Calorie Consumption During Bicycle Work: A Statistical Analysis of an Incomplete Dataset

Nuno Chicoria, Boris Shilov, Murat cem Kose, Yibing Liu, Robin Vermote

18 March, 2018

Contents

1 Introduction

This project aimed to examine data originally gathered by Macdonald (1914) and conveyed to us by Greenwood and TF (1918), consisting of observations on seven people performing work using a bicycle ergometer, although our current dataset appears to include extra values and data not found in Greenwood and TF (1918), though these values may indeed be present in Macdonald (1914), access to which could not be obtained in a timely manner. Hitherto it shall be assumed that every row in our dataset represents a separate individual, giving a total of 24 separate individuals across 24 rows. The dataset includes three separate measurements - weight of the individuals, calories per hour spent by individuals which serves as a measure of workout intensity, and calories spent during the task.

2 Methods and procedure

2.1 Data exploration

First we load and examine the data.

```
##    weight calhour calories
## 1    43.7    19.0       NA
## 2    43.7    43.0      279
## 3    43.7    56.0      346
## 4    54.6    13.0       NA
## 5    54.6    19.0       NA
## 6    54.6    43.0      280
## 7    54.6    56.0      335
```

```
## 8    55.7    13.0        NA
## 9    55.7    26.0       212
## 10   55.7    34.5       244
## 11   55.7    43.0       285
## 12   58.8    13.0        NA
## 13   58.8    43.0       298
## 14   60.5    19.0        NA
## 15   60.5    43.0       317
## 16   60.5    56.0       347
## 17   61.9    13.0        NA
## 18   61.9    19.0       216
## 19   61.9    34.5       265
## 20   61.9    43.0       306
## 21   61.9    56.0       348
## 22   66.7    13.0        NA
## 23   66.7    43.0       324
## 24   66.7    56.0       352
```

And the summary:

```
##      weight         calhour         calories    
##  Min.   :43.7   Min.   :13.0   Min.   :212  
##  1st Qu.:54.6   1st Qu.:19.0   1st Qu.:276  
##  Median :58.8   Median :38.8   Median :302  
##  Mean   :57.5   Mean   :34.0   Mean   :297  
##  3rd Qu.:61.9   3rd Qu.:43.0   3rd Qu.:338  
##  Max.   :66.7   Max.   :56.0   Max.   :352  
##                                NA's   :8
```

Here are some descriptive statistics.

Some exploratory statistics for all individuals:

```
##                 weight  calhour  calories
## nbr.val        24.0000  24.0000   16.0000
## nbr.null        0.0000   0.0000    0.0000
## nbr.na          0.0000   0.0000    8.0000
## min            43.7000  13.0000  212.0000
## max            66.7000  56.0000  352.0000
## range          23.0000  43.0000  140.0000
## sum          1381.0000 817.0000 4754.0000
## median         58.8000  38.7500  302.0000
## mean           57.5417  34.0417  297.1250
## SE.mean         1.3453   3.3396   11.4669
## CI.mean.0.95    2.7829   6.9085   24.4412
## var            43.4338 267.6721 2103.8500
## std.dev         6.5904  16.3607   45.8677
## coef.var        0.1145   0.4806    0.1544
```

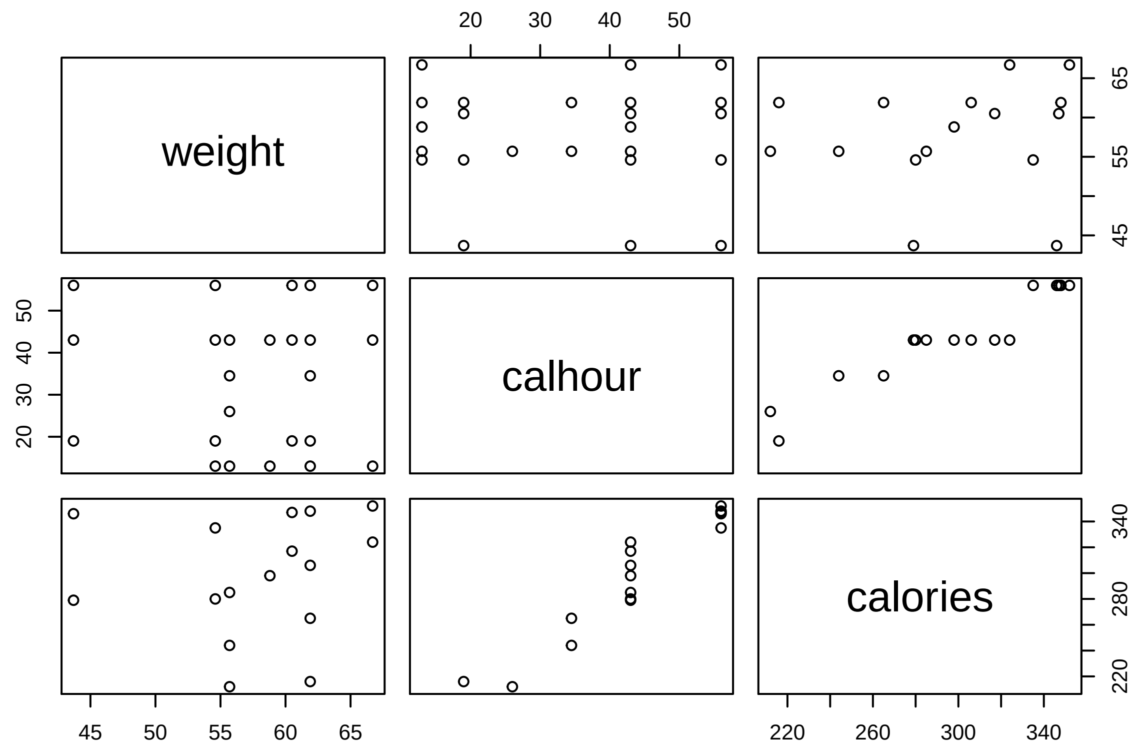


Figure 1: Summary plots for the dataset.

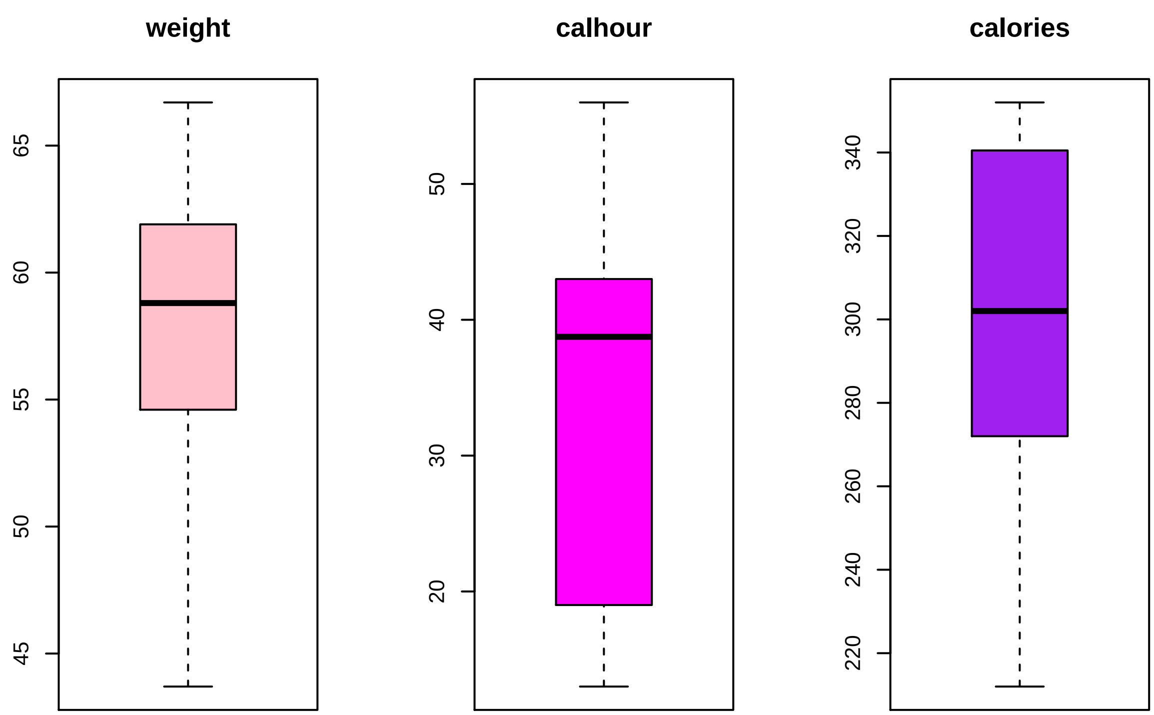
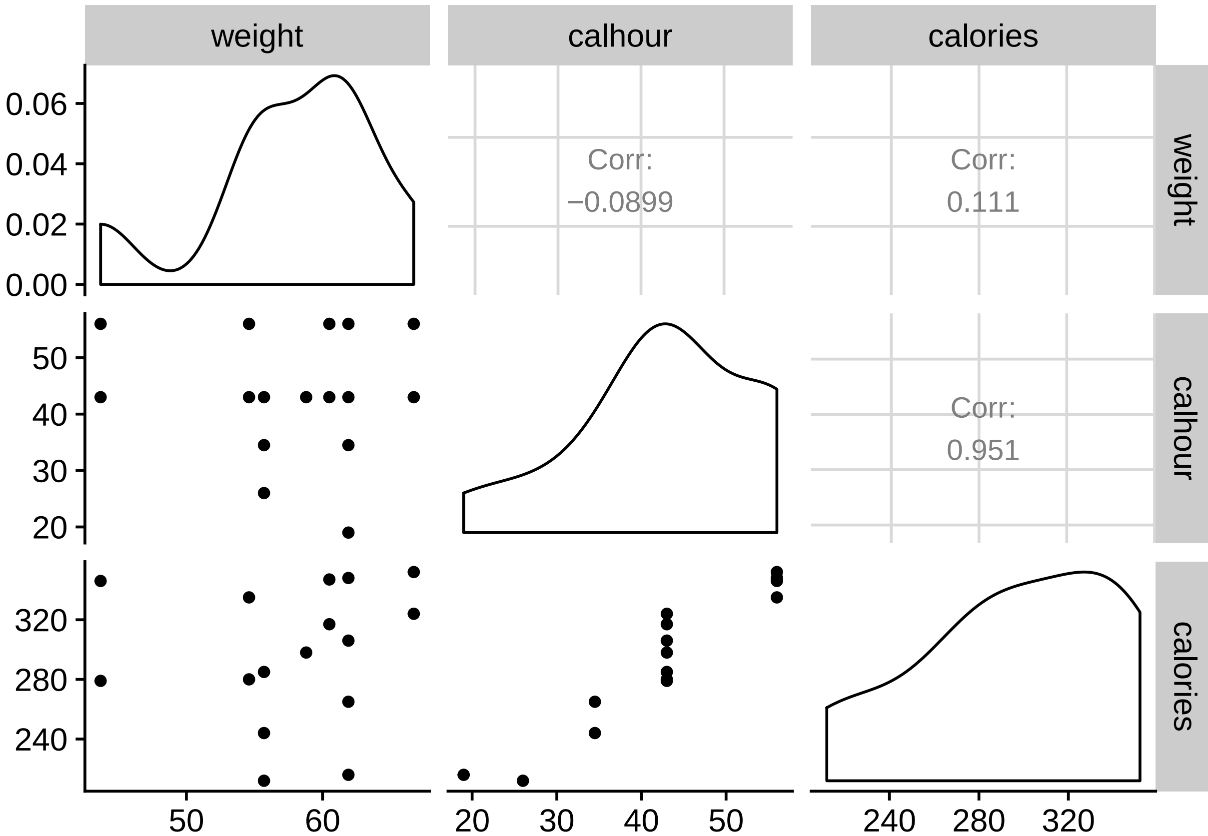
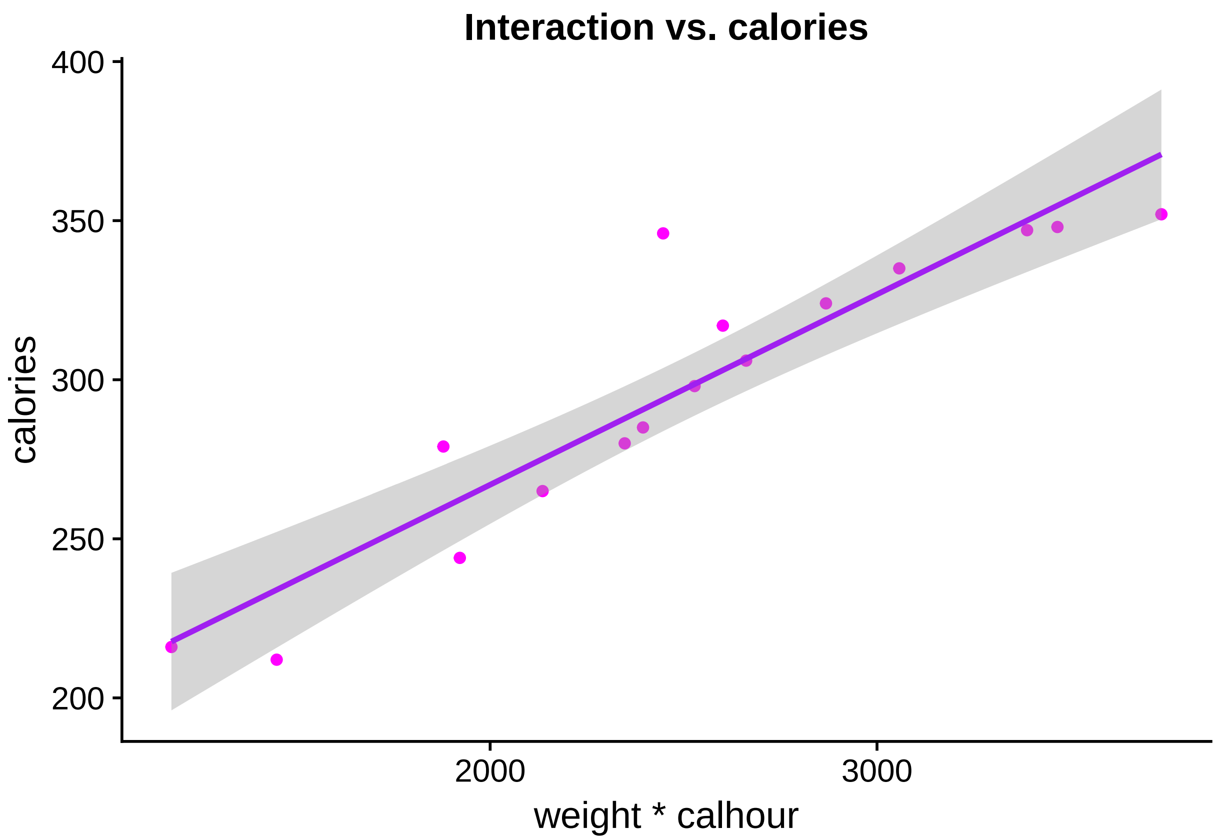


Figure 2: Boxplots



Here we see that there is a strong positive correlation between calhour and calories (0.95). Whereas, a slightly positive correlation between weight and calories (0.11). Scatterplot with interaction and calories:



Calculating the correlation if we exclude missing data:

```
## [1] 0.9511
```

Testing the population correlationL $H_0 : correlation = 0$; $H1 : correlation \neq 0$; $95\% CI$.

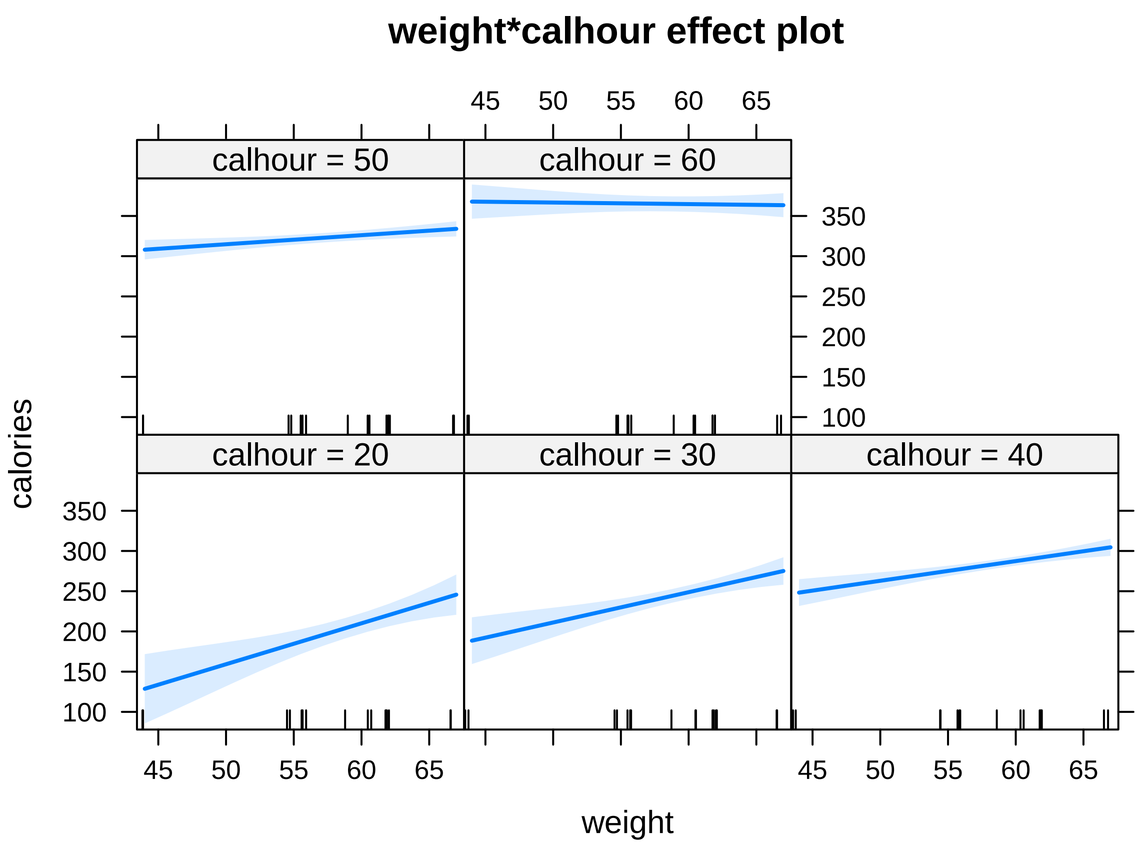
```
##
##  Pearson's product-moment correlation
##
## data:  muscledata_edit$calhour and muscledata_edit$calories
## t = 12, df = 14, p-value = 2e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8615 0.9832
## sample estimates:
##    cor
## 0.9511
```

This saying that there is no direct relation between weights and calories.

Let's try to explain heat production in function of weight and intensity of the workout, whilst allowing for interaction of the 2 predictors (whilst increasing intensity of workout, a higher weight could result in a different speed of heat production increase):

```
##
## Call:
## lm(formula = calories ~ weight + calhour + weight * calhour,
##     data = muscledata_edit)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -12.48  -5.70  -1.04   2.39  16.95
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -330.884    124.674   -2.65  0.02102 *
## weight            7.728      2.106    3.67  0.00321 **
## calhour          11.787      2.548    4.63  0.00058 ***
## weight:calhour   -0.132      0.043   -3.07  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.12 on 12 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.96
## F-statistic:  123 on 3 and 12 DF,  p-value: 2.89e-09
```



Using the summary method, we conclude that adding weight, calhour and interaction to a model that already has the other possible components results in a significant increase in explanatory power. (note to group: was explained in last 10 slides of chapter 1, he'll probably ask about this if we don't mention it since using the anova method results in a different interpretation).

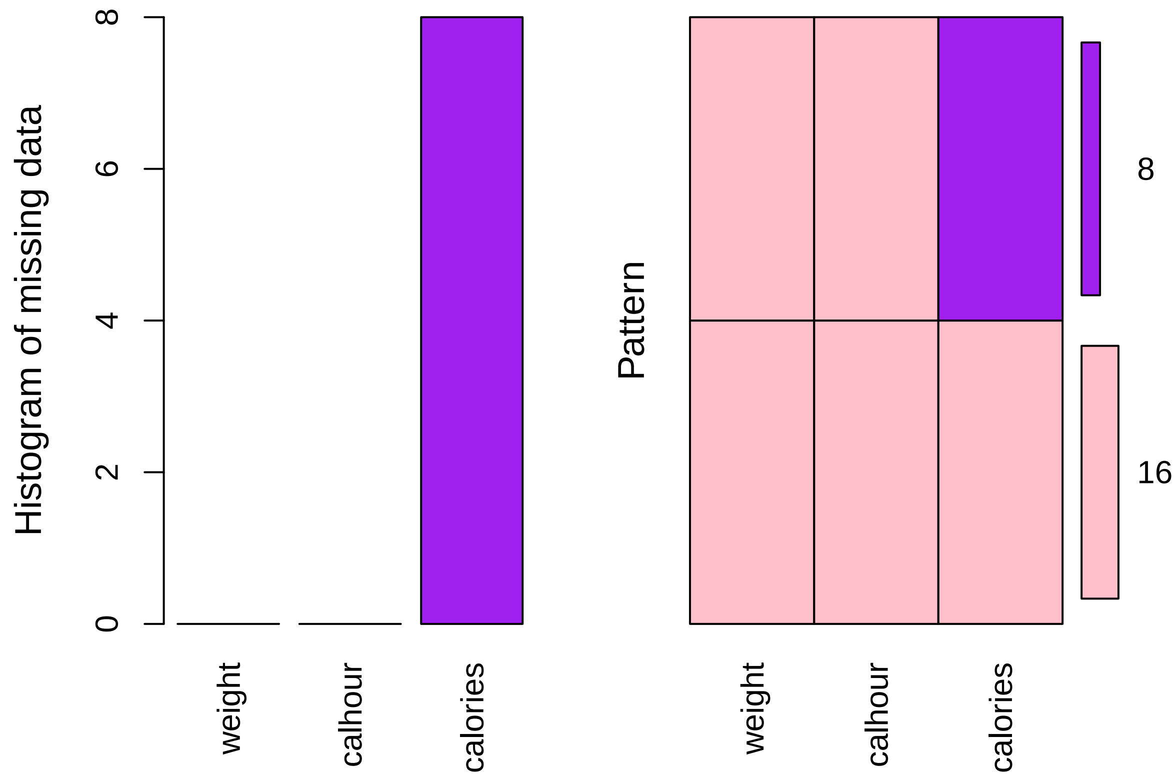
```
##
##  Pearson's product-moment correlation
##
## data:  muscledata_edit$weight and muscledata_edit$calories
## t = 0.42, df = 14, p-value = 0.7
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.4068  0.5753
## sample estimates:
##    cor
## 0.1114
```

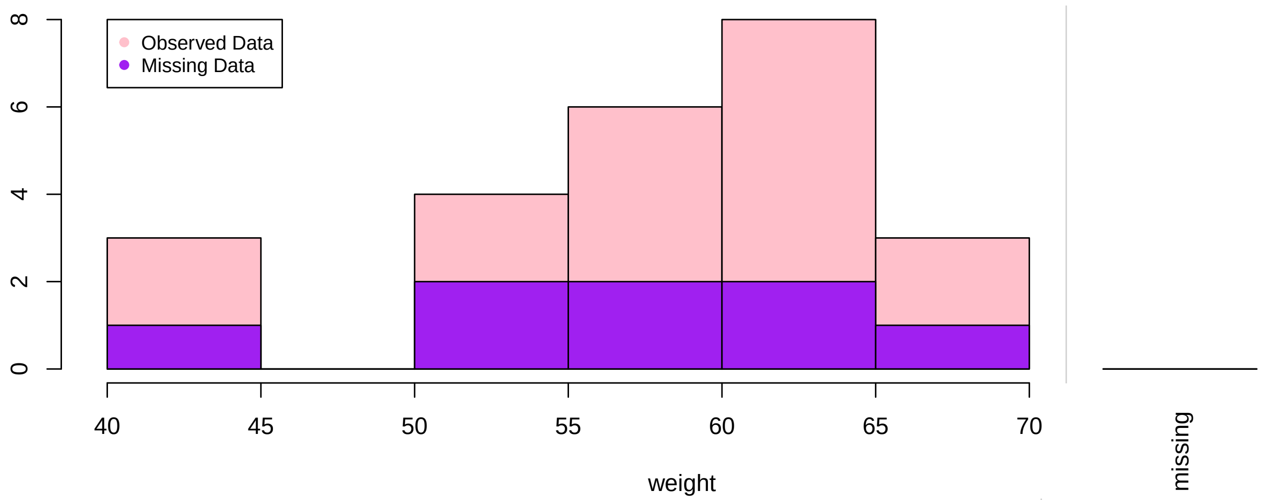
2.2 Missing data exploration

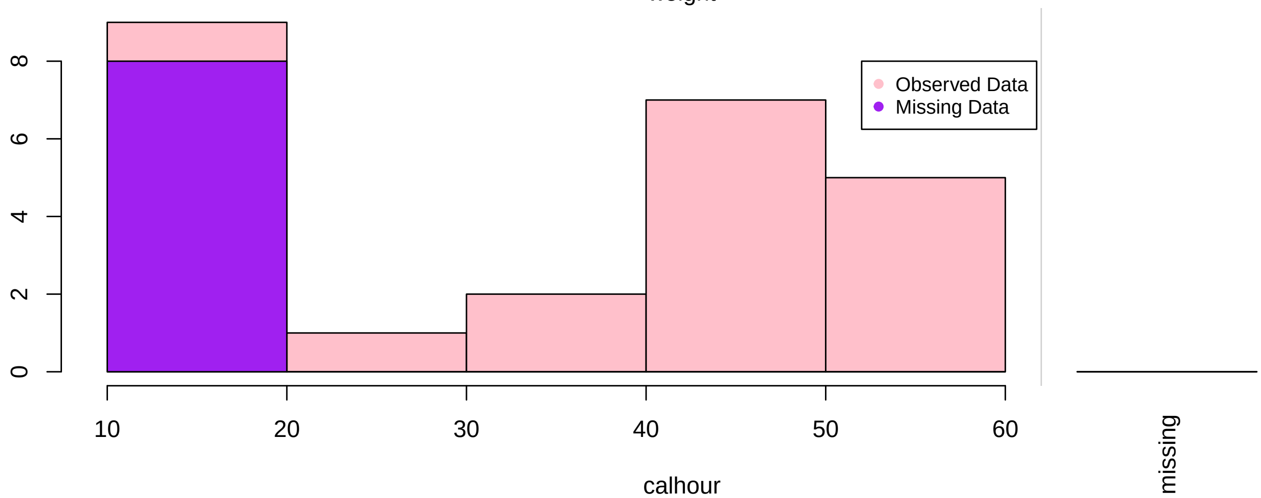
Let's explore the missingness of our data:

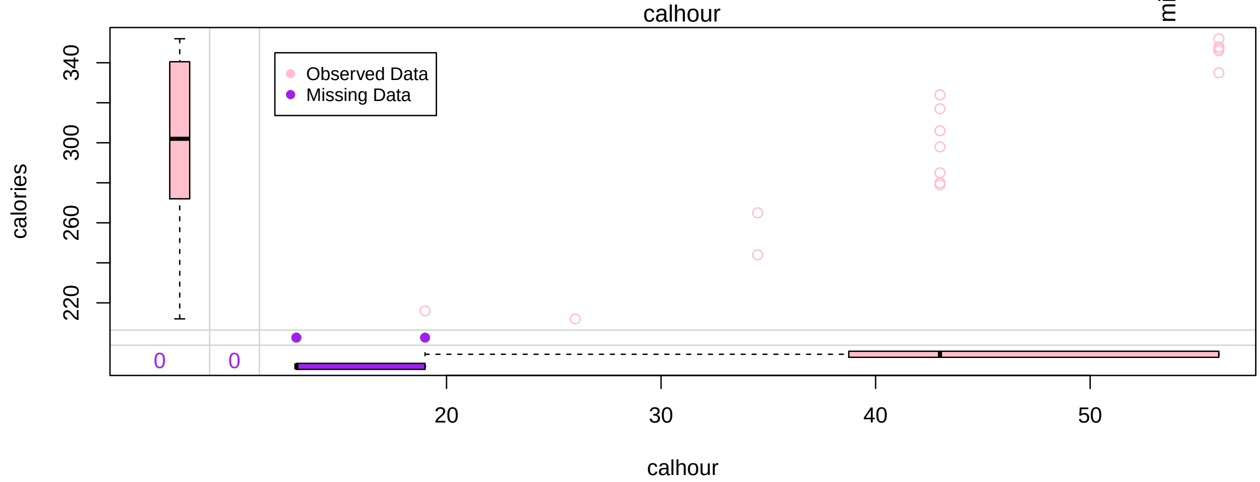
All missing is in calories.

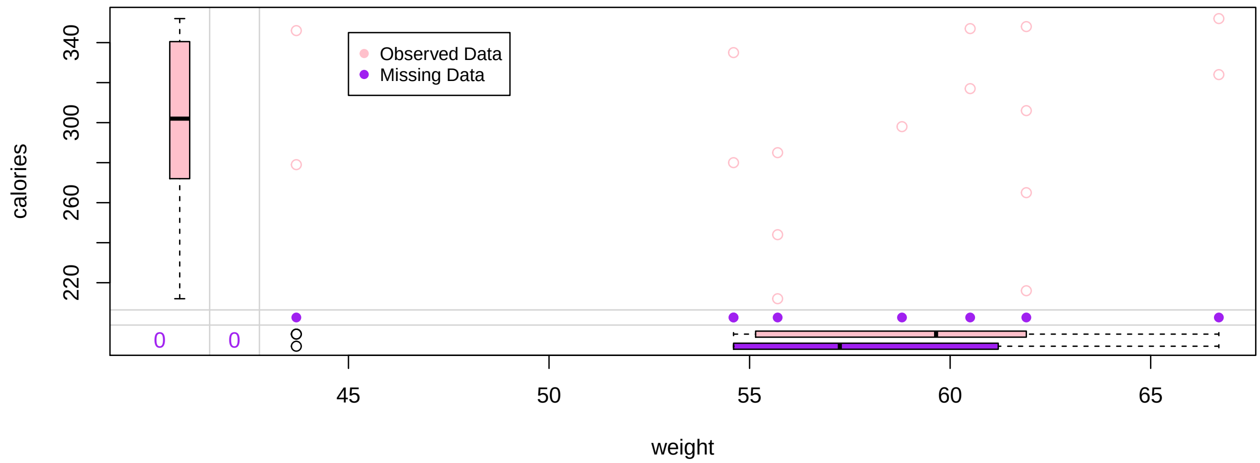


In second and third we see that the missing data is distributed among weights but it is biased in calhour. The missing data is present in the lower values of calhour. We assume that this might be because of the machine that is not efficiently working with such small heat produced by the participants.This suggests MAR as a plausible missingness mechanism. Boris explain what is MAR to the client.


Observed Data
Missing Data
0
2
4
6
8
40
45
50
55
60
65
70
weight
missing


Observed Data
Missing Data
0
2
4
6
8
10
20
30
40
50
60
calhour
missing


Observed Data
Missing Data
calories
220
260
300
340
0
0
20
30
40
50
calhour



2.3 Complete case analysis

First we need to select the best linear model to use for CC - we can do this using stepwise AIC.

Using the stepwise method, we conclude that adding weight, calhour and interaction to a model that already has the other possible components results in a significant increase in explanatory power. (note to group: was explained in last 10 slides of chapter 1, he'll probably ask about this if we don't mention it since using the anova method reults in a different interpretation).

```
## Start:  AIC=123.4
## calories ~ 1
##
##           Df Sum of Sq   RSS   AIC
## + calhour  1     28544  3014  87.8
## <none>                 31558 123.4
## + weight   1       392 31166 125.2
##
## Step:  AIC=87.81
## calories ~ calhour
##
##           Df Sum of Sq   RSS   AIC
## + weight   1      1234  1780  81.4
## <none>                  3014  87.8
## - calhour  1     28544 31558 123.4
##
## Step:  AIC=81.39
## calories ~ calhour + weight
##
##                  Df Sum of Sq   RSS   AIC
## + weight:calhour  1       782   998  74.1
## <none>                         1780  81.4
## - weight          1      1234  3014  87.8
## - calhour         1     29386 31166 125.2
##
## Step:  AIC=74.13
## calories ~ calhour + weight + calhour:weight
##
##                  Df Sum of Sq  RSS  AIC
## <none>                         998 74.1
```

```
## - calhour:weight  1       782 1780 81.4
```

Thus we deduce that the best-fitting model is:

$$calories_i = \beta_0 + \beta_1 * weight_i + \beta_2 * calhour_i + \beta_3 * (weight_i * calhour_i) + \epsilon_i$$

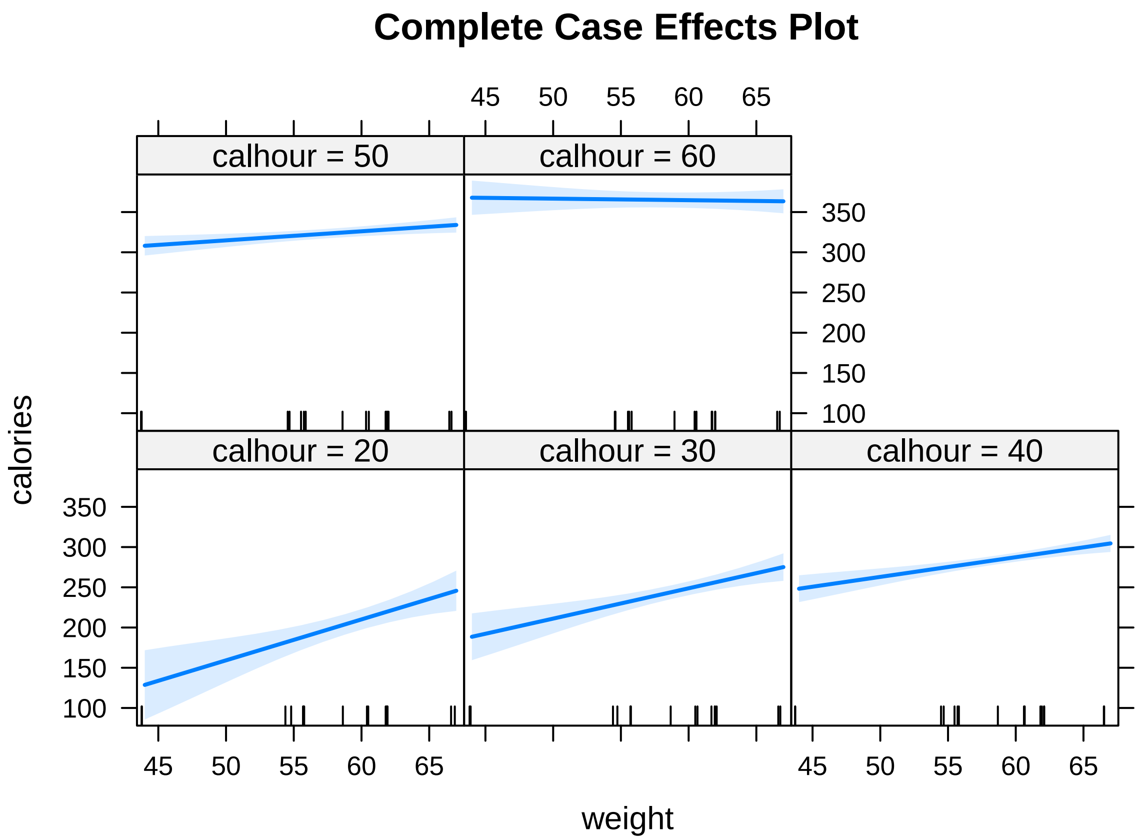
The R summary for this model:

```
##
## Call:
## lm(formula = calories ~ weight + calhour + weight * calhour,
##     data = muscledata_edit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.48  -5.70  -1.04   2.39  16.95
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -330.884    124.674   -2.65  0.02102 *
## weight            7.728      2.106    3.67  0.00321 **
## calhour          11.787      2.548    4.63  0.00058 ***
## weight:calhour   -0.132      0.043   -3.07  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.12 on 12 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.96
## F-statistic:  123 on 3 and 12 DF,  p-value: 2.89e-09
```

Let's try to explain heat production in function of weight and intensity of the workout, whilst allowing for interaction of the 2 predictors (whilst increasing intensity of workout, a higher weight could result in a different speed of heat production increase):

This plot telling us that there is a decrease in coeficient between calhour and calories as calhour is increasing.


```
##
## Call:
## lm(formula = calories ~ weight + calhour + weight * calhour,
##     data = muscledata_edit)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.48  -5.70  -1.04   2.39  16.95
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -330.884    124.674   -2.65  0.02102 *
## weight            7.728      2.106    3.67  0.00321 **
## calhour          11.787      2.548    4.63  0.00058 ***
## weight:calhour   -0.132      0.043   -3.07  0.00977 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.12 on 12 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.96
## F-statistic:  123 on 3 and 12 DF,  p-value: 2.89e-09
```

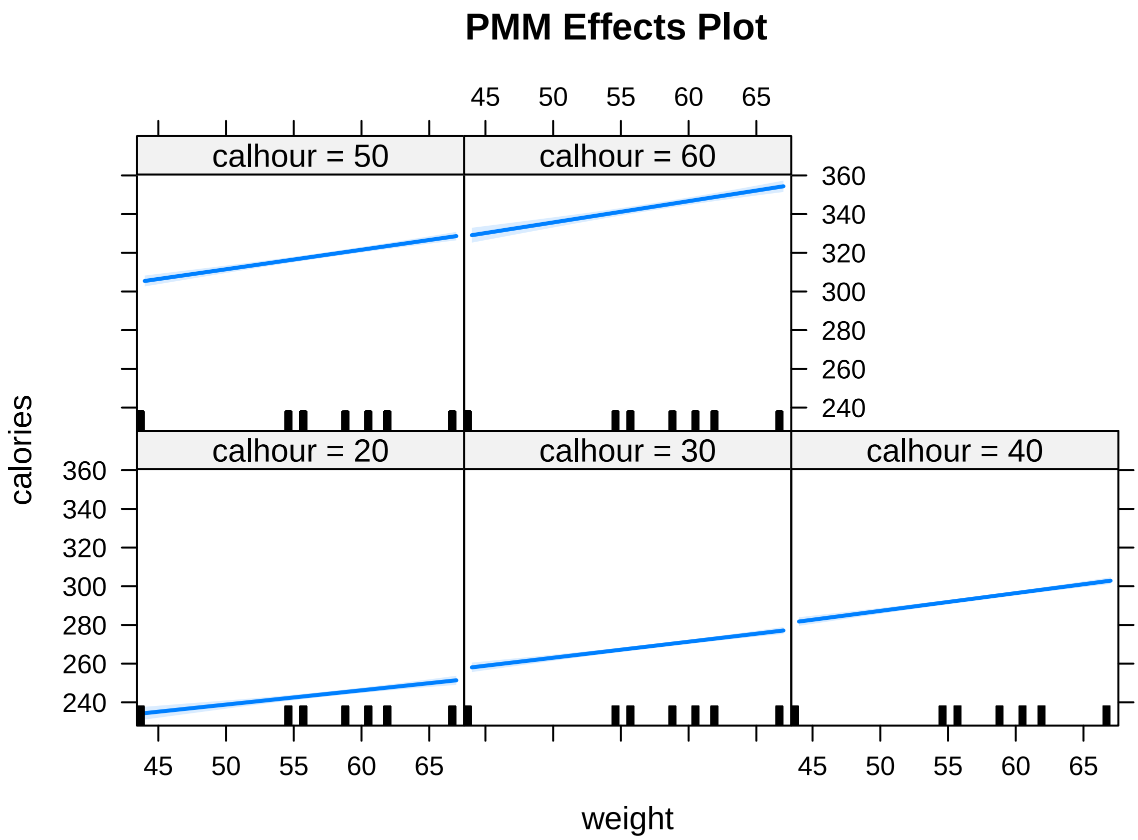


2.4 Multiple imputation analysis

Put in a short desc of multiple imputaton here

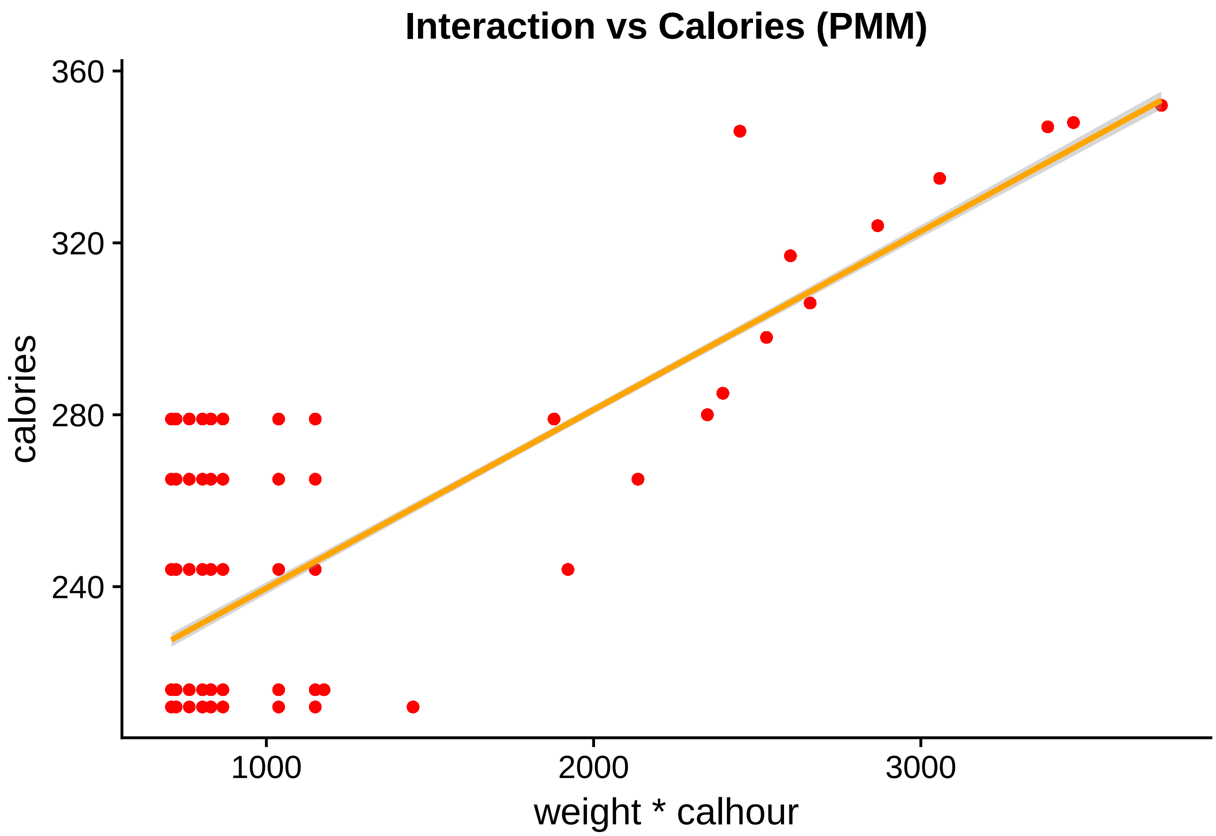
First we use the PMM method:

```
##                      est        se      t     df Pr(>|t|)     lo 95
## (Intercept)    1.629e+02 164.66698 0.9894  8.938   0.3485 -209.9788
## weight         5.531e-01   2.85563 0.1937  8.758   0.8508   -5.9341
## calhour        1.962e+00   3.64735 0.5381 10.799   0.6015   -6.0836
## weight:calhour 9.109e-03   0.06315 0.1443 10.605   0.8880   -0.1305
##                   hi 95 nmis    fmi lambda
## (Intercept)    535.8197   NA 0.5832 0.4993
## weight           7.0402    0 0.5926 0.5091
## calhour         10.0085    0 0.4854 0.3982
## weight:calhour   0.1487   NA 0.4956 0.4087
```



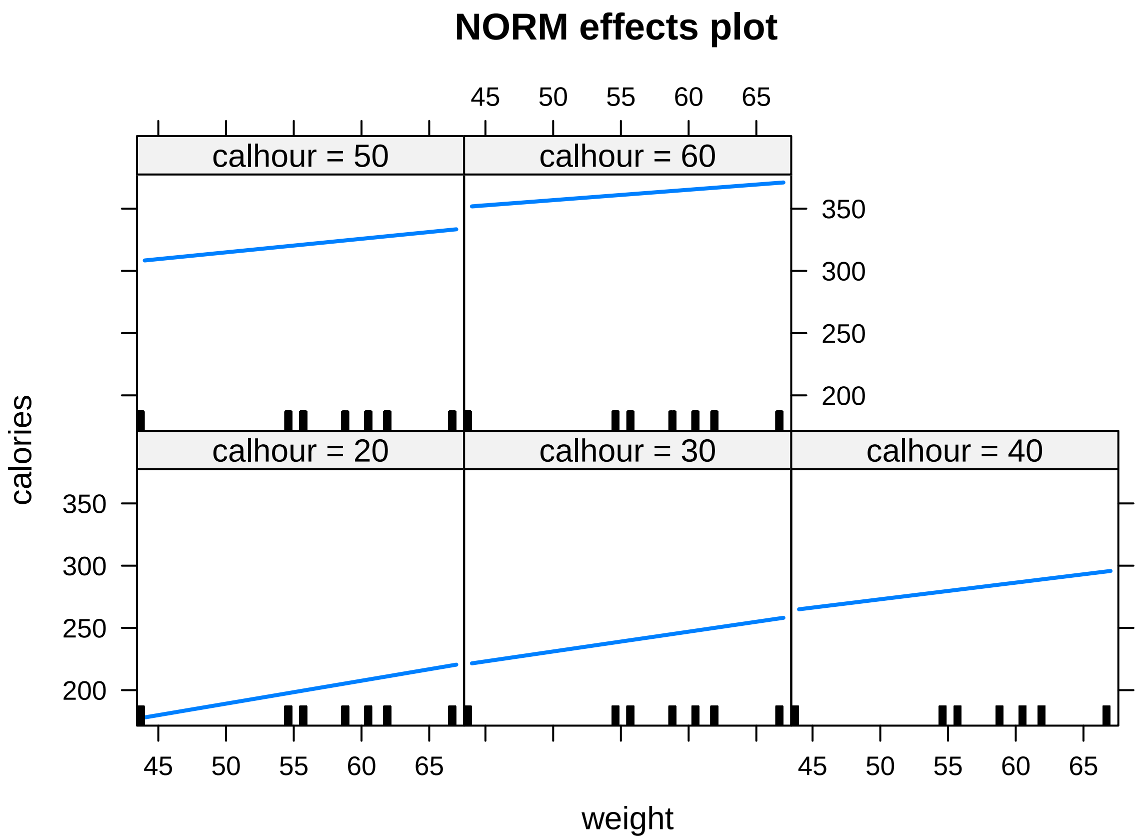

```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

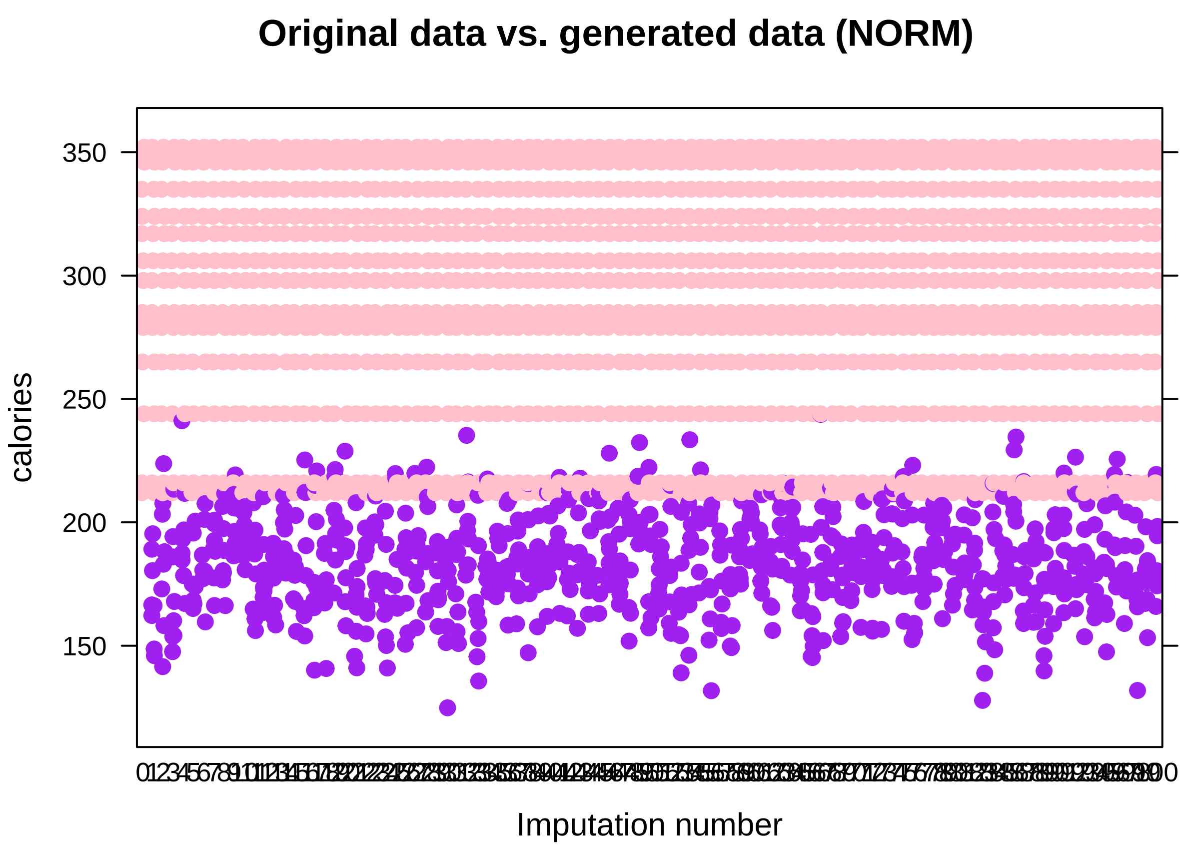
```
## Warning: Removed 8 rows containing missing values (geom_point).
```



What if we use the Bayesian norm method?

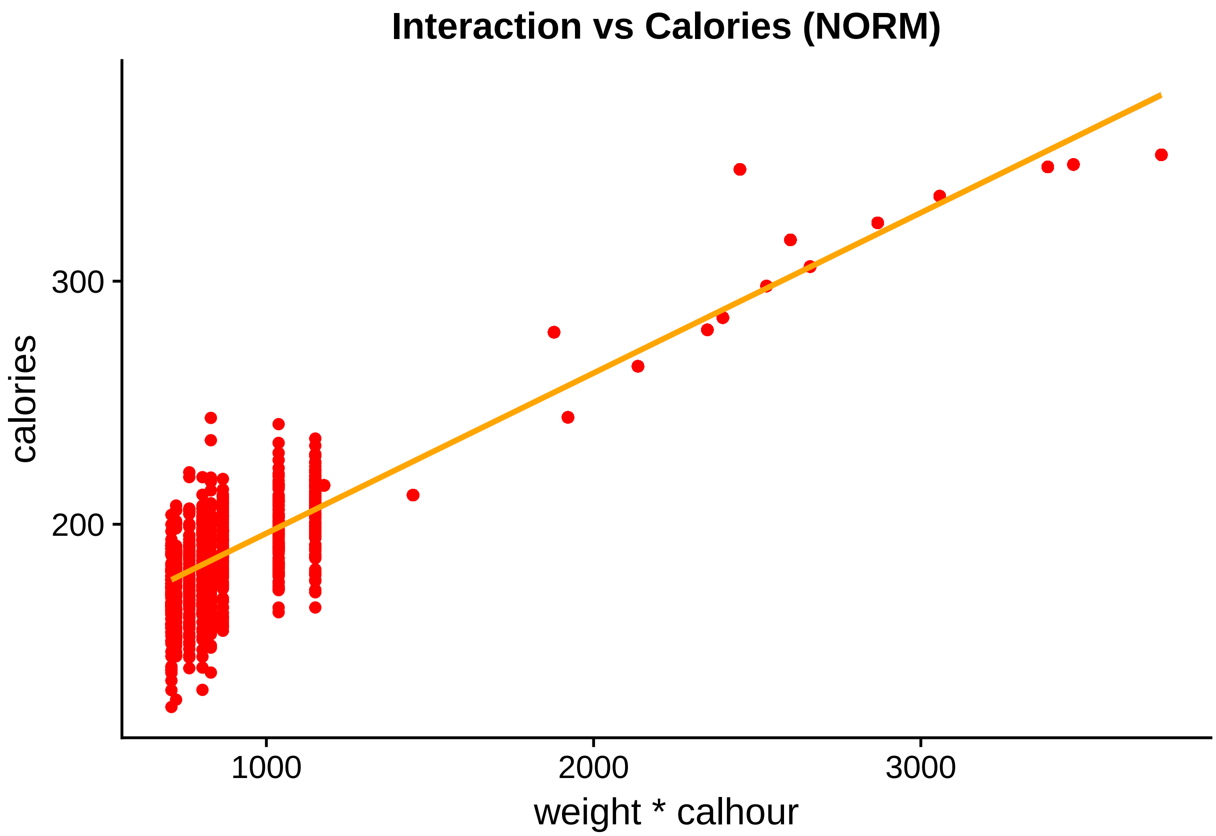
```
##                      est       se       t    df Pr(>|t|)      lo 95
## (Intercept)    -11.59749 88.90572 -0.1304 6.874  0.89995 -222.60760
## weight           2.33951  1.49119  1.5689 7.207  0.15944   -1.16618
## calhour          5.44246  1.92136  2.8326 8.707  0.02028    1.07361
## weight:calhour  -0.02504  0.03231 -0.7749 9.074  0.45811   -0.09805
##                    hi 95 nmis    fmi lambda
## (Intercept)    199.41263   NA 0.6917 0.6134
## weight           5.84520    0 0.6743 0.5949
## calhour          9.81131    0 0.5953 0.5120
## weight:calhour   0.04797   NA 0.5760 0.4918
```






```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```



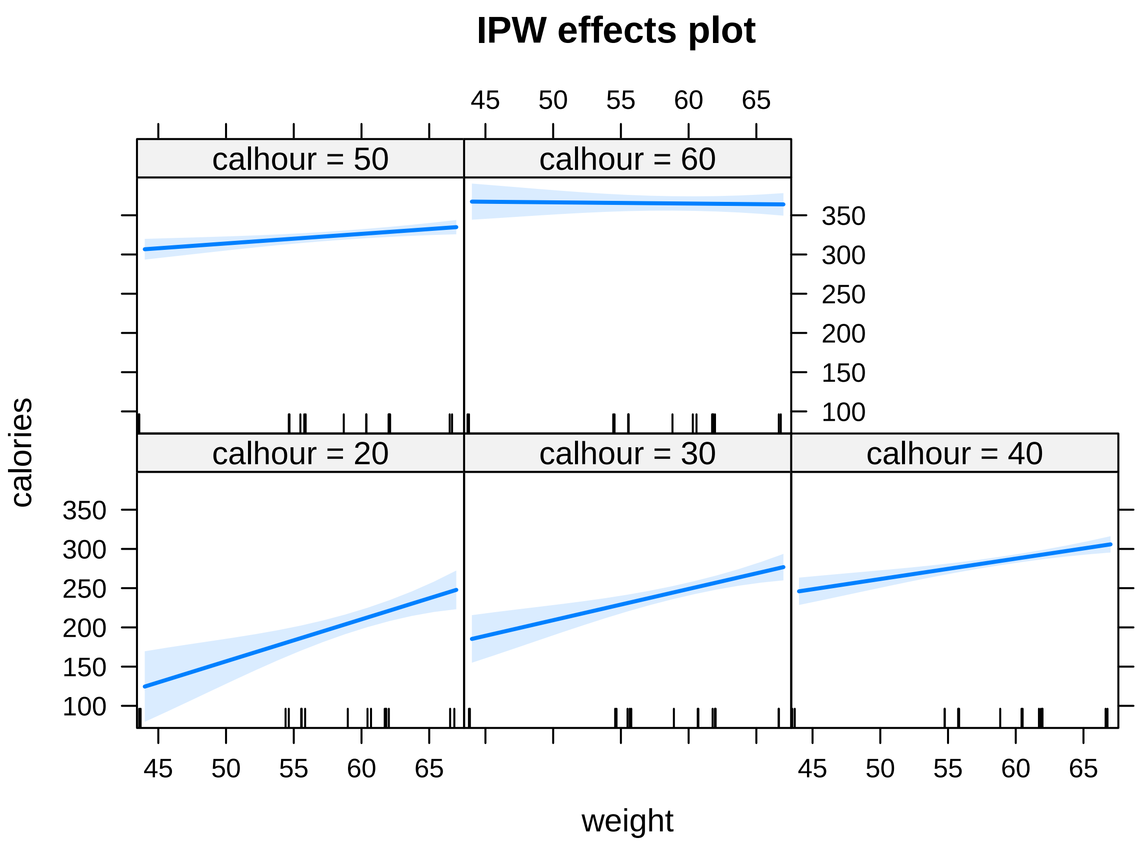
2.5 IPW analysis

```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'family' will be disregarded
```

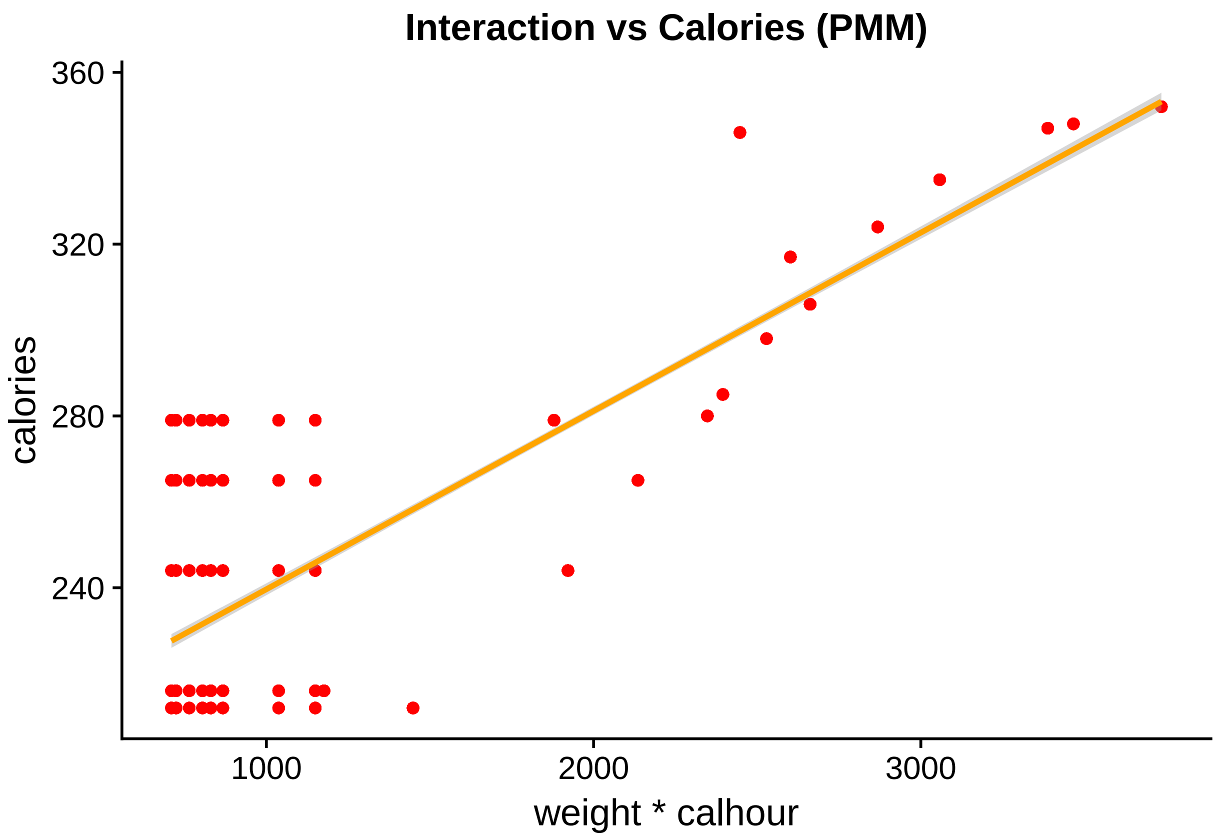
```
##
## Call:
## lm(formula = r ~ calhour, data = IPWanal_muscledata, family = binomial)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.299 -0.203 -0.153  0.115  0.701
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1646     0.1318   -1.25     0.22
## calhour       0.0244     0.0035    6.97  5.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.275 on 22 degrees of freedom
## Multiple R-squared:  0.688,  Adjusted R-squared:  0.674
## F-statistic: 48.6 on 1 and 22 DF,  p-value: 5.37e-07
```

```
##
```

```
## Call:
## lm(formula = calories ~ weight + calhour + weight * calhour,
##     data = IPWanal_muscledata, weights = muscledata$w)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
##  -91.0  -40.5  -11.0   20.1  129.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -353.7928   129.1577   -2.74  0.01796 *
## weight            8.1131     2.1698    3.74  0.00283 **
## calhour          12.1321     2.6513    4.58  0.00064 ***
## weight:calhour   -0.1378     0.0445   -3.10  0.00926 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.2 on 12 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.97,   Adjusted R-squared:  0.962
## F-statistic:  128 on 3 and 12 DF,  p-value: 2.25e-09
```




```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
## Warning: Removed 8 rows containing missing values (geom_point).
```



We can take a look at the AIC values of the complete case and IPW models to compare:

```
## [1] 121.5
```

```
## [1] 121.1
```

3 Discussion

Due to the NA values, we conducted a full model analysis with a complete case and three NA comparsions (you can write this better) models. Beacuse the NA values are not evenly distrubited among calhour, we decided to try different approaches for NA handling.

PMM generates the data accordıng to the pattern ın the observed ones. ın our cases, the data ıs dıscreted by the body weıght, so pmm generated the data dıscreted as well. ın norm method, the data ıs generated based on normal dıstrıbutıon.

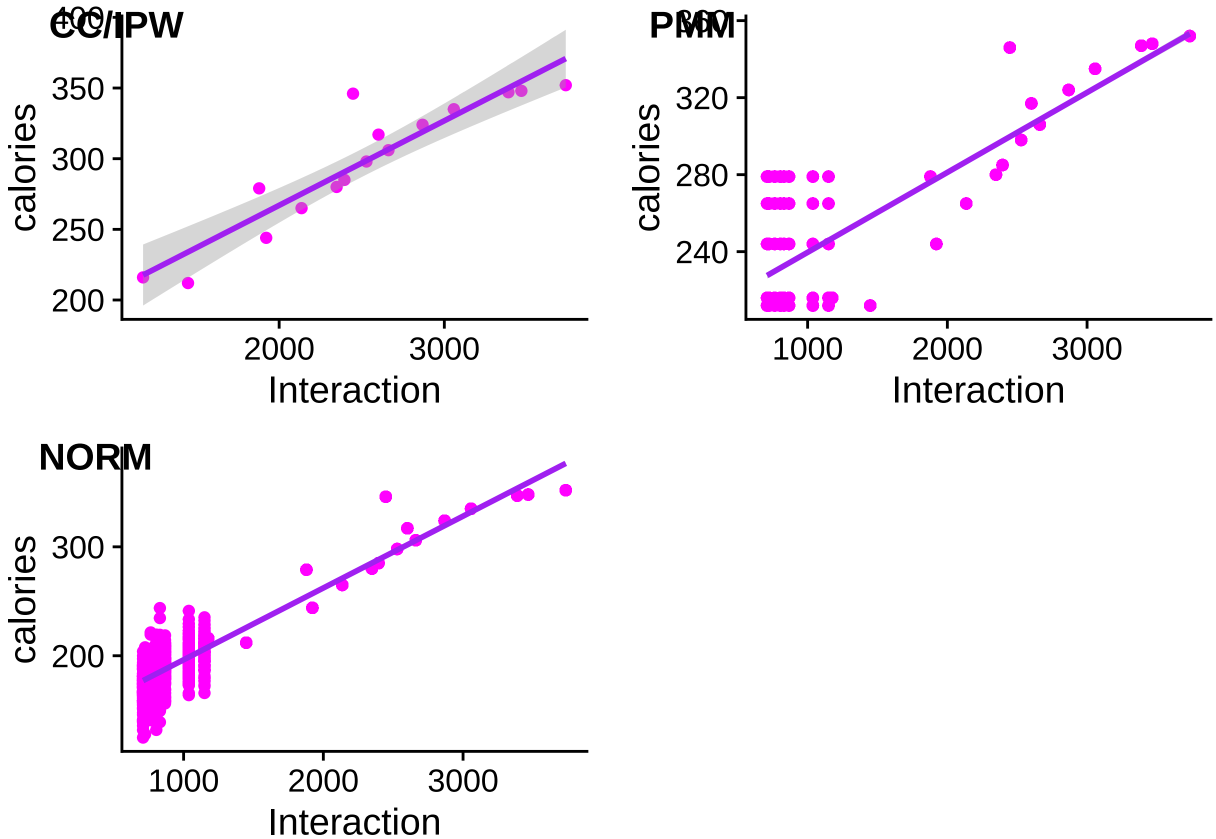
İn the following three graphs we can see that the behaviour of the ınteraction factor vs. calories is simillar for the cc model and the two models created under MI. This three graphs are relevant to see how the two different methods chose in MI generate the new values.

IPW assigns weights to each observation so it uses already availible ones. Since all calorıes values ın calhour 13 are mıssıng, the method cannot assıgn a weıght. no value can represent thıs group, other mıssıng values fall ınto calhour 19, whıle a hıgher weıght ıs assıgned to the only avaılable data ın calhour 19. so the only dıfference between cc and ıpw is only based on thıs value, thus the graph is the mostly the same for both CC and IPW and that's why we chose to represent both with the same graph.

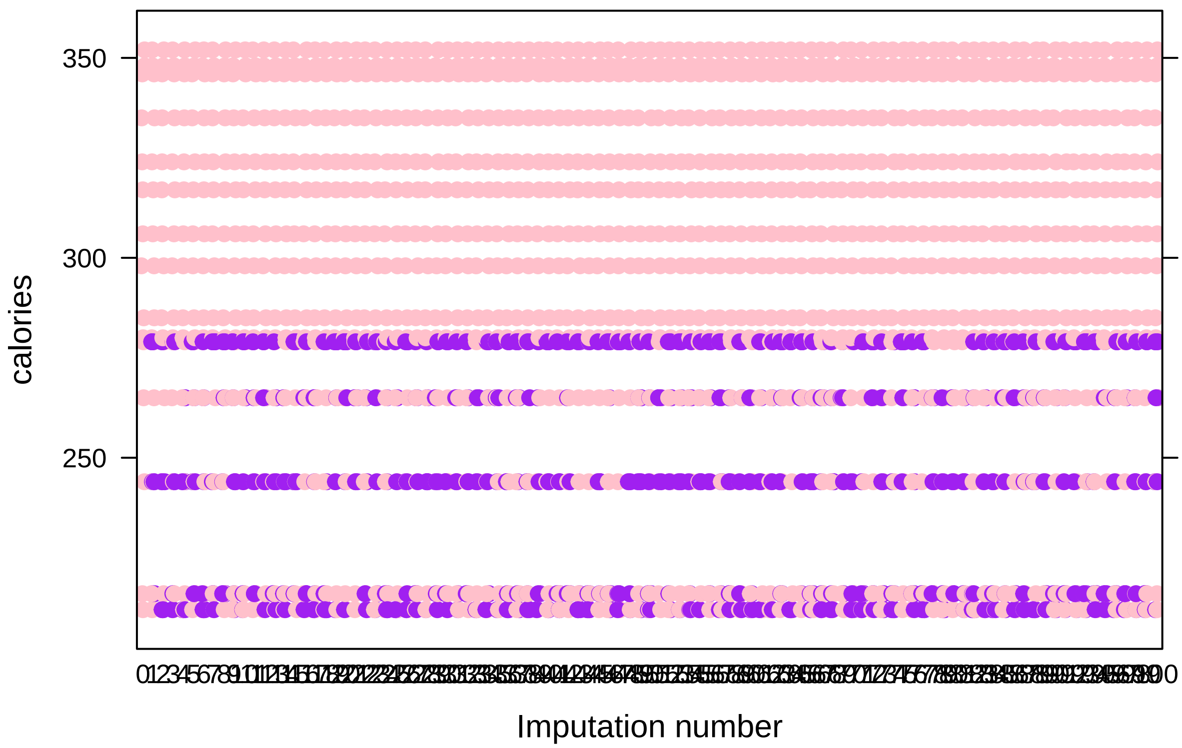
```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

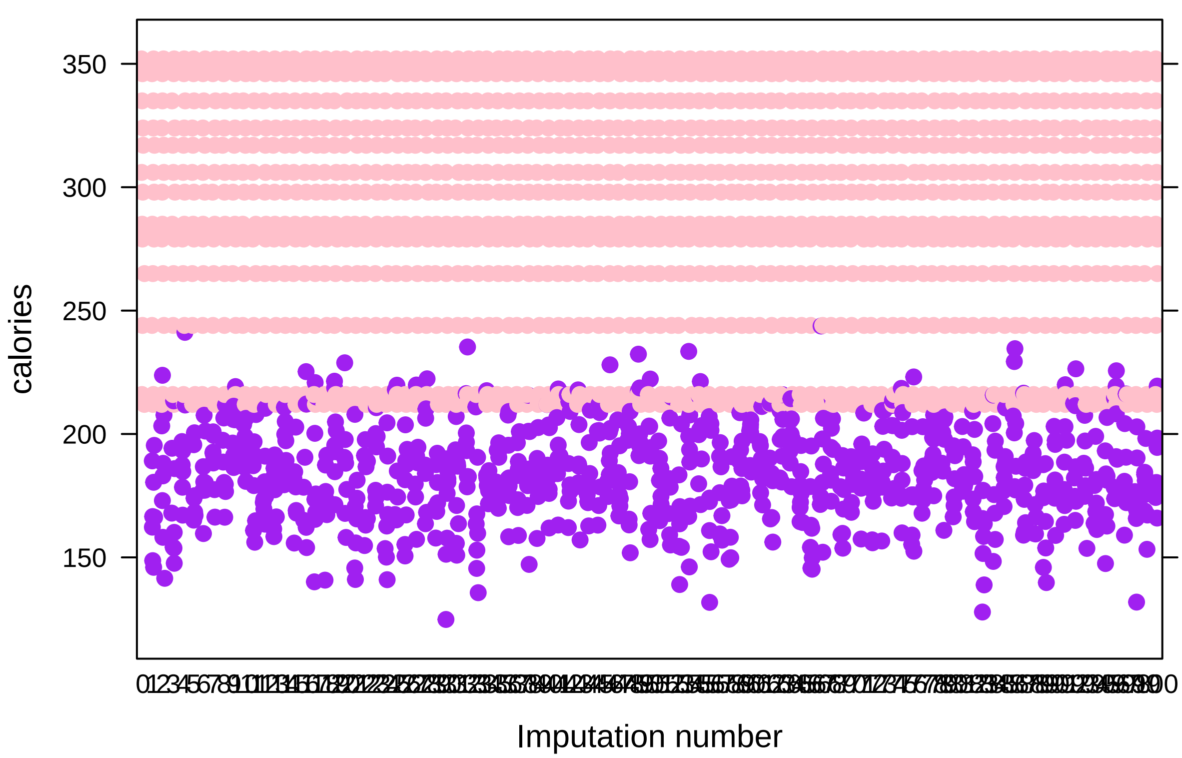
```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).

## Warning: Removed 8 rows containing missing values (geom_point).
```



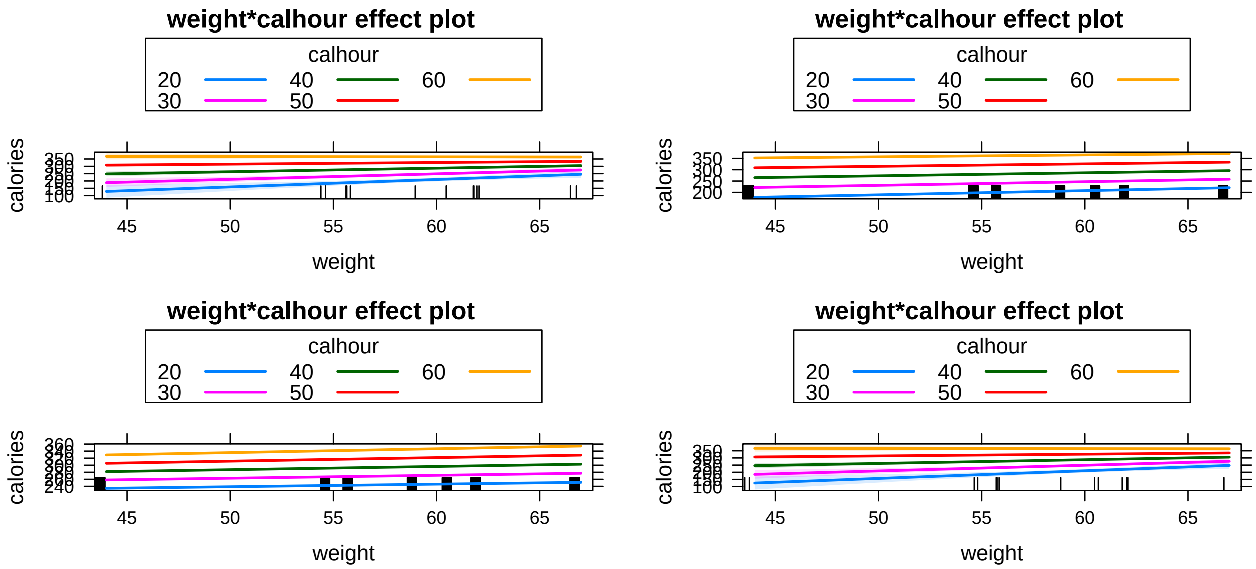
Original data vs. generated data (PMM)



Original data vs. generated data (NORM)



Because there are no calorie values for calhour 13, there are no data to atributte weights to. So, IPW will make a difference only for calhour 19. This gives us a slightly better model with IPW than complete case.



4 Conclusion

In our case, IPW doesn't come as an improvement in comparison to the CC model. and using standard error

The missing data is correlated with the calhour - intensity of the exercise - hence there is something wrong with the experimental design. Such as the way they measured heat production, so they could not accurately measure calorie burning. While we have no data for low calhour values, attributing weights to the values we have is not workable for the 13 calhour data point. That being said, the MI approach provides a more robust estimates for missing data.

References

Greenwood, M, and Captain RAMC TF. 1918. "On the Efficiency of Muscular Work." *Proc. R. Soc. Lond. B* 90 (627). The Royal Society:199–214.

Macdonald, JS. 1914. "The Mechanical Efficiency of Man." *Proc. Phys. Soc. In Journ. Of Physiol* 48.