

STAT-X 498 Practicum 1 SAP

Introduction

We have been given several health measurements of employees working at a call center. Over eight months, the metrics have tracked information including work shift time, exercise time, weight gained (binary yes/no measurement), amount of weight gained, total metabolic minutes, and more. We hope to figure out whether certain factors, specifically total metabolic minutes and shift, play a role in employees' weight gain.

Specific Aims

1. To show what effect, if any, total metabolic minutes has on weight gain.
2. To show what effect, if any, shift has on weight gain.

Analysis Population

We define the call center employee population as the analysis population. In order for individuals to be considered for the analysis, they must have records of weight gained, shift, and total metabolic minutes (or some combination of records that would allow for calculation of total metabolic minutes) on file.

Descriptive Analysis

The raw dataset contains 392 rows and 83 columns, with variables such as member information, body measurement, exercise time, and weight situation. There are 13 variables that we will use for our analysis, including shift, gender, age, height, weight gain, lbs_gained, pounds_gained, bweight, BMI, Vig Ex Time, Mod Ex Time, Walk Ex Time, and Total Met Min. We will rename some of the variables to make the analysis process more efficient and deal with the missing values using proper means.

Table 1

Potential Variables of Interest

<u>Variable Name</u>	<u>Description of Variable</u>	<u>Early Facts about Variable</u>
weightgain	Binary Y/N variable indicating whether the observation gained weight during the eight-month period	Yes: 237/352 (67.3%) No: 111/352 (31.5%) NA: 4/352 (1.14%)
lbs_gained	Numerical variable indicating how much weight the observation gained in pounds; if the observation did not gain weight, then this takes a value	Median: 15.00 Mean: 16.76 Min: 0.0 Max: 70.0 NA: 120 (most of which need to be

	of NA, which needs to be changed	changed, potentially to 0)
Total_Met_Min	Linear combination of minutes of different rigor levels of exercise per week	Median: 35.50 Mean: 58.12 Min: 1.00 Max: 183.00 Other/Missing: 84
Shift	Categorical variable indicating an observation's work shift time	Early morning (shift beginning from 7 to 9am): 202 Late morning (shift beginning from 10 to 11am): 94 Afternoon (shift beginning from 12 to 2pm): 37 Other: 15 NA: 4
Gender	Description of observation's gender	Female: 248/352 (70.5%) Male: 99/352 (28.1%) NA/Missing: 5/352 (1.4%)
Age	Numerical value indicating observation's age	Median: 31.00 Mean: 33.76 Min: 19.00 Max: 64.00 NA: 30
Height	Numerical value indicating observation's height in inches	Median: 66.00 Mean: 66.64 Min: 57.00 Max: 82.00 NA: 21
BMI	Numerical value indicating observation's body mass index	Median: 26.62 Mean: 27.85 Min: 16.31 Max: 48.25 NA: 99

Outcome Measures

The main outcome measure for this analysis is the pounds gained variable for each observation. We realize that specific numerical values are not listed for those who did not gain weight, so we will replace those values with 0 as a way to show that those individuals did not gain weight. We understand that this may not be an accurate representation of actual pounds lost, but this analysis

is one-sided in nature, with the sole focus being on weight gain, not loss. Therefore, we believe that changing these values to 0 will be an effective way to conduct our analysis.

A secondary outcome measure is the binary Y/N weightgain variable. This variable allows us to do more categorical analysis when numerical input is not available.

Analysis Plan

SA1. To show what effect, if any, total metabolic minutes has on weight gain.

1.1 We must first take care of observations that do not have recorded total metabolic minutes values. For the most part, this should be a simple calculation, because we know that

$$\text{Total_met_min} = 8 * \text{Vig_ex_time} + 4 * \text{Mod_ex_time} + 3.3 * \text{Walk_ex_time}.$$

After filling in this missing data, we can ignore observations that still do not have total metabolic minutes values, because there is not really any useful analysis that we can do without this key variable.

1.2 We will create a simple linear regression model using weight gain as the response variable and total metabolic minutes as the independent variable. This seems to be a sound method because both variables are numerical and continuous, albeit with some degree of rounding.

1.3 We will produce a scatter plot to visually represent the trends between total metabolic minutes and weight gain.

1.4 Depending on the appearance of the scatter plot, we may create different nonlinear models to better fit the relationship between the two variables.

SA2. To show what effect, if any, shift has on weight gain.

2.1 We will clear any remaining observations that have missing shift values, as they serve no purpose in shift analysis.

2.2 As the shift variable is categorical by nature, we will not be able to do extensive numerical analysis the way that we can do with the total metabolic minutes variable. Because of this, we will create a generalized linear model using the binomial distribution family. The two possible responses will be Yes and No in regards to whether weight was gained.

2.3 We will also create a scatter plot between shift and weight gained, although with more transparent points than in the total metabolic minutes plot. We want to see if certain shift values tend to cluster at specific weight gained values.

In addition to these outlined aims and respective plans, we may also use some form of cross validation to pick out other variables that may heavily influence weight gain, applying interaction effects when appropriate. We hope to not just answer the questions posed by the

client, but to also provide additional insights in areas that the client may not have initially considered.

If we find that other variables are significantly indicative of weight gain, we will create additional models, possibly including generalized additive models, which will allow us to smooth numerical variables. Doing so may involve further cleaning of data.

Conclusion

We will apply different methods of analysis to determine whether certain factors, including total metabolic minutes and shift time, affect call center employees' weight gain. We plan to build scatter plots, simple linear models, binomial generalized linear models, and potentially generalized additive models to achieve our goals. We will need to do an ample amount of exploratory data analysis, which may lead us to pursue different or additional methods of analysis. Our final product will be a model or multiple models that will predict weight gain as a binary variable and weight gain as a continuous numerical variable, using different employee health metrics as inputs.