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

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

19 March, 2018

Contents

1 Introduction **1**

2 Methods and procedure **2**

2.1 Data exploration . . . 2

2.2 Missing data exploration . . . 4

2.3 Complete case analysis . . . 4

2.4 Multiple imputation analysis . . . 7

2.5 IPW analysis . . . 13

3 Discussion **15**

4 Conclusion **17**

References **18**

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. In the body of this work 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.

```
##    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
```

2 Methods and procedure

2.1 Data exploration

A set of summary statistics for the dataset is presented below. It can be immediately seen that the response calories variable is missing in eight cases and is the only incomplete variable in the dataset. The mean and median values for all variables in the dataset are very similar to each other which indicates a symmetric distribution. A matrix of summary plots for the dataset is presented in Figure 2. We can clearly see what appears to be an extremely strong positive correlation between calories and workout intensity (0.95), and a very small positive correlation between calories and weight(0.11).

The distributions of the values are plotted as boxplots in Fig. 1. Note that the response variable is plotted with missing values excluded in all of these figures, thus despite expecting an approximately similar distribution between workout intensity and calories variables, the calories distribution is shifted upwards due to the missing values.

```
##                 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
```

We check the signifance of the two positive correlations we have found using Pearson's correlation (using Central Limit Theorem as the dataset contains around 20 rows). Here $H_0 : correlation = 0$; $H1 : correlation \neq 0$; $95\% CI$.

```
##
##  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
```

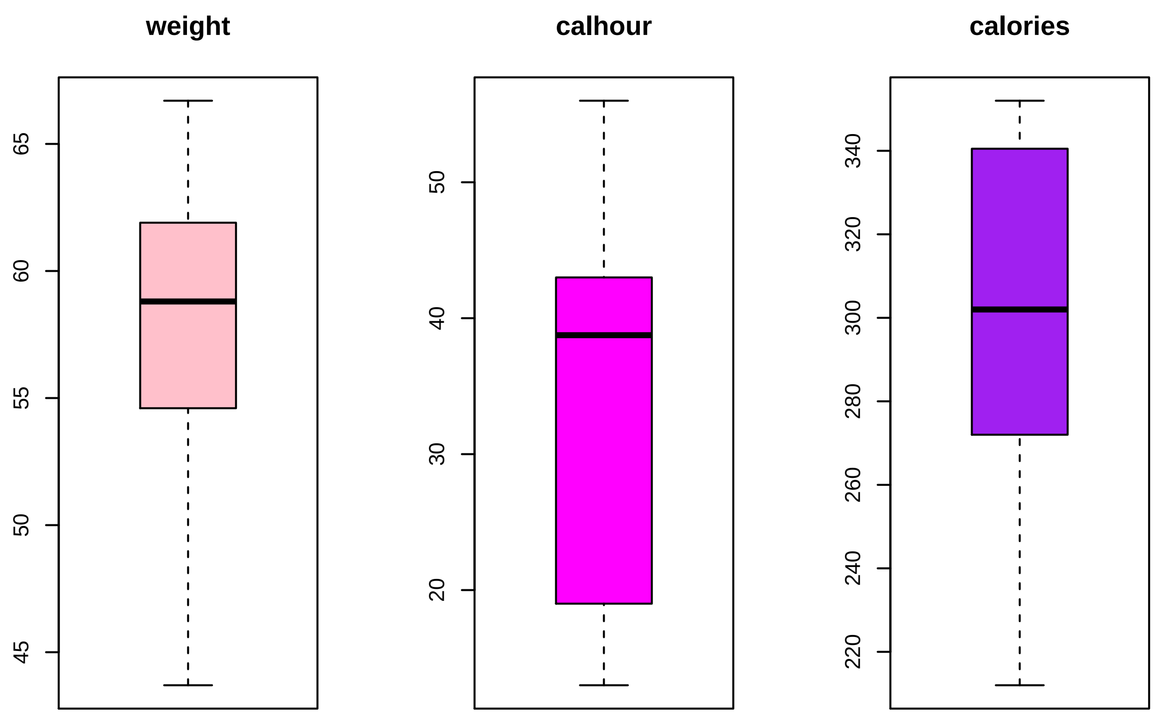


Figure 1: Boxplots for the dependent variables weight, calhour and independent variable calories.

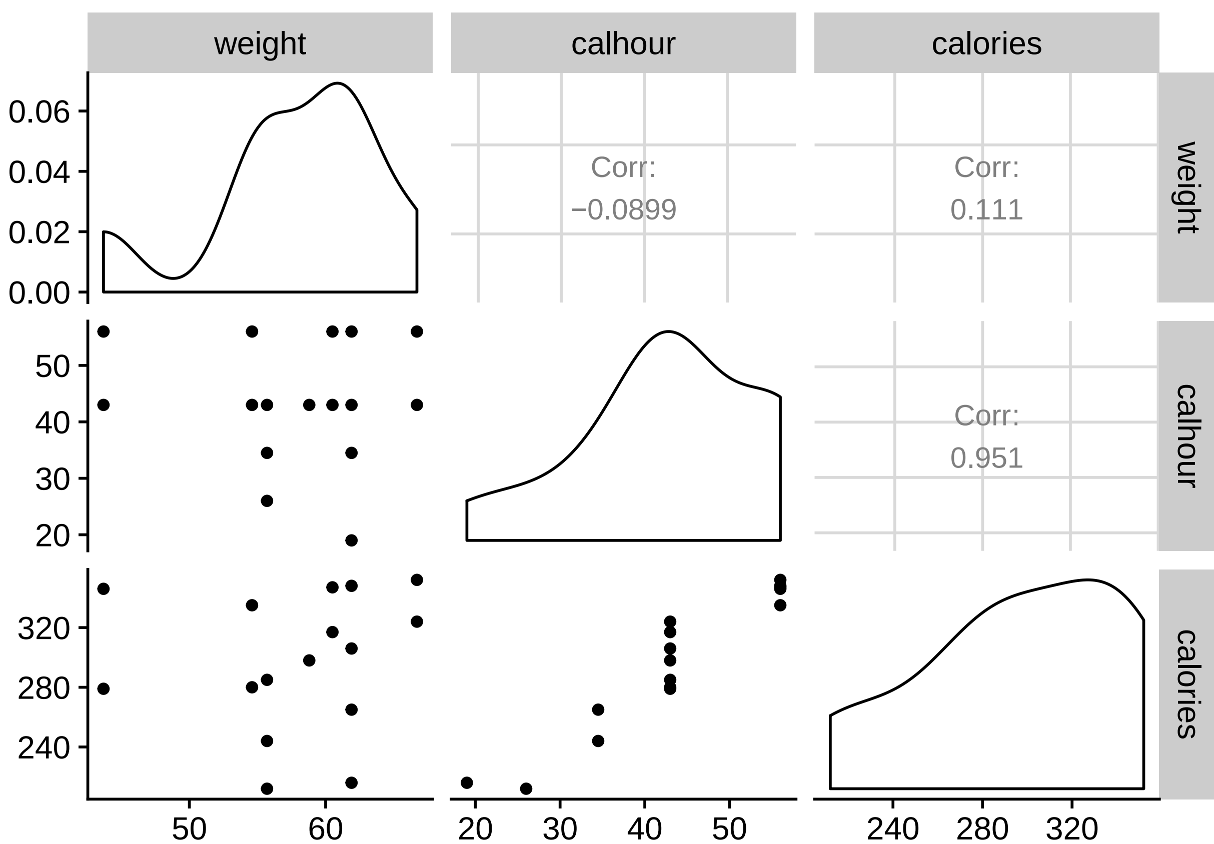


Figure 2: A summary statistics plot of the dataset using the ggplot command.

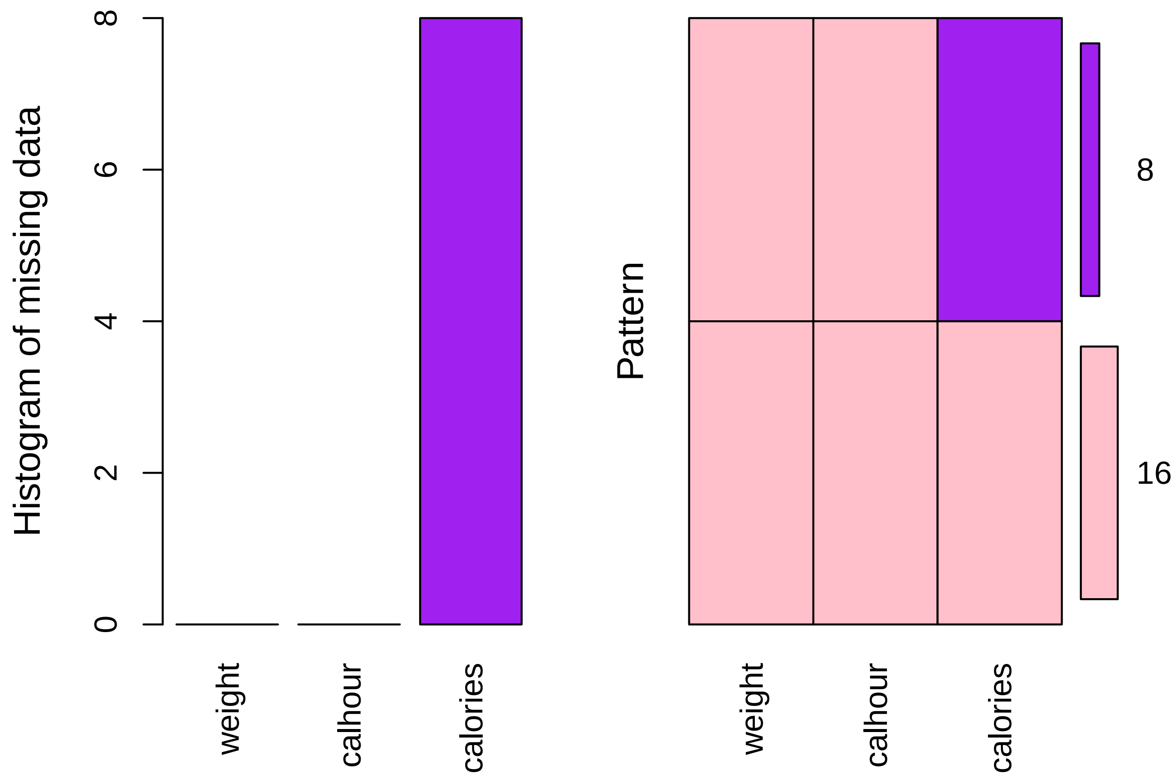


Figure 3: Pattern of missing data across variables

```
## sample estimates:
##    cor
## 0.9511
```

```
##
##  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
```

Clearly we reject the null hypothesis with regards to workout intensity and calories and accept it with regards to weight and calories. This indicates that there is a non-spurious correlation between workout intensity and calories in the population.

2.2 Missing data exploration

A histogram of missing data is shown in Fig. 3. We confirm our previous observation that all the missing values are located in our response variable.

Fig. 4A and C we see that the missing data approximately evenly distributed among the different weight variables. In Fig. 4B and D we see that the missing data distribution is extremely biased towards the lower end of the range with regards to workout intensity. This may be because of the difficulty in measuring heat production at lower exercise intensity - in other words, the missingness is likely systematic due to technical noise. Importantly, the missingness appears to depend only on an observed variable in this study - the calories. Thus, this suggests "Missing-at-Random" as the most probable missing data mechanism, allowing us to proceed with applying missing data strategies - particularly MI and IPW.

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:
```

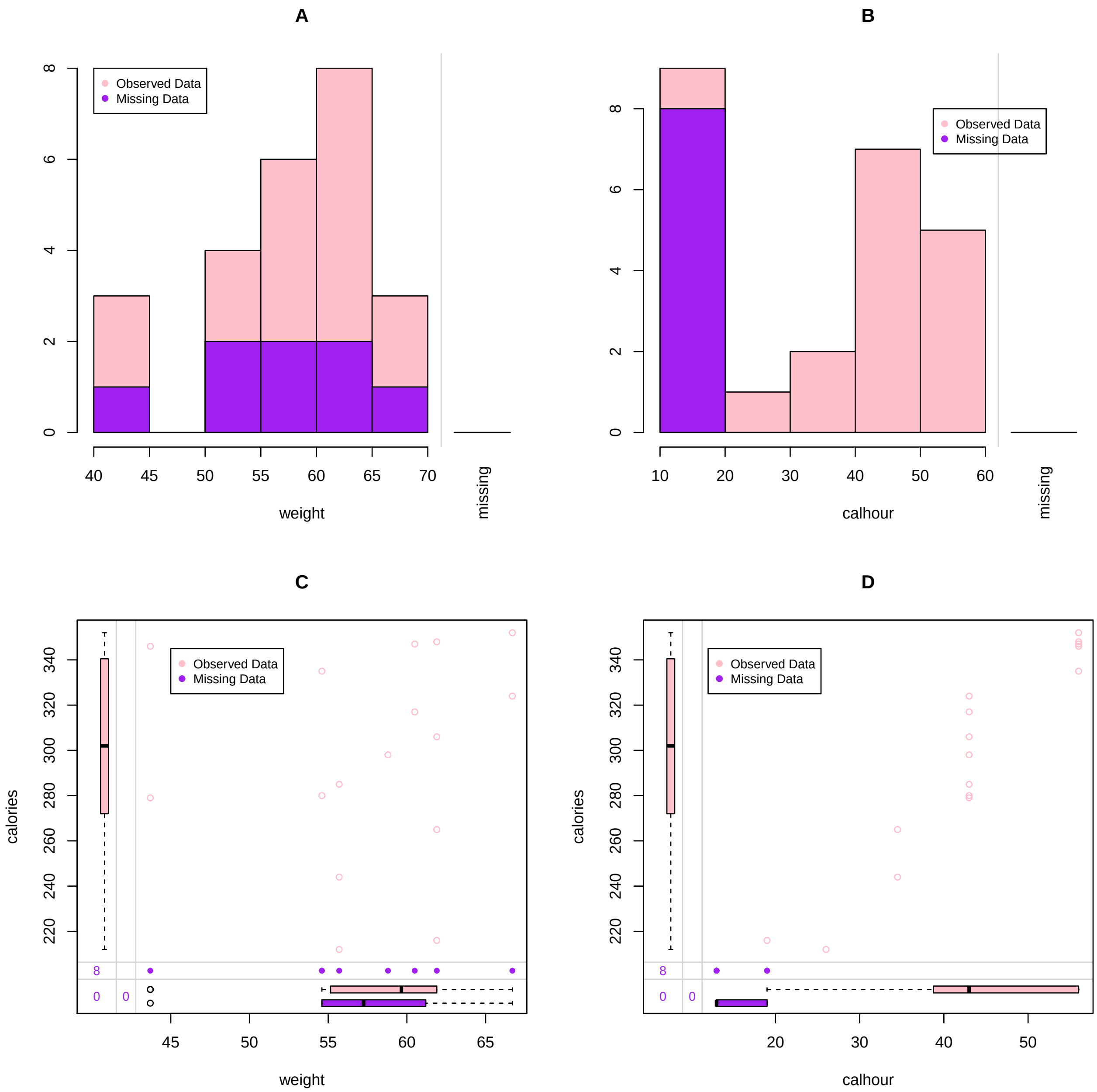


Figure 4: Histograms of the observed and missing data as well as marginplots depicting histograms and correlations.

```
## 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.

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.

```
##
##  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
```


```
##
## 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 **
```

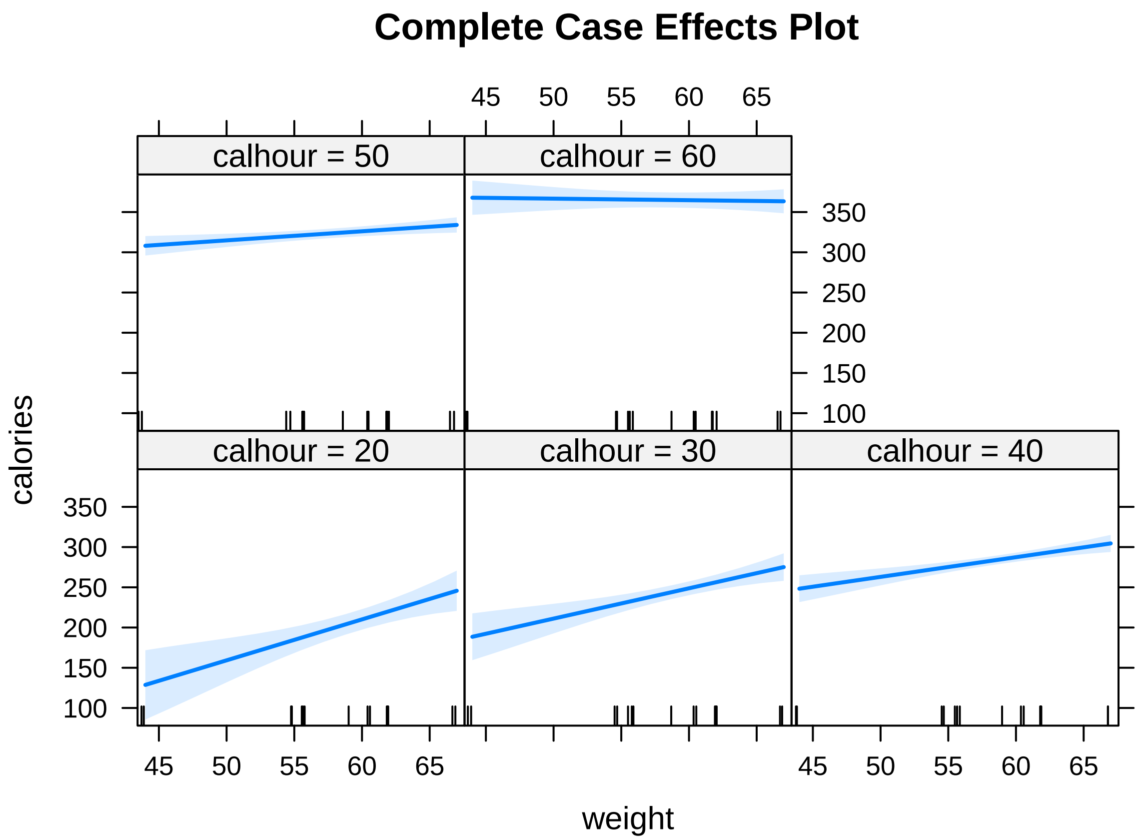


Figure 5: The All Effects plot for the Complete Case linear model.

```
## ---
## 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)    192.09281 160.15785 1.19940  9.293   0.2601 -168.4778
## weight           0.03595   2.74334 0.01311  9.339   0.9898   -6.1358
## calhour          1.39645   3.56650 0.39154 11.109   0.7028   -6.4440
## weight:calhour   0.01919   0.06109 0.31416 11.152   0.7592   -0.1151
##                    hi 95 nmis    fmi lambda
## (Intercept)     552.6634   NA 0.5645 0.4799
## weight            6.2077    0 0.5621 0.4774
## calhour           9.2368    0 0.4692 0.3815
## weight:calhour    0.1534   NA 0.4669 0.3792
```

What if we use the Bayesian norm method?

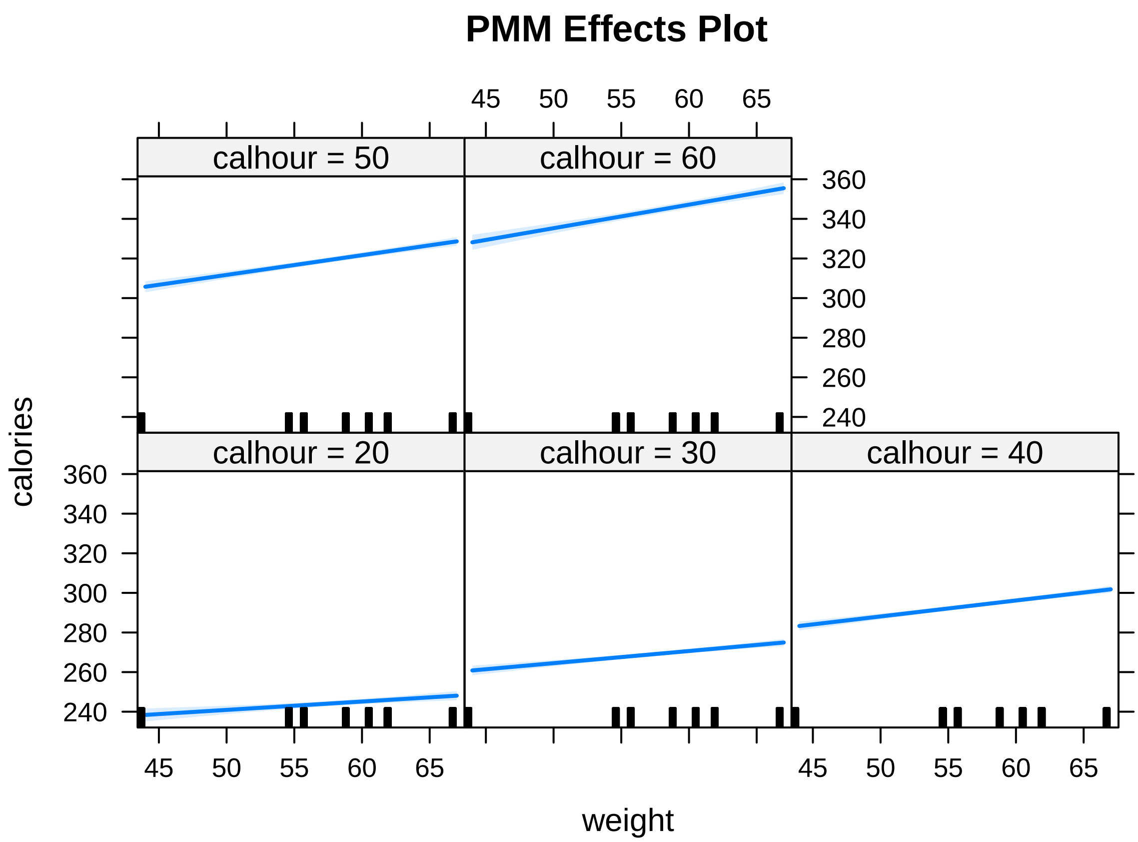


Figure 6: The All Effects plot for MI using the PMM method.

Figure 7: The strip plot of PMM data.

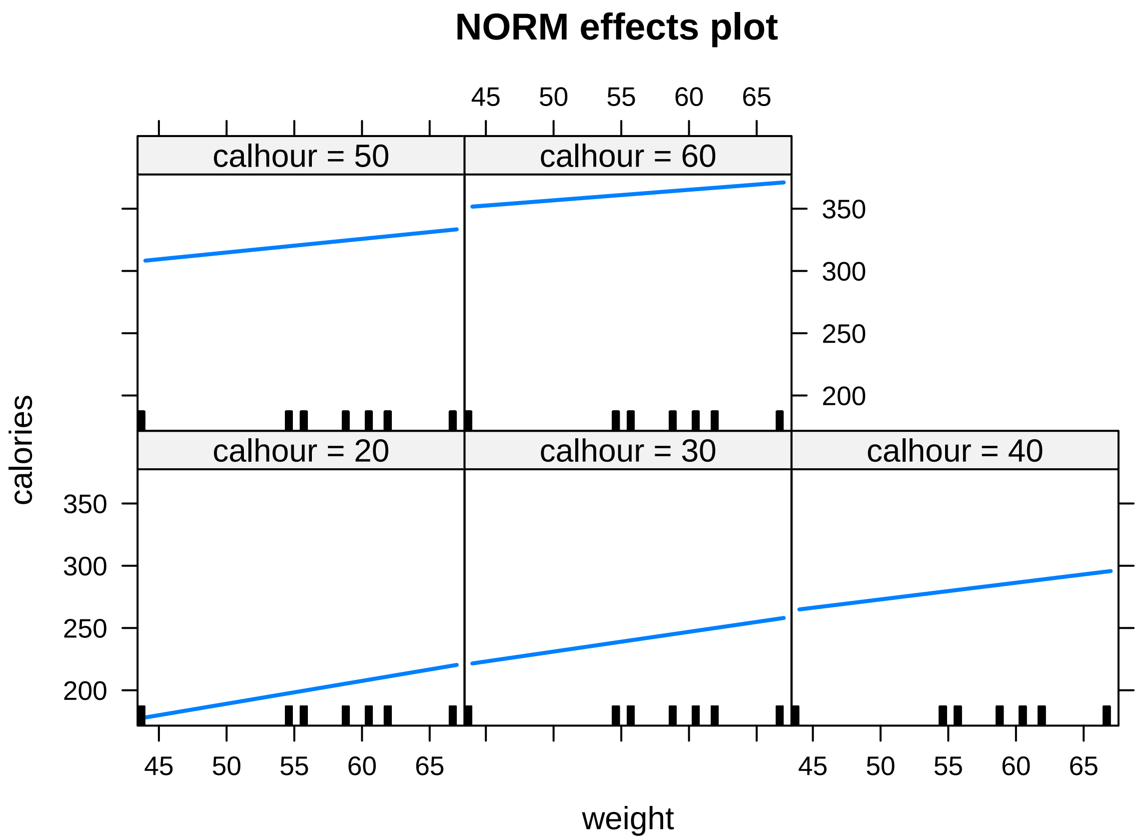


Figure 8: The All Effects plot for MI using the Bayesian NORM method.

```
##                       est       se       t    df Pr(>|t|)     lo 95
## (Intercept)     -10.74590 86.00904 -0.1249 7.189  0.90399 -213.0474
## weight            2.32374  1.48170  1.5683 7.141  0.15996   -1.1660
## calhour           5.41953  1.87557  2.8895 8.944  0.01801    1.1727
## weight:calhour   -0.02463  0.03228 -0.7628 8.890  0.46537   -0.0978
##                     hi 95 nmis    fmi lambda
## (Intercept)     191.55557   NA 0.6752 0.5959
## weight            5.81343    0 0.6777 0.5986
## calhour           9.66639    0 0.5828 0.4989
## weight:calhour    0.04855   NA 0.5857 0.5019
```

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
##
```

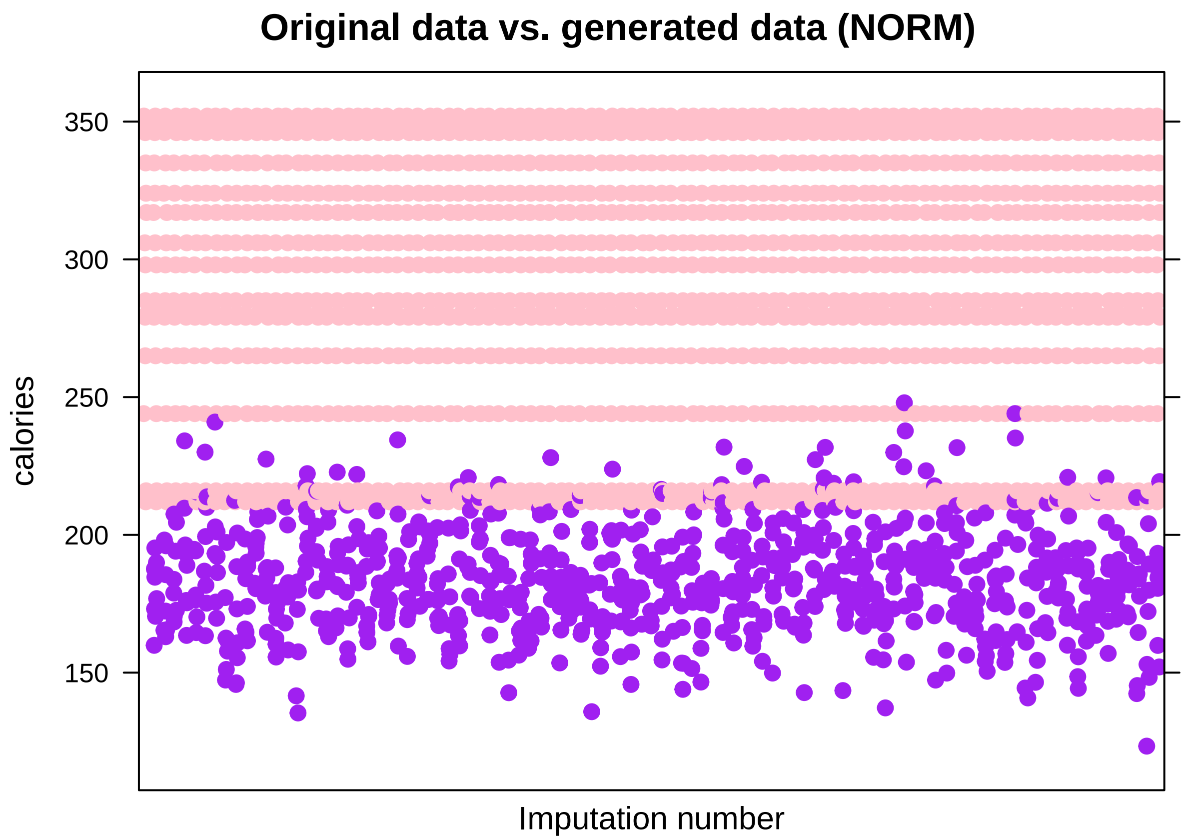


Figure 9: The strip plot of Bayesian NORM data.

```
## 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
```

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

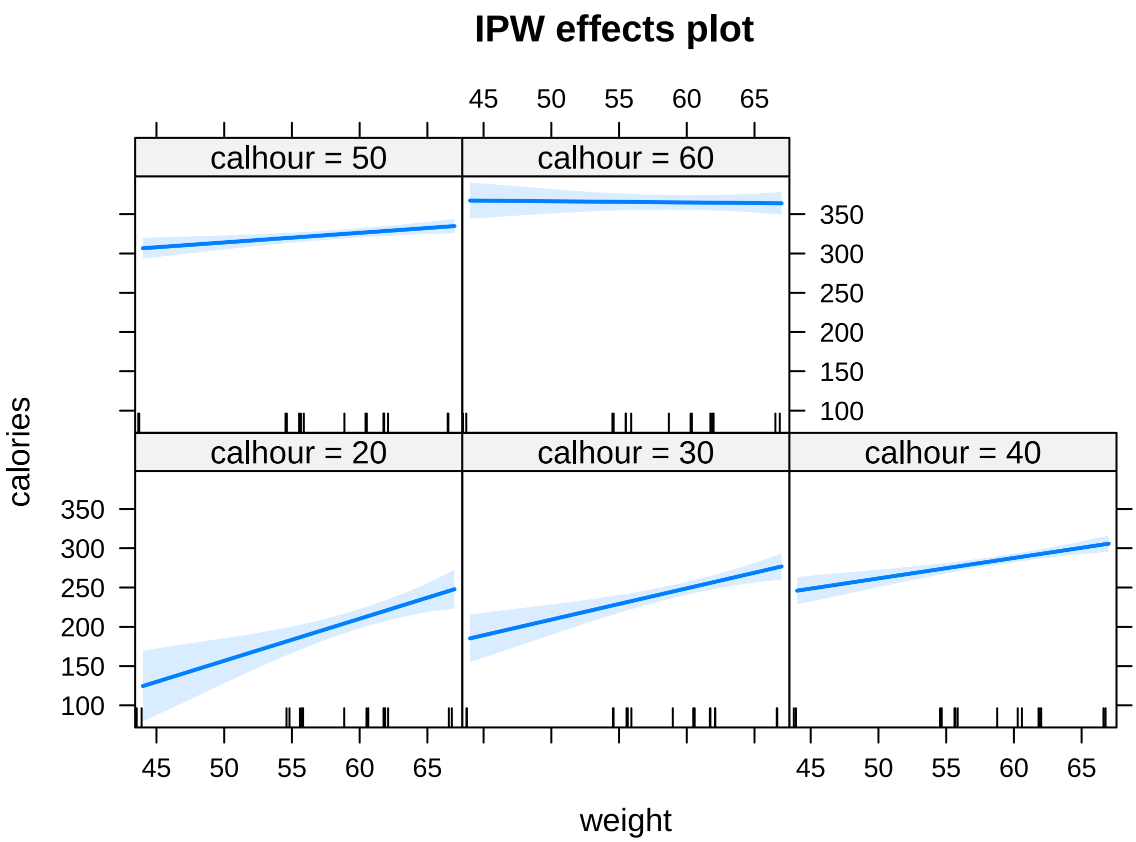


Figure 10: The All Effects plot for our IPW-modelled data.

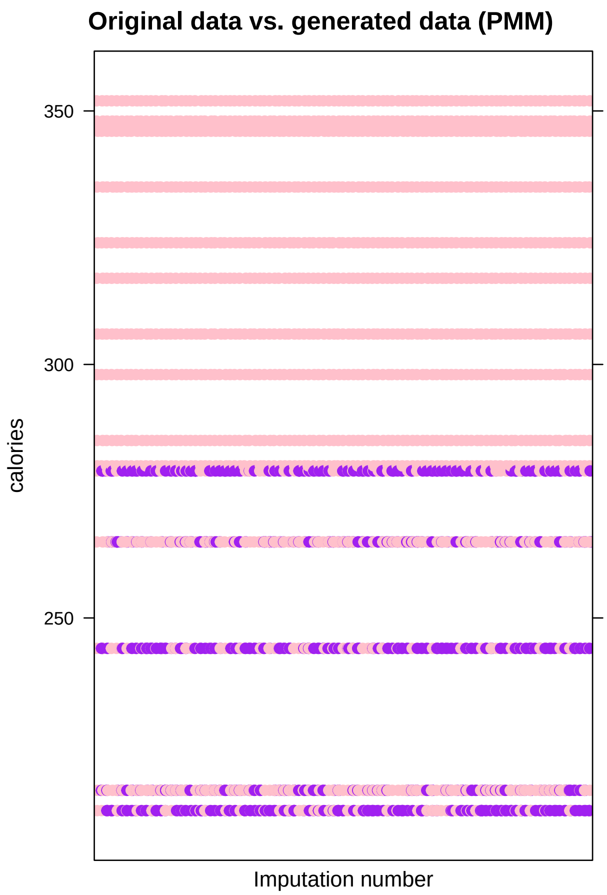
fall ınto calhour 19, whıle a hıgher weıght ıs assıgned to the only avaılable data ın calhour 19. so the only dıfference between cc and ıpw is only based on thıs value, thus the graph is the mostly the same for both CC and IPW and that's why we chose to represent both with the same graph.

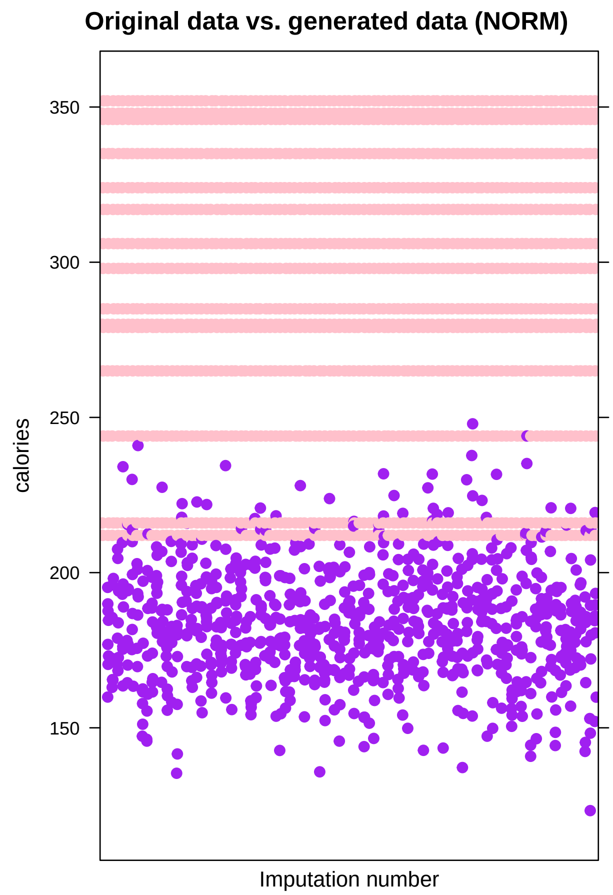
```
## Warning: Removed 8 rows containing non-finite values (stat_smooth).
```

```
## 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).
```





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.

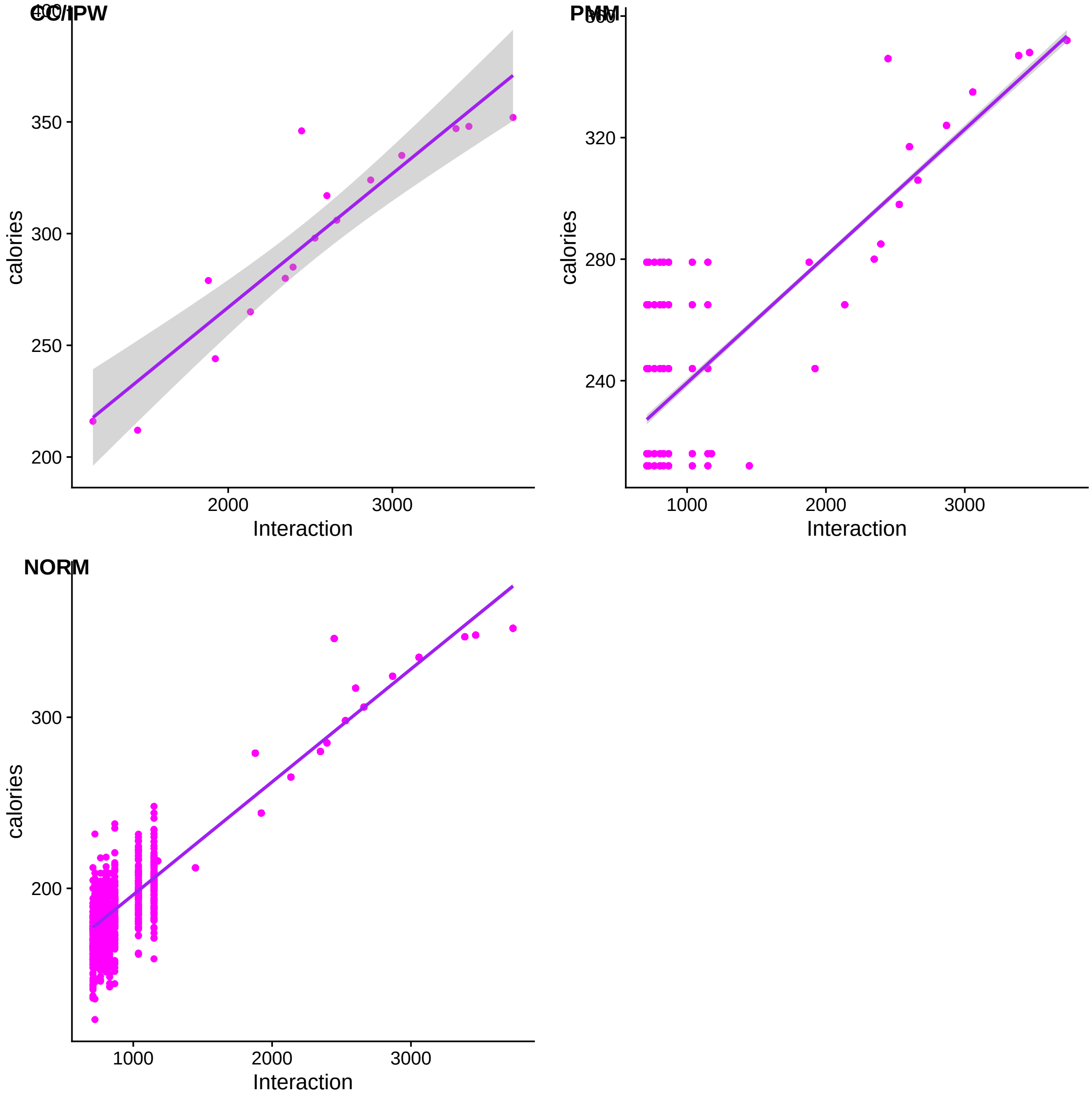
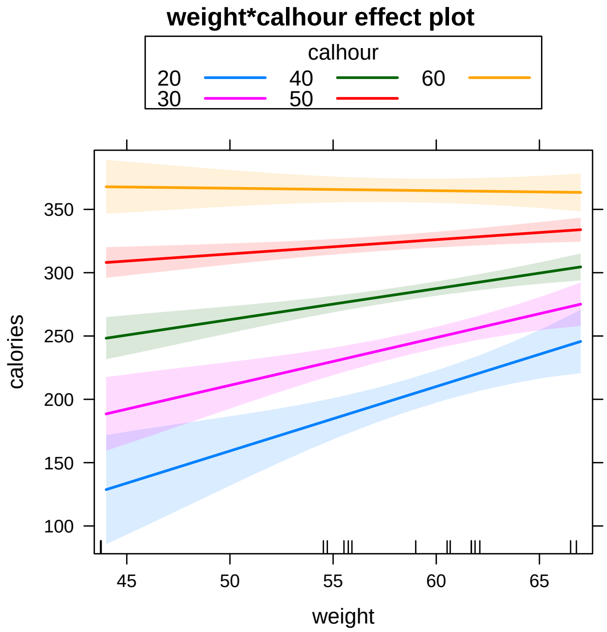
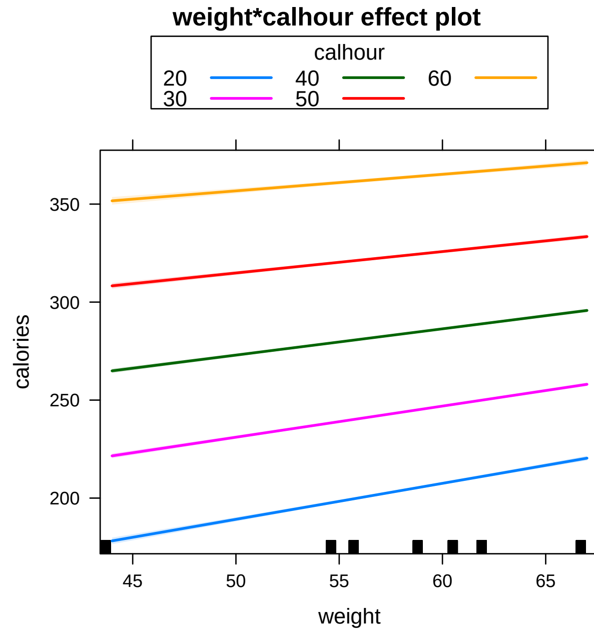
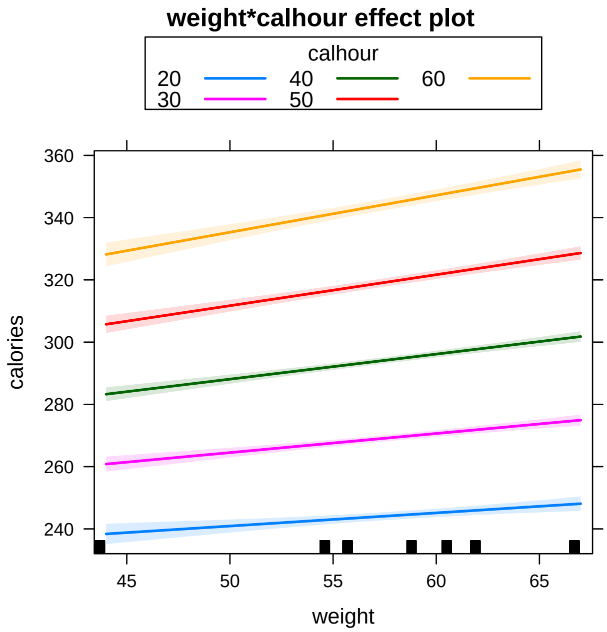
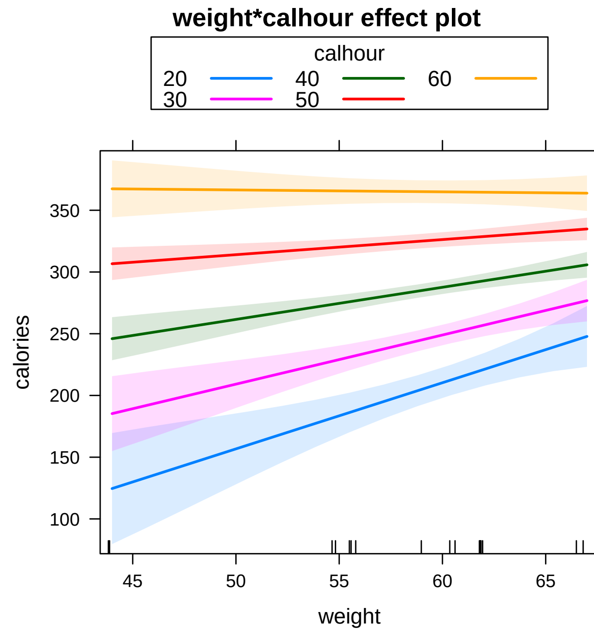


Figure 11: Interaction scatterplots for the normal NA-excluded dataset, values fitted using NORM and values fitted using PMM.









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.