Practicum 1

Peter Claussen Ben Derenge Stephanie Liebl

9/16/2021

Overview

We have been given an Excel data file with data collected for a study from a call center. These data include health metrics from **employees** gathered over an eight month period.

Variables that are of particular interest have been highlighted in the original data. These include variables relating to weight change and demographics. Specifically, we will consider in this preliminary proposal the variables

Highlighted in yellow

- `shift`
- `Total_Met_Min`

Highlighted in orange

- `gender`
- `Age`
- `height`
- `weightgain`
- `lbs_gained`
- `BMI`
- `Vig.ex.Time`
- `Mod.ex.time`
- `Walk.ex.Time`

We will use the convention that text rendered in `sans serif` font denote variable or column names found in the original data set, or derived variables calculated from columns in the original data, while *italics* will denote real-world processes or phenomena of interest. Thus, `weightgain` denotes the data column in the original data file, while *weight gain* denotes some measure of the change in weight by individuals over the study period.

. We have been tasked to provide an analysis to address two specific aims:

- **(SA1)** Does *total metabolic minutes* have an effect on *weight gain*?
- **(SA2)** Does *shift* have an effect on *weight gain*?

We will refer to as **SA1** and **SA2** in further discussion. We address *total metabolic minutes*, *weight gain* and *shift* in the following sections.

Total MET minutes

Total MET Minutes is a composite measure obtained from survey responses (International Physical Activity Questionnaire (IPAQ) short form) to estimate overall physical activity. Physical activity is divided into categories described as *vigorous*, *moderate* and *walking*.
Total MET Minutes can be calculated from 3 data columns. We are given the formula

```
Total_met_min = 8*Vig_ex_time + 4*Mod_ex_time + 3.3*Walk_ex_time
```

to calculate *Total MET Minutes* from the data.

`Total_met_min` contains many missing values. We will create a new data column, `CalcTMM` that is calculated from existing data columns as given in the formula above. To visualize the number of missing values, we plot missing values as 0 in the following figure.

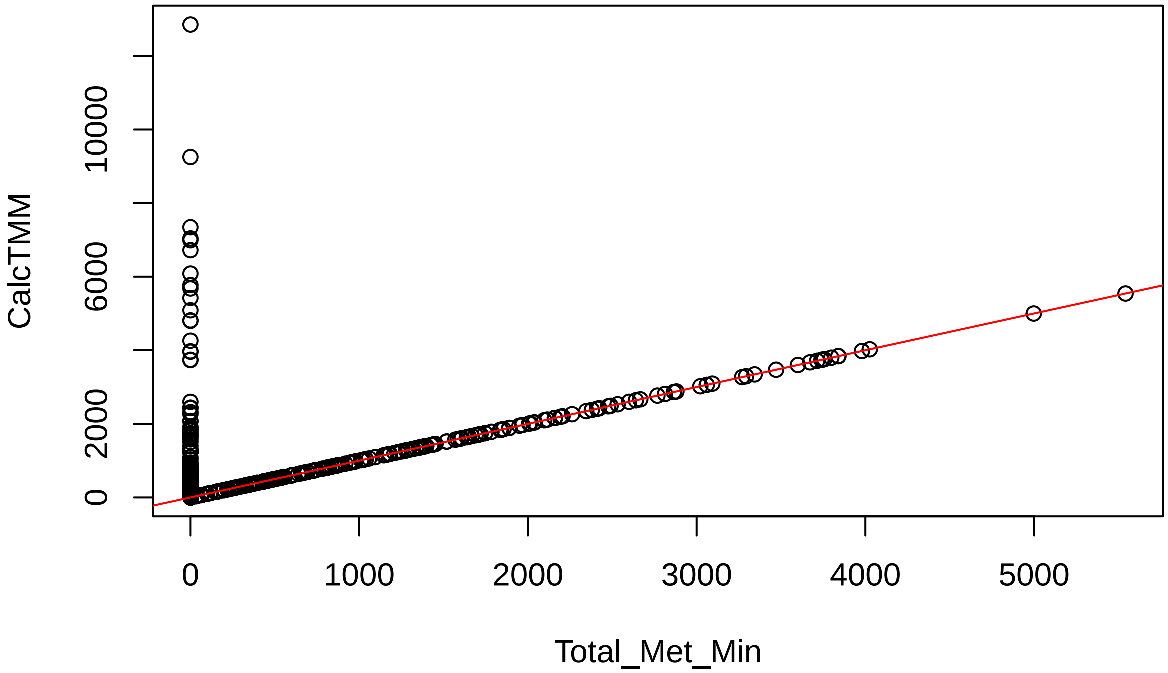


Figure 1: Total MET Minutes calculated from data, plotted against the data column title Total_Met_Min

We can see that the original data column `Total_met_min` has a large number of missing values. If we use this data column as given, we will reduce the number of observations in the analysis. This graph also confirms that the formula given above was used to compute `Total_met_min` from the data.

Shift

Shift takes values of the form `7am`, `8am`, `...`, `2pm`, `other`. There are some columns that have missing values for `shift`. We propose that *shift* be modeled as an ordinal data type, with missing values grouped with `other` and `other` takes an ordinal value greater than `2pm`.

```
metrics.dat$shift[metrics.dat$shift==''] <- 'other'
shift.levels <- c(paste(c(7:11),'am',sep=''),paste(c(12,1:2),'pm',sep=''),'other')
metrics.dat$shift <- factor(metrics.dat$shift,shift.levels)
summary(metrics.dat$shift)
```

```
##   7am   8am   9am  10am  11am  12pm   1pm   2pm other
##    31   115    56    50    44    14     8    15    19
```

Weight Gain

While the data include a variable (data column) named `weightgain`, we are asked to consider other response variables, including change in weight (`lbs_gained`) and change in BMI. Thus, our first task is to determine the appropriate response variable. The choice of response variable will dictate both choice of statistical method (i.e. logistic regression vs linear regression) and methods for data cleaning.

weightgain (Binomial Response)

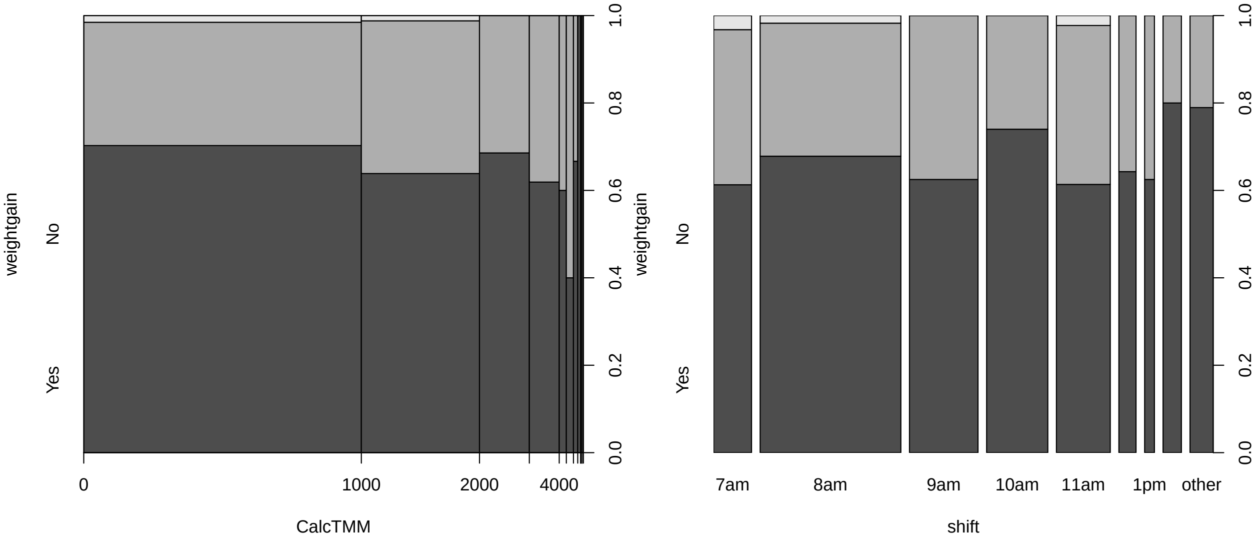


Figure 2: `weightgain` versus Total MET Minutes (calculated) and Shift

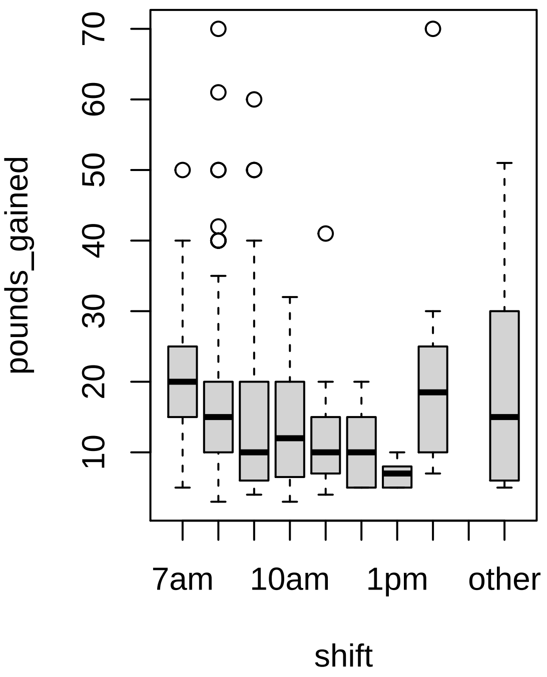
At first glance, this appears to be a logistic regression - the response variable (weightgain) should be binary (Yes/No), ignoring a small number of missing values represented as light gray blocks in the plot. However, we are asked to consider alternate response variables (BMI, pounds gained) as markers for the general response "weight gain’, as opposed to the specific data variable ‘weightgain’

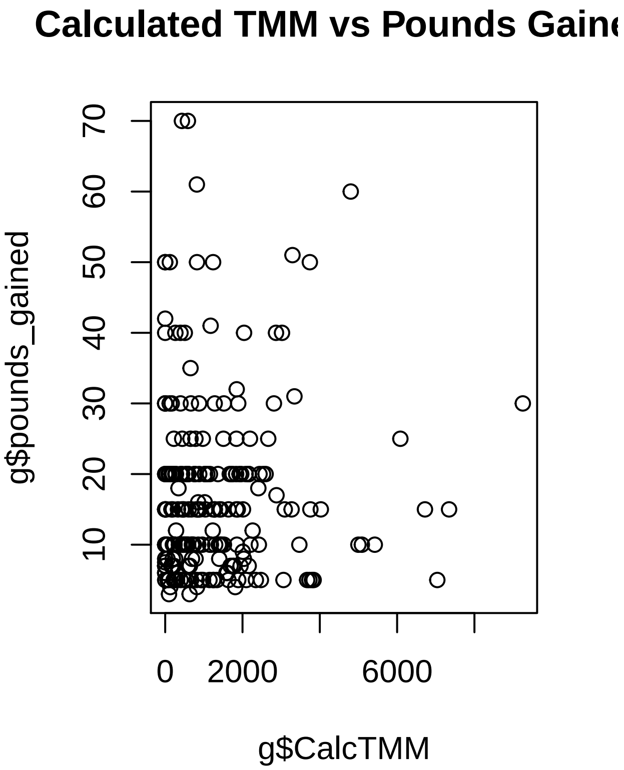
Pounds gained (Continuous Response)

The `pounds_gained` columns contains only non-negative integer values and missing values. This implies every known observation has gained weight. Checking corresponding values in the `weightgain` column confirms all the NA in `pounds_gained` are “No” in `weightgain`. This will affect the interpretation of SA.1 and SA.2. To use `pounds_gained` as the response variable, we would be ignoring all observations who lost weight.

Plots of Pounds_Gained

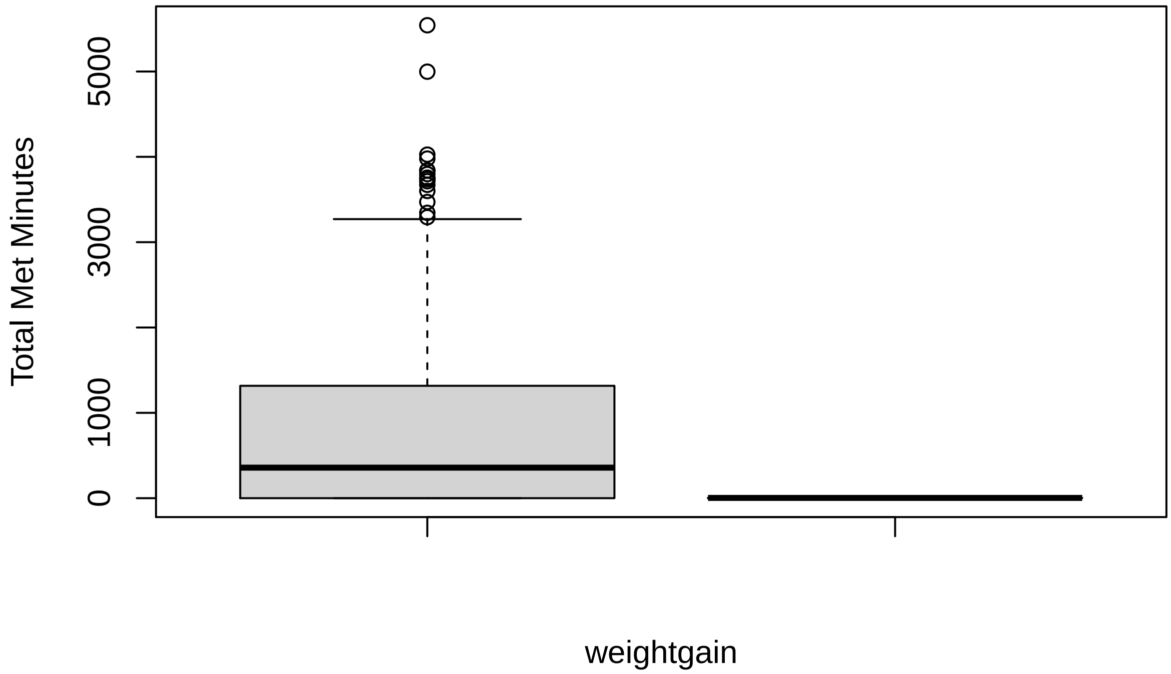
120 missing rows and 1 row zero need to be excluded for the following plots.





The bar plot above suggest shift does have potential to predict pounds_gained. The column 'pounds_gained' seems to decrease on average as shifts move later into the day, with an exception for the 2pm shift. If we don't classify Missing shifts into the "other" category, the missing group has the highest gained weight of all the shifts. As the plot shows some relation around adjacent shifts, we may rank the shifts as opposed to treating shifts like independent factors.

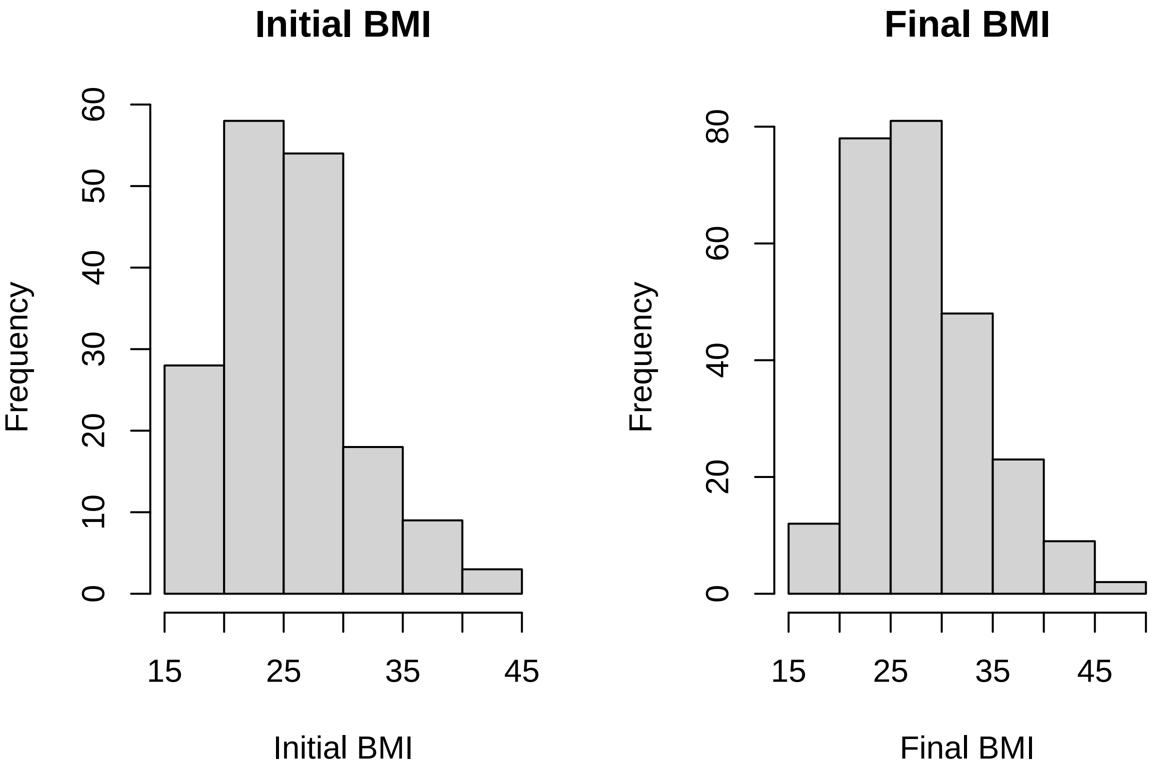
Targeting Binary outcome with `weightgain`



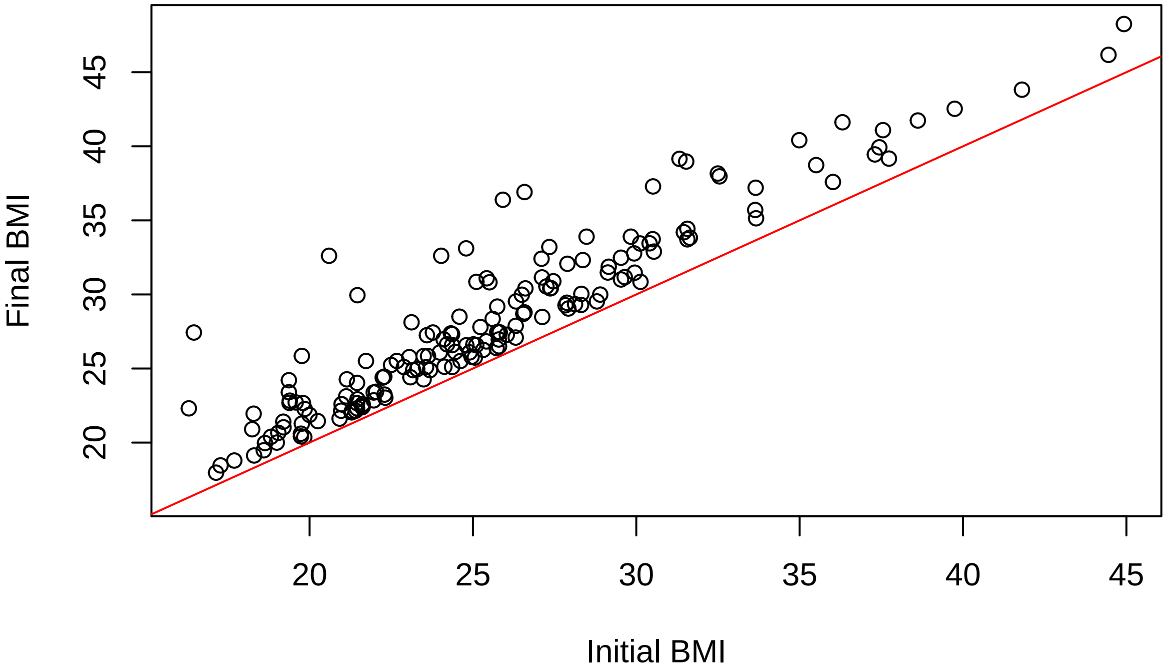
If desired, `weightgain` can be predicted from Total_met_minutes, as most of the “No”s from `weighgain` have NAs or zeros from `total_met_minutes`. This might change if we predict the missing TMM values using the equation from earlier.

Change in BMI

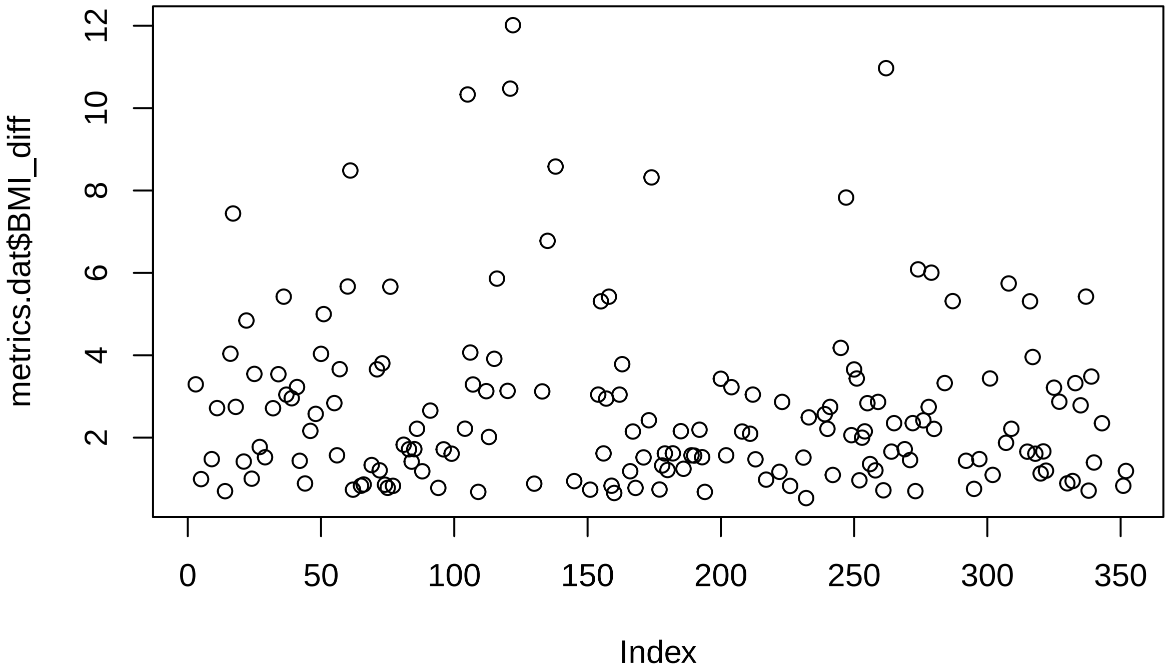
BMI at the end of the study period has been recorded, while initial BMI has not. We can, however, calculate an initial BMI from other data columns. We calculated the initial BMI using the following calculation: `initial_BMI=bweight/(height^2) * 703`



In the scatter plot visualizing (`initial_BMI`) versus final BMI (‘`BMI`’), it appears that there is a strong linear relationship between the variables. The scatter plot also suggests that (`metrics.dat$BMI`) is frequently greater than (`metrics.dat$initial_BMI`).



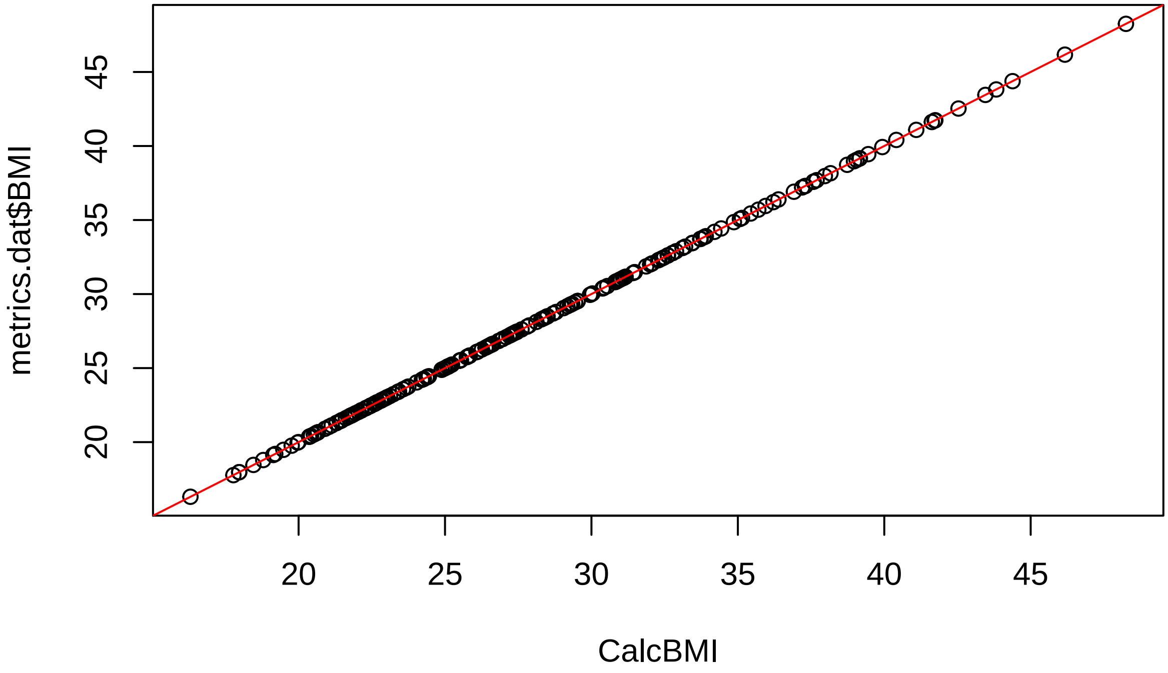
To look at the difference between the initial BMI and the final BMI for call center employees, we created a new column (`BMI_diff`) by subtracting the initial BMI from the final BMI. Thus, in this column, positive values represent increases in BMI over the eight month period.



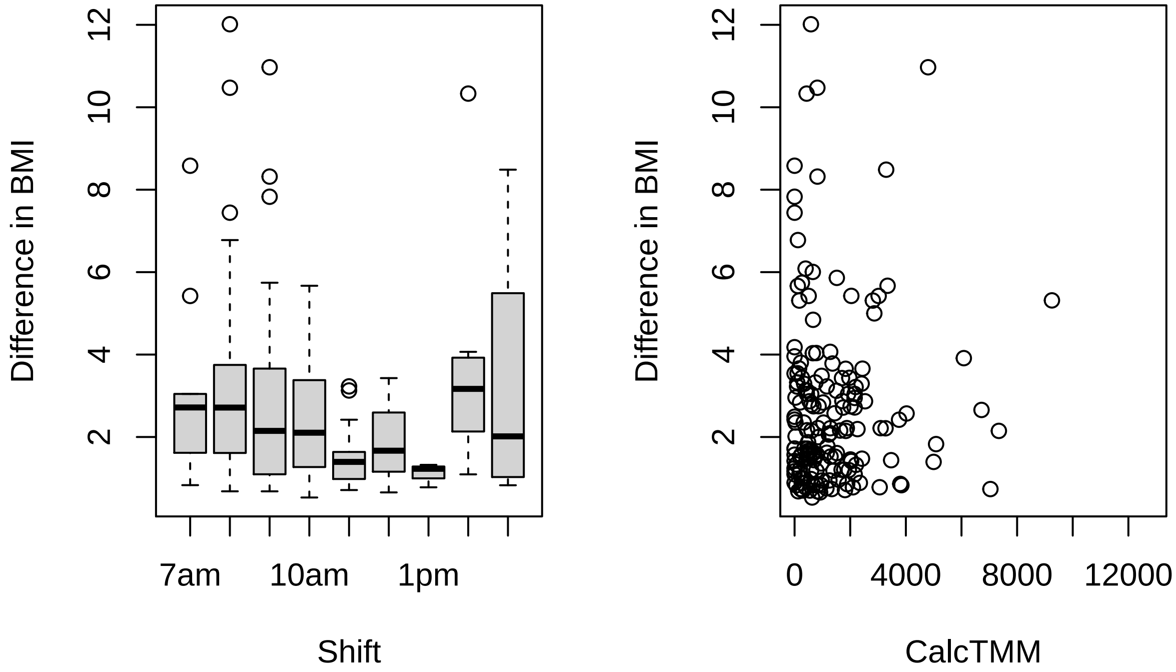
The BMI calculation is dependent on whether weight is a measure of pounds or kilograms. To ensure that the BMI values in the given dataset are correct, we manually calculated the BMI using the following formula: `bweight`/(`height`^2)*703. We then plotted our calculated BMI against the given BMI with a y=x line to verify the values are all the same.

```
# Were the BMI calculations correct?
# BMI = (weight/height^2)*703
CalcBMI <- ''
CalcBMI <- ((metrics.dat$bweight)/(metrics.dat$height)^2)*703

plot(CalcBMI, metrics.dat$BMI)
abline(0,1,col='red')
```



If BMI is chosen to be the response variable, two relationships to be examined are `shift` and `BMI_diff` as well as `CalcTMM` and `BMI_diff`.



```
## Warning in `[<-.factor`(`*tmp*`, metrics.dat$weightgain == "", value =
## structure(c(2L, : invalid factor level, NA generated
```

```
## [1] "Female"  "Male"    "Missing"
```

```
## [1] No   Yes  <NA>
## Levels:  No Yes
```

Points for consideration

Selection of response variable.

We have proposed three response variables of interest as surrogates for weight gain, specifically `weightgain`, `pounds_gained`,and `initial_BMI`.

Pounds Gained

The skew of `pounds_gained` should be addressed. If this is our response variable, we would consider either data transformation, or a generalized linear model of the poisson family.

Using `Pounds_gained` would result in 121 fewer responses.

The box-whisker plot `pounds_gained` vs `shift` shows a curvilinear response. We should be able to create a straightforward model predicting `weight_gained` from shift. The relationship between weight_gained and Calculated `Total Met Minutes` would require further exploratory analysis, as any relationship is not apparent from the scatterplot.

BMI (Difference and Initial)

Similar to the variable lbs_gained, the box-whisker plot of shift vs BMI_diff shows a relationship between the variables. However, this relationship is not as skewed nor as linear as that of shift and lbs_gained. The scatter plot of CalcTMM vs BMI_diff also looks similar to CalcTMM vs lbs_gained, however the relationship between CalcTMM vs BMI_diff appears less linear.

Given our findings, our preference is to use the initial BMI as a predictor variable rather than response for either lbs_gained or weightgain. We would like feedback from the client as to which is preferred moving forward.

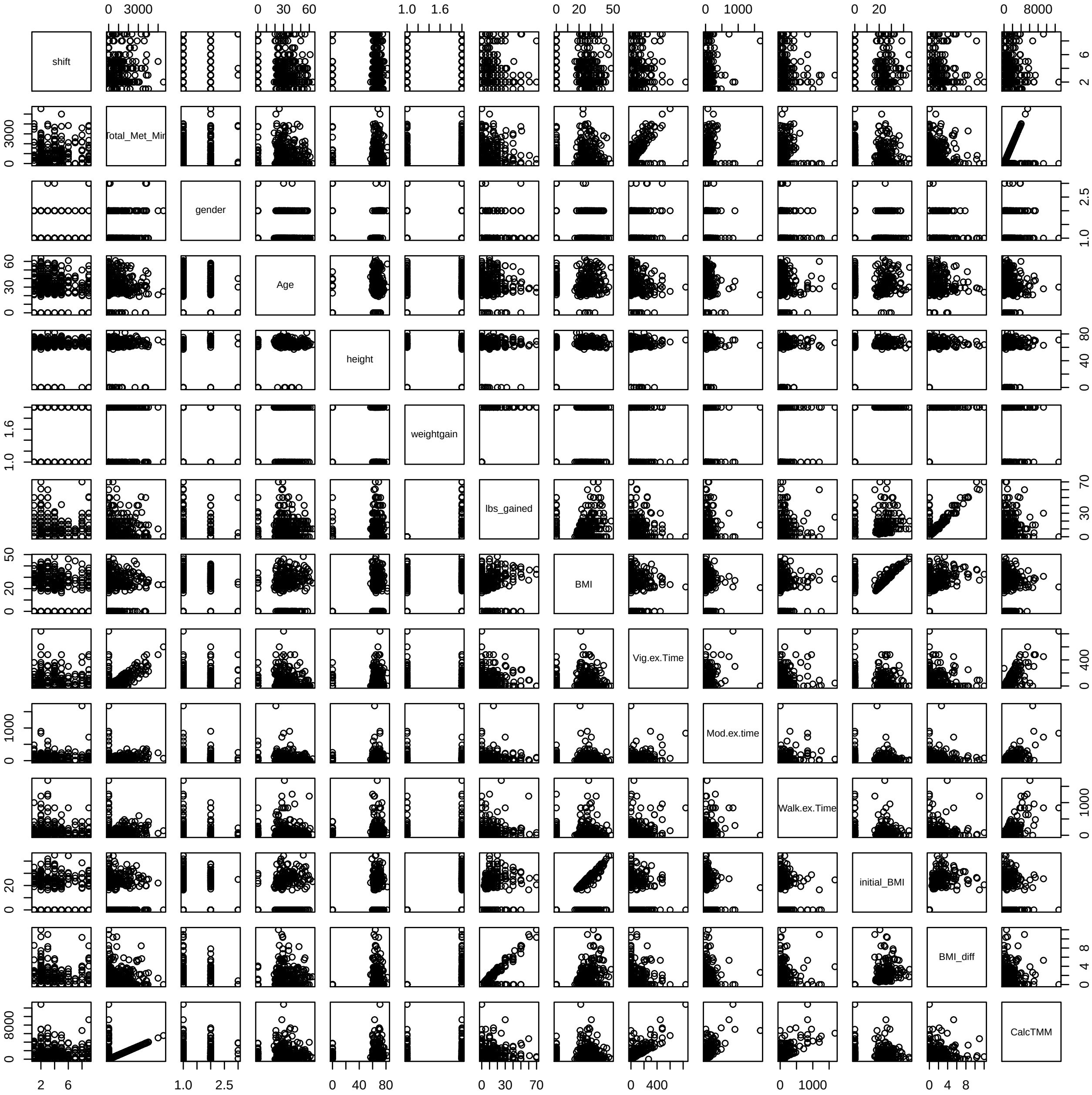
Additional predictor variables.

SA1 and **SA2** include only one predictor variable each (Total_Met_Min and shift, respectively). This implies a simple regression of one-way ANOVA analysis. However, the data includes other demographic variables that may be confounded with the primary ("yellow") predictors

- Does the client wish us to perform multiple regression and variable selection with the "orange" highlighted variables?
- Does the client wish us to perform multiple regression and variable selection with any additional variables in the original data file?

We include, for further discussion, a pairs plot of the variables described in this document:

```
## Warning in `[<-.factor`(`*tmp*`, thisvar, value = 0): invalid factor level, NA
## generated
```



This provides a visualization of the scope of the variable selection problem, if restricted to the highlight variables and additional variables described in this document. We also provide as an addendum a summary of the original data and calculated columns, as used in this document.

```
##       shift      Total_Met_Min        gender          Age             height     
##  8am     :115   Min.   :   0.0   Female :248   Min.   : 0.00   Min.   : 0.00  
##  9am     : 56   1st Qu.:   0.0   Male   : 99   1st Qu.:25.00   1st Qu.:63.00  
##  10am    : 50   Median : 357.8   Missing:  5   Median :30.00   Median :66.00  
##  11am    : 44   Mean   : 797.4                 Mean   :30.88   Mean   :62.67  
##  7am     : 31   3rd Qu.:1315.5                 3rd Qu.:39.25   3rd Qu.:69.00  
##  other   : 19   Max.   :5542.0                 Max.   :64.00   Max.   :82.00  
##  (Other): 37                                                                  
##  weightgain   lbs_gained          BMI          Vig.ex.Time      Mod.ex.time    
##  No  :111   Min.   : 0.00   Min.   : 0.00   Min.   :  0.00   Min.   :   0.00  
##  Yes :237   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:  0.00   1st Qu.:   0.00  
##  NA's:  4   Median : 7.75   Median :24.12   Median : 27.00   Median :  30.00
##             Mean   :11.05   Mean   :20.02   Mean   : 76.16   Mean   :  73.64
##             3rd Qu.:16.25   3rd Qu.:29.46   3rd Qu.:120.00   3rd Qu.:  90.00
##             Max.   :70.00   Max.   :48.25   Max.   :840.00   Max.   :1680.00
##
##   Walk.ex.Time      initial_BMI       BMI_diff         CalcTMM
##  Min.   :   0.00   Min.   : 0.00   Min.   : 0.000   Min.   :    0
##  1st Qu.:   9.75   1st Qu.: 0.00   1st Qu.: 0.000   1st Qu.:  263
##  Median :  60.00   Median : 0.00   Median : 0.000   Median :  822
##  Mean   : 122.92   Mean   :12.36   Mean   : 1.311   Mean   : 1302
##  3rd Qu.: 136.25   3rd Qu.:24.66   3rd Qu.: 2.107   3rd Qu.: 1746
##  Max.   :1680.00   Max.   :44.93   Max.   :12.014   Max.   :12852
##
```

Finally, a brief note about typesetting. This document was produced in RMarkdown. The original `.Rmd` with R code and additional details of our preliminary analysis is available upon request.