Factors Influencing Body Weight: Modelling and Analysis

1. Introduction

This report presents an analysis of variables affecting body weight. The dataset used is from 2111 individuals aged 14-61 from Mexico, Peru and Colombia and considers variables of a biological nature (height, age, gender) as well as lifestyle choices such as eating habits and preferred methods of transportation. The aim of this report is to identify key factors influencing body weight, and to understand and attempt to model their interactions.

2. Distribution of Weight Data: Patterns and Trends

2.1. Weight, Family History and Number of Meals Eaten

Weight data was initially plotted by examining a family history of being overweight ('family') and number of main meals eaten per day (NCP)

Figure 1

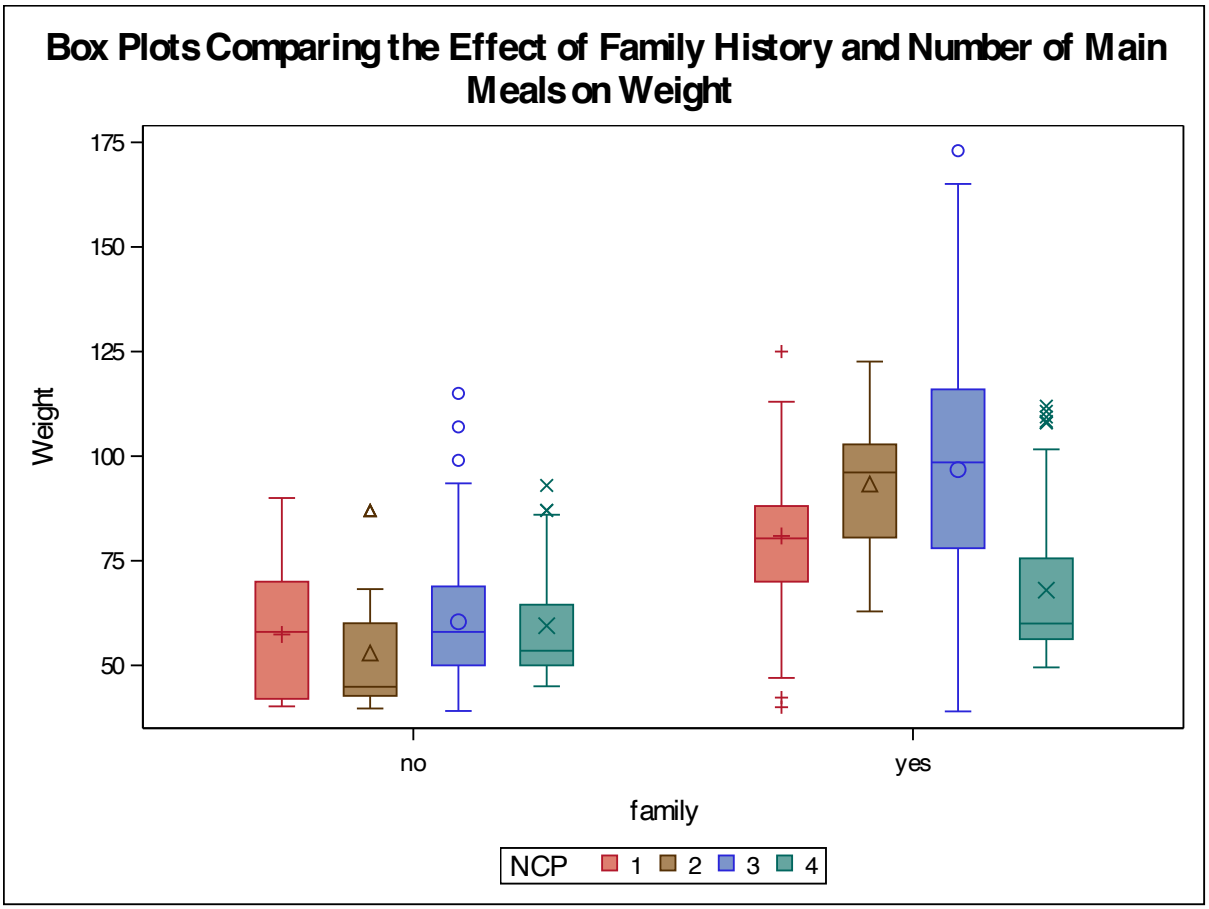


Figure 1 suggests that a family history of being overweight is likely a significant factor in determining an individual's weight: the medians of the 'no' group are all lower than those in the 'yes' group. For

all but one of the NCP groups (4 meals per day being the outlier) the difference is likely to be significant as the medians in the 'no' group are all lower than the first quartiles of the 'yes' group.

With respect to the number of meals eaten (NCP), this appears to have a negligible effect on weight. The differences between NCP with no family history of being overweight do not appear to be statistically significant with similar means and medians and no notable differences in distribution observed. Where a family history exists ('yes'), some potentially significant differences may exist where those eating 4 meals may have a statistically significant difference in weight to those eating 1, 2 or 3 meals per day.

Not all data appeared symmetrical. Positive skew was observed in 'NCP = 4' data, with longer tails on the right side and a mean greater than the median regardless of family history. 'NCP = 2' also had a positive skew for 'family = no' but had a negative skew in individuals that did have a family history of being overweight, suggesting differing weight distributions based on family history.

Outliers were observed in the 'NCP = 3' and 'NCP = 4' population, regardless of family history. Small numbers of outliers were also observed in 'NCP =1' and 'NCP =2' groups, all of which suggests variations in weight distribution.

2.2. Weight, Family History and Number of Consumption of Calorific Food

Weight data was examined by plotting a family history of being overweight ('family') and whether high calorie food was frequently consumed (FAVC).

Figure 2

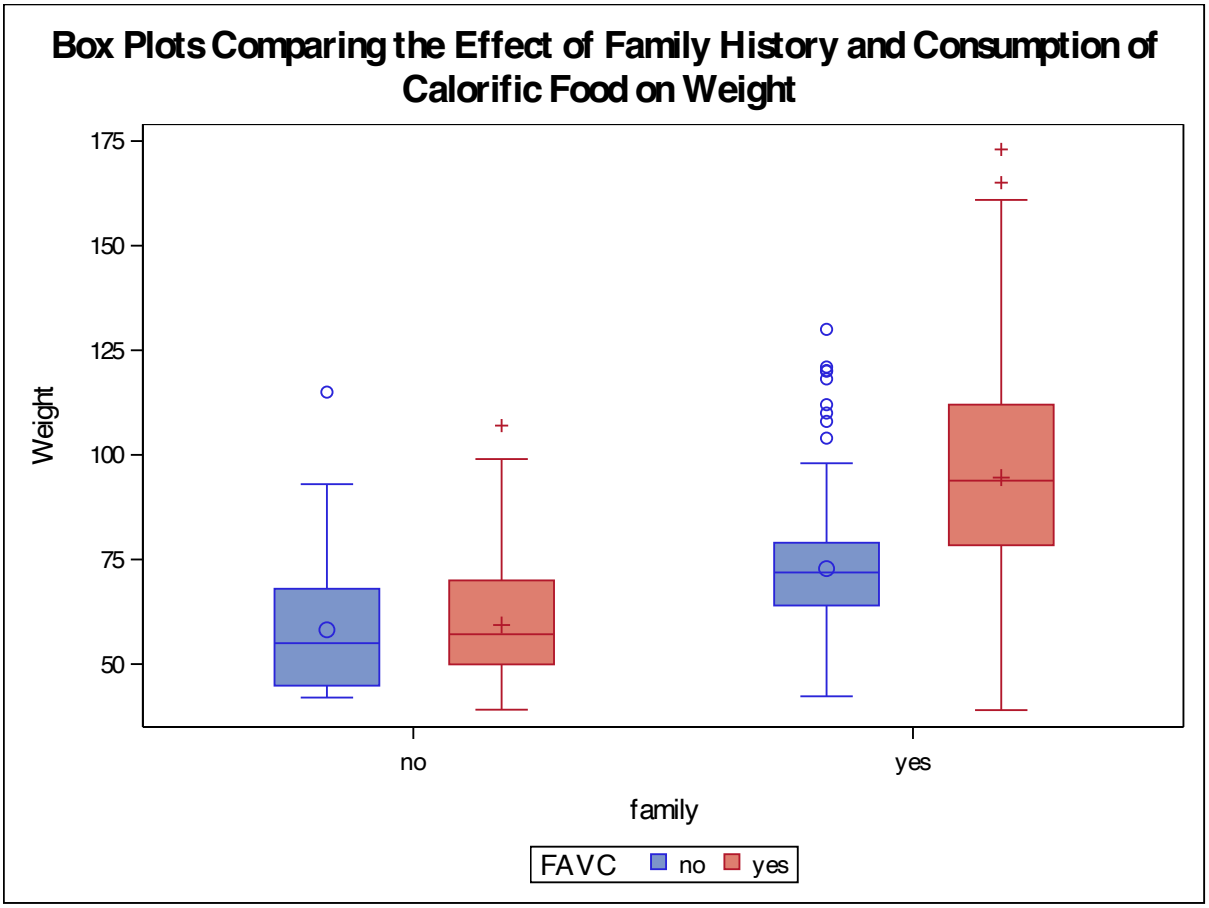


Figure 2 supports the observation in the previous analysis that family history affects an individual's weight as the groups with a family history of being overweight appear likely to have a statistically significant difference to those without.

FAVC appears to have no notable impact on weight of individuals who have no family history of being overweight ('family = no'), with similar means, medians and distributions (both have a slight negative skew). It appears to have minimal impact on variability as the tail lengths and number of outliers are similar, regardless of FAVC group.

For individuals who have a family history of being overweight, FAVC status does appear likely to have a statistically significant difference as the median of the 'FAVC = no' group is lower than Q1 of the 'FAVC = yes' group. Distributions of these groups appear relatively symmetrical in shape and possess similar means and medians. Multiple outliers are observed, suggesting variability within the populations, particularly in the 'FAVC = no' group.

2.3. Summary

The two analyses above suggest that family history likely plays a statistically significant role in determining a person's weight.

Regarding the number of meals consumed, NCP, the results were mixed. NCP appeared to have no notable impact on weight in individuals with no family history of being overweight. This appears to be the same for individuals with a family history of being overweight, except for individuals who consume 4 meals a day; these individuals appear likely to weigh less than those who consume fewer meals; however it should be noted that a number of outliers were noted in this group suggesting there may be other relevant factors influencing weight not captured here.

Consumption of calorific food (FAVC) appeared to play no notable role in determining weight in individuals with no family history of being overweight, with the converse being true in individuals who did. This suggests that calorific food consumption plays a role in weight but only in the presence of other relevant factors (such as family history).

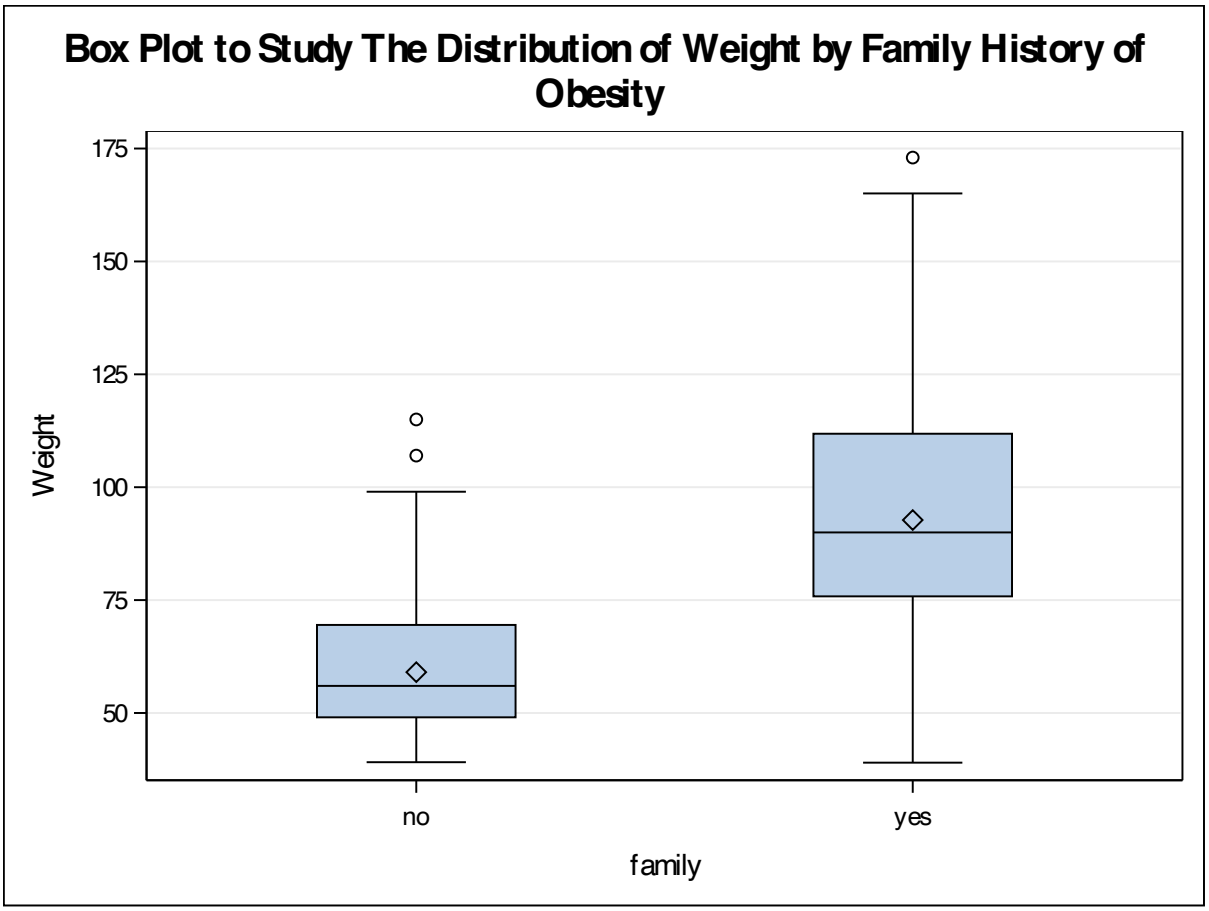
3. Distribution of Weight by Family History

The previous analysis suggests that family history likely plays a significant role in determining an individual's weight and is investigated further below.

Table 1 – Dispersion, Skewness and Kurtosis of Weight by Family

Analysis Variable: Weight							
family	N Obs	N	Mean	Std Dev	Skewness	Kurtosis	%CV
no	385	385	59.0411410	14.1815452	0.7613633	0.1148801	24.1
yes	1726	1726	92.7302024	24.2321899	0.1713414	-0.6152659	26.1

Figure 3



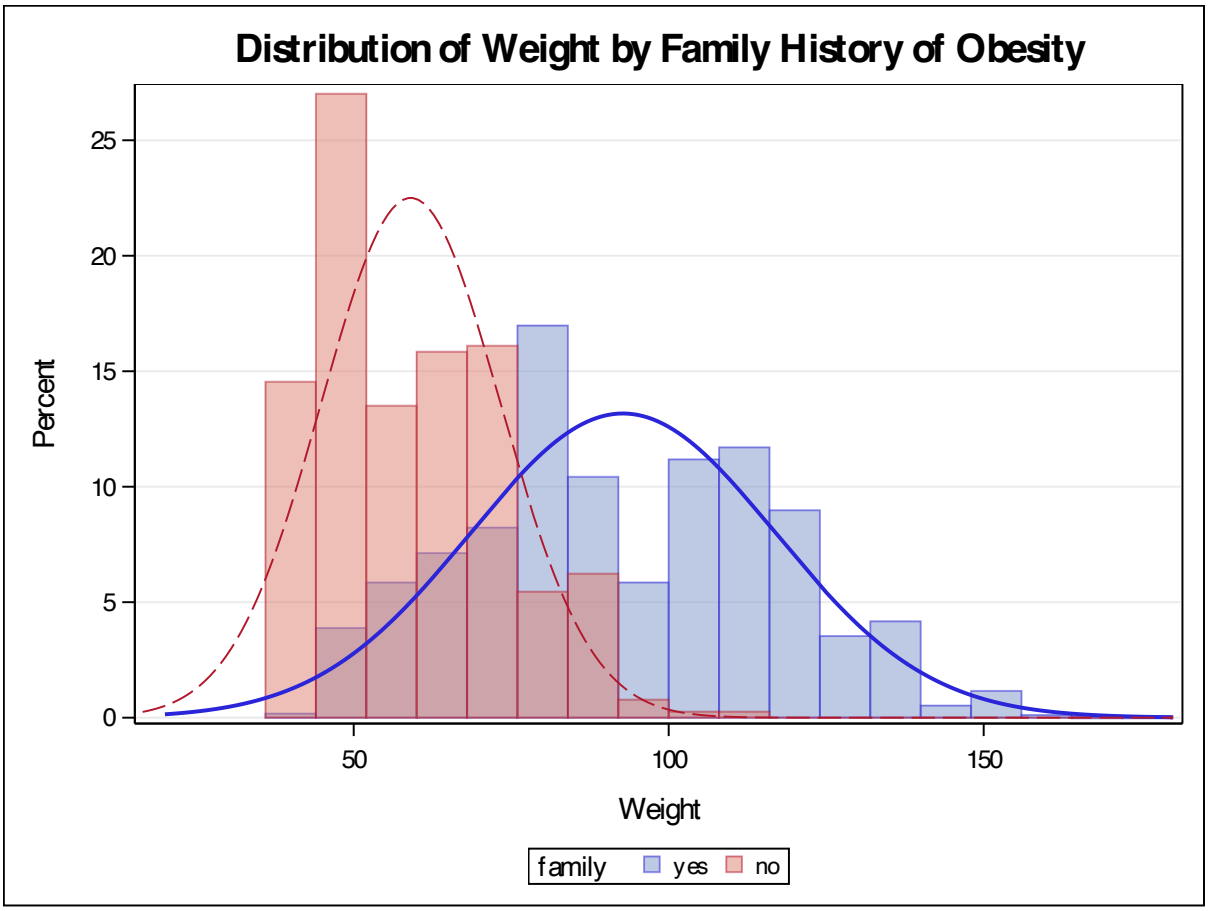
Individuals with no family history of being overweight displayed lower mean and median averages than those that do. The differences appear likely to be statistically significant as they are lower than the 'yes' groups' Q1.

Variability, expressed as Std Dev is higher in the yes group than the no group, however when this is normalised by the mean the relative variability, expressed as %CV, is similar (24.1% and 26.1%).

The distributions appear close to normal. Both groups have a positive but exhibit low skewness (<±1), suggesting no significant departure from normality. A positive skew suggests there are individuals with higher-than-average weights in each group, which can be seen in the slightly longer upper tails and outliers.

Kurtosis values are low suggesting that the distribution has tails similar to a normal distribution.

Figure 4



The positive skew is evident in the histogram (Figure 4) for both groups. For the no group, this appears to be driven by a slightly longer tail, whereas for the yes group this appears to be driven by a potential second population as it exhibits a bimodal distribution.

The bimodal distribution for 'no' is not evident in the density plot however which its exhibits slight platykurtic (kurtosis <0) distribution.

Figure 5

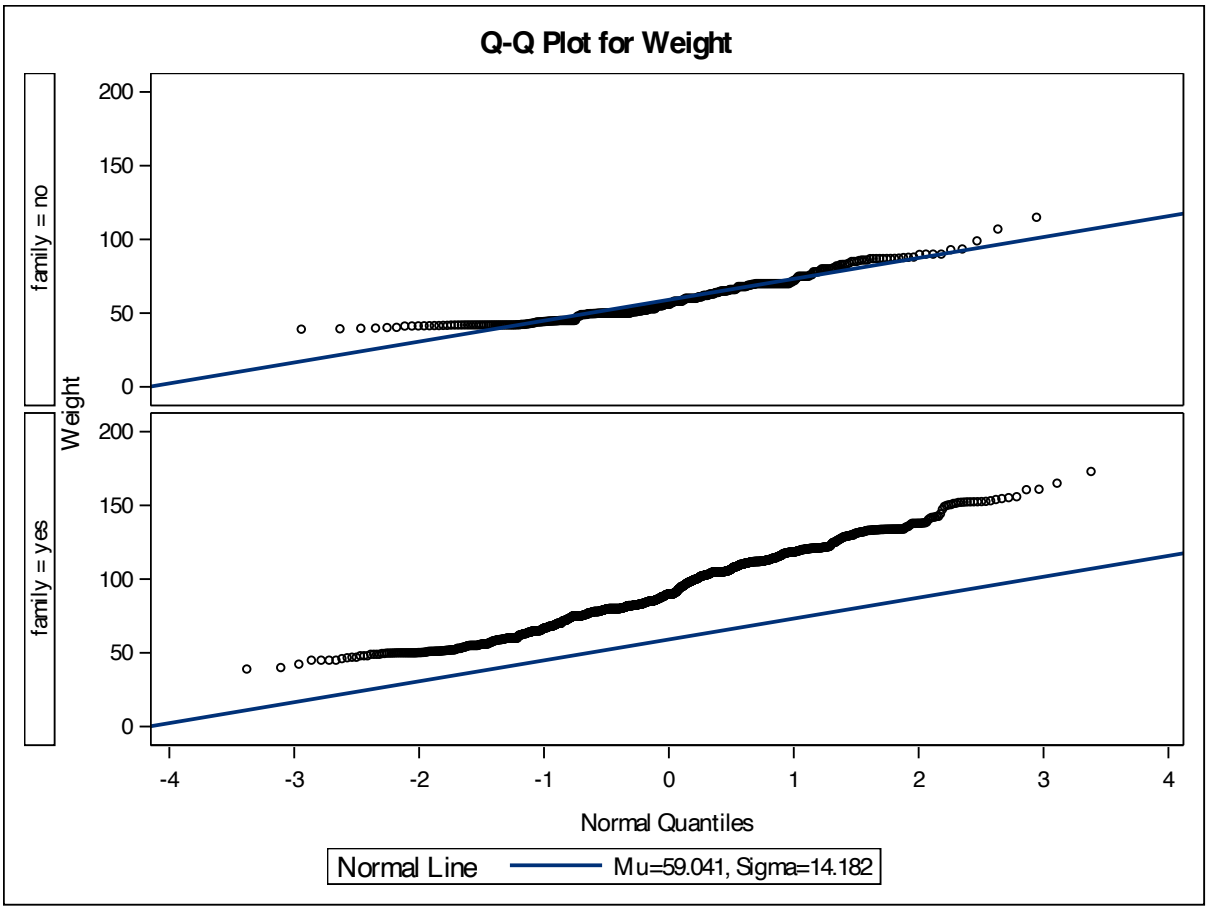


Figure 5 shows no significant departure from normality for either group as no curved pattern is observed.

Deviations from normality have been observed in each group however none appear significant enough to conclude a non-normal distribution without performing statistical test.

Table 2

Tests for Normality (family = no)				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.932971	Pr < W	<0.0001

Table 3

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.983246	Pr < W	<0.0001

The null hypothesis for the Shapiro-Wilk test is that the data follows a normal distribution. Testing both populations returned a p-values less than 0.0001 meaning that the null hypothesis is rejected, and the data therefore significantly deviates from a normal distribution.

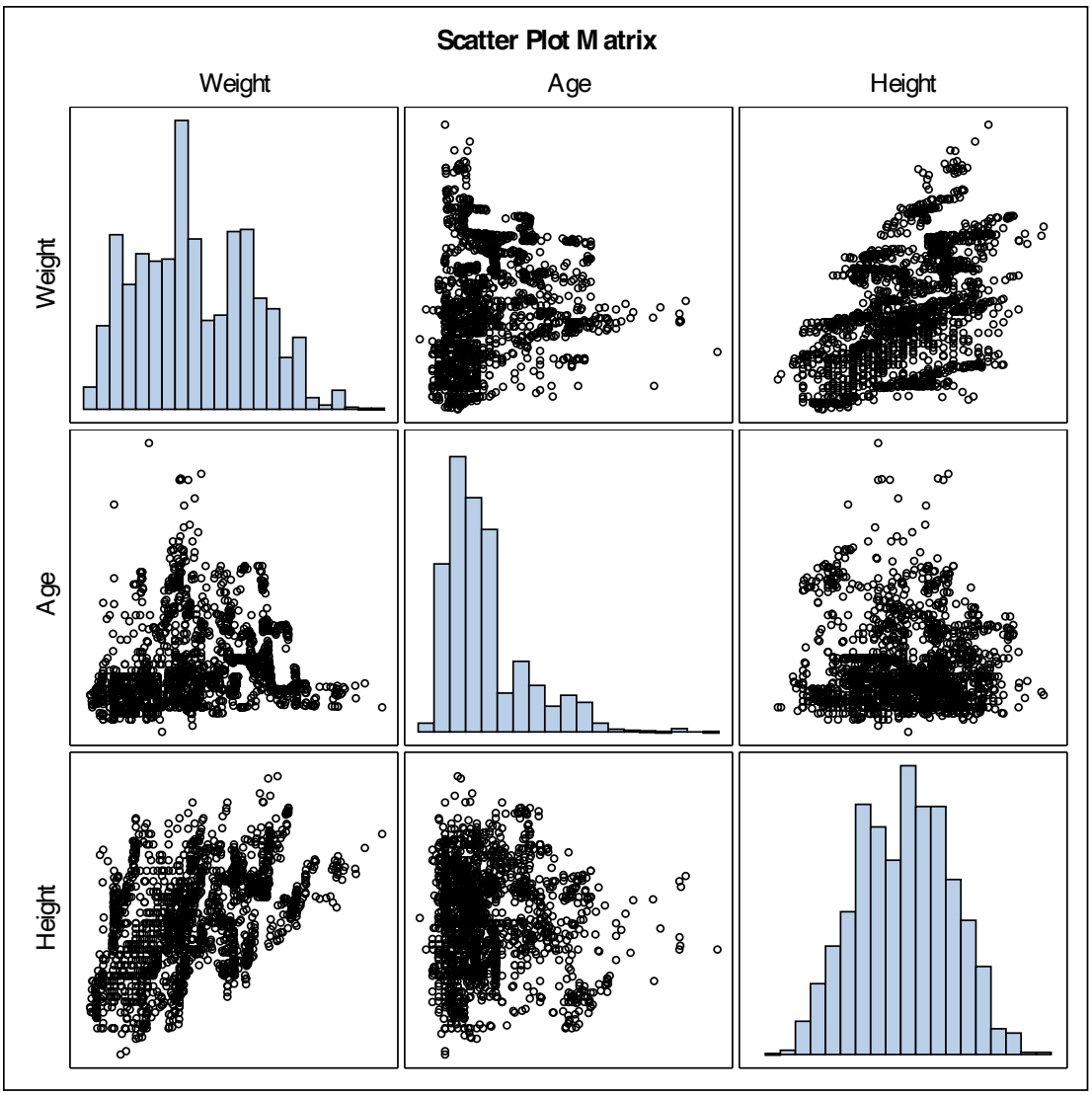
4. Correlation Between Weight, Age and Height

Age and height are biological variables that could reasonably be believed to contribute to weight, which are investigated using a correlation analysis.

Table 4

Pearson Correlation Coefficients, N = 2111 Prob > |r| under H0: Rho=0			
	Weight	Age	Height
Weight	1.00000	0.20256 <.0001	0.46314 <.0001
Age	0.20256 <.0001	1.00000	-0.02596 0.2332
Height	0.46314 <.0001	-0.02596 0.2332	1.00000

Figure 6



The weight and age data both show departures from normality: weight exhibits a bimodal distribution, while age has a notable positive skew. Height data shows no significant departures from normality.

Positive, linear correlations can be observed between weight and height and weight and age, with correlation coefficients 0.20 (small effect) and 0.46 (medium effect), respectively. These correlations are statistically significant with p-values <0.0001. These correlations suggest that taller people and older people weigh more.

There appears to be no correlation between age and height. Scatter plots do not appear random suggesting a non-linear relationship, while the p-value for the correlation is 0.2 which is greater than the 5% significance level and therefore not significant.

Given the lack of normality observed in the weight and age variables, data should be normally distributed for a Pearson's correlation analysis. A Fisher's transformation was obtained, and the data reassessed to confirm the results.

Table 5

Pearson Correlation Statistics (Fishers z Transformation)								
Variable	With Variable	N	Sample Correlation	Fisher s z	Bias Adjustment	Correlation Estimate	95% Confidence Limits		p Value for H0:Rho=0
Weight	Age	2111	0.20256	0.20540	0.0000480	0.20251	0.161244	0.243077	<.0001
Weight	Height	2111	0.46314	0.50130	0.0001097	0.46305	0.428859	0.495916	<.0001
Age	Height	2111	-0.02596	-0.02596	-6.1512E-6	-0.02595	-0.068539	0.016729	0.2332

After Fishers Z transformation the conclusions regarding the correlation analysis reaffirm that weight, age and height are correlated.

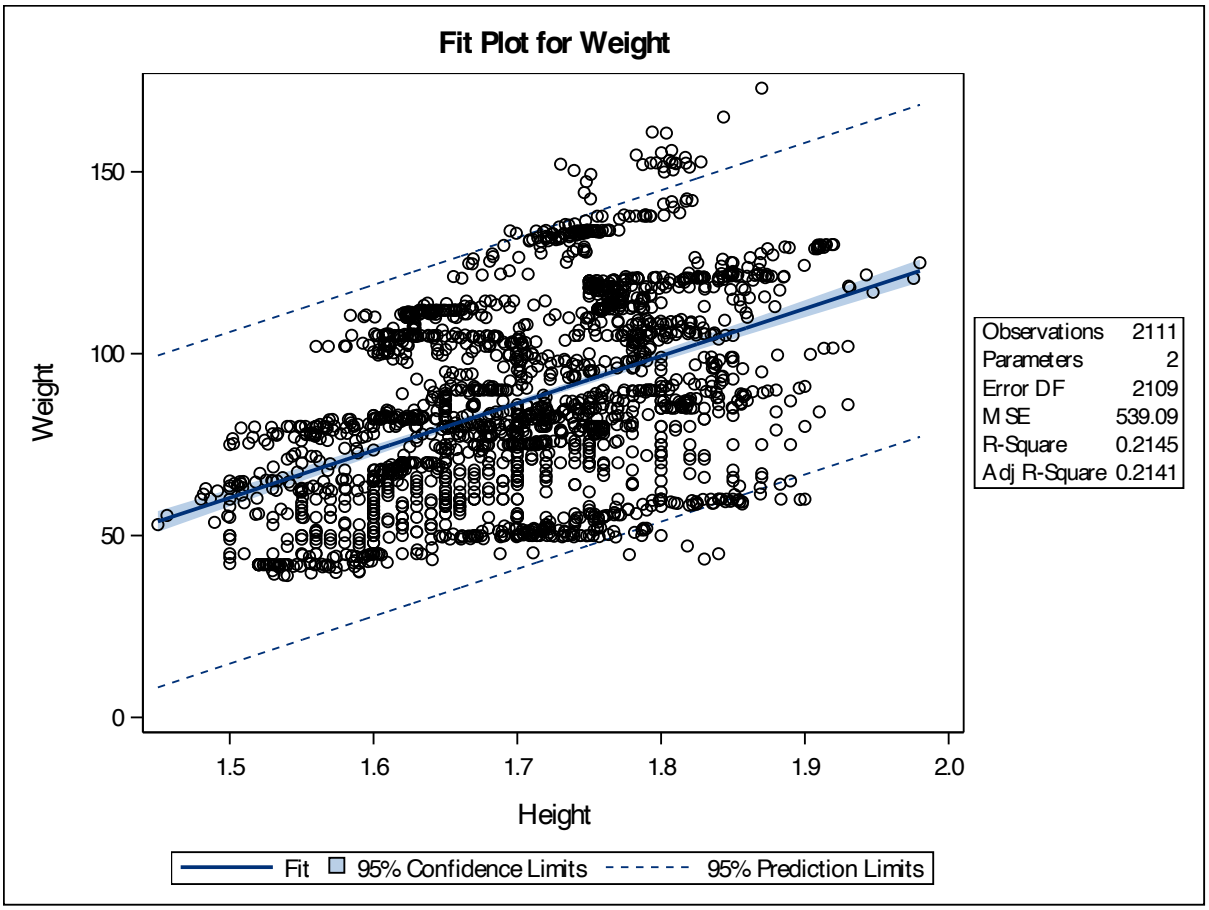
For weight vs age and weight vs height, correlation estimates are only marginally different, and the p-value is highly significant suggesting a significant positive correlation.

For age vs height, the correlation estimate also remains largely unchanged and the p-value is not significant, suggesting there is no correlation.

5. Simple Linear Regression of Weight and Height

As height showed a medium effect correlation with weight, it was investigated further using a simple linear regression.

Figure 7



From the plot (Figure 7) the positive relationship is apparent. The 95% confidence limits are narrow and evenly distributed suggesting that this is a reasonable estimate of this relationship in the entire population.

The R^2 is 0.2, suggesting the quality of the regression model is weak as height only explains approximately 20% of the variability in an individual's weight.

Table 6

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-134.64022	9.23243	-14.58	<.0001	-152.74583	-116.53461
Height	1	130.00483	5.41735	24.00	<.0001	119.38092	140.62874

The intercept parameter (B0) estimate is statistically significant ($p < 0.0001$) but has little real-world value as an individual's height cannot be zero.

The slope (Height/B1) is significantly different to zero ($p < 0.0001$), meaning the probability of observing a t-value as extreme as the one calculated, assuming the null hypothesis is true, is practically zero. As such, the null hypothesis that there is no linear relationship between these variables is rejected.

The slope suggests that a 1-meter increase in height would result in a change in weight of approximately 130 kg and it can be said with 95% confidence that the population value would be between approximately 119 and 141 kg.

Table 7

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	310463	310463	575.90	<.0001
Error	2109	1136950	539.09419		
Corrected Total	2110	1447412			

The analysis of variance (Table 7) quantifies what variability can be explained by the dependent variable ('Model' i.e. height) and what error is unexplained ('Error'). The F-statistic, the ratio of explained to unexplained variance, is high. The probability of observing an F-statistic as extreme if the null hypothesis was true, is practically zero (<0.0001), suggesting that using height as a predictor of weight is statistically significant.

Analysis of the fit diagnostics will assess the suitability of using a linear regression model on this data.

Figure 8

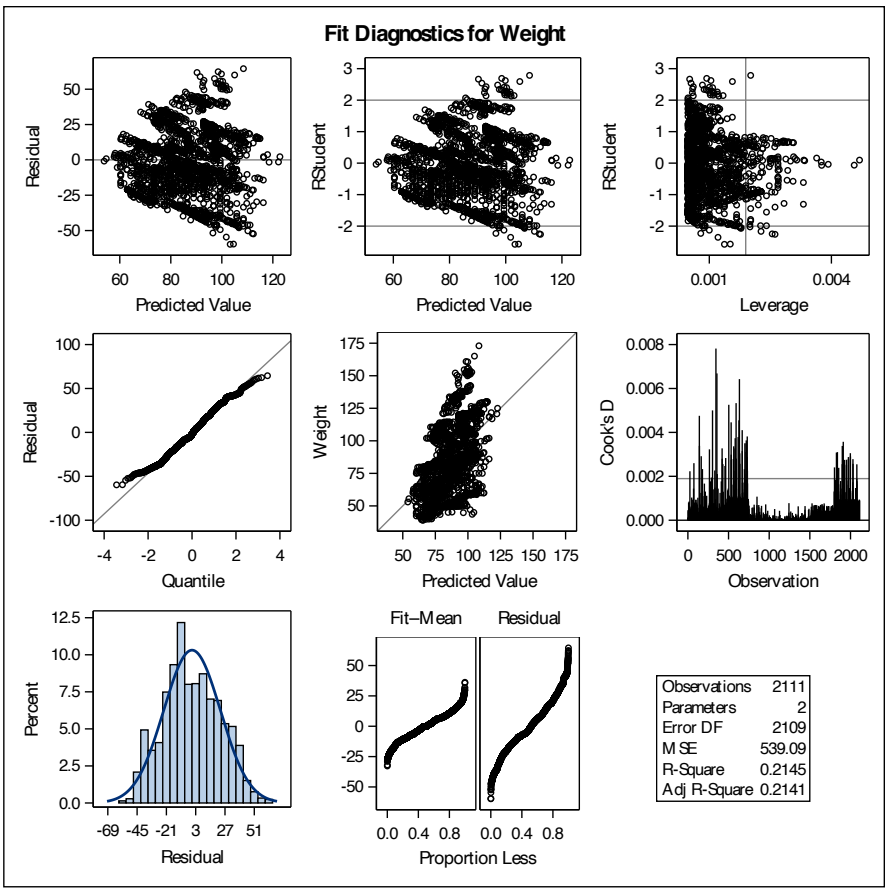


Figure 9

Analysis of the residuals reveals several important findings:

- Heteroscedasticity: A fan-shaped pattern in the residuals versus predicted values plot indicates non-constant variance. This is further evident in the studentised residuals, where numerous observations fall outside the ±2 range, indicating the presence of both positive and negative outliers.
- Influential Observations: The plot of leverage against R-student highlights a notable number of influential observations. These are data points whose removal can significantly alter the parameter estimates. The Cook's D plot also suggests the presence of several high-leverage observations.
- Normality: The histogram and QQ plot do not show any significant deviations from normality. The histogram appears normally distributed with short tails on either side, which can be observed in the QQ plot of the residuals which is largely straight with the tails deviating slightly at the end.

This model does not meet several key assumptions required for linear regression. While the residuals appear to exhibit linearity and independence, they do not possess equal variance. The model may potentially be improved using transformations of the data (e.g., log), a weighted least squares regression to account for varying levels of variance or using robust standard errors. The presence of both outliers and influential observations may be distorting the results of the regression. The impact of outliers may be reduced using a transformation.

This model is suitable for drawing conclusions about the sample but cannot be generalised beyond the sample to draw conclusions about the population.

6. Multiple Regression Model for Weight in Individuals with a Family History

Additional variables (including height) were assessed in a multiple regression model in individuals who have a family history of being overweight using weight as the response variable. The aim was to find the simplest model that could be used to predict a person's weight.

The model was obtained using a stepwise regression model to find the simplest model using Mallow's Cp (Cp ≤ p+1) and highest R^2.

All variables available were assessed except NObeyesdad. This variable was excluded because it is derived from an individual's weight and would therefore provide no new or additional information to the model. Dummy variables were created from categorical variables to include them in the regression.

The use of a stepwise allowed the inclusion of all variables in the model without the risk of confounding its predictive power. The stepwise method iteratively adds new variables to the model and the model is assessed at each new iteration to see if redundant variables can be removed.

Table 8

Number in Model	C(p)	R-Square	Variables in Model
9	8.4389	0.4278	Height Age FAVC_dummy FCVC CAEC_dummy SCC_dummy FAF CALC_dummy MTRANS_dummy

The model was selected as it was the model with the lowest number of parameters to meet Mallow's Cp criteria and included the variables listed in Table 8. Selecting the most parsimonious model is desirable as increasing model complexity (i.e., increasing the number of variables) increases the risk of overfitting. This model represents the best balance between simplicity and predictive power.

The current model produced an R^2 of 0.43 suggesting the above combination of variables can account for approximately 43% of the variability in weight. This is an improvement from the previous model (simple linear regression using only height) which could account for approximately 21% of the variance ($R^2 = 0.21$).

Table 9

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	-151.37292	10.36317	-14.61	<.0001	0	-171.69870	-131.04714
Height	1	110.63623	5.59623	19.77	<.0001	1.27899	99.66007	121.61240
Age	1	0.93655	0.09228	10.15	<.0001	1.74936	0.75555	1.11755
FAVC_dummy	1	10.20500	1.66441	6.13	<.0001	1.09637	6.94051	13.46948
FCVC	1	9.70522	0.78056	12.43	<.0001	1.02634	8.17427	11.23617
CAEC_dummy	1	-14.04585	1.14774	-12.24	<.0001	1.02534	-16.29696	-11.79474
SCC_dummy	1	-14.10467	2.79196	-5.05	<.0001	1.05539	-19.58067	-8.62867
FAF	1	-2.86786	0.54723	-5.24	<.0001	1.18884	-3.94117	-1.79455
CALC_dummy	1	7.64020	0.90268	8.46	<.0001	1.09764	5.86972	9.41068
MTRANS_dummy	1	5.01596	0.44297	11.32	<.0001	1.67956	4.14715	5.88477

All variables have a p value <0.0001 suggesting that they are all significant for this model with respect to predicting weight.

The parameter estimates indicate the positive or negative contribution the variable has on weight, all other variables being equal.

The FAVC dummy variable suggests that consumption of calorific food ('yes') leads to an increase of approximately 10.2 kg in weight.

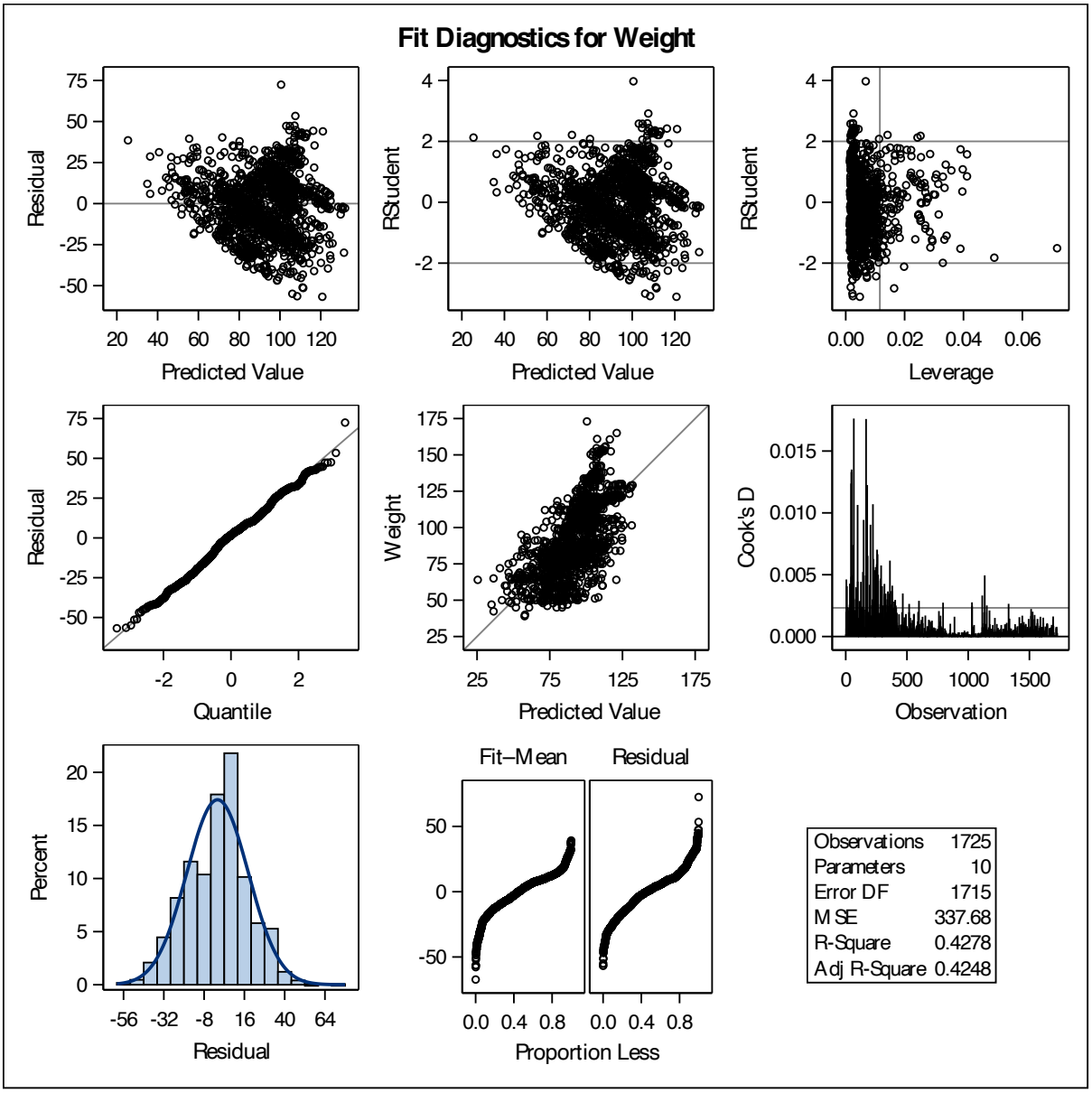
The CAEC and SCC dummy variable suggests that eating less food between meals and calorie consumption monitoring can lead to lower weight estimates.

The model was assessed for multicollinearity between variables using variance inflation and no causes for concern (>10) were noted.

Table 10

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	433025	48114	142.48	<.0001
Error	1715	579124	337.68		
Corrected Total	1724	1012149			

Figure 10

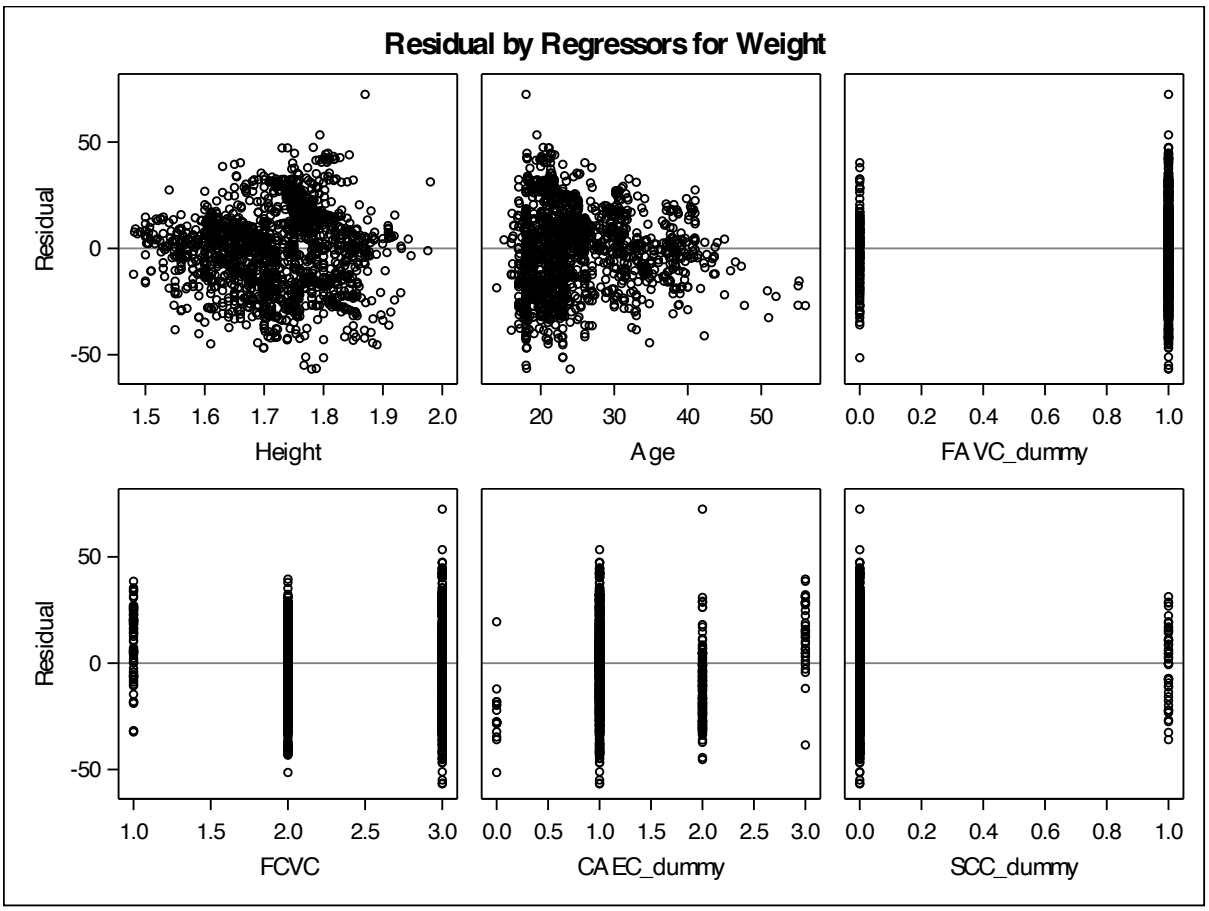


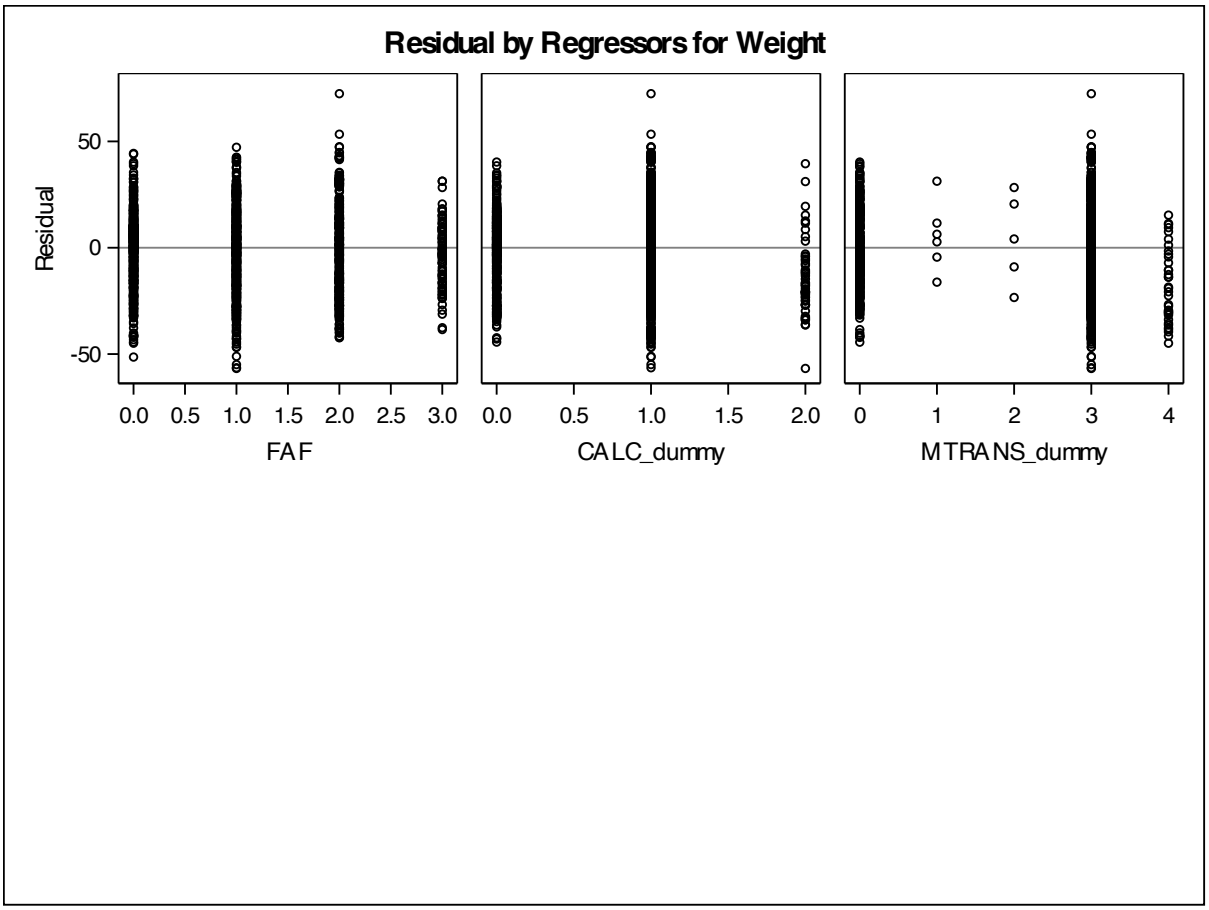
Analysis of the residuals reveals several important findings:

- Heteroscedasticity: As with the previous model, a fan-shaped pattern in the residuals versus predicted values plot indicates non-constant variance. The studentised residuals still suggest a number of outliers although the proportion appears reduced compared to the simple linear model used in Section 5 of this report.

- Influential Observations: The leverage and Cook's D plot indicate the presence of a notable number of influential data points that require further inspection due to their influence on the model.
- Normality: The histogram and QQ plot do not show any significant deviations from normality.

Figure 11





The residual plots for the individual regressors give some insight into weight residuals:

- Age has distinct shape suggesting a degree of non-constant error, but the inverse shape of the weight residuals.
- For CAEC (food between meals), the residuals do not appear to be independent and trend upwards. CAEC was coded with a dummy variable on scale from 0 ('no') to 3 ('always') and the upward trend suggests the model is missing an additional variable(s) to explain the relationship between CAEC and weight, or that the relationship is possibly not linear.
- FCVC and SCC appear to demonstrate heteroscedasticity.

In summary, the multiple linear regression model included the following variables: Height, Age, FAVC, FCVC, CAEC, SCC, FAF, CALC, and MTRANS, and performed well, accounting for approximately 43% of the variability in weight. This is an improvement from a simpler linear regression model (Section 5) using only height. All predictor variables in the model were highly significant ($p < 0.0001$), with no evidence of multicollinearity observed, and ANOVA confirming the model's overall significance in explaining the variation in the dependent variable. Assumptions of normality and independence were met; however, issues were noted around heteroscedasticity of the residuals and the presence of outlier and influential observations. The model is therefore a reliable estimator for weight in the sample, but it is not suitable for generalising the population in its current iteration.

7. Multiple Regression Model for Weight in Individuals Without a Family History

Additional variables (including height) were assessed in a multiple regression model for individuals who do not have a family history of being overweight with the same aims, methodology and selection criteria as the previous model (Section 6).

Table 11

Number in Model	C(p)	R-Square	Variables in Model
7	7.8998	0.5260	Height Age FCVC CAEC_dummy SCC_dummy TUE MTRANS_dummy

Compared to the previous model, this model appears more parsimonious (7 variables as opposed to 9) and a higher R^2 (0.53 vs 0.43) suggests it explains a greater proportion of the variance.

In the new model, FAVC, FAF and CALC have been lost while TUE has now been included.

Table 12

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	40619	5802.73179	59.76	<.0001
Error	377	36610	97.10745		
Corrected Total	384	77229			

The second model is based on a smaller data set (1726 vs 377), which may be a challenge for such a complex model and more data may be required to improve this model.

The model for individuals without a family history of being overweight has a lower sum of squares than the model for people who do, 40619 compared to 433025, which indicates that the first model explains more variance in the response variable. However, despite a lower sum of squares, the second model explains a greater proportion of the total variance (higher R^2) suggesting that the more complex model is capturing additional variability (i.e., noise).

8. Summary

A simple linear regression model found a positive relationship between height and weight. While this makes logical sense, height only explained approximately 20% of variation in weight, suggesting it is not alone is not a strong predictor of weight and that other variables play a role in determining weight.

Having determined the variables relating to weight were likely complex, multiple regression models in individuals who do and do not have a family history of being overweight were created. Models were selected to strike a balance between simplicity and predictive power and the resulting models included various biological factors (e.g., age, gender, height) dietary factors (e.g., consumption of high calorie food, consumption of vegetables, number of meals) and lifestyle factors (e.g., smoking, alcohol consumption).

The resulting model for individuals with a family history of being overweight was an improvement. This is because predictability over the simple model looking at height improved, explaining approximately 43% of variability in weight. The model for individuals without the same family history accounted for more variability, 53%.

The model for individuals with a family history of being overweight was more complex and contained more variables than those for individuals without the same family history yet resulted in reduced predictive power. This suggests that the determinants of an individual's weight when there is a family history of being overweight is more complex and likely includes stronger influence from additional factors that were not included in this study such as biological and genetic factors or factors relating to socioeconomic status. More data investigating these variables may provide an opportunity to improve the models.

All the models created in the analysis were significant and were shown to work well with the sample but had limitations that mean the models are unlikely to be suitable for generalising beyond the sample data to the population.

This report has demonstrated that multiple regression models have potential to be used as predictors for weight. However, to reliably predict weight in the population, refinement of the models is required, and more data is needed, particularly when individuals have a family history of being overweight.