

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	IPW analysis	8
3	Discussion	13
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 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 significance of the two positive correlations we have found using Pearson's correlation (using Central Limit Theorem as the dataset contains around 20 rows). Here $H_0 : correlation = 0$; $H_1 : correlation \neq 0$; 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:
```

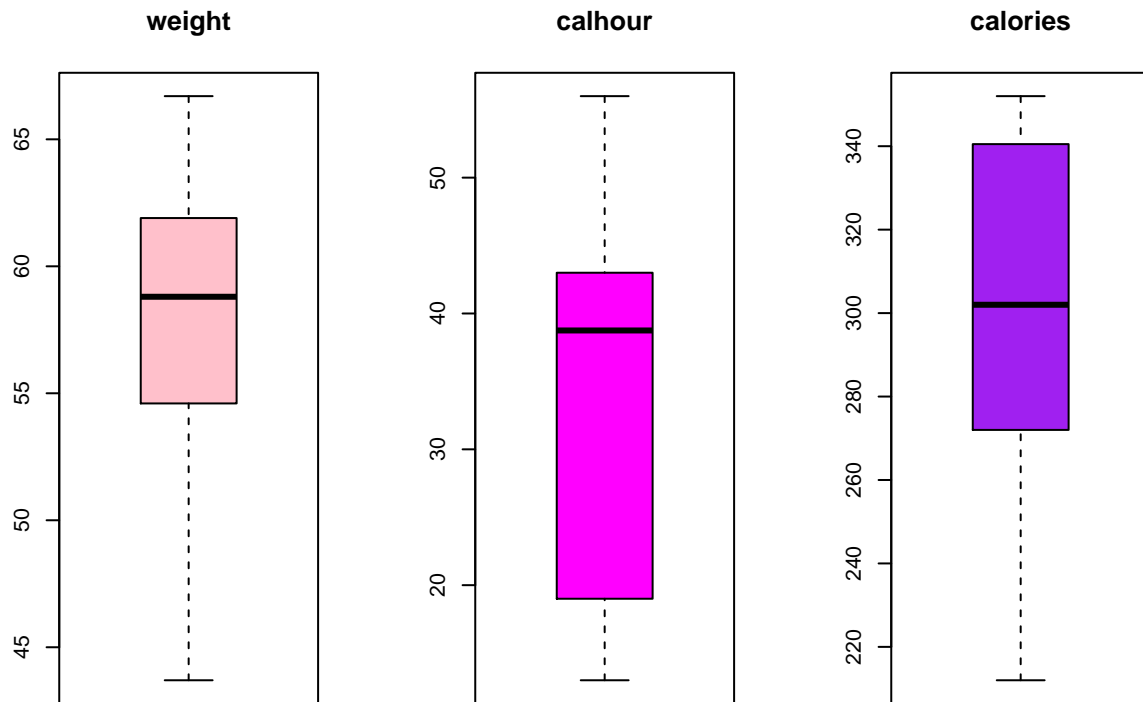


Figure 1: Boxplots for the dependent variables weight, calhour and independent variable calories.

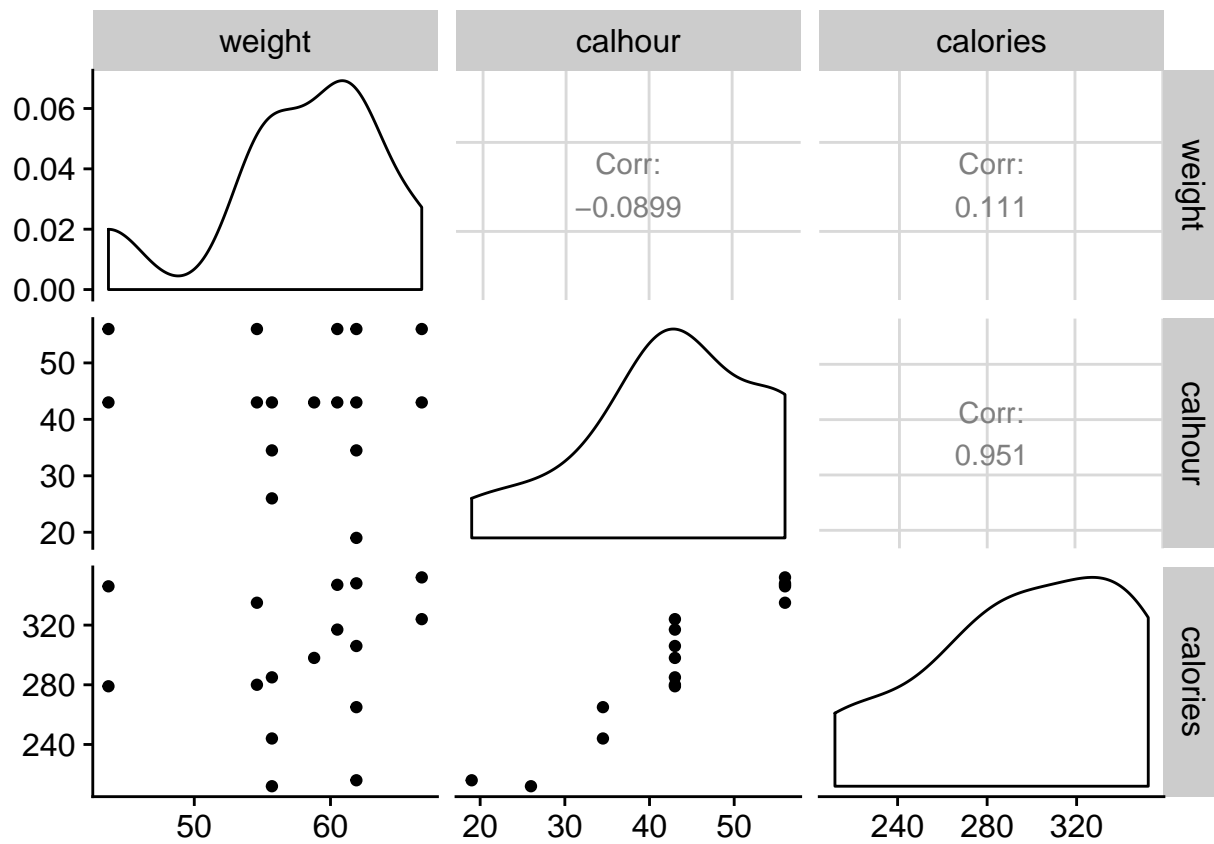


Figure 2: A summary statistics plot of the dataset using the ggplot command.

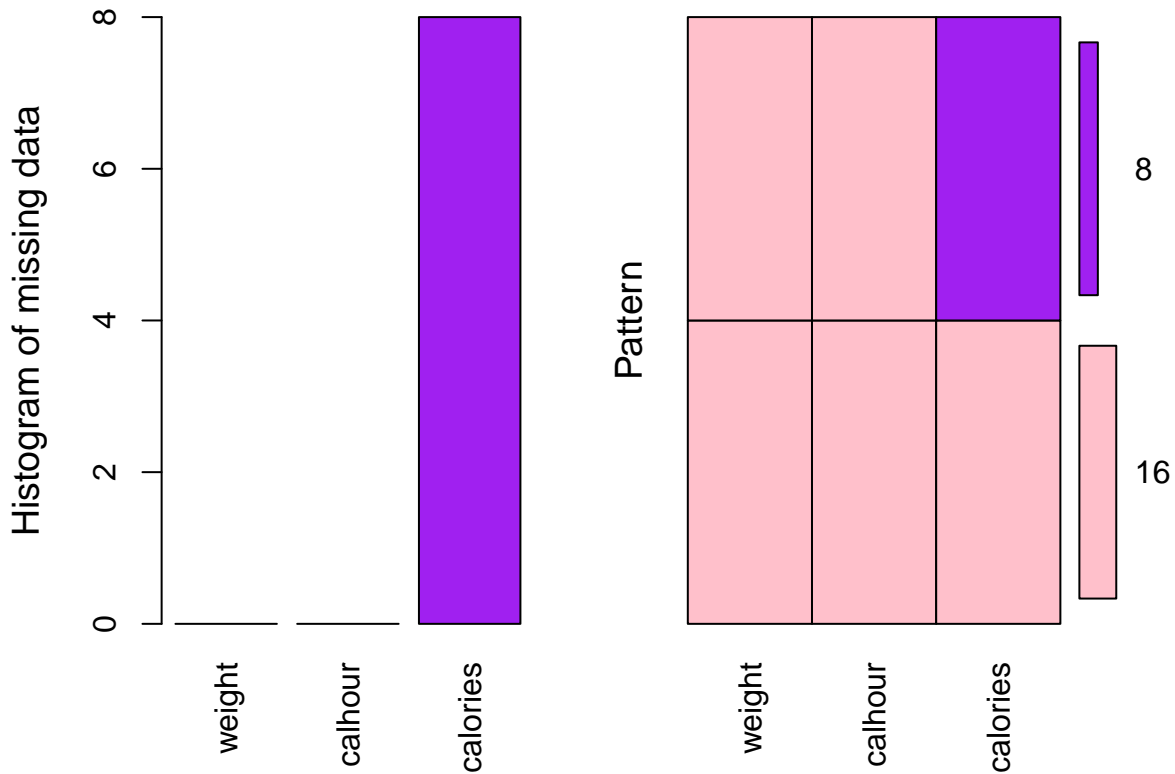


Figure 3: Patterns of missing data across variables.

```
## 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:

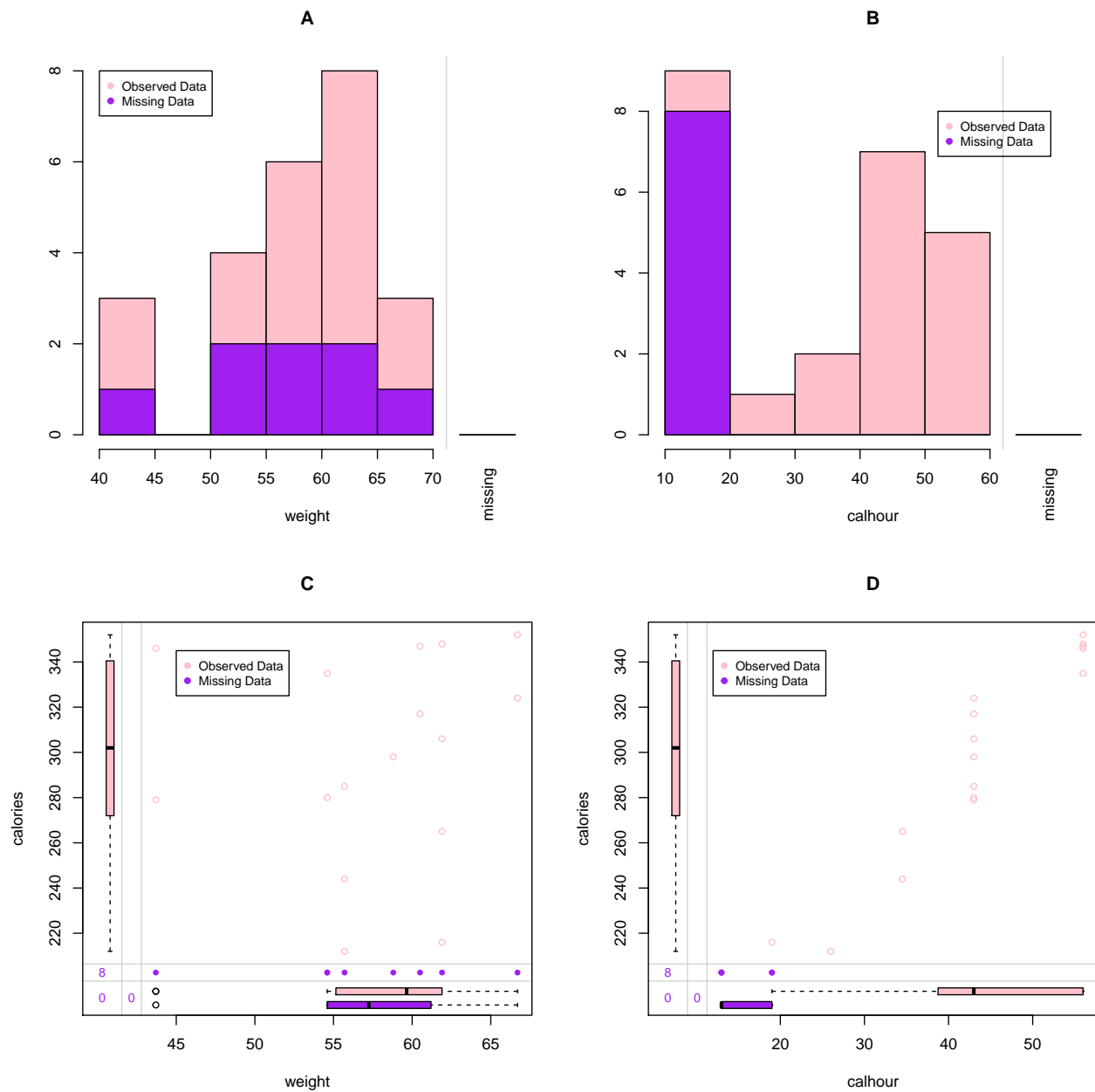


Figure 4: Histograms of the observed and missing data as well as marginplots depicting histograms and correlations.

```
##
## 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 dealing 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 occurred. 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. 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:

```
##              est          se          t    df Pr(>|t|)      lo 95
## (Intercept)  1.563e+02 151.38514 1.0322 10.40  0.3254 -179.3146
## weight       6.354e-01  2.59370 0.2450 10.44  0.8112  -5.1107
## calhour      2.101e+00  3.41041 0.6161 12.15  0.5492  -5.3194
## weight:calhour 7.408e-03  0.05846 0.1267 12.18  0.9012  -0.1198
##
##              hi 95 nmis      fmi lambda
## (Intercept)  491.8273   NA 0.5065 0.4199
## weight       6.3814    0 0.5041 0.4174
## calhour      9.5220    0 0.4150 0.3260
```

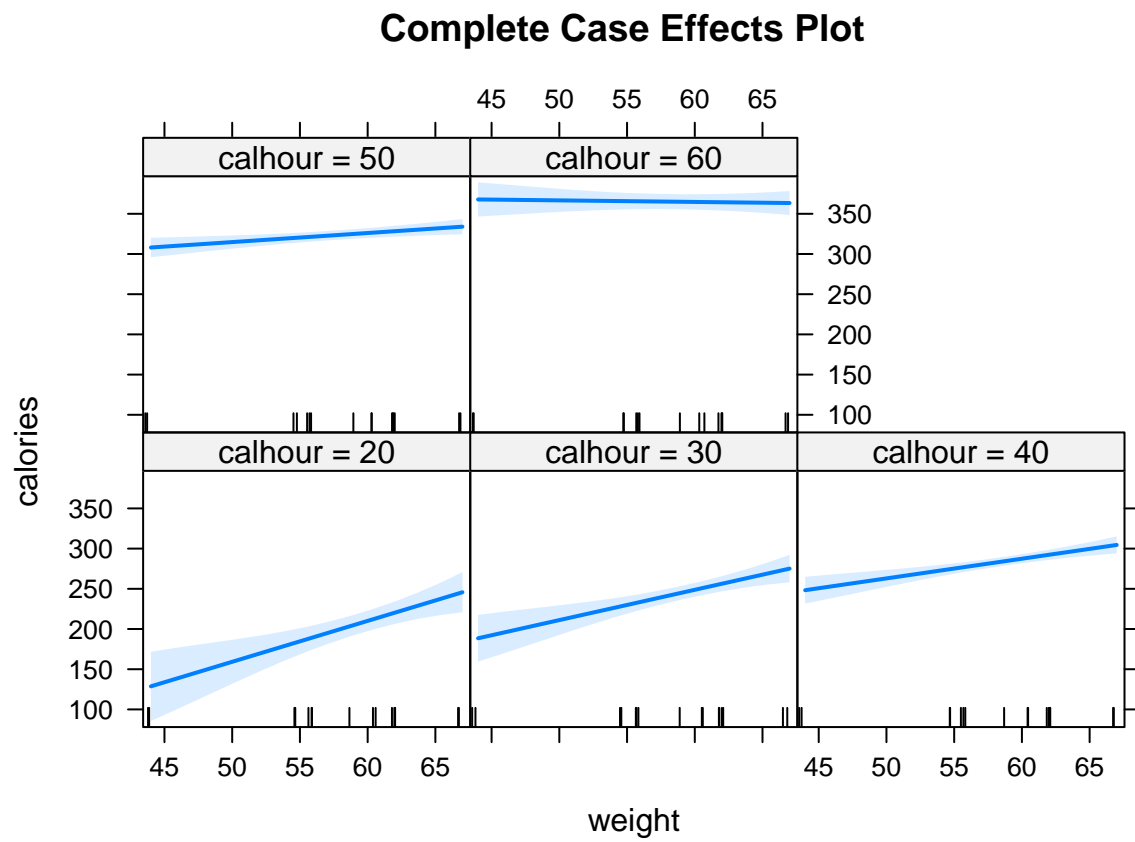


Figure 5: The All Effects plot for the Complete Case linear model.

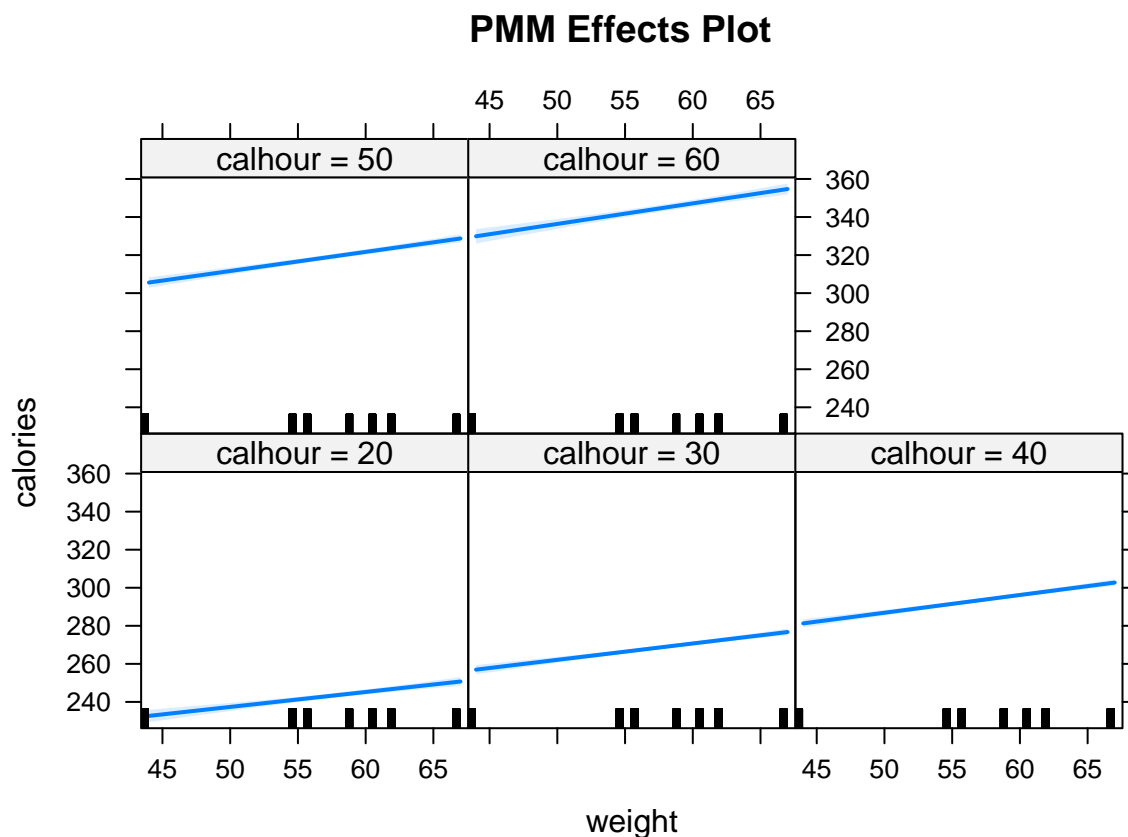


Figure 6: The All Effects plot for MI using the PMM method.

```
## weight:calhour    0.1346    NA 0.4135 0.3245
```

What if we use the Bayesian norm method?

```
##               est      se      t    df Pr(>|t|)      lo 95
## (Intercept)  -6.56646 86.90787 -0.07556 7.114  0.94185 -211.40727
## weight       2.21926  1.44880  1.53179 7.551  0.16633  -1.15657
## calhour      5.33508  1.88510  2.83013 8.946  0.01983   1.06680
## weight:calhour -0.02248 0.03154 -0.71284 9.425  0.49322  -0.09335
##               hi 95 nmis    fmi lambda
## (Intercept)  198.27434    NA 0.6792 0.6001
## weight       5.59509     0 0.6562 0.5758
## calhour      9.60336     0 0.5827 0.4988
## weight:calhour 0.04838    NA 0.5576 0.4727
```

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:
```

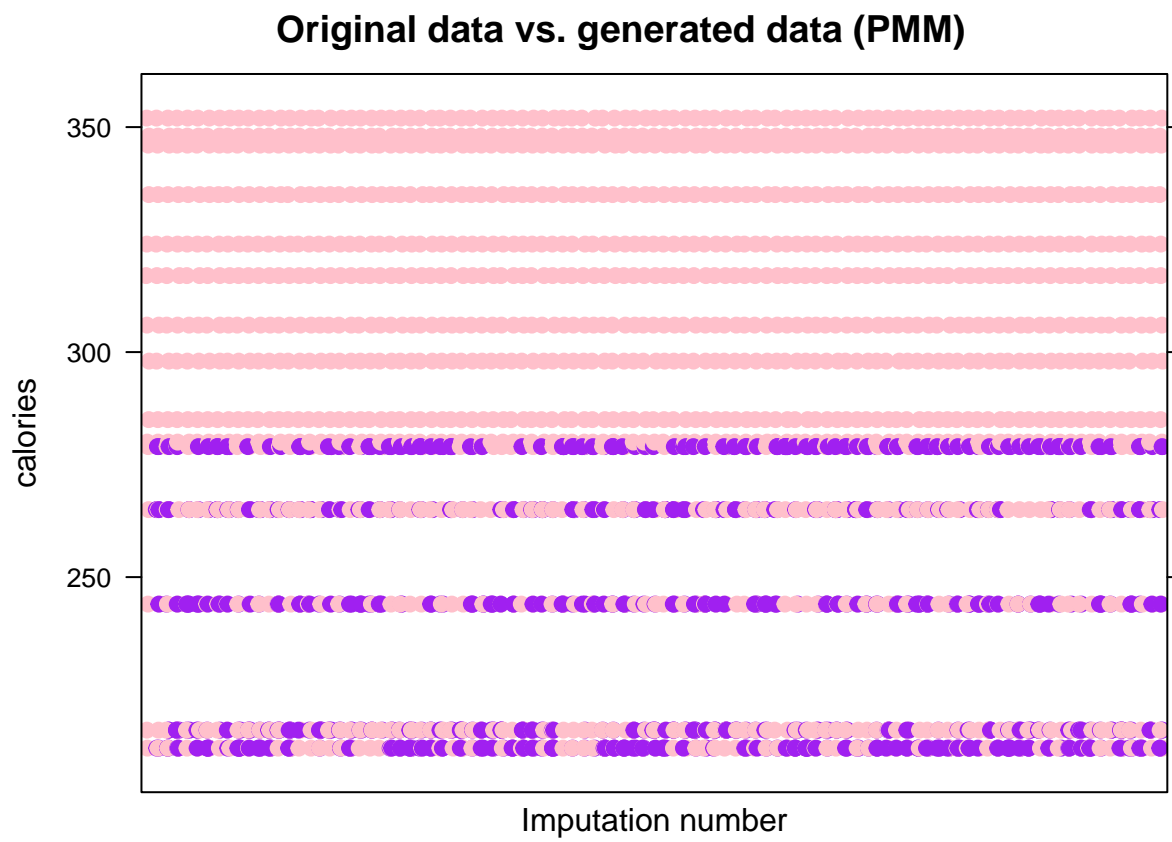


Figure 7: The strip plot of PMM data.

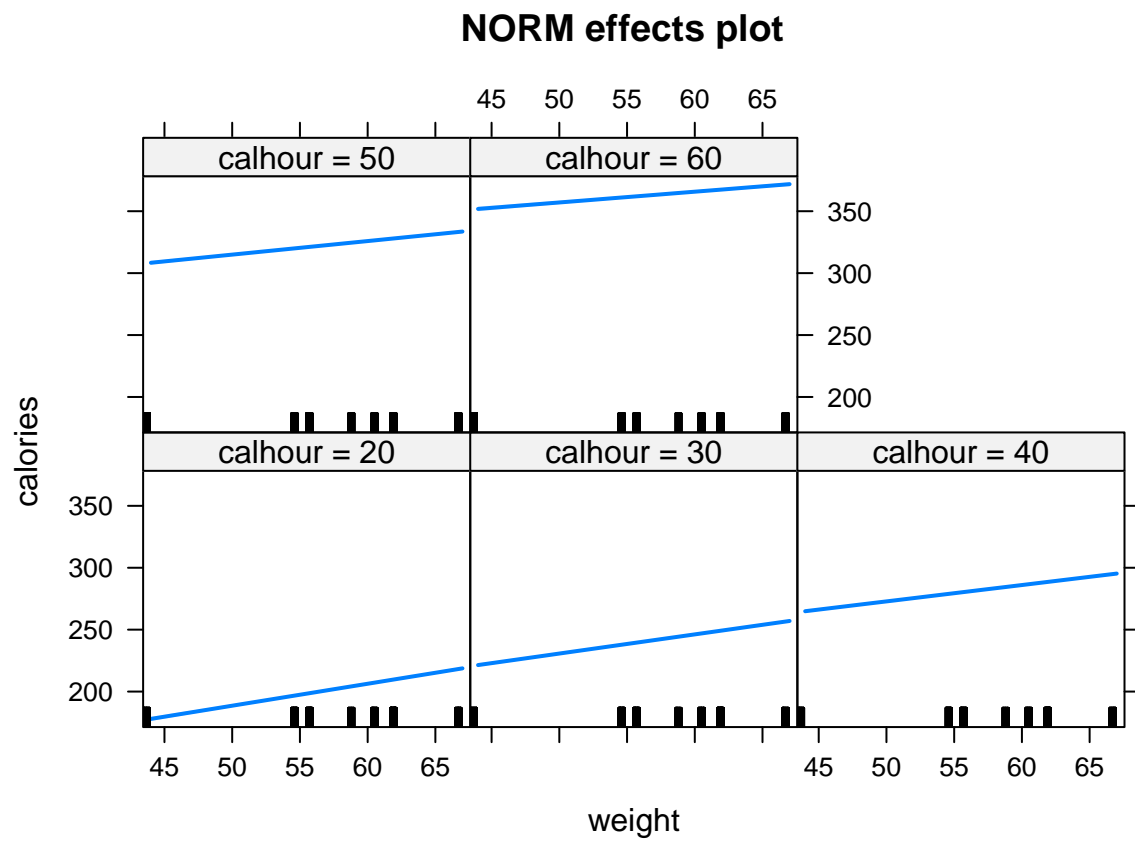


Figure 8: The All Effects plot for MI using the Bayesian NORM method.

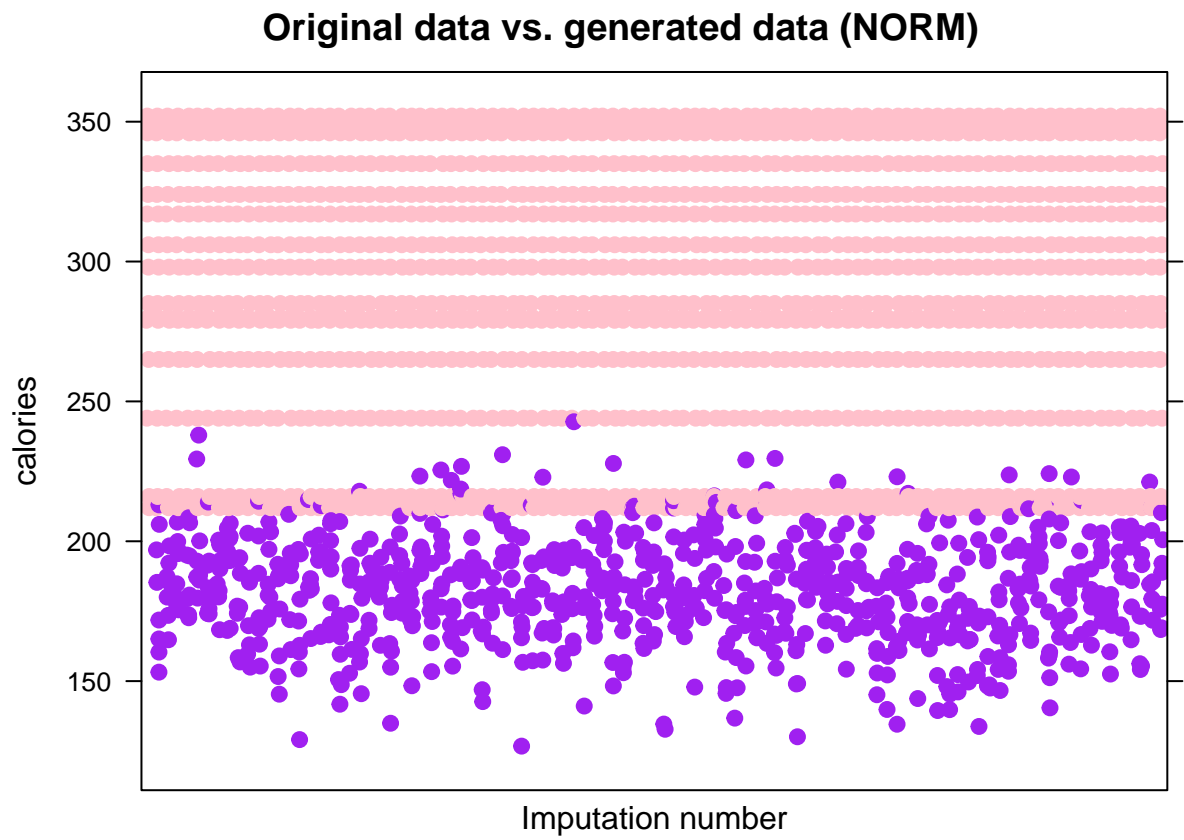


Figure 9: The strip plot of Bayesian NORM data.

```

##      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

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 comparisons (you can write this better) models. Because the NA values are not evenly distributed among calhour, we decided to try different approaches for NA handling.

IPW assigns weights to each observation so it uses already available ones. Since all calories values in calhour 13 are missing, the method cannot assign a weight. no value can represent this group, other missing values fall into calhour 19, while a higher weight is assigned to the only available data in calhour 19. so the only difference between cc and ipw is only based on this 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.

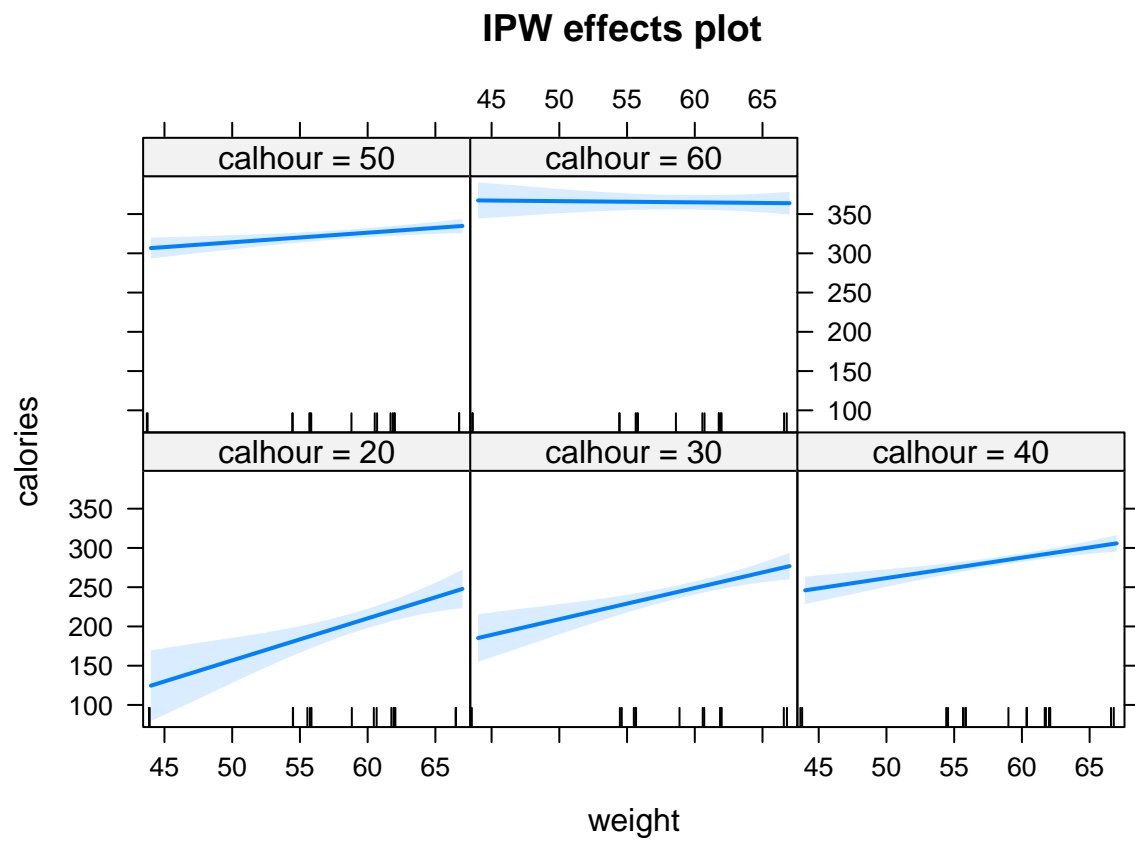


Figure 10: The All Effects plot for our IPW-modelled data.

for the MI method, we tried two ways to generate the data corresponding to the missing data, PMM and NORM. PMM generates the data according to the pattern in the observed ones. in our cases, the data is discretized by the body weight, so pmm generated the data discretized as well. in norm method, the data is generated based on normal distribution.

In our MI models, there is no changes in slope between each calhour and weight interactions. by checking the fitting model generated, the p-value corresponds to the weight and the interaction is higher than 0.05, which indicates these terms are not significant. only the calhour effects the value of calories in this estimated model, which make nonsense. in common sense, person with larger body weight generates more heats during the same intense level of workout comparing to a person with lower body weight. In the following three graphs we can see that the behaviour of the interaction factor vs. calories is similar 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.

```
## 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 attribute weights to. So, IPW will make a difference only for calhour 19. This gives us a slightly better model with IPW than complete case.

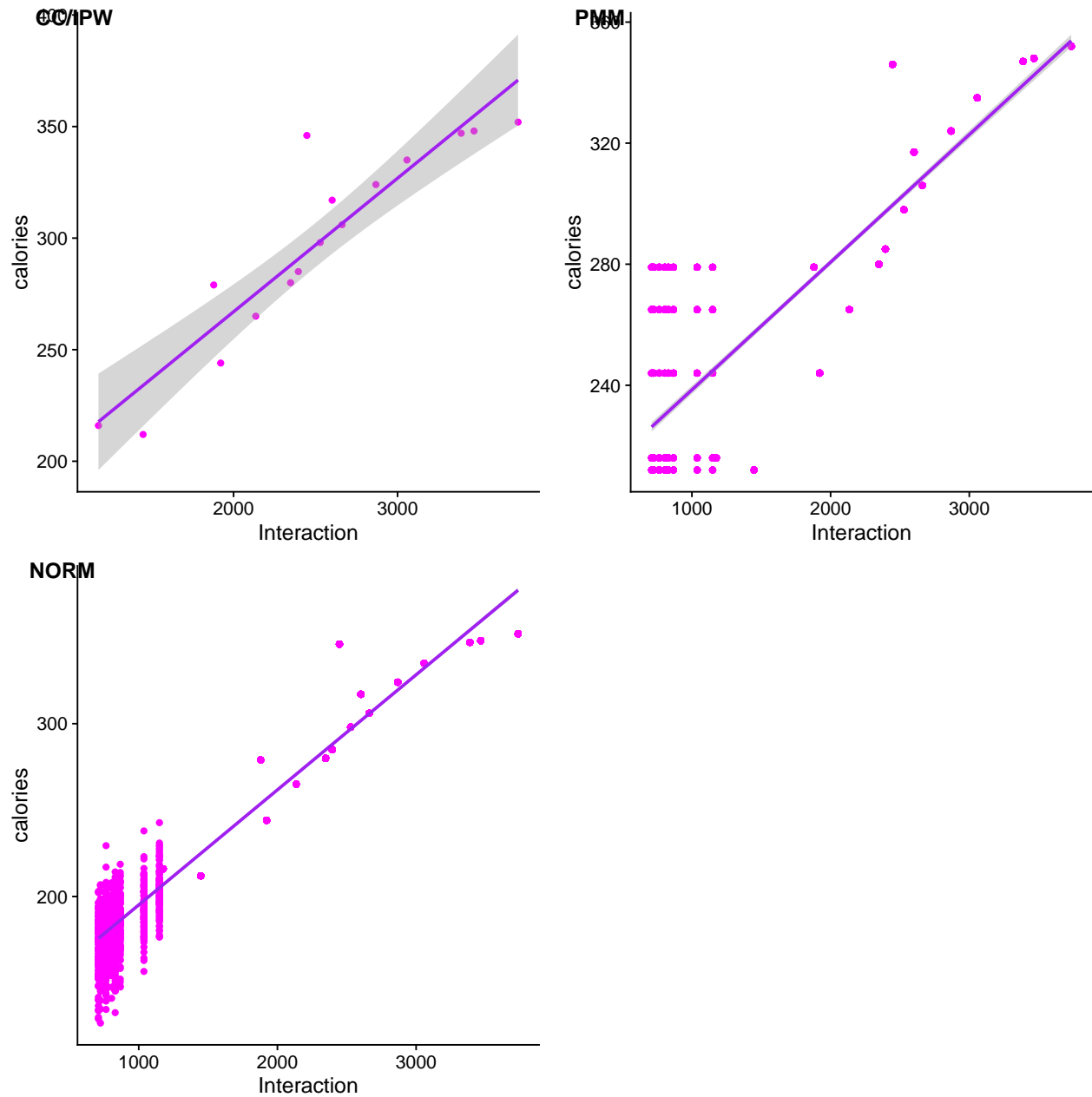
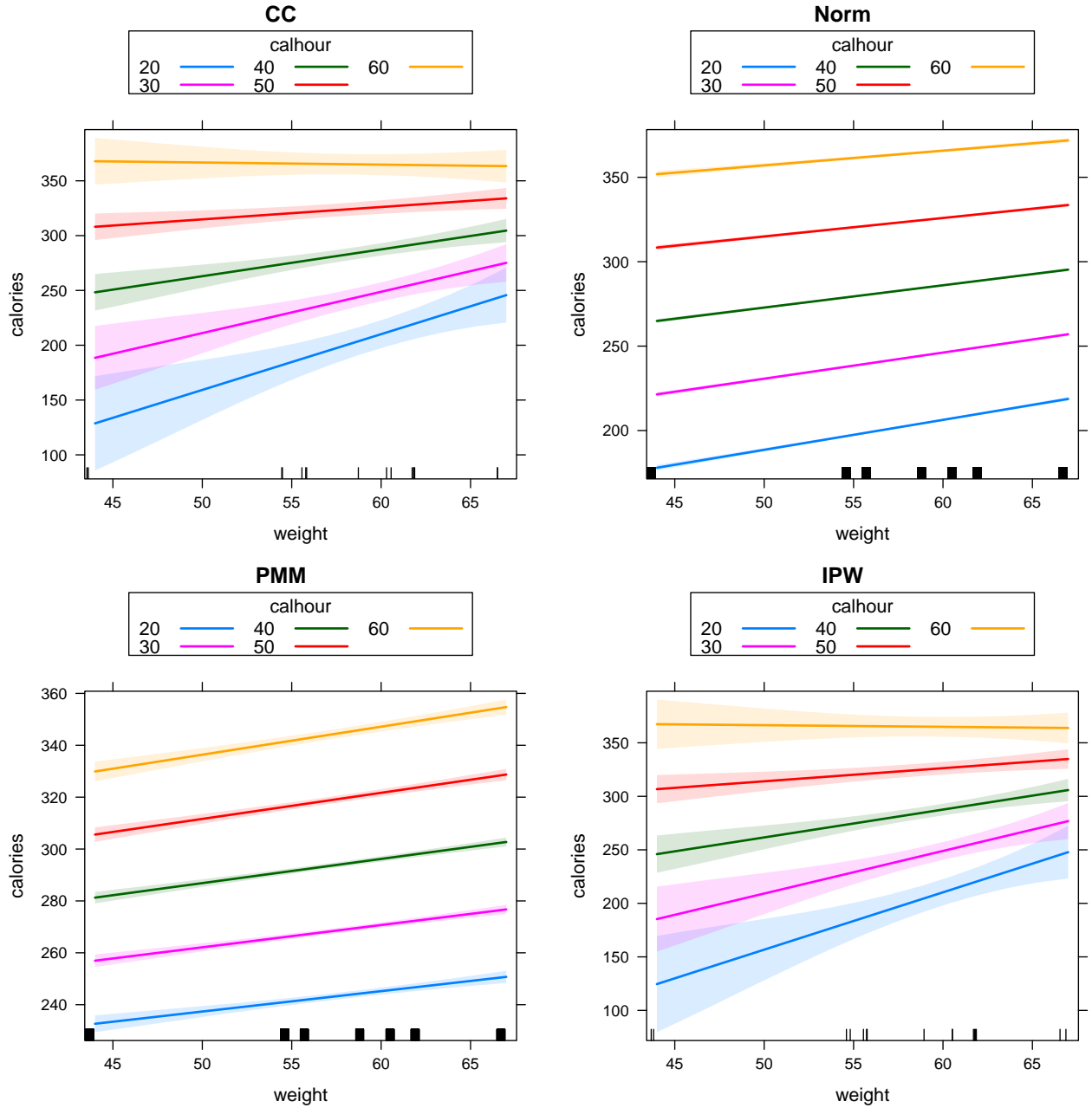


Figure 11: Interaction scatterplots for the normal NA-excluded dataset, values fitted using NORM and values fitted using PMM.



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.