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 Results **2**

2.1 Data exploration . . . 2

2.2 Missing data exploration . . . 4

2.3 Complete case analysis . . . 6

2.4 Multiple imputation analysis . . . 8

2.5 Inverse Probability Weighting analysis . . . 12

3 Discussion **12**

4 Conclusion **16**

References **16**

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 Results

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 scatterplot further indicates that the correlation between calories and workout intensity is very likely to be linear.

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:
```

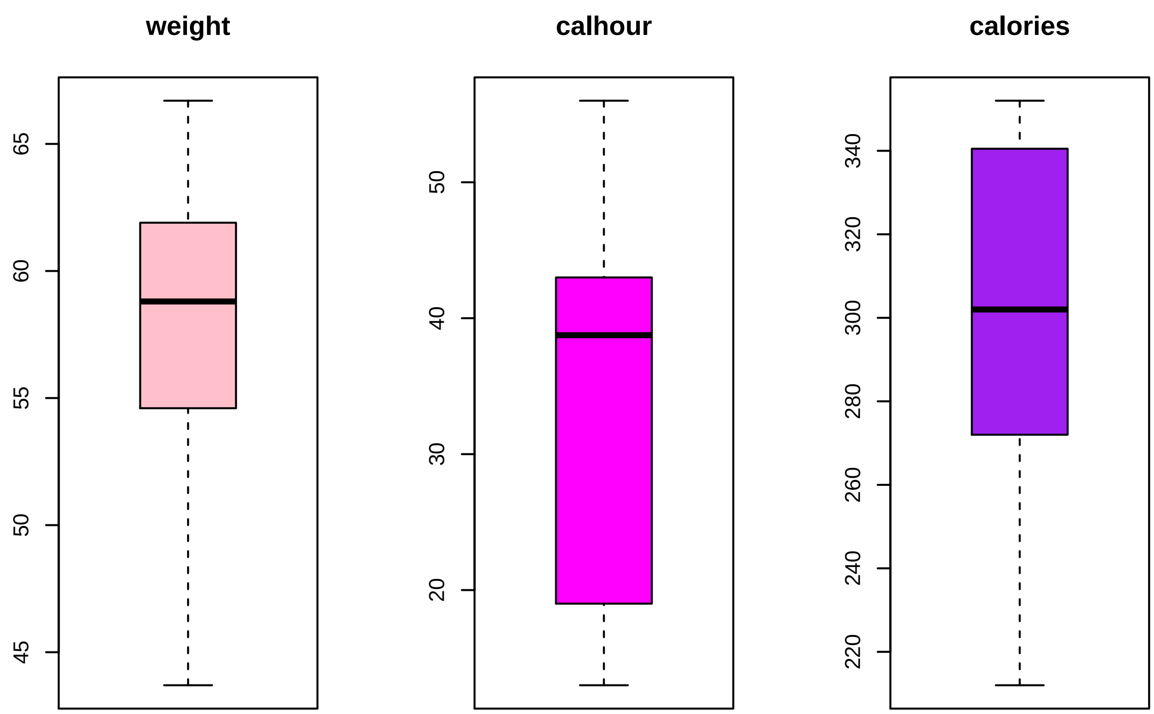


Figure 1: Boxplots for the dependent variables weight, calhour and independent variable calories.

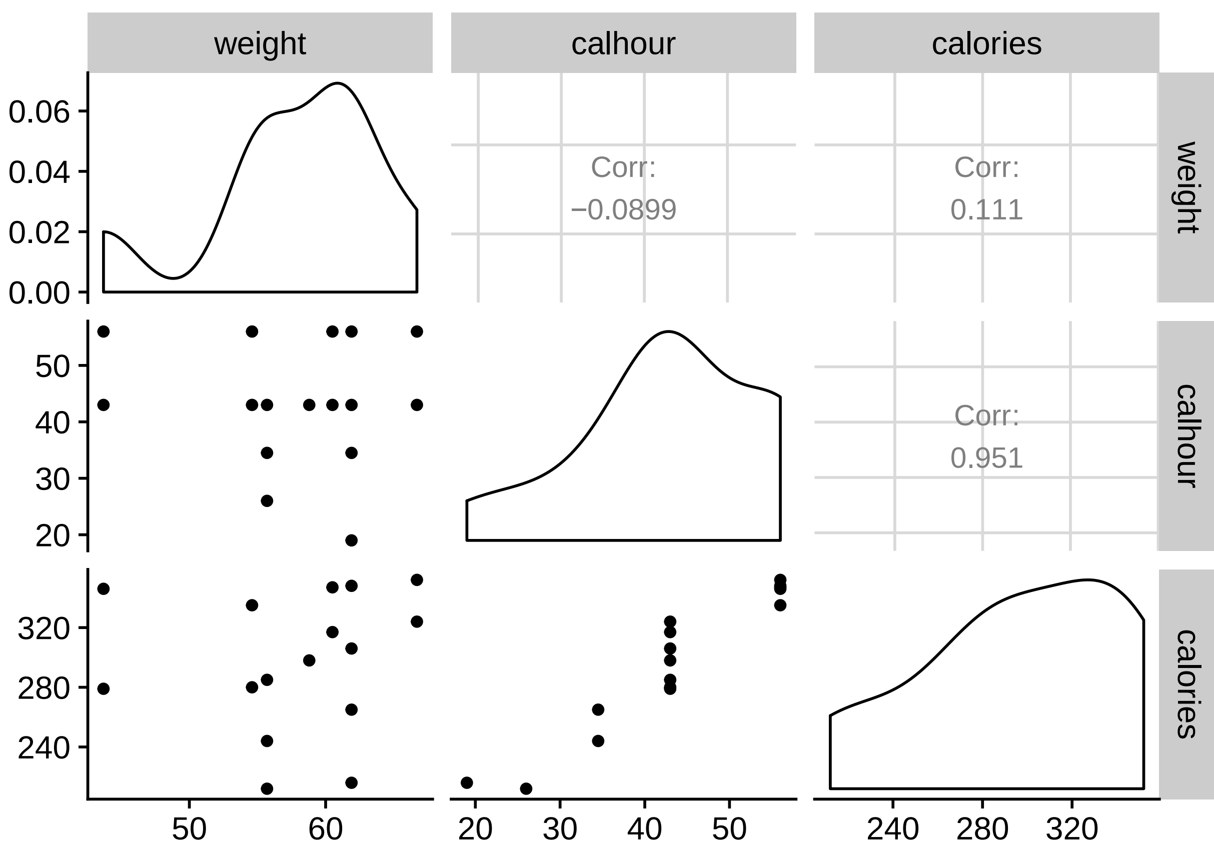


Figure 2: A summary statistics plot of the dataset using the ggplot command.

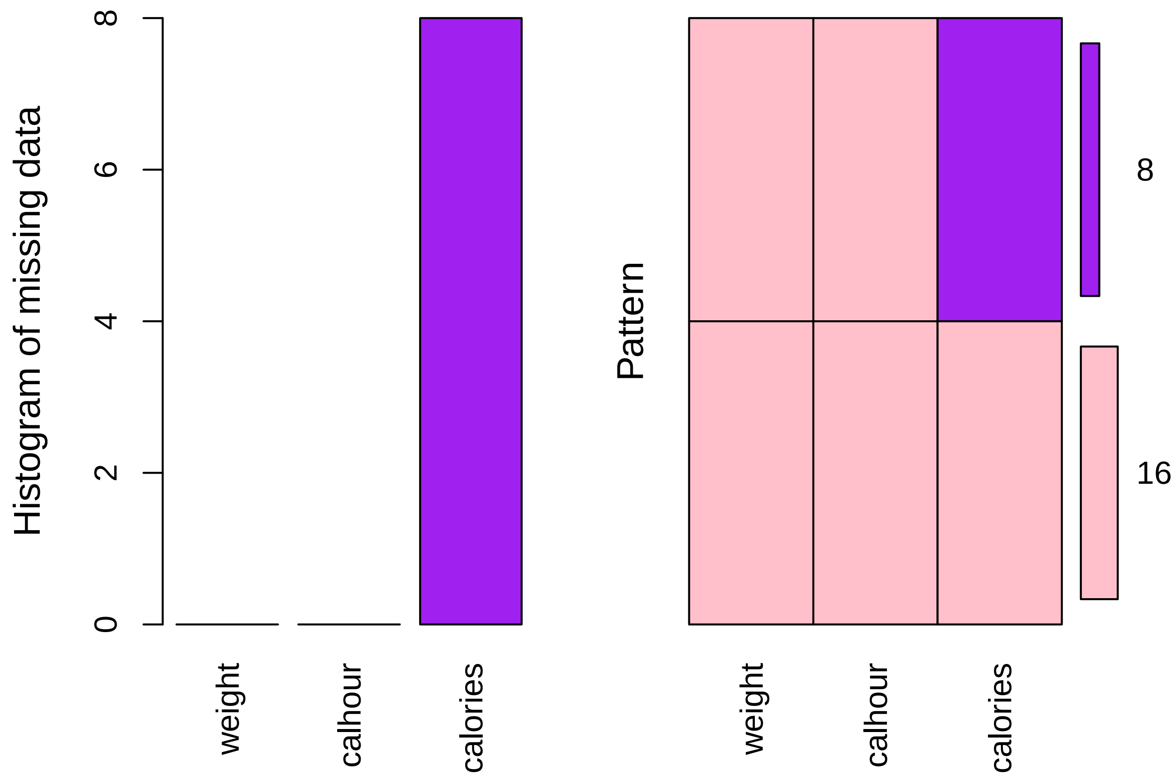


Figure 3: Patterns of missing data across variables.

```
##  0.8615 0.9832
## 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. We will nonetheless evaluate some of the more common methods as well.

2.3 Complete case analysis

Complete case analysis relies on removing rows of our dataset that have missing values - giving us a restricted sample of 16 rows to work with.

We select the best linear model to use for complete case analysis using stepwise Akaike's Information Criterion - a measure of the quality of our statistical models relative to each other, which indicates the amount of information lost by excluding or including model terms.

```
## 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
```

Lower AIC is better, thus we conclude that a model incorporating weight, calhour and an interaction term is the most explanatory linear model available given our exploratory analysis. In mathematical terms:

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

The summary for this model:

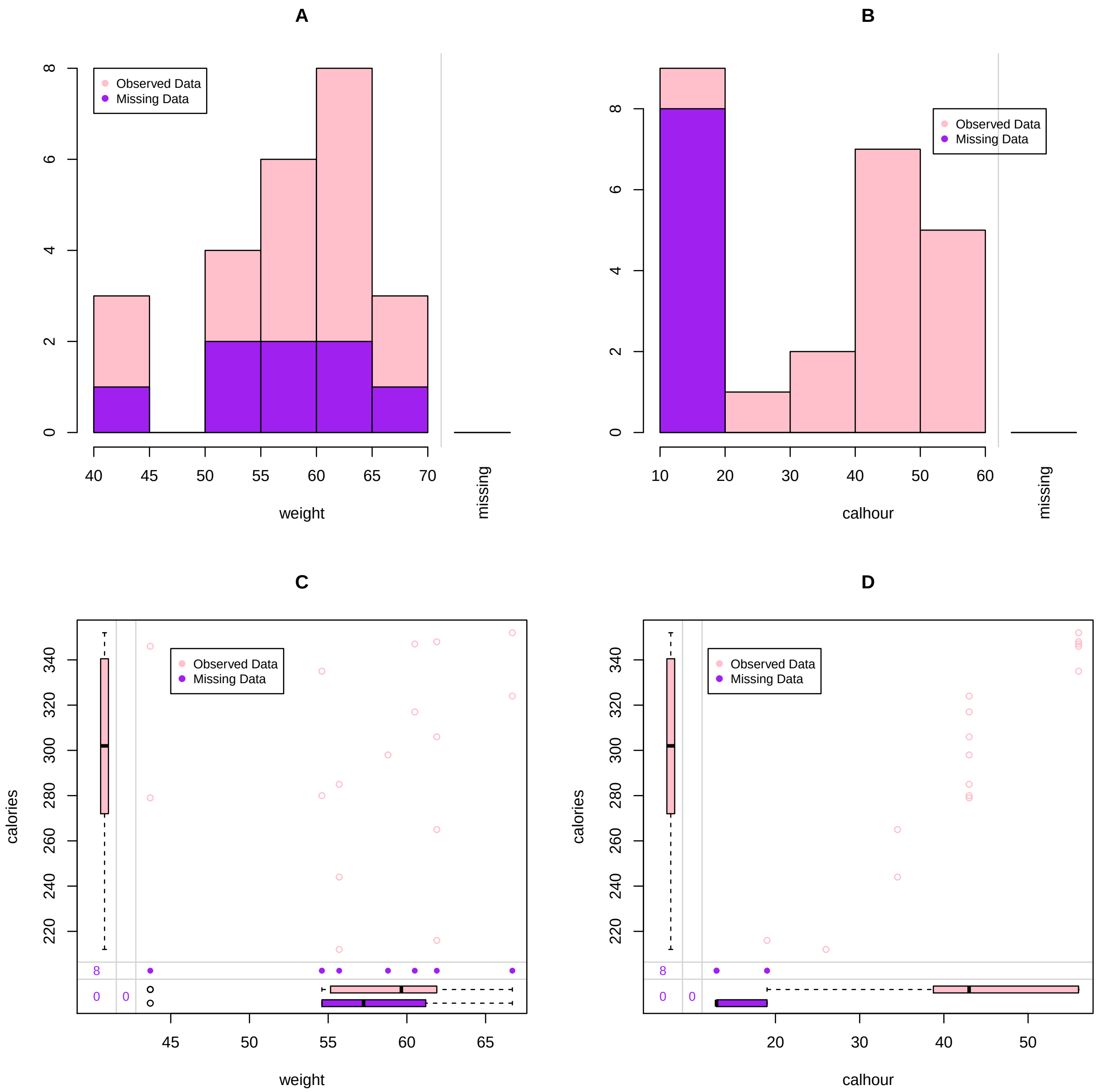


Figure 4: Histograms of the observed and missing data as well as marginplots depicting histograms and correlations.

```
##
## 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
```

All four terms appear highly significant in this model. However, the intercept term β_0 does not have a physical meaning in this model. The significance of the interaction term means that the weight has an influence on the effect of workout intensity on calories.

An effects plot is presented in Fig. 6. This plot indicates that there is a decrease in slope, determined by workout intensity and calories, as workout intensity is increasing.

2.4 Multiple imputation analysis

Multiple imputation is an approach to deal with incomplete data that can be applied to univariate or multivariate data. The technique replaces missing values with two or more imputed values. Unlike simpler single imputation methods where only a single value is imputed, such as mean imputation, multiple imputation as the name suggests replaces each missing value with multiple imputed values, in effect generating a number of datasets. Practically, this allows us to represent a variety of theoretical mechanisms for why the nonresponse occured. These differing datasets are known as multiply imputed datasets. These datasets are used to generate a matrix of regression coefficients, in essence building a regression model. We generate as many regression coefficient matrices as there are multiply imputed datasets. We then pool the regression coefficients into a single estimate which can be used to estimate variance (Rubin 2004). There are several methods of imputation available.

First we use the predictive mean matching (PMM) method. This is one of the "default" methods and it faithfully reproduces the relations present in the original data even if they happen to be nonlinear. The results of such a simulation with 100 imputed value datasets:

According to the resulting PMM model, none of the possible dependent variables have any statistically significant influence on our response variable. The effects plot in Fig 6 shows that there is no influence of the interaction term on the response variable since the slope does not change. Notice also the contraction of the 95% confidence limit due to the imputation process.

The strip plot is shown in Fig. 7 showing original data in pink and generated data in purple. We thus indeed confirm PMM-generated data follows very similar relations to the original dataset.

An alternative to PMM with our dataset is Bayesian linear regression, which can create univariate missing data. This reformulates our linear model in probabilistic terms. In this method we assume a prior distribution

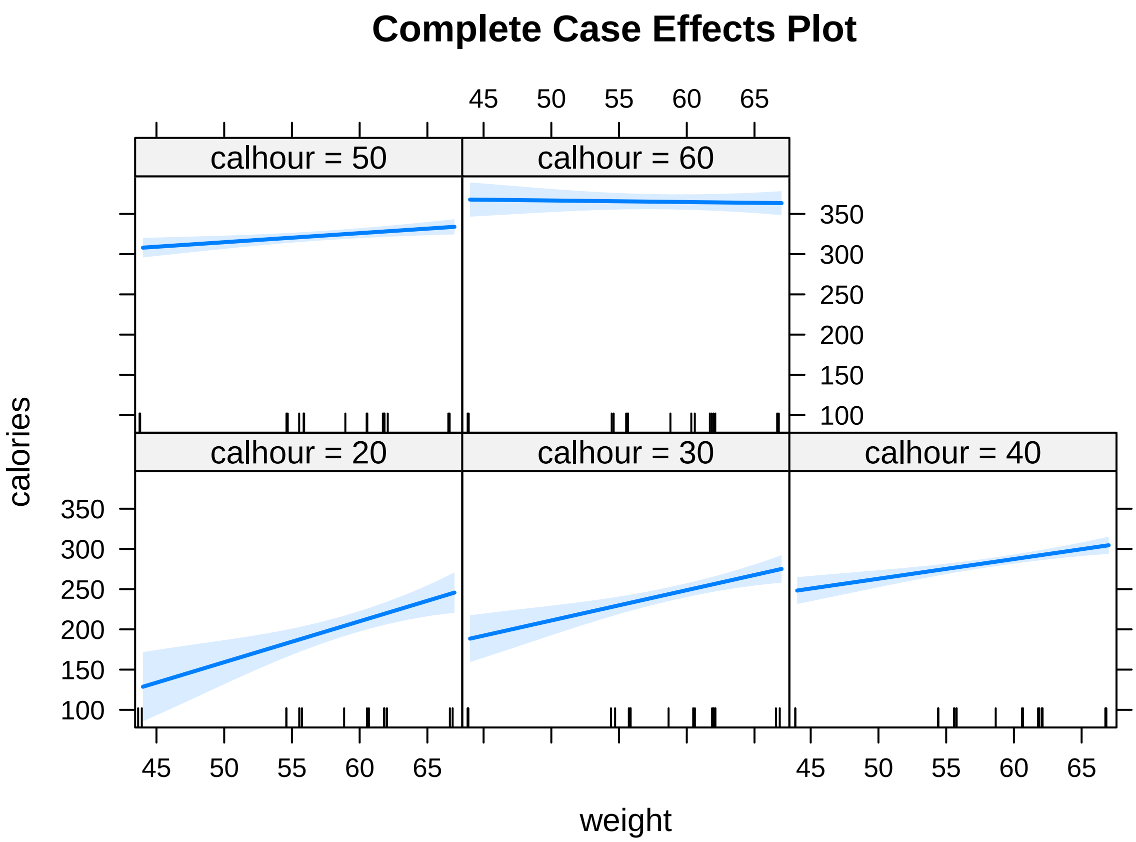


Figure 5: The All Effects plot for the Complete Case linear model.

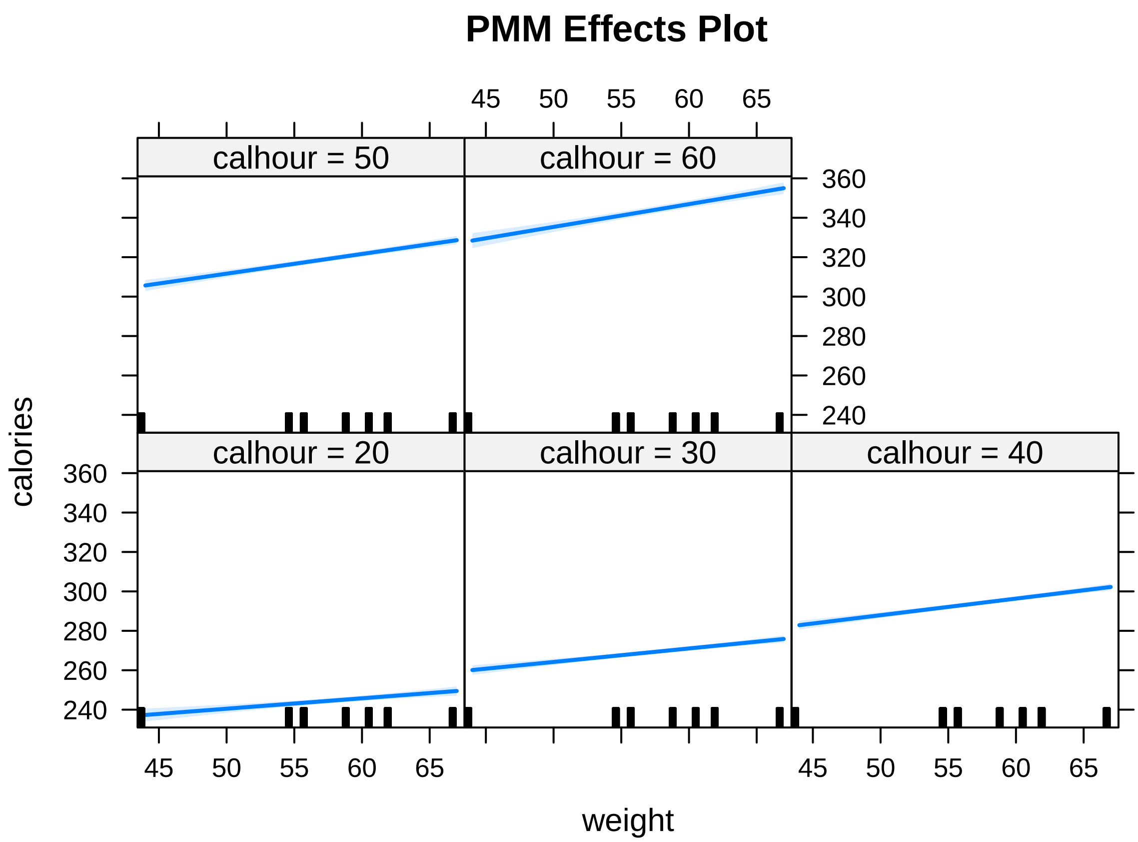


Figure 6: The All Effects plot for MI using the PMM method.

Figure 7: The strip plot of PMM data.

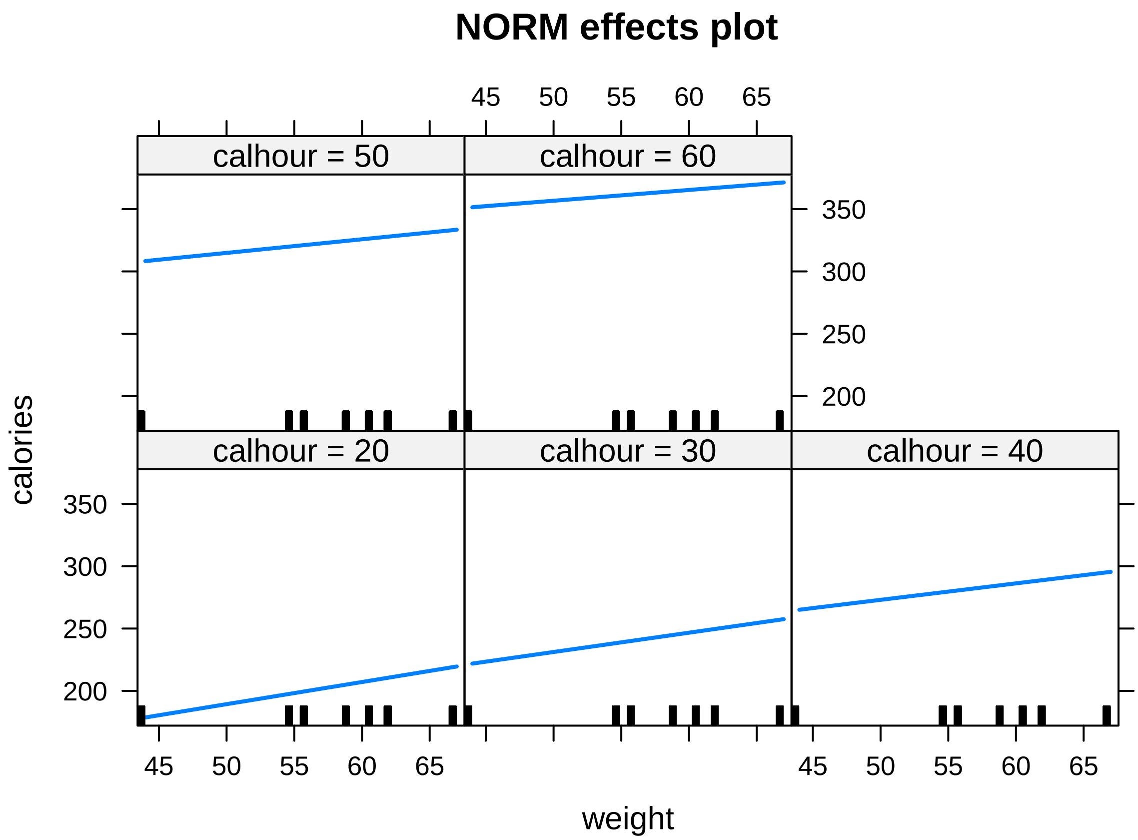


Figure 8: The All Effects plot for MI using the Bayesian NORM method.

for the parameters of the regression (Rubin 2004). In our method the prior distribution is assumed to be normal (the normal model). The result is:

```
##                      est       se         t    df  Pr(>|t|)      lo 95
## (Intercept)     -5.66533 86.02692 -0.06586 7.214   0.94928 -207.8670
## weight           2.22718  1.45323  1.53257 7.456   0.16664   -1.1670
## calhour          5.31666  1.87909  2.82937 8.946   0.01986    1.0619
## weight:calhour -0.02268  0.03181 -0.71284 9.200   0.49363   -0.0944
##                    hi 95 nmis    fmi lambda
## (Intercept)    196.53637   NA 0.6739 0.5945
## weight           5.62137    0 0.6612 0.5810
## calhour          9.57137    0 0.5827 0.4988
## weight:calhour   0.04905   NA 0.5694 0.4850
```

According to our Bayesian model, the workout intensity is the only variable that has a statistically significant influence on our response variable. The effects plot in Fig. 8 demonstrates a similar finding to the PMM effects plot in that due to preserving the slope remaining the same the interaction variable clearly does not have much effect. The strip plot in Fig. 9 highlights the probabilistic nature of the Bayesian imputation algorithm, the imputed data points being sampled from the obtained posterior distribution of our parameters.

2.5 Inverse Probability Weighting analysis

The Inverse Probability Weighting method attemps to mitigate the bias introduced by complete case studies if the excluded population appears to be systematically different from the complete cases. The complete cases

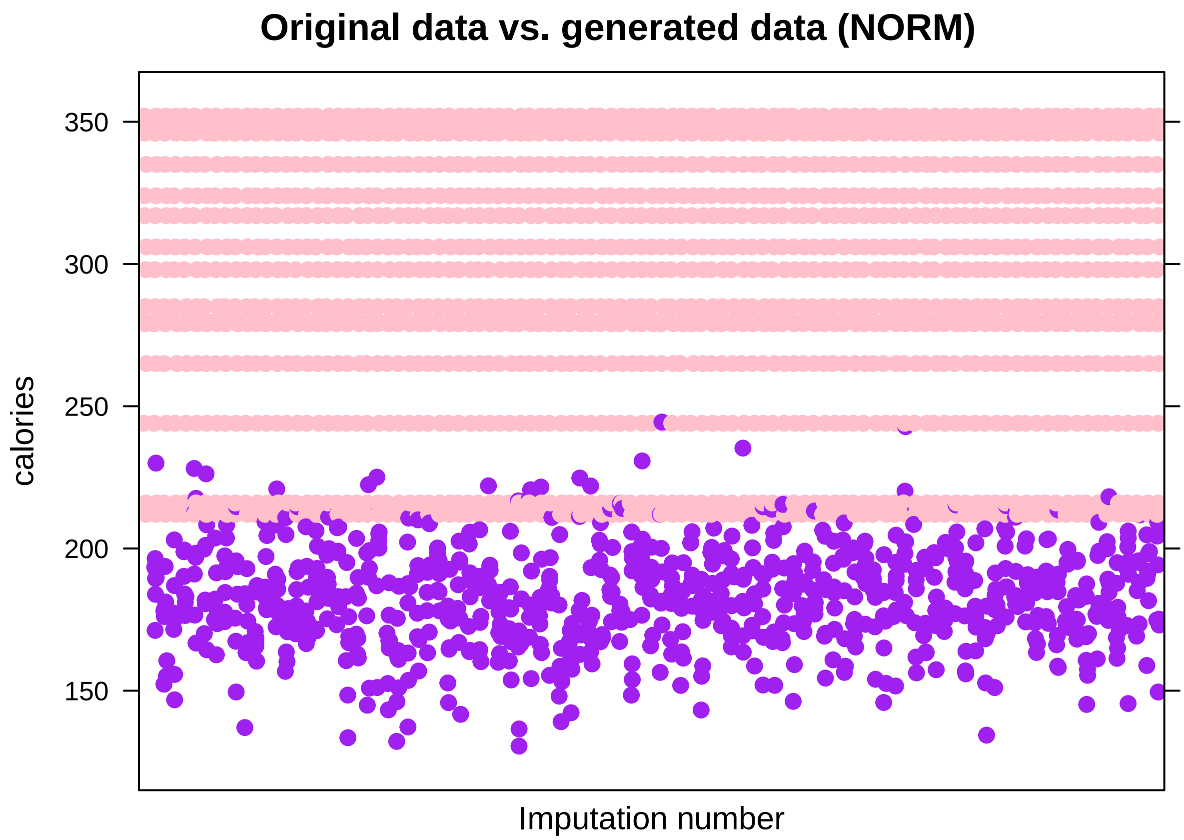


Figure 9: The strip plot of Bayesian NORM data.

are weighted using the inverse probability of their being a complete case. As you may recall, this is intuitively a very plausible model for our data since a mechanistic hypothesis for the MAR present is due to technical noise. To emphasize, in IPW the analytical model is only fitted to complete cases. The results are thus:

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

```
##
## 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
```

All the parameters are highly significant, similarly to the complete cases analysis as can be expected. Fig. 10 effects plot is also highly similar.

The relative AIC values of the complete case and IPW models can be used for comparison to validate that our IPW model is indeed better than simple CC:

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

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

We can observed that IPW yields only a miniscule improvement over CC in this case.

3 Discussion

Due to the NA values, we conducted a full model analysis with a complete case and two different methods for NA values replacement (MI and IPW). Beacuse the NA values are not evenly distributed among calhour, we decided to try different approaches for NA handling.

At Fig. **??** we are comparing the 4 effect plots previously shown. In our MI models, there are no significant changes in the slope which means that the interaction factor plays no role over the weight and calories as we change the workout intensity values. This supports the p-value we observed while building the model that showed us no significance for the interaction term. Nevertheless, to be able to compare all our models, and since the MI models with less parameters showed no overall improvement, we decided to use the same parameters to generate all models. For both the Complete Case and IPW model we can observe a change in the slope for higher values of workout intensity. This leads us to conclude that the interaction term is significant for this models as shown by their respective p-values.

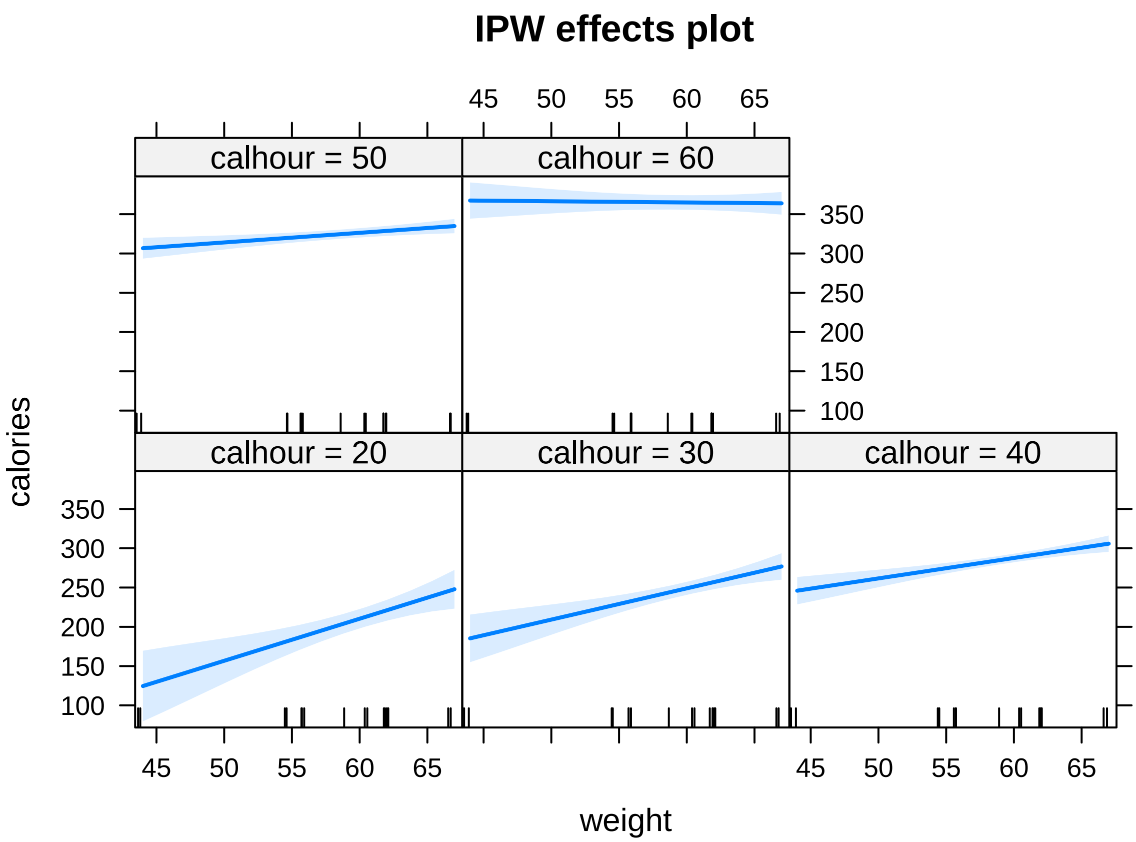
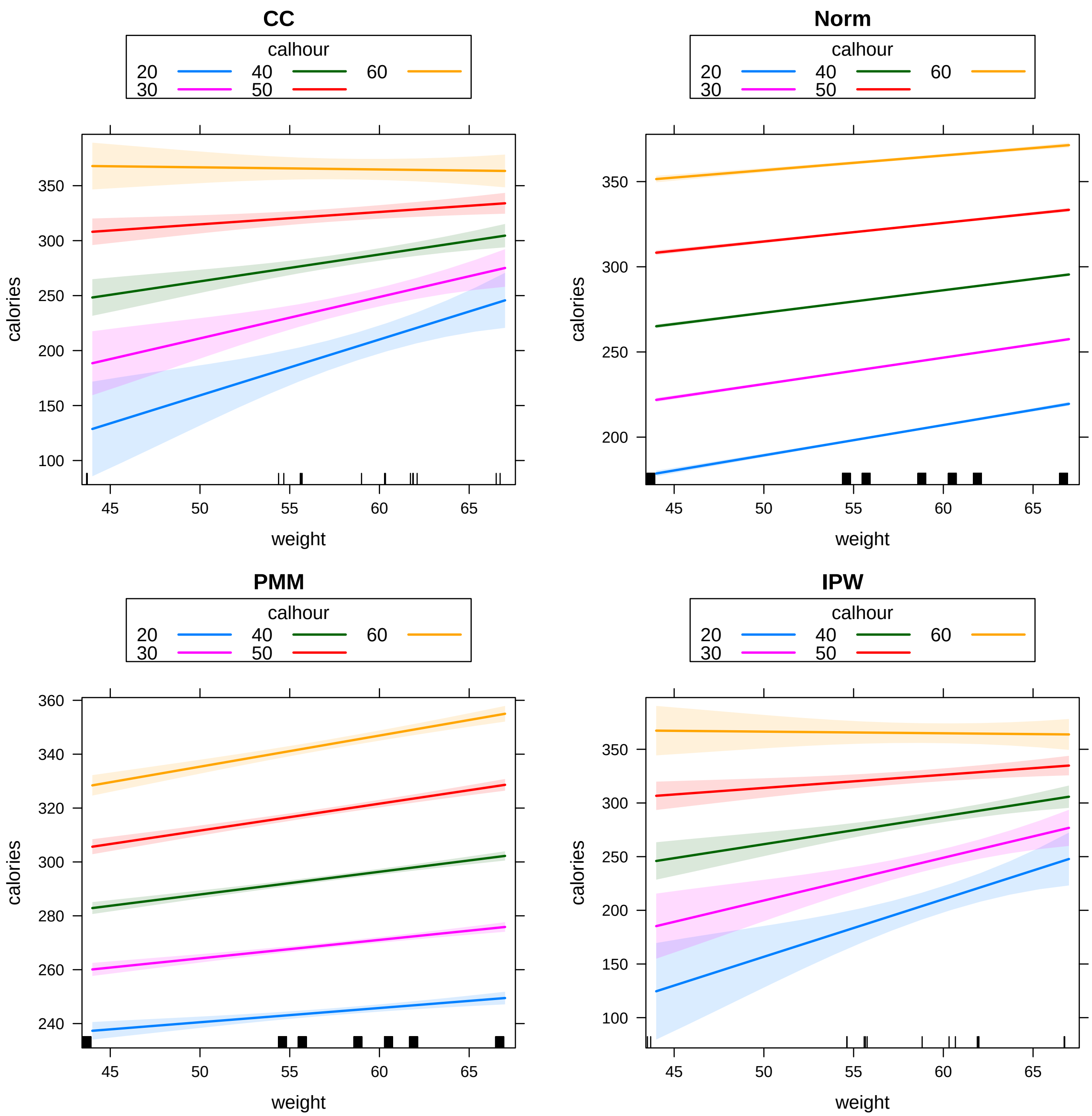


Figure 10: The All Effects plot for our IPW-modelled data.

Coefficient	Explanation	CC			IPW		
		Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
β0	Intercept	-330.884	124.674	0.02102	-351.223	110.4449	0.00792
β1	Weight	7.728	2.106	0.00321	8.0343	1.8406	0.00092
β2	Calhour	11.787	2.548	0.00058	12.205	2.344	0.00022
β3	Weight*Calhour	-0.132	0.043	0.00977	-0.1382	0.0391	0.0041
		MI-PMM			MI-NORM		
β0	Intercept	171.837	158.033	0.387	-2.572	84.992	0.96315
β1	Weight	0.384	2.731	0.7772	2.175	1.436	0.19008
β2	Calhour	1.775	3.525	0.5408	5.258	1.854	0.2313
β3	Weight*Calhour	0.013	0.061	0.9508	-0.022	0.031	0.54274

Figure 11: The coefficients for the different chosen models.



In the following three graphs in Fig. 12 we can see that the behaviour of the interaction factor vs. calories is somewhat simillar for the CC model and the two models created under MI. These three graphs are relevant to see how the two different methods chose in MI generate the new values. We can see in the graph for the PMM method that the line is more deviated from the original data than in the NORM graph. So, the PMM method way of generating the new values is actually bringing our model away from the original data.

```
## 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).
```

Finally, IPW assigns weights to each available observation. In our case, all calories values corresponding to workout instensity 13 are missing. Hence, the method cannot assign a weight to values that not exist. So, the

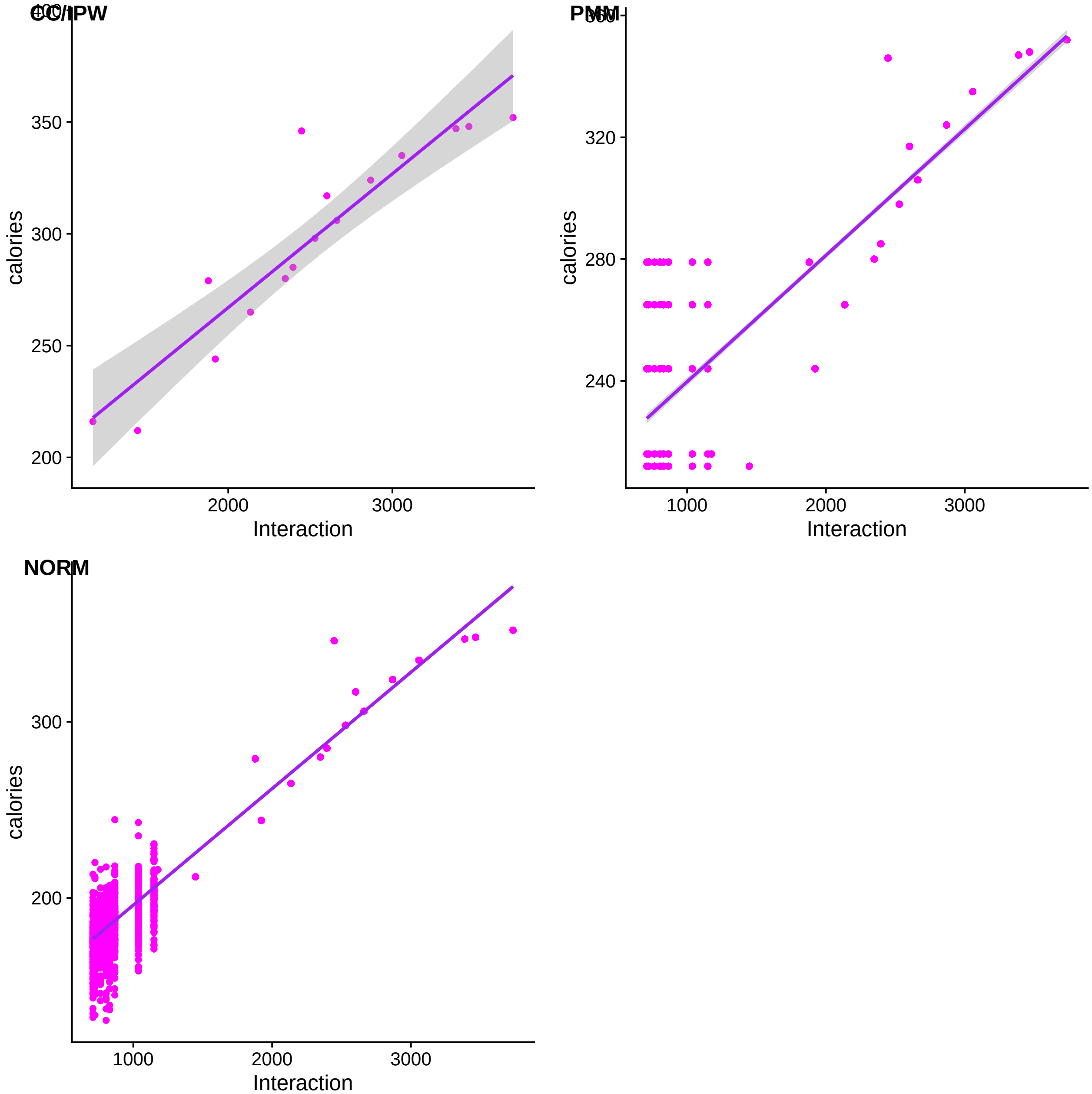


Figure 12: Interaction scatterplots for the normal NA-excluded dataset, values fitted using NORM and values fitted using PMM.

only difference between the CC and IPW model is based on value generated for the workout intensity 19 (the only other entry with an NA value). This supports all our previous graphs coefficient values for both models that are always simmilar.

4 Conclusion

Based on our discussion, and since the IPW model presented a lower AIC value than the CC model, we chose as a final model the IPW one.

$$calories_i = -351.223 + 8.0343 * weight_i + 12.205 * calhour_i - 0.1382 * (weight_i * calhour_i) + \epsilon_i$$

Analysing the model we conclude that both weight and workout intensity have a positive impact in the heat production for the individual. On the other hand, the interaction term has a slight negative impact in the heat production that is shown in previous graphs when we start to arrive at a plateau for higher values of workout intensity for variable weights. Also of importance is the intercept value -351.223 that has no physical significance as heat production is a strictly positive value.

```
##                      2.5 %    97.5 %
## (Intercept)    -635.2033 -72.38233
## weight            3.3855  12.84065
## calhour           6.3555  17.90875
## weight:calhour   -0.2347  -0.04081
```

Looking at the confidence intervals, we can affirm with a 95% confidence that our values will fall inside the presented intervals. Hence, weight and workout intensity will be positive and the interaction factor negative. This is a good parameter to estimate how the population falls under our model and not just a certain individual.

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.

Rubin, Donald B. 2004. *Multiple Imputation for Nonresponse in Surveys.* Vol. 81. John Wiley & Sons.