent which leads to less income and less quality of food.

It also shows that people that monitor their calories are less likely to obese because they are watching what they eat so they're less likely to be obese.

Lastly it shows that who eat food between meals are more likely to be obese this is probably do to due to more overall food consumption.

The project goal was to test whether our variables have a statistically significant impact on obesity. We expected that snacking outside of meals has the most significant effect on obesity. In the leftover predictors, age and height were statistically significant as predictors of obesity; however, they are variables that cannot be controlled. We concluded that gender, age, height, food between meals, main meals, monitoring calories, and transportation significantly impact obesity. This model is limited because socio-economic factors skew the result if attempted to be applied outside of South and Central America compared to Central Europe and the United States, for example, which are in general wealthier.

As the data set had already been purged of missing data before this project, all results should be consistent when attempting to infer data surrounding countries like Mexico , Peru, and Columbia.

References:

Data set for estimation of obesity levels based on eating habits and

physical condition in individuals from Colombia, Peru and Mexico(3 May 2019)

Fabio MendozaPalechor, Alexis de la HozManotas

Universidad de la Costa, CUC, Colombia

https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub#tbl1