Statistics Project

Project: calorie burning

Erim Erdal (r0772570) - Jaldert Francois (r0653709) - Pinar Onat (r0772819) - Alicia Hernandez Gimenez (r0734014)

This study explores the relationship between heat production, body mass and workout intensity. Measurements of heat production (variable calories), at different body Masses (variable weight, measured in kgs) and Work levels(variable calhour, measured in calories/hour) on a stationary bike were taken from a group of 24 volunteers. The aim of this study is to compare the use of complete Case analysis, Multiple imputation analysis and Inverse probability weighting in building a model to predict heat production given body mass and workout intensity.

Exploring the data and descriptive statistics

Descriptive statistics and graphical summaries were carried out on this dataset before multiple imputation or ipw. As such only complete cases (16) were taken into account. This can induce bias and wrong conclusions, therefore interpreting these graphs and numbers can only give some rough perception of the data. A sample of the dataset is shown at Table 1.

	weight	calhour	calories
2	43.7	19	NA
3	43.7	43	279
4	43.7	56	346
5	54.6	13	NA
6	54.6	19	NA
7	54.6	43	280

Table 1. Sample of the dataset.

The dataset contains measurements of calhour and weight which can be split up into groups (measurements at 43.7kg, 54.6kg,...). All variables were analysed as continuous variables since weight, calhour or calories can take on any (positive) value. The response variable (calories) has 8 missing values.

Some descriptive statistics are shown in the following table. It can be seen that the mean and median of each variable are close to each other, indicating a symmetric distribution of the values.

	weight	calhour	calories
median	58.8000000	38.7500000	302.0000000
mean	57.5416667	34.0416667	297.1250000
SE.mean	1.3452670	3.3396114	11.4669362
CI.mean.0.95	2.7828969	6.9085125	24.4411959
var	43.4338406	267.6721014	2103.8500000
std.dev	6.5904355	16.3606877	45.8677447
coef.var	0.1145333	0.4806077	0.1543719

Table 2. Descriptive statistics per variable.

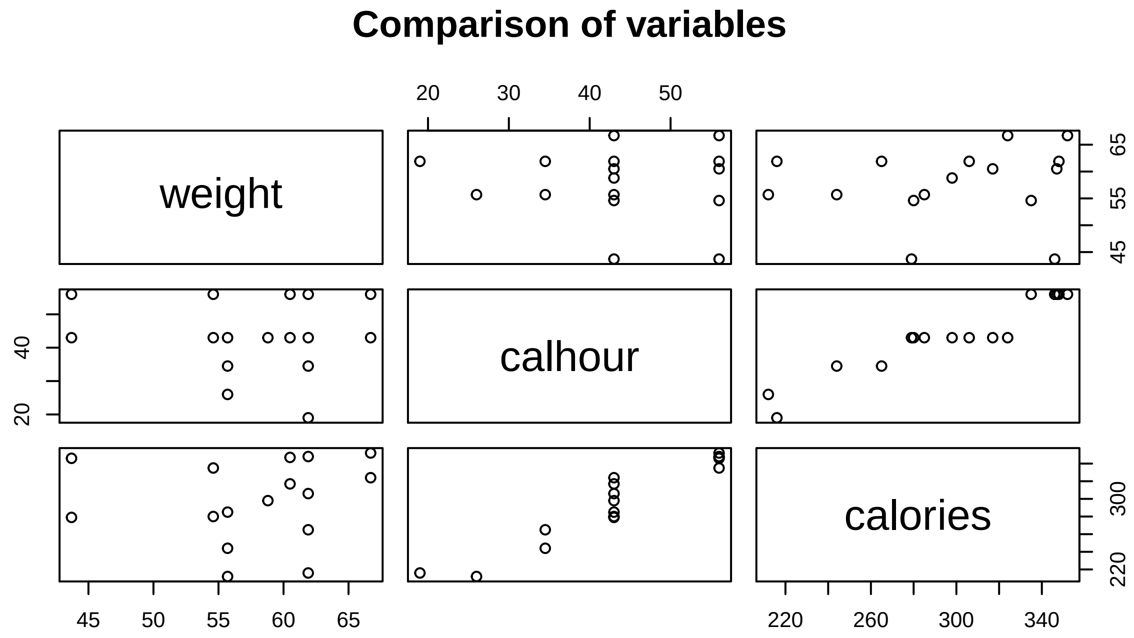


Figure 1. Comparison of variables.

The plot at figure 1 shows that there might be some positive correlation between calhour and calories. Increasing calhour seems to increase the resulting calories measured. The variable weight does not show a clear pattern.

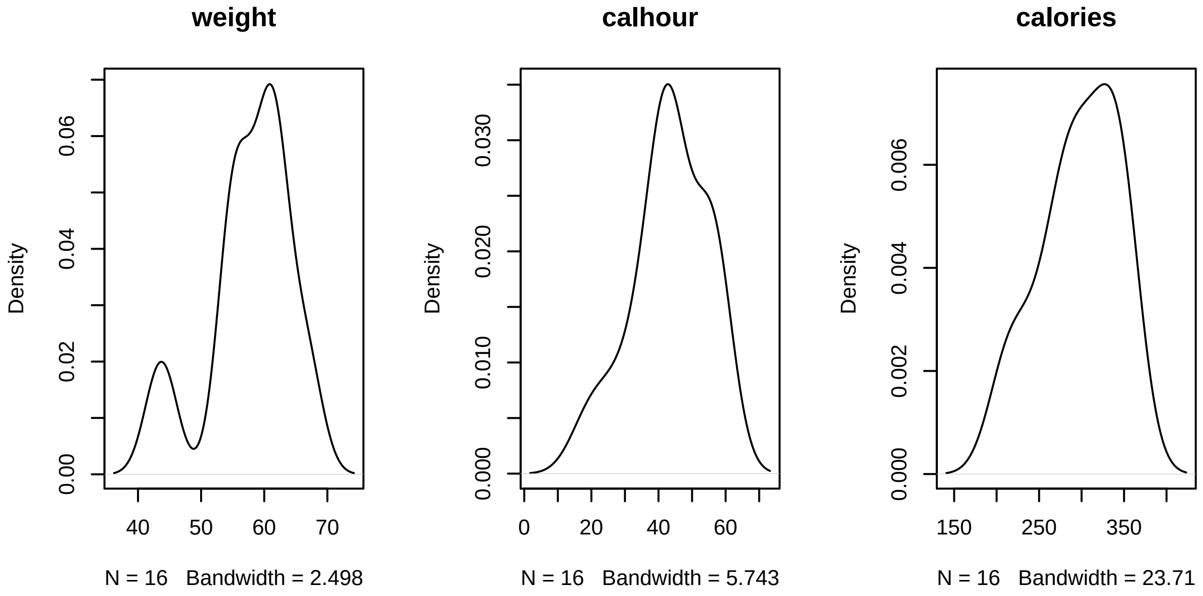


Figure 2. Density plots per variable.

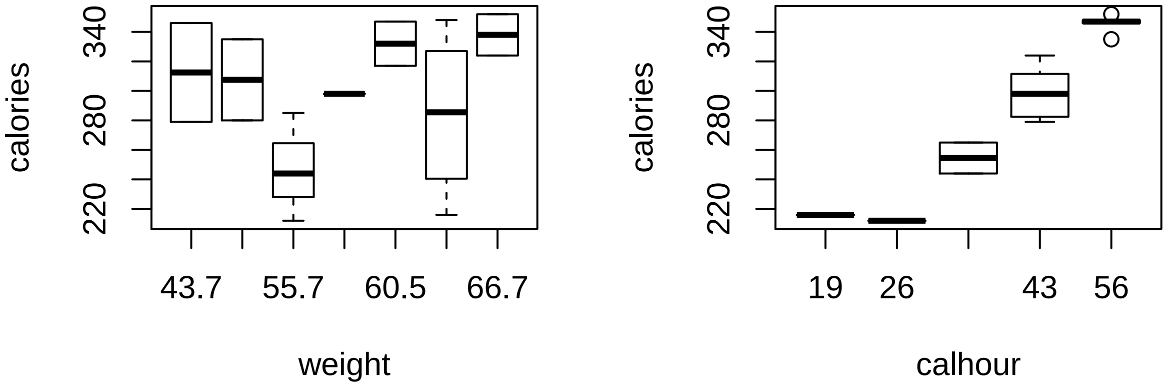
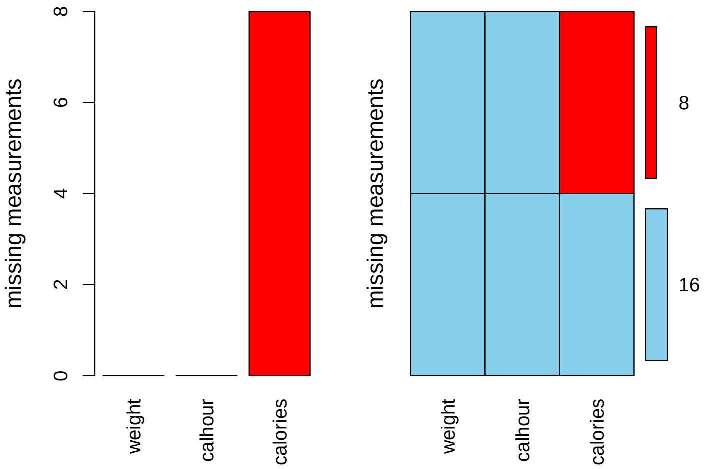


Figure 3. Boxplot distribution per variable weight (left) and calhour (right).

The boxplot (Figure 3) shows that there is no clear correlation/increase in calories burned with an increasing weights size. Moreover, it shows also a great variability within each weight group for calories values, which is as expected since the calhour (workout intesnity) differs within each weight group. In the calhour boxplot, there seems to be an increasing trend of calories used when increasing the workout intensity. Variability within each calhour group (variability caused by weight) is lower.

Graphical summary of missing data

The amount of missing data corresponds to the 33% of the total data, which belongs to the measurements for the variable calories in 8 subjects (Figure 4).



```
##
## Missings in variables:
## Variable Count
## calories     8
```

Figure 4. Missing measurements by group

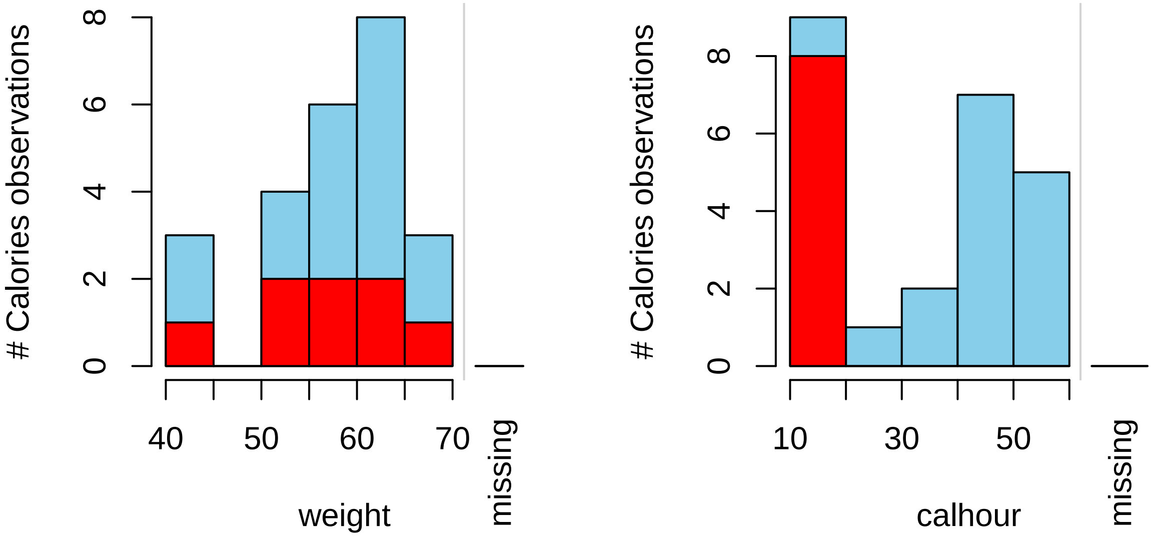


Figure 5. Missing values distribution within groups

It is observable at figure 5 that the missigness seems to follow a random distribution between all weight groups. Whereas in the calhour variable, most measurements for calories were missing in the lower (13, 19) calhour groups (low workout intensity).

Complete Case Analysis

The data has been analyzed by complete case analysis. Thus, 33% of the data, which corresponds to the missing values, have been dropped.

	Estimate	CI (lower)	CI (upper)	Std. Error	t value	Pr(>|t|)	
(Intercept)	-330.8839554	-602.5261356	-59.2417751	124.6743991	-2.653985	0.021	*
weight	7.7275346	3.1390714	12.3159978	2.1059465	3.669388	0.003	**
calhour	11.7874671	6.2359004	17.3390339	2.5479778	4.626205	0.001	***
I(weight * calhour)	-0.1320162	-0.2258023	-0.0382301	0.0430446	-3.066963	0.01	**

Table 3. Summary of the CC final model.

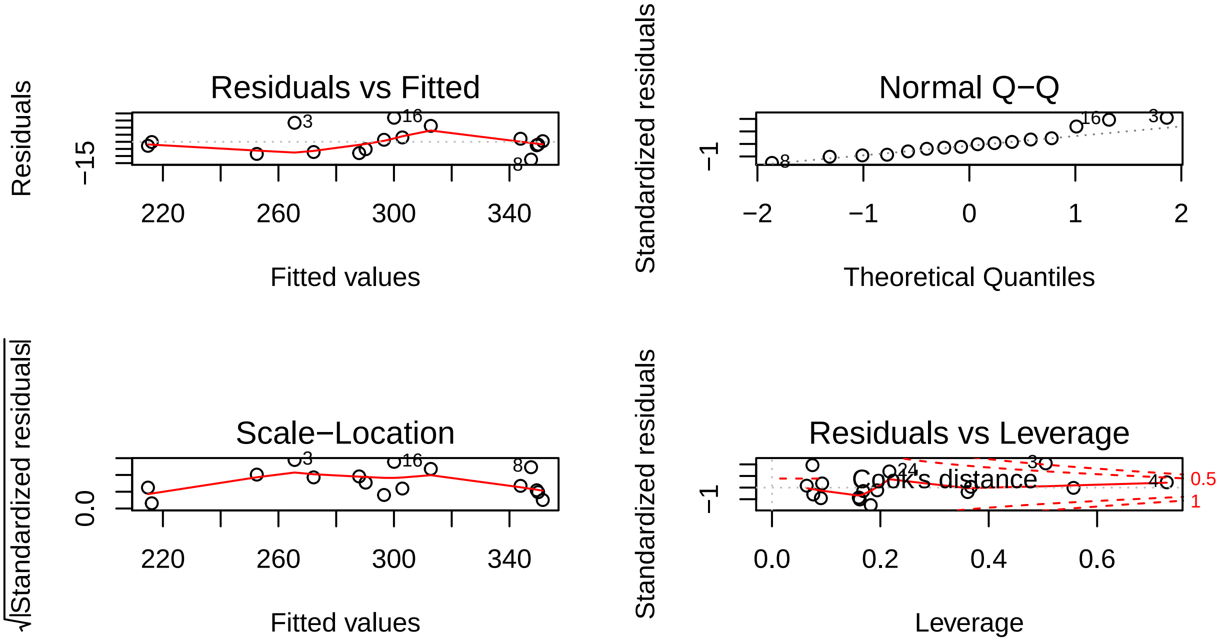
Variables calhour, weight and the interaction term are significant in the complete case analysis (with a 95% confidence interval). There are no terms that can be dropped. The wald test is used to look if a model without interaction is significantly worse than the model with interaction.

```
## Wald test
##
## Model 1: calories ~ weight + calhour + I(weight * calhour)
## Model 2: calories ~ weight + calhour
##   Res.Df Df      F   Pr(>F)
## 1     12
## 2     13 -1 9.4063 0.009772 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

		Res.Df	Df	F	Pr(>F)	
2	2	13	-1	9.406263	0.01	**

The wald test confirms that the interaction term is significant in our model.

The assumptions of this regression model (linearity, normality of residuals and constant variance) were checked (figure 6).



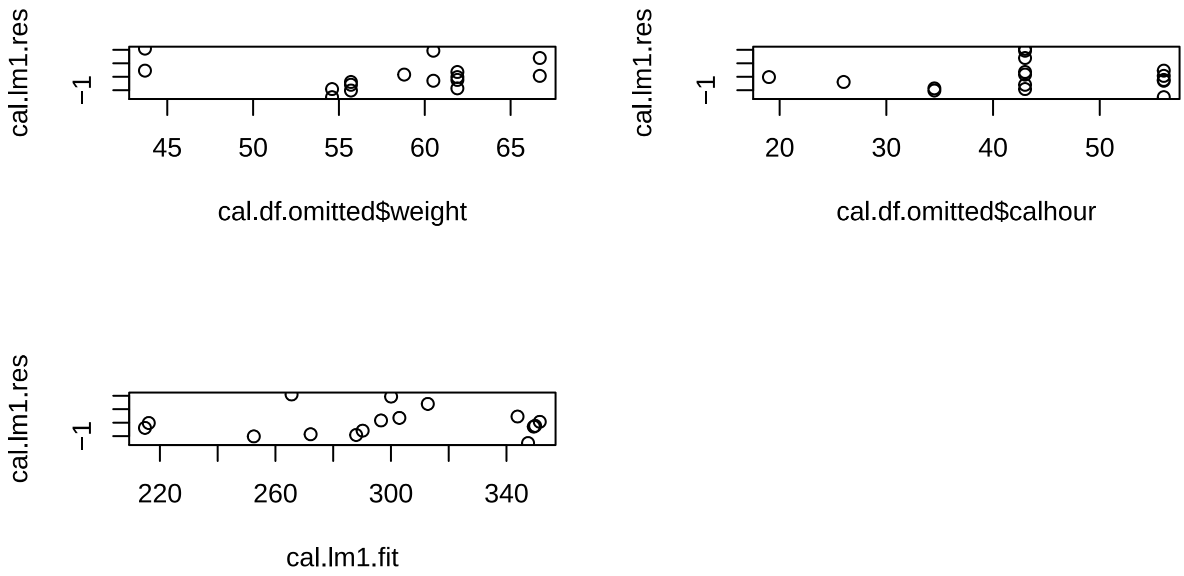


Figure 6. Plots related to the CC regression model in order to check its assumptions.

The residuals are more or less distributed at random, there is no other apparent trend in the data.

```
##
##  Shapiro-Wilk normality test
##
## data:  cal.lm1.res
## W = 0.92924, p-value = 0.2371
```

Normality of residuals holds. Using a linear regression model is possible. Our final model with complete case analysis:

calories = -330.88 + 7.73 * weight + 11.79 * calhour - 0.13 I(weight*calhour)

Notice that in this model calhour is slightly more significant and has a higher influence on the response variable.

Multiple Imputation

Since our data has a MAR missing mechanism, the previous approach (Complete Case analysis) can introduce extreme bias. Completing the data with mean values would also introduce severe errors in the final analysis. Here multiple imputation approach was used to create 300 datasets with imputed data.

We opted for a bayesian multiple imputation approach. Here the regression line estimate is taken into account, however drawing a value from the regression estimate would introduce errors since even the best values may differ from the actual unkown values. Therefore an uncertainty of the predicted values has to be taken into account. Bayesian mutliple imputation adds noise to the predicted values and, in contrast to stochastic regression imputation, also adds a parameter uncertainty. This parameter uncertainty is important since our predictions requires intercept, slope and standard deviation of the residuals to be known for imputing the missing data, while these values are unkown and estimated from the sample. Including parameter uncertainty takes into account that these values might differ when a new sample would have been drawn from the population. Moreover, bayesian multiple imputation was used to predict missing values, taking noise and parameter uncertainty into account.

The data was further explored to make sure no impossible values were imputated (negative calories). Furthermore, graphs were made to assess whether the imputated data are plausible and close to the real dataset (figure 7).

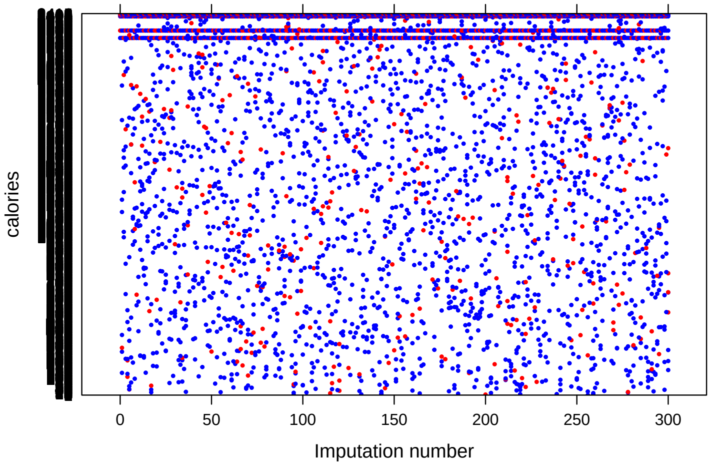


Figure 7. Imputations made by 'norm' method.

Neither the data shows impossible values nor the graphs shows any skewed pattern. The imputations were succesfull and are usefull to further build a model. After that, a wald test was made in order to compare different models taking into account the imputed data. We compared the different models using the D1 multivariate wald test to assess which variables are insignificant and can be dropped.

```
## [1] "Model with interaction - model without interaction"
```

```
##    test statistic df1      df2 df.com   p.value       riv
##  1 ~~ 2  0.568348   1 14.86615     20 0.4626907 0.8158122
```

```
## [1] "model without interaction - model without calhour"
```

```
##    test statistic df1      df2 df.com      p.value     riv
##  1 ~~ 2  207.7932   1 9.724311     21 6.965307e-08 2.33573
```

```
## [1] "model without interaction - model without weight"
```

```
##    test statistic df1      df2 df.com     p.value       riv
##  1 ~~ 2  8.931672   1 17.30548     21 0.008133107 0.4643361
```

```
## [1] "model with calhour - intercept model"
```

```
##    test statistic df1      df2 df.com      p.value     riv
##  1 ~~ 2  176.6561   1 14.08303     22 2.318595e-09 1.48252
```

There is a non-significant p.value when comparing a model with and without interaction term, the interaction term can be removed. The analysis shows that the variables calhour and weight are significant. Our final model thus has calhour and weight, which performs better than a model with only intercept or either of the variables alone (table 5).

For illustration we show the full model with interaction term (table 4).

	term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
1	(Intercept)	-8.3224791	83.4513377	-0.0997285	7.853579	0.9230585	-201.3879090	184.7429509
2	calhour	5.3746767	1.8279392	2.9402929	9.665233	0.0153006	1.2825712	9.4667821
3	weight	2.2542825	1.4047469	1.6047606	8.173092	0.1464146	-0.9731670	5.4817320
4	calhour:weight	-0.0232679	0.0308638	-0.7538886	9.988769	0.4683149	-0.0920473	0.0455115

```
##            est     lo 95     hi 95 fmi
## R^2 0.9743215 0.9300775 0.9907069 NaN
```

Table 4. Complete model with interaction, which shows that there are insignificant variables.

	term	estimate	std.error	statistic	df	p.value	2.5 %	97.5 %
1	(Intercept)	42.712052	30.4078046	1.404641	10.683427	0.1885293	-24.457834	109.881938
2	calhour	4.027498	0.2793957	14.415033	5.716753	0.0000103	3.335546	4.719450
3	weight	1.373404	0.4595492	2.988590	13.088030	0.0103996	0.381287	2.365522

```
##            est     lo 95     hi 95 fmi
## R^2 0.9719451 0.9247012 0.9897088 NaN
```

Table 5. Final model with calhour and weight variables.

All regressors are significant.

Final model:

calories = 43.69 + 4 * calhour + 1.37 * weight

Inverse Probability Weighting

With inverse probability weighting we are not predicting/estimating missing values, but assign weights to variables in certain groups. Measurements of data in a certain group receive a weight depending on the amount of missing data belonging to that group. IPW eliminates bias that might occur when data for certain groups is much more incomplete.

First a logistic regression was made to a variable R, which has a value 0 if the measurement was missing or 1 if there was a measurement for that group. The inverse of this probability (of missingness) was used to create a new variable which holds the "variables weight". The most complex model (calhour+weight+interaction term) was used to assign weights to the groups.

	Estimate	Odds Ratio	CI (lower)	CI (upper)	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1790.2057145	0.000000e+00	0	Inf	647070.7502	-0.0027666	0.998	
calhour	6.1921964	4.889188e+02	NA	Inf	13650.8897	0.0004536	1	
weight	20.9330165	1.233370e+09	0	Inf	8832.0226	0.0023701	0.998	
calhour:weight	0.3366737	1.400282e+00	0	NA	266.5876	0.0012629	0.999	

Note that the p.values for this logit regression model are extremely high (insignificant). The problem here is that the dataset contains the group calhour (13) which has no measured values and a group calhour (19) which only has 1 measurement that we can assign a weight to.

Fitting the complete model (calories~weight+calhour+I(calhour*weight)) taking the weights into account:

	Estimate	CI (lower)	CI (upper)	Std. Error	t value	Pr(>|t|)	
(Intercept)	-330.8839554	-575.2412873	-86.5266234	124.6743990	-2.653985	0.021	*
weight	7.7275346	3.5999553	11.8551138	2.1059465	3.669388	0.003	**
calhour	11.7874671	6.7935224	16.7814119	2.5479778	4.626205	0.001	***
I(calhour * weight)	-0.1320162	-0.2163821	-0.0476503	0.0430446	-3.066963	0.01	**

```
## Wald test
##
## Model 1: calories ~ weight + calhour + I(calhour * weight)
## Model 2: calories ~ weight + calhour
##   Res.Df Df      F   Pr(>F)
## 1     12
## 2     13 -1 9.4063 0.009772 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the wald test shows that removing the interaction yields a significant worse model. We keep the interaction term. The final model has weight, calhour and interaction term as predictor variables.

The big difference between the null deviance and residual deviance shows that the final model performs a lot better than the model with only the intercept. Furthermore, we observe a high standard error of our intercept estimate.

Final model:

calories = -330.88 + 7.72 * weight + 11.79 * calhour - 0.13 * (calhour*weight)

Note that this final model is the same final model we obtained from complete case analysis. This is not a suprise since all measurements for the group calhour 13 are missing and thus not included in this model (no real weight could be assigned).

IPW, Multiple Imputation and Complete Case comparison

Both the complete case and IPW methods gave a model with all explanatory variables and the interaction term. Both models were almost identical with equal slope parameters. Only the confidence interval was smaller for our analysis with IPW. It is not a suprise that both models are almost identical: since all data was missing from the lowest calhour group (calhour=13), IPW fails to assign "weights" and takes these missing values into account. Although the advantage of IPW that no data is created (only seen data is used), it's clear disadvantage is that it seems to fail if complete groups are missing. Complete case analysis is a valid approach under MCAR, but the problem here is that data misses following a MAR principle: data is absent based upon the calhour value.

Multiple Imputation on the other hand gives a final model with calhour and weight as independent variables. This method has less trouble with all data missing in the lowest calhour group. The missing values were estimated by Bayesian Multiple Imputation taking the regression line, an extra noise term and a parameter uncertainty into account. The advantage of Multiple imputation in this case is that completely missing groups are taken into account by means of data generation where IPW failed to do so. Although the drawback is that we create data based upon the observed dataset, in this case it is better than the other approaches but it remains an estimate and can have considerable bias. Using different methods for imputation can have different outcomes and in the end give different models.

Conclusion

Final model after all the analysis: calories = 43.69 + 4 * calhour + 1.37 * weight

In relation to the obtained model, we can see that with every 1 unit of increased weight in kg's when calhour is kept stable, calories burned will increase 1.37 calories. Also when weight is kept stable, every 1 unit of increase in calhour variable will result in an increase of 4 calories burned. Thus, both of these independent variables have positive correlation with the response variable, calories burned. Finally, model argues that interaction term is insignificant, meaning two different individuals with different weights will burn the same amount of calories while conducting the exercise with the same workout intensity. Neverthless, the data collected from the experiment had too little observations, and 33% of that data were missing, which caused some faulty implementations of missing data analysis. More data would be necessary to make more plausible conclussions.

It is also questionable on how the Multiple Imputation model behaves when an entire group of data is missing, in our case happens to be 13.0. Multiple Imputation will try to generate a model in order to estimate these variables that are missing. However, in our case, the estimated variables are the bottom-most values of the dataset and no single observation from that group exist. This degrades the ability of the model generated by Multiple Imputation, creating bias. Confidence intervals of the final model offered by Multiple Imputation shows us these coefficents have high variability, while calhour being the one predictor with the smallest variability. This shows us once again that our dataset is too small to make accurate, stable predictions for the coefficients of the final model.

Our final model argues that the more you weight, the more calories you will burn when you do the same exercise with your peers in the study. Model also argues that the higher calhour workout that you conduct, you will be burning more calories than your peers at the same level of weight. It is also argued that workout intensity (calhour) is a more determining factor for the calories burned than weight of the individual (weight).

References

Greenwood, M. (1918). "On the efficiency of muscular work." Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character, 90(627), 199-214.

Stef Van Buuren, (2018). "Flexible Imputation of Missing Data" Second Edition. Taylor & Francis Group, p. 415