Exploring predictors of impaired fasting glucose during adolescence - a longitudinal study

by

Nathan Wayne Eastwood

School of Mathematics and Statistics
University of Sheffield

Thesis submitted as part of the requirements for the award of the MSc in Statistics, University of Sheffield, 2015–2016

Abstract

Abstract of: Exploring predictors of impaired fasting glucose during adolescence - a longitudinal study

Author: Nathan Wayne Eastwood

Date: August 2017

Background

The prevalence of childhood overweight and obesity are rising and have multiple causes. Consequences for health include the development of diabetes and it is therefore important to identify individuals at risk. Simple measures of adiposity such as BMI are limited due to their inability to distinguish fat from lean mass whereas dual-energy X-ray absorptiometry (DEXA) is considered a gold standard technique for body composition assessment. A novel measure of DEXA-derived body shape (trunk to leg volume) been proposed in adults [1] but this has not been assessed in children.

Objective

The aim of this project was to investigate how total and regional body composition changes during adolescence, particularly in relation to gender and pubertal timing, can be used to predict the onset of impaired fasting glucose.

Methods

347 normal healthy children (173 male; 174 female) from the EarlyBird prospective study were considered for this study to model their prevalence of impaired fasting glucose (5.6mmol/L $\geq$ glucose $<$ 7mmol/L) between ages 9 and 16. Missing data were substantial and were imputed where necessary to facilitate reasonable models using last observation carried forward and mean imputation methods. Separate models for total body measures, regional body measures and trunk to leg volume ratio were considered. Measures included both lean and fat indices (or percentage for the regional models), physical activity, age, gender and socio-economic status. Generalised linear mixed effect models were used assuming a random effect for each child to adjust for individual trend effects and a Lasso approach was used to identify important independent variables.

Results

Only 198 children (95 male; 103 female) were considered suitable for the study due to data completeness after imputation, of which 36 (23 male; 13 female) had detectable impaired fasting glucose. For the models considering only total body lean and fat indicies, the Lasso algorithm found total body lean mass index to be highly significant ($p < 0.001$) and positive whereas total body fat mass index was removed (the parameter was shrunk to zero). The algorithm also found age to be positive and highly significant ($p < 0.001$) and the interaction between age and lean mass index to be highly significant ($p < 0.001$) and negative; and all other parameters were removed from the model. The interaction effect between age and lean mass index is complex; children with lower lean mass indices at a younger age have lower odds of developing impaired fasting glucose but have a higher risk closer to adulthood than those children with higher lean mass index. The interaction parameter was much smaller than age and lean mass index.

Regional models focussed on lean or fat mass percentages of compartmental areas of the body:

trunk, arms and legs. The lean mass percentage model found only the trunk lean mass percentage to be significant ($p < 0.05$). The interaction between the lean masses and age were important, though non-significant ($p > 0.05$), in the model. The regional fat percentage model showed that trunk fat percentage ($p < 0.05$) and its interaction with gender ($p < 0.05$) are indicative of impaired fasting glucose though many other non-significant ($p > 0.05$) variables were retained in the model. Converse to the total body and regional lean percentage models, gender and age remain in the model though are non-significant ($p > 0.05$).

No evidence was found to suggest that the novel trunk to leg volume ratio is indicative of impaired fasting glucose in children as this variable was removed from all models by the Lasso.

Conclusion

Lean mass index and age are indicative in the prevalence of impaired fasting glucose in children along with their interaction. This is in contrast with other research which typically suggests [2] that it is fat levels which are indicative and not lean levels, though this is usually measured in adults. Given the complexity of the interaction, it may well be that lean mass index is important in the prevalence of pre-diabetes in children but it is fat mass index that is important for adults.

Declaration

This dissertation is a presentation of my original research work. Wherever contributions of others are involved, every effort is made to indicate this clearly, with due reference to the literature, and acknowledgement of collaborative research and discussions.

This work was completed under the guidance of Dr. Timothy Heaton of Sheffield University and Dr. Rana Moyeed and Dr. Joanne Hosking, both of Plymouth University.

Nathan Wayne Eastwood

Date: 08-Aug-2017

Acknowledgements

To my parents: because I owe it all to you! Many thanks for all your continued love and support through the years!

Particular thanks go to my supervisors Jo, Rana and Tim for your perseverance, patience and support, I thank you.

A very special gratitude goes out to all the children of the EarlyBird study, the study trustees, patrons and donors for opportunity to work with this data.

Last but by no means least, a very special mention is reserved for Susan Ball, without you I would have never made it through this degree. My eternal thanks go out to you for all your love, support and kindness.

Thank you everyone, for all your encouragement!

Contents

List of Figures

List of Tables

1 Study Objectives

1.1 Introduction

The aim of this project will be to investigate how total and regional body composition change during adolescence, particularly in relation to gender and pubertal timing, and whether these changes can be used to predict the onset of impaired fasting glucose. The association between changes in body composition and other factors in childhood (e.g. socio-economic status and total physical activity) will also be considered. Furthermore, a novel measure of dual-energy X-ray absorpitiometry (DEXA)-derived body shape, trunk to leg volume ratio, has been proposed in adults [1,4] as a way of identifying the risk of diabetes, however this has not been assessed in children. Using data collected from the National Health and Nutrition Examination Surveys (NHANES) in the United States [5], Joseph Wilson et. al [1] claim to show that the ratio of trunk to leg volume has a strong association to diabetes which is independent of total and regional fat distributions. This claim will be investigated within this report. None of the children in the EarlyBird study that this report links to, and whose data it uses, were diabetic though some children were defined as having impaired fasting glucose which will be used as a proxy for diabetes. These aims are structurally defined in the following hypotheses.

Hypothesis 1 *DEXA related weight variables are predictive of impaired fasting glucose.*

Hypothesis 2 *Trunk to leg volume ratio is independent of total and regional fat distributions and predictive of impaired fasting glucose.*

Hypothesis 3 *Growth and development of children as well as the outcome of impaired fasting glucose in adolescence are not mutually exclusive issues. Body composition changes over time and so too does the risk of impaired fasting glucose.*

The objectives will be investigated using appropriate statistical models, in particular generalized linear mixed models (GLMMs) will be fitted to the longitudinal EarlyBird data where the dependent variable will be whether or not the participant has impaired fasting glucose (otherwise known as pre-diabetes). However it is recognised that in the presence of many predictors, GLMMs yield unstable estimates [6] and so the use of a Lasso-type approach using an ℓ_1-penalized algorithm [7] to identify relevant predictors will be used.

1.2 An Overview of Diabetes

The prevalence of childhood overweight and obesity are rising and have multiple causes; particularly under-activity and over-nutrition are both believed to contribute [8]. Obesity is of concern since

it is thought to cause the insulin resistance that underlies diabetes and cardiovascular disease [2]. Consequences for health include the development of diabetes and it is therefore important to identify individuals at risk. Diabetes and its complications are fast becoming one of the UK's biggest health threats, outstripping smoking-related diseases, cancer and drugs. Type 2 or so-called 'adult' diabetes, is by far the commonest form of diabetes. It's hugely on the increase - teenagers and even some children are now getting it. According to Diabetes UK there are 4 million people living with diabetes in the UK [9], or more than one in 16 people (diagnosed or undiagnosed); of which it is estimated that 31500 are children and young people under the age of 19, though it is believed this may be an underestimate [10]. This figure has more than doubled since 1996, when there were 1.4 million people living with diabetes in the UK and by 2025, it is estimated that 5 million people will be affected. Around 700 people a day in the UK are being diagnosed with diabetes [9] which is equivalent to one person every two minutes.

1.3 An Overview of DEXA Data

Simple measures of adiposity such as BMI are limited due to their inability to distinguish fat from lean mass whereas dual energy X-ray absorptiometry (DEXA) is considered a gold standard technique for body composition assessment. When undergoing a DEXA scan, a radiation source is aimed at a radiation detector placed directly opposite the site to be measured. The patient is placed on a table in the path of the lose-dose radiation beam (or X-ray) and a large scanning arm then moves across the measurement region. Typically bone density varies in different parts of the skeleton and so different parts of the body are usually scanned [11]. Some of the X-rays that are passed through the body are absorbed by tissue, such as fat and bone. An X-ray detector inside the scanning arm will measure the amount of X-rays that have passed through your body. This information will be used to produce an image of the scanned area. More formally, the attenuation of the radiation beam is determined and is related to the bone mineral density (BMD) in the center of the skeleton [12].

Several different types of DEXA systems are available, but they all operate on similar principles [12]. Original DEXA scanners used what was known as pencil-beam technology but this has since been replaced with fan-beam technology and systems are still being upgraded [13]. During the upgrade process of DEXA scanners, The International Society for Clinical Densitometry (ISCD) recommend that cross-calibration should be performed in order to maintain continuity of measurement performance [14] since BMD changes slowly over time and even small differences can affect the reliability of results [15,16]. DEXA scanners have been shown that they can be highly concordant in single BMD measurements however they have also been shown to be discordant for long-term precision depending on the manufacturer [16].

1.4 The EarlyBird Study

EarlyBird was a unique 12-year non-intervention prospective cohort study observing the health and lifestyle of ~300 normal healthy children caught up in the obesity epidemic of the 21^{st} century. The children were recruited at age 5y and followed up annually until they were 16y. Data collected include demographics (several measures of socio-economic status), various anthropometric measurements (e.g. height, weight, BMI, skinfolds, circumferences) and whole body DEXA scans which include regional body composition and bone mineral density. Food frequency questionnaires were used to assess quality of diet and physical activity assessed objectively using accelerometers. The metabolic health of the children was also assessed from annual fasting blood samples (full blood count, lipid profile, insulin resistance). Several measures of pubertal status are also available (hormonal profile, height velocity, Tanner Stage).

The study has published over 60 peer-reviewed articles, many in top medical journals such as Diabetes, Diabetes Care, Pediatrics, BMJ, International Journal of Obesity. The principal aim was to help parents and teachers understand the preventable factors in childhood that are responsible for the current epidemics of diabetes and heart disease.

The study also underpins the Accelerator Prevention Trial now under way by providing the natural history of metabolic change in contemporary children and standards against which the impact of the intervention can be assessed.

Without such knowledge, rational attempts at prevention, whether through the environment or through medication, will not be possible. Findings from the study have already helped inform the public and politicians about the early causes of diabetes and heart disease in our children and how they, as individuals, might avoid them.

The data from this study will be used to test the hypotheses stated in section 1.1.

1.5 Variables of Interest

A table of variables that will be considered in this report from the EarlyBird study are given in table 1.1. Unfortunately many variables from the EarlyBird study were not available during the analysis that takes place herein. The available variables of interest will be studied in detail in the exploratory data analysis section (section 2) and used in the models discussed in sections 5, 6 and 7.

Code	Description
ID	Child's ID Number
pre_diabetes	Logical value. Does the child have impaired fasting glucose?
visityear	Categorical age at visit (ie. 9, 10, 11, 12, 13, 14, 15, 16)
age	Exact age (years)
sex	1 = Boys, 2 = Girls
imd2004	IMD score (measure of socio-economic status)
fat_mass_arms; lean_mass_arms	Fat and Lean mass (kg), region: arms
fat_mass_legs; lean_mass_legs	Fat and Lean mass (kg), region: legs
fat_mass_trunk; lean_mass_trunk	Fat and Lean mass (kg), region: trunk
fat_mass_total; lean_mass_total	Fat and Lean mass (kg), region: total body
fat_massTBLH; lean_massTBLH	Fat and Lean mass (kg), region: total body less head
APHV	Age at peak height velocity (years)
glucose	Glucose (mmol/l)
waist_sds; BMI_sds; height_sds	Waist, BMI and height sd scores: expression of the child's measure relative to their peer group
TPA	Total physical activity (counts)
trunk_leg_vol	Trunk to leg volume ratio
trunk_leg_mass	Trunk to leg fat mass ratio

Table 1.1: Variables of interest; their codes and descriptions

1.6 Summary

This project was inspired by the results of Joseph Wilson et. al [1]. They stated that the novel measure they created [4], Trunk to Leg Volume Ratio, is predictive of diabetes. The EarlyBird study followed ~300 normal healthy children in the Plymouth area and took numerous measurements in a longitudinal study, including DEXA scans, meaning we can test whether whether mass measurements are indicative of pre-diabetes (Hypothesis 1) and whether the novel measure, Trunk to Leg Volume Ratio, is applicable to children (Hypothesis 2). It has also been hypothesised that children's growth and development as well as pre-diabetes prevalence are not mutually exclusive (Hypothesis 3) and therefore there is a time dependency to the risk of developing pre-diabetes.

2 An Overview of the Variables of Interest

2.1 Introduction

To investigate Hypothesis 3, that body composition as well as the odds of becoming diabetic are not mutually exclusive issues, an initial exploratory analysis will take place.

2.2 Exploration of Diabetes

The original idea was for this study to focus on diabetes, but there were no participants in the whole EarlyBird study for whom diabetes was detected (glucose levels ≥ 7 mmol/L). In this case, children who are detected as having impaired fasting glucose, or pre-diabetes, will be used as a proxy for diabetes. A child is determined to have impaired fasting glucose if their glucose levels are greater than 5.6 mmol/L and less than 7 mmol/L. There were 111 (3.999%) time points where impaired fasting glucose was recorded; this was recorded for a minimum of one time point in 60 (17.29%) participants, however they are not necessarily consecutive time points - such is the nature of using cut-off points. Breaking this down by gender, the number of males and females identified as having pre-diabetes is shown in tables 2.1a and 2.1b.

Age	Negative	Positive	Percentage
9	144	3	2.041
10	140	3	2.098
11	140	1	0.709
12	131	8	5.755
13	124	15	10.791
14	116	18	13.433
15	119	16	11.852
16	123	15	10.870

(a) Males

Age	Negative	Positive	Percentage
9	145	0	0.000
10	144	1	0.690
11	140	1	0.709
12	130	3	2.256
13	125	8	6.015
14	124	7	5.344
15	126	7	5.263
16	129	5	3.731

(b) Females

Table 2.1: The number of children with impaired fasting glucose recorded in the EarlyBird data

2.3 An Exploration of DEXA Data

2.3.1 Body Composition Changes Over Time

Figure 2.1 shows the fat mass and lean mass of the children's trunks for both males and females. This plot highlights the increase in the variance of the data as the children age which is particularly

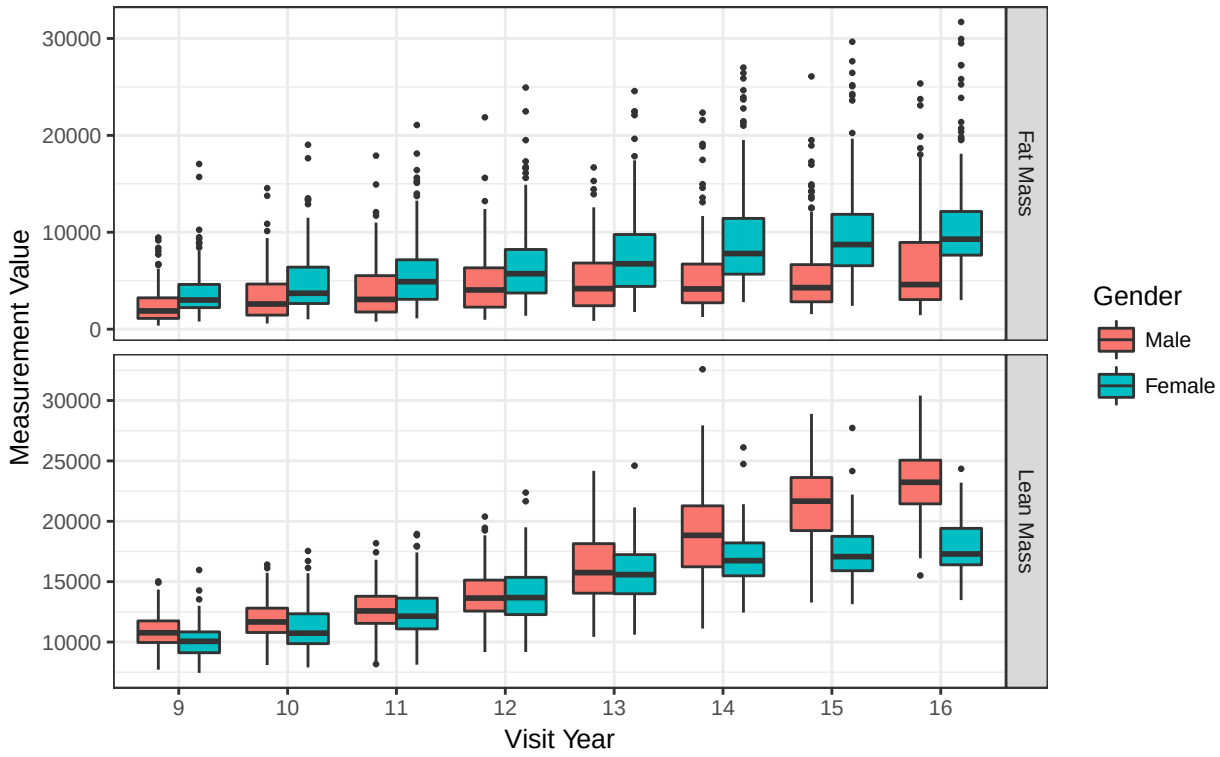

Figure 2.1: Boxplots of the fat and lean masses (grams) of the children's trunks for both males and females

prominent for males in the lean mass variable and females in the fat mass variable. As a result, the distribution of the data becomes much less peaked.

Using t-tests (not adjusted for multiple testing) to test for a difference in mean levels between genders of the log of variables fat mass and lean mass in the children's trunk sizes; p-values are calculated as $p < 0.001$ for ages 9 and 16. Hence we can conclude that at both ages, there is overwhelming statistical evidence of a difference in both mean trunk fat mass and mean trunk lean mass between genders. Similar results were seen in the arms and legs of the children. These results go somewhat towards supporting the theory of hypothesis 3 since we also saw an increase in pre-diabetic levels in the children in section 2.2.

2.3.2 Correlation Between Lean Mass and Bone Mass

Both the lean mass and bone mass variables are highly correlated (figure 2.2), as are the fat mass and bone mass variables, particularly in the earlier years of childhood. This relationship is complex as the spread becomes more pronounced as the children age. Many studies [17] have linked either fat mass or lean mass, or in some cases both, to the size of bone mass. Therefore since these variables are so closely linked, bone mass will not be considered in the models to avoid problems of multicollinearity.

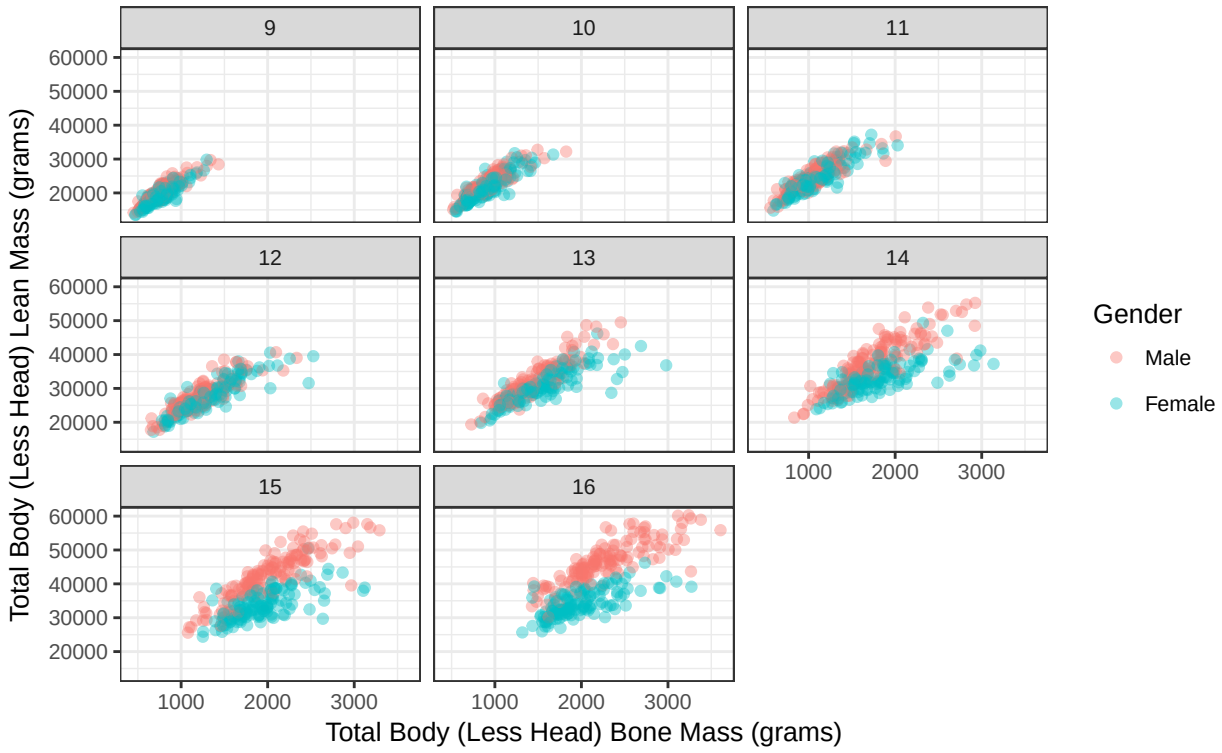

Figure 2.2: Scatter plot of the total (less head) bone and less mass (grams) variables

2.3.3 Correlation Between Weight Related Variables

Traditional measures of fatness such as BMI and waist circumference are highly correlated with several of the fat mass DEXA variables (figure 2.3). As the focus of this report is on DEXA related variables, BMI and waist circumference will be excluded from any models to avoid issues of multicollinearity.

Hypothesis 2 questions whether trunk to leg volume ratio is independent of total and regional fat distributions. We can see in figure 2.3 that the correlation between trunk to leg volume ratio and other weight related measures is low. Although this isn't conclusive evidence, it does point towards this part of Hypothesis 2 being true.

2.3.4 The Influence of Height

Fat mass as an absolute measure is not necessarily a good measure of adiposity since it is correlated with lean mass and height. We are only interested in the effect of the amount of fat a child may have relative to their overall size. As an example, a taller child will have a larger mass but this isn't necessarily a dangerous level of fat, it could simply be because they are taller; therefore we must ensure the masses are proportionate to height. Hence to remove the potential influence a child's height may have on the prediction of impaired fasting glucose, we will use both fat mass

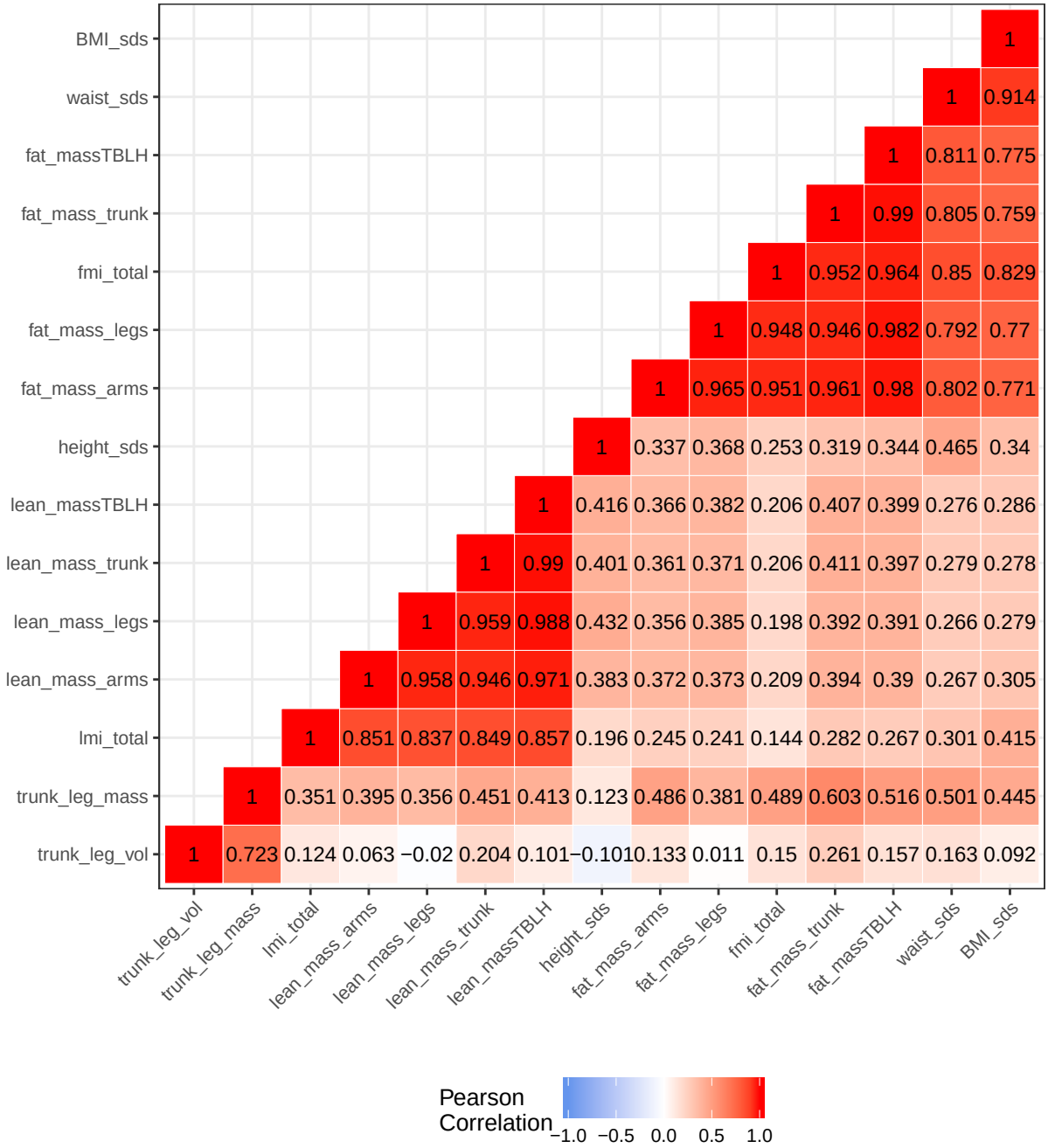

Figure 2.3: Correlation between weight related variables

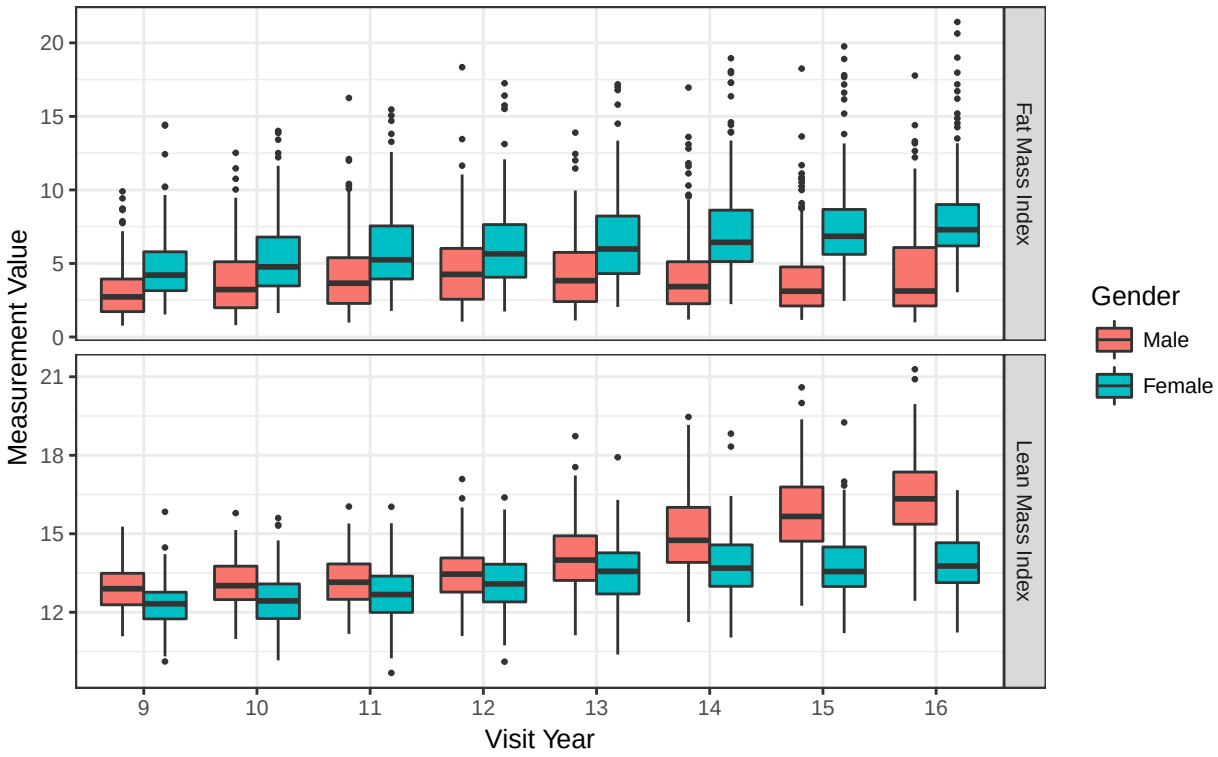

Figure 2.4: Boxplots of the fat and lean mass indices of the children split by gender

index and lean mass index in the modelling process. These are calculated as

$$\text{Size Index} = \frac{\text{mass total (kg)}}{\text{height (m)}^2}$$

Note that this issue has caused some debate in the literature [18], with some suggestion that height $(\text{m})^2$ shouldn't necessarily be used in children. Though there is evidence to suggest that height can be influential in the incidence of diabetes [19]. For the regional compartments of the body, the mass percentage will be used instead, i.e. the mass value (fat or lean) divided by the total mass of the respective part of the body, multiplied by 100.

As we are now modelling the mass indices and not the mass (kg), it is worth taking a look at how these variables change over time. Figure 2.4 shows that these variables behave similarly to the mass (kg) variables as seen in figure 2.1 in that they increase with age and the mass indices of each gender grow at different rates.

2.3.5 Principal Components Analysis

In order to place a more statistically robust emphasis on the results discussed in sections 2.3.1, a principal components analysis was performed on the weight related variables. This can also infer

potential variable reduction. The standard deviations for each of the variables are different; we do not want these variables dominating the analysis and hence we scale the data. This also alleviates the issues of comparing variables of different units.

The first two principal components for each of the age groups cover 81.72%, 81.5% and 75.53% of the variance in their respective data; this is represented visually by the screeplots in figure A.1 of appendix A.1. The variable loadings for the first two principal components are given in table 2.2 and show that at age 9, PC1 is an overall measure of size. PC2, however appears to be a comparison between the lean mass variables and height, which are all negative, and fat related variables, which are all positive; with males having a much higher lean mass to fat mass ratio when compared with females. The separation between males and females can be seen in the biplot in figure 2.5. However by age 12, whilst PC1 still measures the whole body, the separation between males and females is very muddied. Finally at age 16 the separation between genders is at its maximum. Males exhibit larger PC1 values and females exhibit larger PC2 values where PC1 for 16 year olds is a comparison between fat mass and lean mass, however the component is dominated by fatness. PC2, on the other hand, is a measure of overall size. As was seen from the DEXA analyses in section 2.3.1, differences in both males and females is seemingly linked to the mass type.

	PC1	PC2	PC1	PC2	PC1	PC2
lean_mass_arms	0.266	-0.263	0.257	0.300	0.221	-0.333
lean_mass_legs	0.258	-0.329	0.254	0.333	0.233	-0.330
lean_mass_trunk	0.259	-0.306	0.261	0.322	0.222	-0.351
fat_mass_android	0.293	0.200	0.297	-0.206	-0.313	-0.235
fat_mass_gynoid	0.306	0.150	0.303	-0.147	-0.334	-0.178
lean_mass_android	0.256	-0.306	0.270	0.280	0.165	-0.379
lean_mass_gynoid	0.273	-0.294	0.277	0.291	0.205	-0.362
BMI_sds	0.298	0.106	0.299	-0.113	-0.234	-0.309
waist_sds	0.304	0.078	0.305	-0.124	-0.271	-0.288
height_sds	0.243	-0.223	0.233	0.276	0.052	-0.223
trunk_leg_vol	-0.060	0.156	0.058	-0.067	-0.111	-0.105
trunk_leg_mass	0.153	0.319	0.187	-0.157	-0.086	-0.210
arms_fat_perc	0.262	0.316	0.250	-0.349	-0.376	-0.037
legs_fat_perc	0.260	0.298	0.246	-0.333	-0.372	-0.032
trunk_fat_perc	0.271	0.322	0.272	-0.319	-0.371	-0.100

Table 2.2: Variable loadings for princinpal components 1 and 2 for all children aged 9, 12 and 16

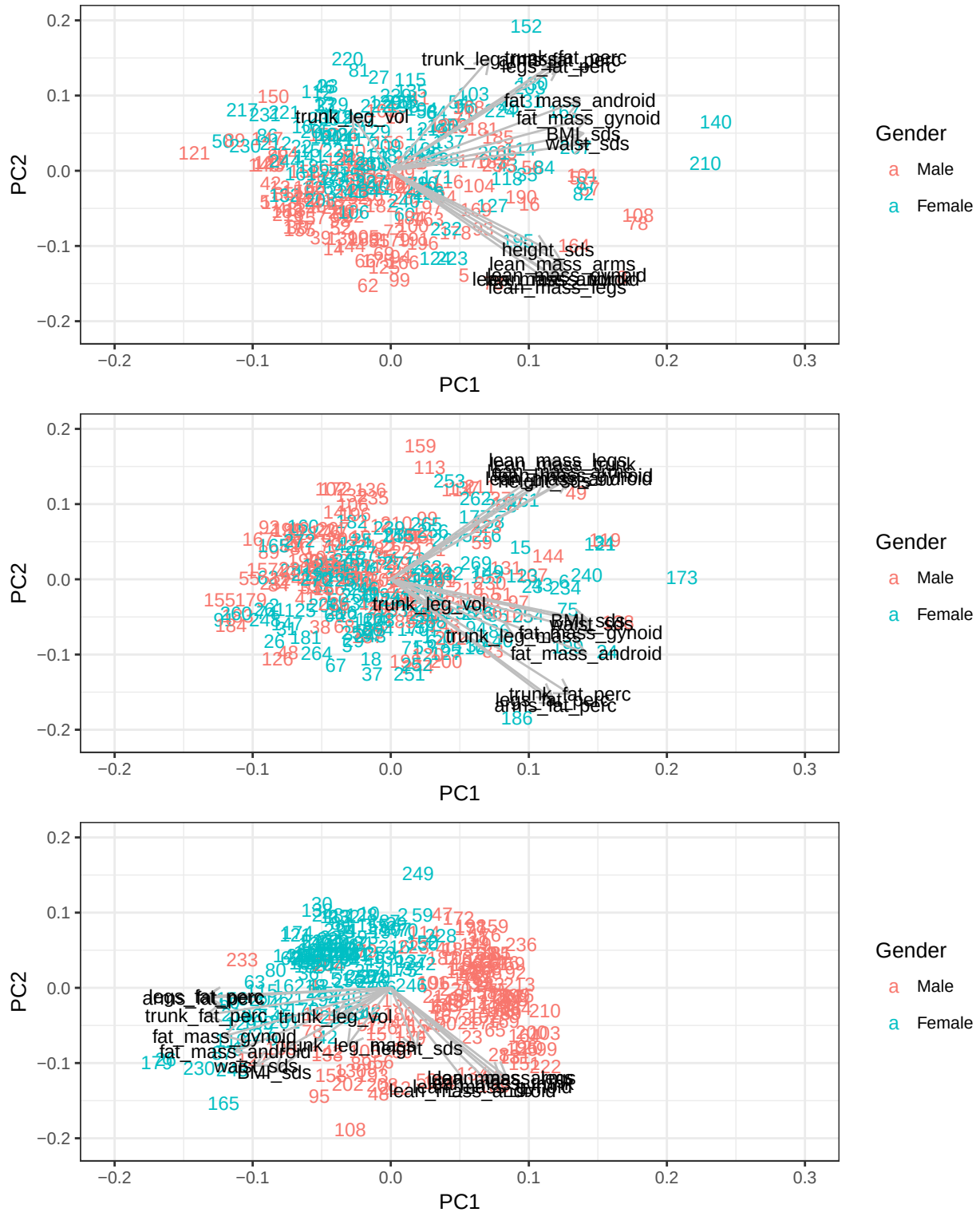

Figure 2.5: Biplots for ages 9 (top), 12 (middle) and 16 (bottom)

2.4 Summary

No children in the EarlyBird study were determined to be diabetic, however some children were classed as pre-diabetic and so this will be used as a proxy for any modelling.

There are large statistically significant differences between genders for each of the body compartments as measured by DEXA. These body compartment measurements are highly correlated with traditional measures of weight such as BMI and waist circumference; as this study is interested in the former, the latter will not be considered in the modelling section to avoid problems of multicollinearity. In addition, bone mass will not be considered either since this data is highly correlated with both fat and lean mass.

Hypothesis 3 states that body composition and the odds of becoming diabetic are not mutually exclusive events. This initial analysis has not tested this hypothesis directly but it shows that there is an increase in pre-diabetes (section 2.2) as time goes on changes in body composition between gender become more pronounced (section 2.3.5).

3 Data Imputation

3.1 Introduction

Missing information are inevitably ubiquitous in longitudinal studies, participants and examiners often cannot make appointments, mistakes are made in the data collection process and participants drop out over time. All of this can result in biased estimates and a loss of power. Pertinent to this report, ways of dealing with missing data are required since as seen in figure 3.1, there are large amounts of missing data in the EarlyBird study; especially for the primary variables of interest from the DEXA scans (23.78%). The approach to modelling taken in this report, as detailed more in section 4, is to use the **glmmLasso** package to find a suitable model and identify important variables which are predictive of the prevalence of impaired fasting glucose. The main function, `glmmLasso`, however requires that all variable lengths be the same, i.e. that no data are missing, otherwise the algorithm does not complete. Hence, as so much data is missing for the children, we must impute some to gain meaningful insights into the data.

For the purposes of this report, different data imputation methods have been considered. Imputation by horizontal mean is discussed in section 3.3 and last observation carried forward and next observation carried backwards are discussed in section 3.2. Additional methods were considered but not used in this report and are briefly mentioned in section 3.4. For all data imputation methods, where age is missing, the visit year was used as a proxy for age.

Due to the large amounts of missing data in this study, only children with no more than 2 missing time points for any given variable were included in the statistical modelling in an effort to reduce the amount of potential bias introduced through imputation. In total there are only 52 children with complete data for the variables of interest. Hence, by including children with up to two missing time points for any given variable, and imputing this missing data, is a worthwhile trade off to use more data from the study and gain more worthwhile insights into the data. This decision to impute missing data increases the amount of children in the model to 198 children out of the original 347. Hence these 198 children have 8 complete data points for all variables, of which 95 are male and 103 are female.

Tests on the missing data, such as testing whether data are missing at random [20] were not performed for this study. The main aim of this study is to model the data that were available but further work should be undertaken to peform such tests.

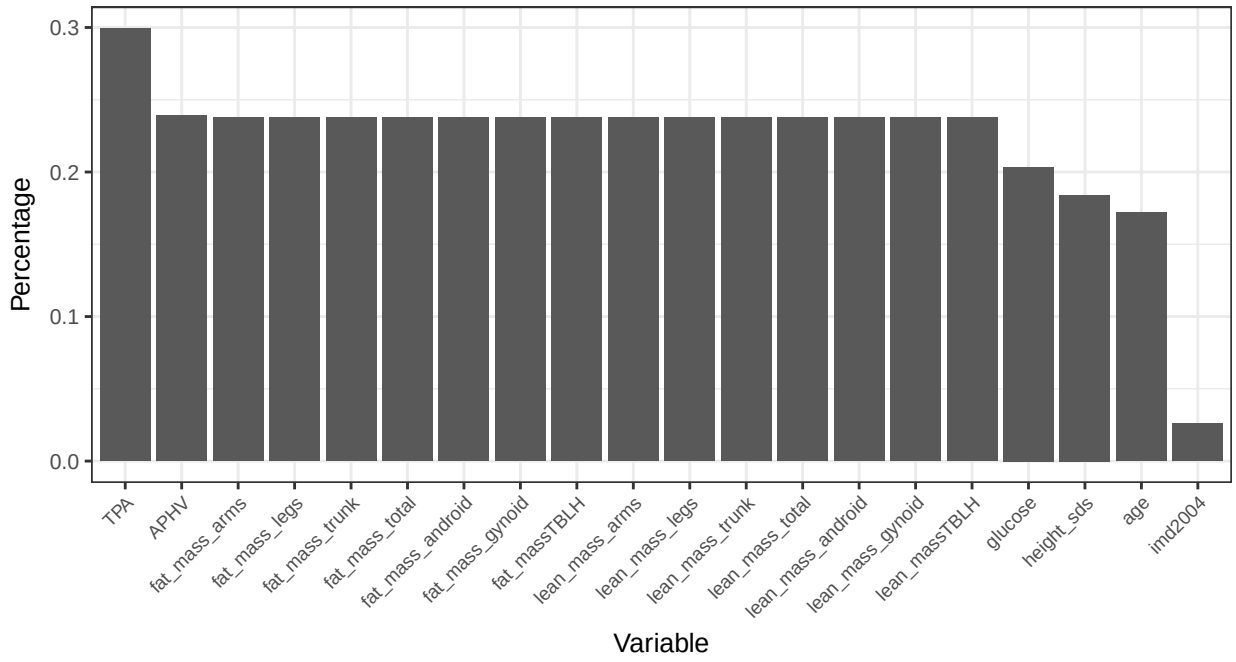

Figure 3.1: Percentage of missing data within the study

3.2 Last Observation Carried Forward

A simple method which was considered for data imputation is the Last Observation Carried Forward (LOCF) approach [20]. This method replaces the missing values with the last available measurement. For example, if a child is missing their third time point, the second time point would be used in its place, assuming this data point were not missing as well. Of course a variable may well be missing data from the start of the study, in this case the first available data point is carried backwards. There are several fallacies to this method, the first being that we are extrapolating data which don't make sense in the context we are using them. These children are growing and changing each year and we would therefore not expect their data to be the same for consecutive years for relevant body measurement variables. Using this method will also seriously alter how the variables relate to one another in a generalised linear mixed effects model given that potentially one quarter of an individual's data for a single variable, or two time points, are effectively mendacious. Finally, variability of data is also reduced, artificially narrowing confidence intervals and increasing the possibility of a Type I error [21]. As a method of data imputation, it relies on the assumption that the data are missing completely at random [21,22].

Similarly the method Next Observation Carried Backwards was considered. This works in a similar way to LOCF whereby it fills in any missing data with the next available data point and fills any missing data in the final years by carrying the previous data forward.

3.3 Mean Imputation

This method aims to alleviate some of the issues caused when using the last observation carried forward technique. We are dealing with body measurement data in adolescents which are continually growing in size, hence carrying a previous year's data point forward does not take this growth into account. We can look to take the mean value of the previous and the next points to the missing value. In doing so, we ensure that we are calculating a more local imputation.

Of course it may be the case that the first or the last observations are missing and so we cannot calculate a mean using data surrounding this time point. Hence we use the last observation carried forward method to fill in the final data points and the next observation carried backwards method to fill in the first data points.

3.4 Other Considerations

The simplest method of data selection is to simply remove all missing data and analyse what is left. If this were a seriously considered method for this study, there would be 38 individuals left in the study. Of course this is very wasteful of data and there are nowhere near enough data for a serious study as it would leave the study very under-powered and would more than likely introduce some very serious bias since the data cannot be assumed to be a random sample [20]. Consequently, this sub data set will not be considered for analysis. Finally, rather than replacing the data with a horizontal mean as discussed in section 3.3, we could use a vertical mean, whereby missing data is replaced by the mean of all children, or gender specific, at the given age where data are missing. However, given the large amount of between variation in the children's data, this is simply not feasible. Other, more complex, methodologies are available but are beyond the scope of this report.

3.5 Summary

Large amounts of data are missing within the EarlyBird study and in order to produce any kind of effective model, we must impute these missing data points in some way so as to avoid a lack of power in our predictions. Some variables are missing up to 23.78% of their data. Two different data imputation methods were considered for this study; Last Observation Carried Forward and Horizontal Mean Imputation. By limiting imputation to those children who were missing up to 2 data points, this gives us 198 children out of the original 347 for whom we can use in our model. It is acknowledged that there are additional imputation methods available.

4 The Modelling Approach

4.1 Introduction

The main objectives of this report (hypotheses 1 and 2) are to discover whether DEXA related measurements can be predictive of impaired fasting glucose in children. However we are not interested in the effects of the children themselves. Recognising that the data are collected into groups of eight observations per subject, we therefore place a random effects term to allow for within-subject correlation.

In addition, there is a time dependent component since the children's physiques are changing drastically over time. Assuming each child grows at a different rate, we should also place a random effect which is an additive modification for each subject, to the slope parameter, i.e. we assume each child has its own trend line.

Of course models with a binary dependent variable require a logistic regression model. The goal of logistic regression is to find the best fitting model to explain the relationship between one dependent dichotomous variable of interest (in our case whether a child is pre-diabetic or not) and one or more independent variables. It does this by calculating the log-odds of an event, i.e. the log odds of the binary variable occurring. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a logit transformation of the probability of presence of the characteristic of interest.

$$\text{logit}(p) = \log\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_{i2} + \cdots + \beta_p x_{ip}$$

where p is the probability of presence of the characteristic of interest and the β terms are the coefficients of the independent variables, x_{ip}. Introducing a random effects component to a logistic regression model requires us to use a generalised linear mixed effects model (GLMM).

There are many potential variables of interest in the EarlyBird study, all of which may or may not be predictive of diabetes and/or pre-diabetes. It is in our interest to find the variables which are most predictive. Many methods for variable selection exist and one particular method, the Lasso [23], uses an ℓ_1-penalty on the regression coefficients which has the effect of shrinking parameter values - some are shrunk to exactly zero; therefore removing the parameter from the model. The idea is to maximise the log-likelihood, $l(\boldsymbol{\beta})$, of the model whilst constraining the ℓ_1-norm of the parameter vector $\boldsymbol{\beta}$. Hence the Lasso estimates are obtained as follows

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \Big[l(\boldsymbol{\beta}) - \lambda ||\boldsymbol{\beta}||_1 \Big]$$

with $\lambda \geq 0$ and with $||\cdot||$ denoting the ℓ_1-norm. λ is a tuning parameter and therefore must be determined, for example by information criterion or cross validation; discussed in section 4.2.2. The regularization problem extends for GLMMs to be

$$\hat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\delta}} \left[l^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) - \lambda \sum_{i=1}^{p} |\beta_i| \right]$$

where $l^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}})$ is the approximate likelihood; $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \boldsymbol{b}^T)$; $\hat{\boldsymbol{\gamma}} = (\phi, \boldsymbol{\psi}^T)$ and ϕ is the dispersion parameter and $\boldsymbol{\psi}^T$ is the covariance structure.

The **glmmLasso** package [24] implements this optimisation problem using a gradient scale algorithm, of which the details are omitted here.

4.2 Computation Details of `glmmLasso`

4.2.1 Starting Values for the Algorithm

In order to use the `glmmLasso` function, starting values for $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\boldsymbol{b}}^{(0)}$ and $\hat{\boldsymbol{Q}}^{(0)}$ (where $\boldsymbol{Q}$ is the covariance structure of the random effects) must be given. Initial values are obtained by fitting the simple global intercept model with random effects [7] using the `glmmPQL` function in the **MASS** library.

4.2.2 Optimising the Value of λ

The Lasso works by penalising and shrinking the coefficients, $\hat{\boldsymbol{\beta}}$, and the **glmmLasso** package [24] provides no way to automate this specification. Therefore two grid search methods are considered in this report which are detailed in the following sections using pseudocode. Both consider a vector of possible values for λ, and then calculate the Bayesian information criterion (BIC) or Deviance at each model iteration, eventually narrowing down to find the optimal value of λ. Using two methods to calculate the optimal tuning parameter provides a form of sensitivity analysis.

4.2.2.1 BIC Grid Search

1. Define a sequence of possible values of λ from 100 to 0, decreasing by 5.
2. For each value of λ, perform the Lasso algorithm and calculate the model's BIC.

3. If there is not much difference in the BIC value for all of the models, go to step 6. Otherwise, go to step 4.
4. Identify the group of models with the lowest BIC, create a new range for λ to be between the largest and smallest values of λ from these models, lowering the interval window from 5 to 1.
5. Repeat step 2 for these new values of λ.
6. Stop; the model with the lowest BIC is the final model.

4.2.2.2 5-Fold Cross Validation Grid Search

1. Define a sequence of possible values of λ from 100 to 0, decreasing by 5.
2. For each value of λ, perform the 5-fold cross validation Lasso parameter search and calculate the deviance. Whichever model has the lowest deviance is the choice model.
3. If there is not much difference in the values of the deviance between choice models, go to step 6. Otherwise, go to step 4.
4. Identify the group of choice models with the smallest deviance, create a new range for λ to be between the largest and smallest values of λ from these models, lowering the interval window from 5 to 1.
5. Repeat step 2 for these values of λ.
6. Stop; the model with the lowest deviance is the final model.

4.2.3 Choosing the Sequence Values of λ

The initial sequence between 100 and 0 by -5 seems rather arbitrary, however an initial investigation took place whereby when the value of λ was set to 100, all models would have no parameters left in the model regardless of the data imputation method and model search method. In other words, the model was over penalised; therefore if the variables in the model are predictive of impaired fasting glucose then the optimal value of λ must be smaller than 100.

4.3 Limitations

There are several limitations to the approach used for this project which are detailed in the following sections.

4.3.1 Grid Search Limitations

The very nature of using grid search limitations means we will not get an accurate view of the 'true' model. Only certain values of λ will ever be considered and it would be computationally expensive

to fine tune λ to its local minimum. This has implications for any model parameters described in this report in that they will never be 'correct'. Though of course,

> All models are wrong, some are useful. –George Box [25]

Of course this also means that there is potential for some parameters to never be shrunk to zero as they should and therefore may remain in the model.

4.3.2 Assuming Individual Trends

Including a random intercept term in the model will allow us to model the substantial differences in general distance between children. However this data has a time dependency and each child grows at a different rate, regardless of age or gender. Furthermore, the rate of children with detectable impaired fasting glucose rates increases over time (table 2.1) but again, the levels of glucose vary between children. It is therefore sensible to assume that each child's level of glucose follows a different trend to all other children. In order to model this feature we need to introduce a random effect which is an additive modification for each subject to the slope parameter of the underlying regression, i.e. `age | ID`.

Unfortunately, the **glmmLasso** and **MASS** packages are computationally expensive. R stores its data in memory [26] and many of the matrices calculated during the algorithm are large in memory terms; therefore using the EarlyBird data, we quickly run out of memory when running models of this type. This in turn causes R to crash as the models are unable to converge and we are unfortunately unable to obtain the results we require. Therefore the models in this report do not assume individual trend effects.

4.3.3 Standarising Variables

Similar issues with memory occur when trying to find the starting values for the algorithm with standardised versions of the variables. Initially it was hoped that we could standardise the variables so that each variable contributes equally to the analysis but when running the `glmmPQL` function on the standardised data, it caused R to crash, thus we couldn't obtain starting values. Standardisation would also allow the interpretation of the model to be more natural by placing emphasis on the variables and not the intercept term.

4.3.4 Including Main Effects and Interaction Effects

Typically when developing a statistical model we would like to leave main effects in the model if interaction effects are significant, though this isn't always necessary. It is possible for the `glmmLasso`

function to produce results containing interaction effects but not main effects. One would hope for a parameter within the `glmmLasso` function to ensure that where interaction effects remain in the model, so too should main effects but this is not the case. The parameter `index` exists to ensure certain parameters always remain in the model but in this case we are interested in finding the significant variables which explain the risk of developing impaired fasting glucose, and not how certain parameters affect this risk. Corresponding with the package author revealed that research is currently under way to implement this feature, though it is not currently available.

4.4 Summary

The study objectives require us to use a generalised linear mixed effects model (GLMM), however as there are so many variables of interest within the EarlyBird study, it is imperative that we discover which of these is indicative of pre-diabetes - if any. Therefore a Lasso approach is considered for this report whereby the log-likelihood is maximised whilst constraining the ℓ_1-norm of the parameter vector, $\boldsymbol{\beta}$ for a given value of λ. This in turn shrinks parameter estimates, sometimes down to zero, removing them from the model.

To determine the optimal value for λ, two separate algorithms are employed for this report which allows us to perform a sensitivity analysis on the found models. The first calculates the BIC of each model for each value of λ to find the optimal model whereas the second uses a 5-fold cross validation at each value of λ and chooses the model with the lowest deviance.

One issue discovered with this approach is that the packages available in R which perform GLMMs are limited. Packages which provide a Lasso parameter search for GLMMs are even more so. The **glmmLasso** package was chosen for this report but it is limited in that it is computationally expensive and will not allow us to perform the algorithm for individual trend effects without running out of memory.

5 Total Body Mass Index Models

5.1 Introduction

A model was fit using the whole body measures of both fat and lean mass - fat mass index and lean mass index. The model also included measures of age, gender, age at peak height velocity, the child's home's IMD score and their total physical activity. In addition, due to the differences in the children's bodies between genders over time seen in section 2.3.1, interactions between the mass index variables, sex and age were all included in the model.

For each of the data imputation methods, mean imputation and LOCF, and model search algorithms, BIC and CV, the same model was found by the lasso algorithm for which the resulting model fit can be seen in table 5.1. Values for both the AIC and BIC for each model are given in table 5.2 and though these cannot be compared between imputation methods, they are very similar for each model.

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-47.017	0.253	-185.488	< 0.001
age	2.816	0.737	3.822	< 0.001
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
lmi_total	2.898	0.762	3.802	< 0.001
fmi_total	0.000			
age:sexFemale	0.000			
age:lmi_total	-0.185	0.052	-3.536	< 0.001
sexFemale:lmi_total	0.000			
age:fmi_total	0.000			
sexFemale:fmi_total	0.000			

Table 5.1: Fixed effect model results from the total value model with mean data imputation and BIC grid search

Imputation	Algorithm	AIC	BIC
Mean	BIC	511.67	892.28
LOCF	BIC	511.56	891.95
Mean	CV	511.67	892.29
LOCF	CV	511.56	891.96

Table 5.2: AIC and BIC values for each of the four total models.

5.2 Inference

The algorithm found age, lean mass index and the interaction between them to all be overwhelmingly significant ($p < 0.001$). It was evidenced in the exploratory analysis that fat mass and lean mass were very different from one another and that these differences changed over time (section 2.3.1). It was also noted that they varied with gender and so it is surprising to see that the algorithm has removed gender from the model.

The final total body mass index model is given as follows

$$\begin{aligned}\text{log odds of pre-diabetes} = &-(47.02 + u_i) \\ &+ 2.82 \times \text{age} \\ &+ 2.9 \times \text{total lean mass index} \\ &- 0.18 \times \text{age} \times \text{total lean mass index}\end{aligned}$$

Where u_i is the random intercept effect for child i.

Therefore interpreting the coefficients tells us that a one unit increase in age is associated with a 2.82 unit increase in the expected log-odds of pre-diabetes; similarly for each unit increase in lean mass index, the expected log-odds increases by 0.29. However for each unit increase in both age and lean mass index, the log-odds of being pre-diabetic decreases by 0.018. This relationship between lean mass index and age is therefore complex. The model tells us that younger children with a larger lean mass index is more likely to be pre-diabetic, and this continues to be the case until children are older; older children with a lower lean mass index are more likely to be pre-diabetic. Hypothesis 3 stated that the prevalence of impaired fasting glucose was dependent on the development of children as they age as this model seems to suggest that this hypothesis is correct.

The coefficients are more interpretable by calculating the odds of being pre-diabetic. To do this, we simply take the exponential of the coefficients, remembering that any coefficients that are 0 in table 5.1 are effectively removed from the model. Therefore we get the following odds ratios

Parameter	OR
(Intercept)	0.00
age	16.72
lmi_total	18.14
age:lmi_total	0.83

Table 5.3: Odds ratios from the total value model with mean data imputation and BIC grid search

Therefore the odds of being a pre-diabetic child can be defined as

$$\begin{aligned}\text{odds of pre-diabetes} = & u_i \\ & + 16.72 \times \text{age} \\ & + 18.14 \times \text{total lean mass index} \\ & + 0.83 \times \text{age} \times \text{total lean mass index}\end{aligned}$$

We should also take note of the standard errors for each of the variables which are relatively small and do not cause any concern.

It is interesting that lean mass index comes out as the significant variable in describing the prevalence of impaired fasting glucose as one might suspect that the important variable here would be fat mass index. Yet running the same model without lean mass index yields models containing only age for 3 versions of the model (BIC + mean imputation, BIC + LOCF imputation, CV + mean imputation). The only model that produces any meaningful results is the model with LOCF data imputation and CV grid search which can be seen in table A.1 of Appendix A.2. Here, fat mass index itself is actually removed from the model and the interactions are highly non-significant (`age:fmi_total`: 0.667, `sex:fmi_total`: 0.9999).

5.3 Model Diagnostics

Figure 5.1 shows the Pearson residuals against the fitted values. This plot appears to show a few outliers which are not dealt with in this report, but overall there are no major causes for concern. Observation 1121 belongs to a 9 year old male. Looking at his data, he did a very small total physical activity (123152.1 counts/minute) compared to the rest of the cohort; the next smallest amount of TPA, for a male this age, is over double 252081.4 whereas the mean is much larger 550231.8.

The residuals have also been plotted against the fitted values for each of the variables within the model to look for any variables which may need an exponent term. Figure 5.2 shows no reason to consider any, however. This figure also highlights that the residuals are largest for those children who were classified as pre-diabetic at least once during the course of the study.

5.4 Random Effects

The variance of the random effect, `ID`, is 1.588, showing there is a large amount of variance between children and therefore we were correct to include a random intercept for each child. Plotting the random effects for each individual shows a vast difference between those children who have had at least one detectable pre-diabetes time point and those who haven't (figure 5.3). There almost

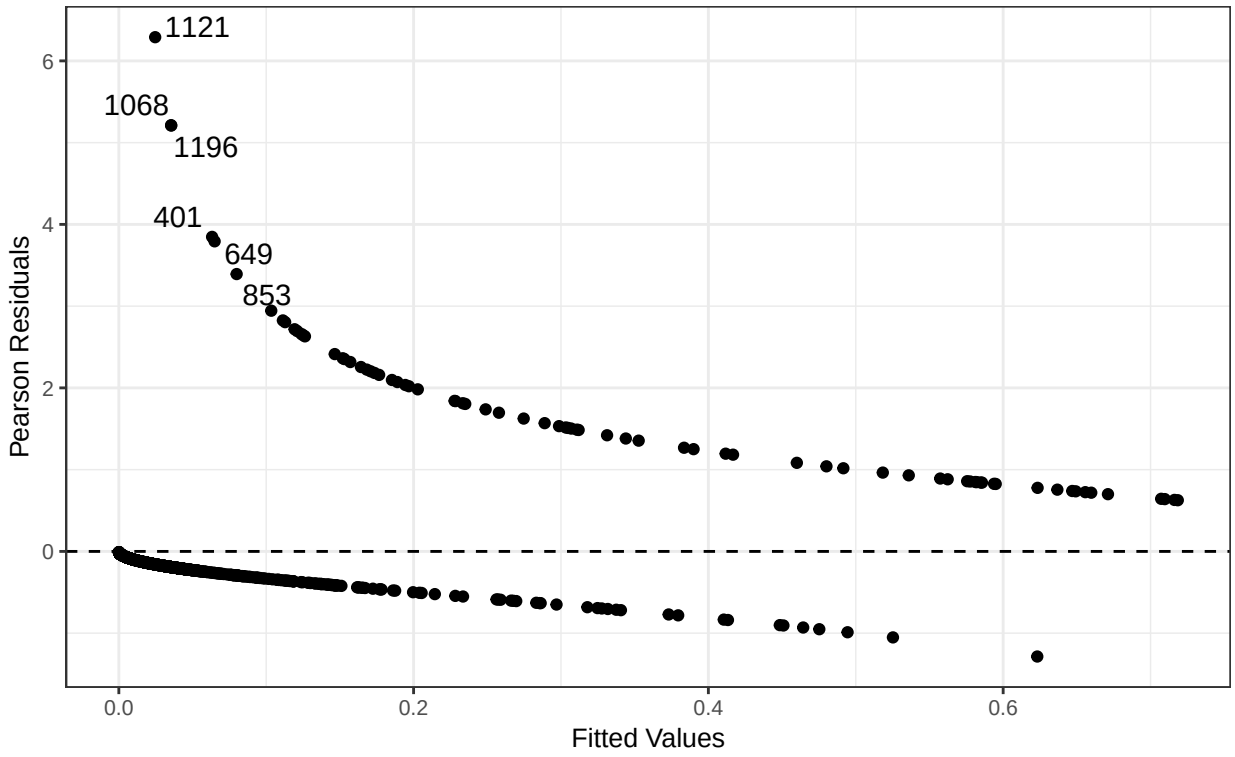

Figure 5.1: Residuals vs. fitted values for the Total Mass Index model

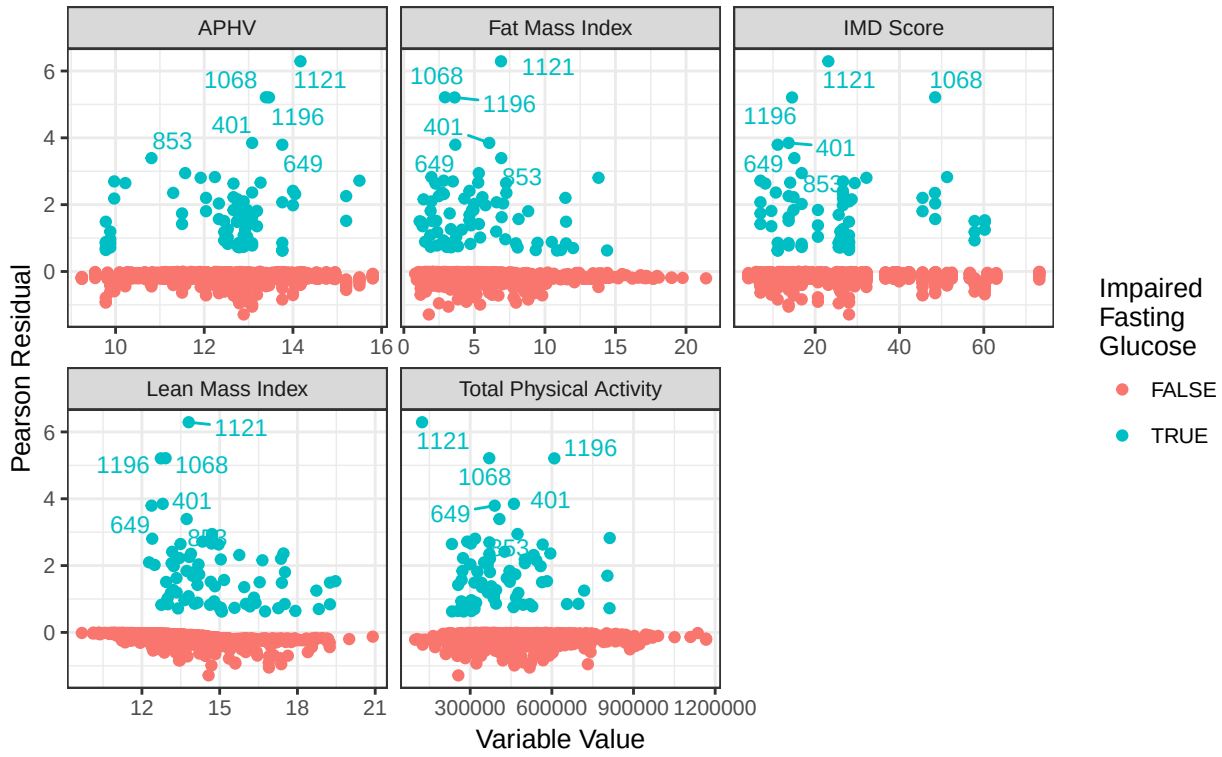

Figure 5.2: Residuals vs. variable values for the Total Mass Index model

appears to be three distinct groups of random effect values; those who were never identified as being pre-diabetic; and two groups of children who were at some point detected as being pre-diabetic, some with a random effect of around 1 and others greater than 1.9.

We can plot histograms of the expected probabilities of impaired fasting glucose from our model for our entire sample, holding the random effects at specific values. Figure 5.4 shows the predicted probabilities at random effect values -0.5, 0, 0.5, 1, 2 and 3. This gives us a sense of how much variability in pre-diabetes can be expected by each individual (the position of the distribution) versus by fixed effects (the spread of the distribution within each graph) [27].

What we can see is that for our sample of children, both the position and the shape of the distribution changes dramatically as the value of the random effect increases. This tells us that there is variability both within and between children and therefore we are correct to include a random effects term in the model, however we also need random effects on one or more of the fixed effects. Yet, as was discussed in section 4.3.2, this is not computationally feasible with the computer at hand.

5.5 Prediction

We can calculate the expected probability of a child being pre-diabetic by calculating the logistic function. As an example, the expected probability of a 9 year old child with mean lean mass index of 12.65 gives a predicted probability of 0.24%[1].

$$\begin{aligned}\text{Prediction} &= \frac{\exp(-47.02 + 2.82 \times 9 + 2.9 \times 12.65 - 0.18 \times 9 \times 12.65)}{1 + \exp(-47.02 + 2.82 \times 9 + 2.9 \times 12.65 - 0.18 \times 9 \times 12.65)} \\ &= 0.0024\end{aligned}$$

If a 9 year old child has a random intercept effect of 3, their predicted probability increases to 4.7%.

$$\begin{aligned}\text{Prediction} &= \frac{\exp(-(47.02 + 3) + 2.82 \times 9 + 2.9 \times 12.65 - 0.18 \times 9 \times 12.65)}{1 + \exp((-47.02 + 3) + 2.82 \times 9 + 2.9 \times 12.65 - 0.18 \times 9 \times 12.65)} \\ &= 0.047\end{aligned}$$

A 14 year old child with mean lean mass index of 14.39 will have a predicted probability of 4.47%. If that child has a random intercept effect of 1.5, it will see their predicted probability increase to 17.3%. These figures fall in line with the results we saw in table 2.1.

[1]Note the exact coefficient values were used in this calculation and so using the rounded values shown in table 5.1 will produce slightly different results.

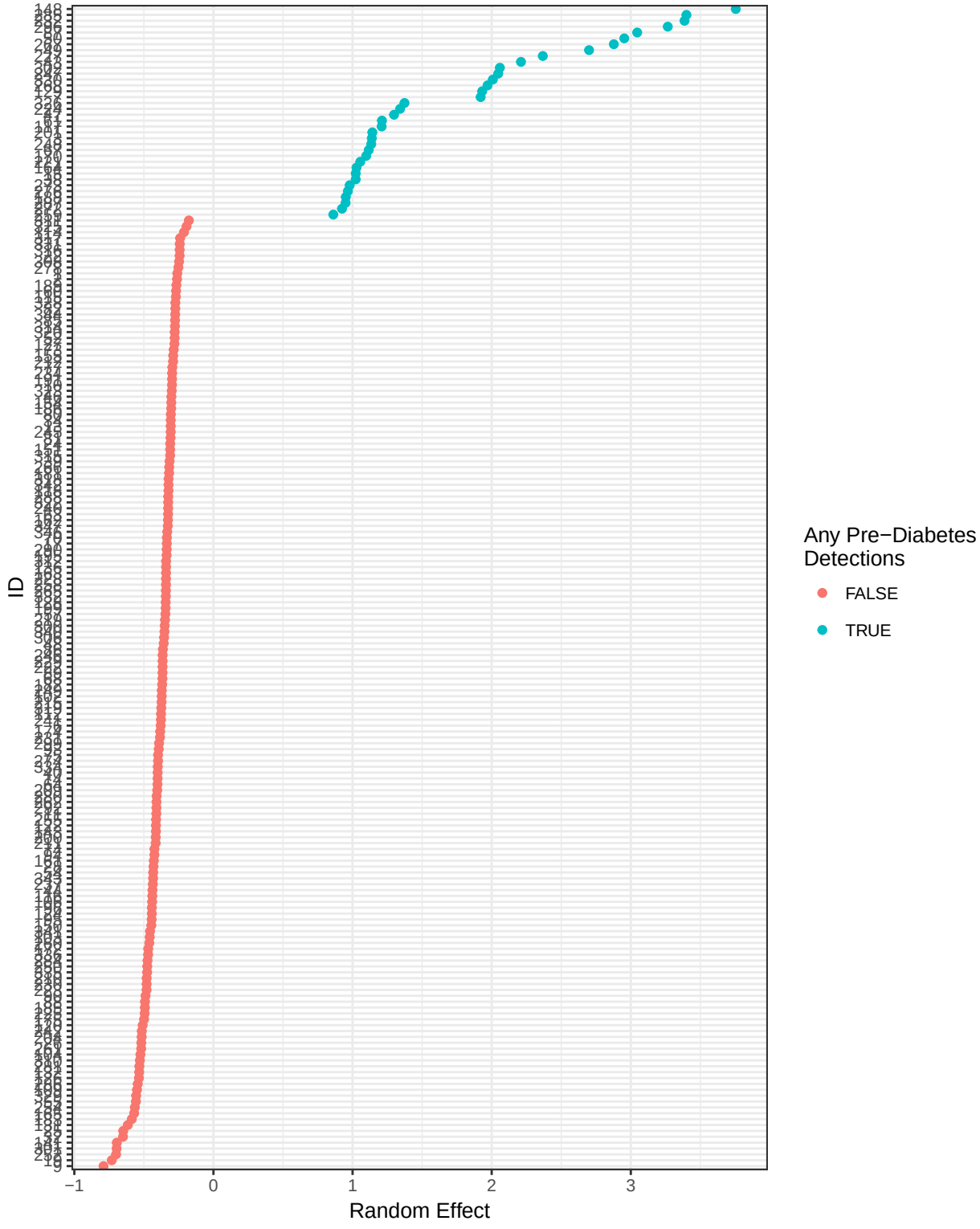

Figure 5.3: Random effects of each child as calculated by the Total Mass Index model

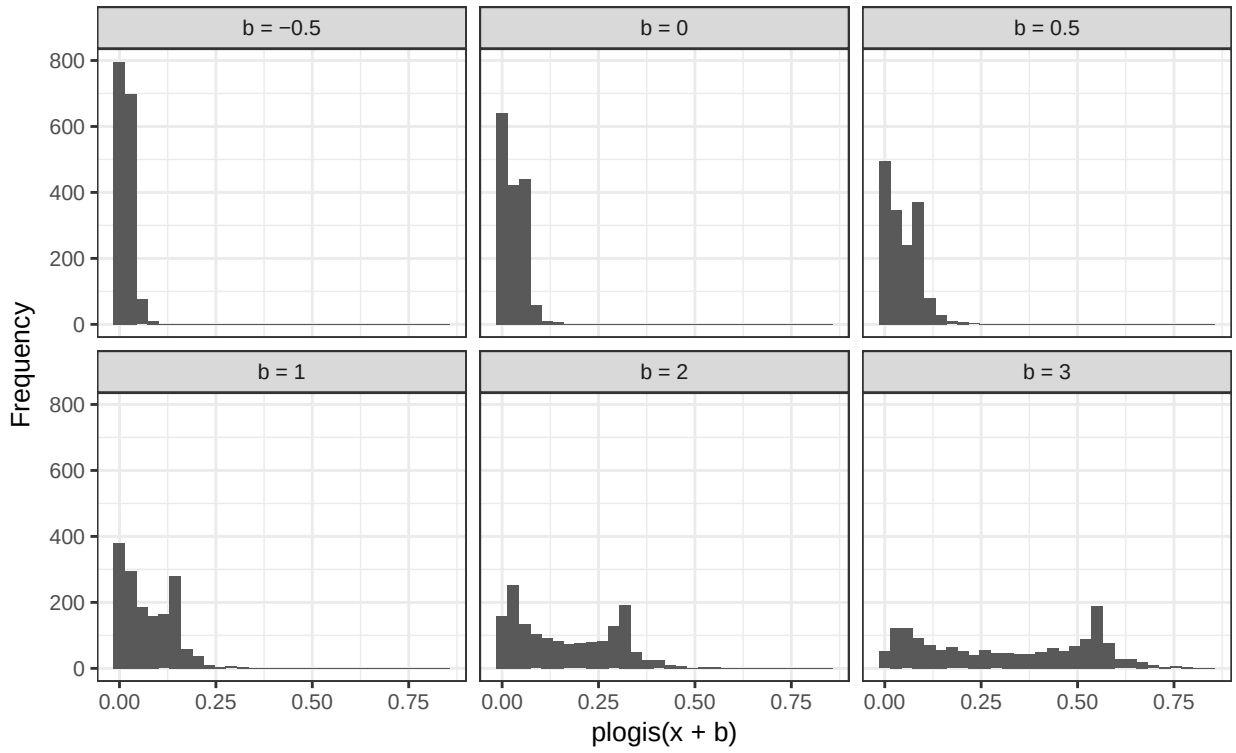

Figure 5.4: Histograms of predicted probabilities of impaired fasting glucose for the children for varying random effect values

Figure 5.5 provides a valuable insight into the interaction between lean mass index and age. This figure shows predicted probability (y-axis) for a child of a given age (x-axis) for increasing quantiles of lean mass index. This is a very useful way to visualise the interaction between age and lean mass index as we can see that children with a higher lean mass index are more likely to be pre-diabetic when they are younger, but as they age, this begins to change. At age 14, those with the highest levels of lean mass index (90^{th} percentile) begin to see a reduced risk of impaired fasting glucose and the same is seen at age 15 for those children with the top 75% lean mass index. Conversely the risk of impaired fasting glucose for those children who have smaller lean mass indices begins to rapidly increase as they reach their mid-teens. This can be seen again is figure 5.6 where the slope for children aged 16 is negative; the plot this time shows the predicted value against increasing quantiles of lean mass index. The slope for age 15 increases along with the rest, though for those in the highest quantile, it is not as steep.

One theory as to why lean mass index is indicative of impaired fasting glucose in adolescents is that their bodies are very different to that of an adult. Typically fat mass is prevalent in the prediction of diabetes [2] but this is typically tested in adults. We see that as the children age, the the risk of pre-diabetes is beginning to decrease for those children with a higher lean mass index and it may well be that fat mass becomes the important factor as the children reach adulthood. This particularly links with hypothesis 3 as we can see that the prevalence of impaired fasting glucose is

changing over time.

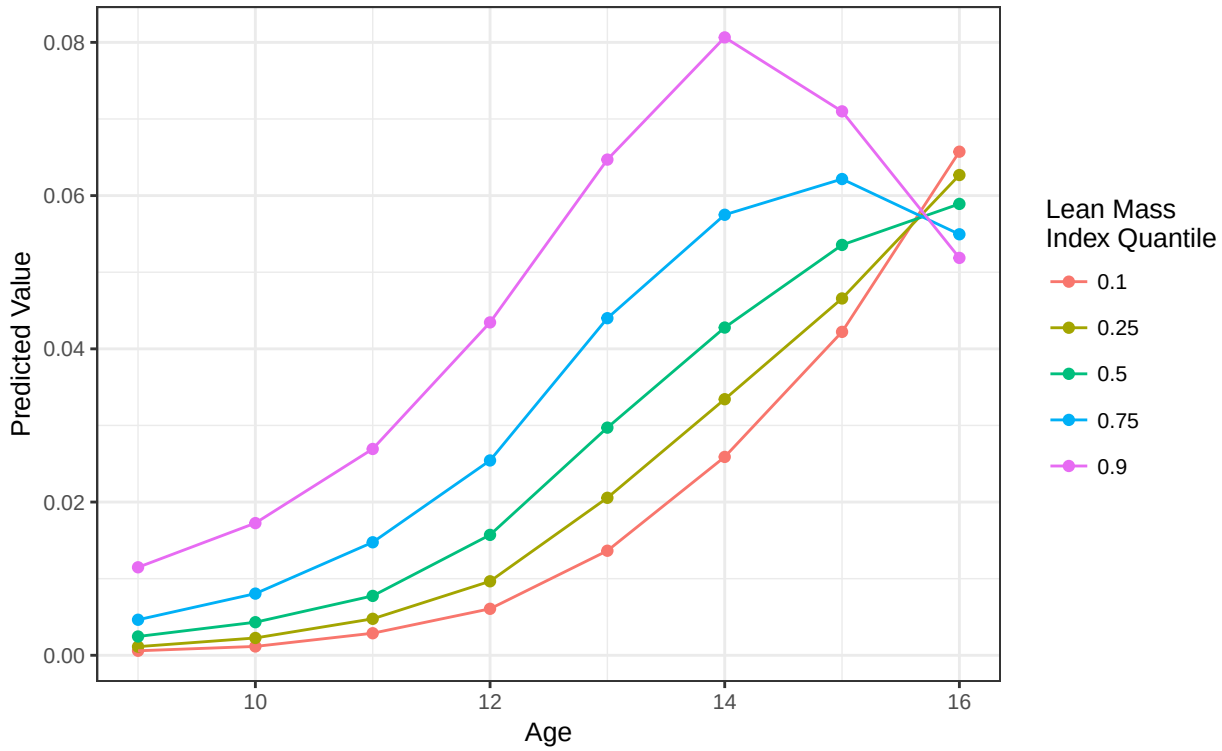

Figure 5.5: Predicted probabilities of impaired fasting glucose in children for varying ages at varying quantiles of lean mass index values

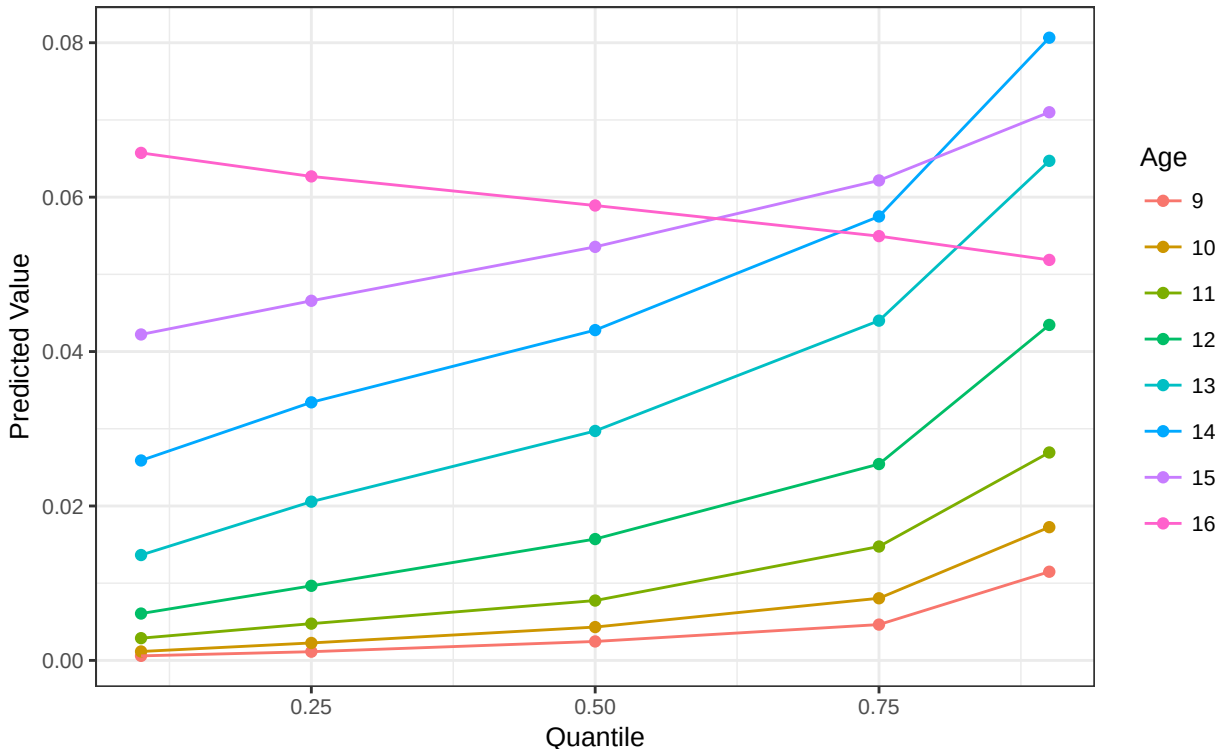

Figure 5.6: Predicted probabilities of impaired fasting glucose in children for varying quantiles of lean mass index for different ages

5.6 Summary

For each of the model search algorithms and data imputation methods, the same model was found whereby, the algorithm removed all variables from the model except for age, total lean mass index and the interaction between the two. Therefore the model is telling us that the prevalence of childhood impaired fasting glucose is dependent on the child's age and their total lean mass index (which takes both lean mass (kg) and height into account) and the interaction between the two. The interaction between age and lean mass index is complex; children with lower lean mass indices at a younger age have lower odds of developing impaired fasting glucose but have a higher risk closer to adulthood than those children with higher lean mass index. One theory for this is that children's bodies are much different to that of an adult, particularly during puberty and therefore as the children reach adulthood it may be that lean mass index becomes less important and fat mass index, which is typically associated with a risk of diabetes [2], becomes more important.

The model diagnostics showed no real causes for concern other than a few outlying values though the random effects values showed that there is clearly a lot of variation both within and between children. There is a real need to study these random effects further in a future study to try to identify which of the variables is causing so much variation.

Finally the model suggests that the risk of developing impaired fasting glucose is increased for children with larger levels of muscle mass at a younger age.

6 Regional Body Mass Percentage Models

6.1 Introduction

The advantage of using DEXA measures over traditional weight related variables is the breakdown of different parts of the body it provides you with. For example in the EarlyBird study, we have access to bone, lean and fat masses for the arms, legs, trunk, android and gynoid as well as the head in each of the children. Therefore whilst the model discovered in section 5 is rather revealing, it is worth knowing whether certain parts of the body are more strongly associated with impaired fasting glucose within children. Given the clear distinction between fat and lean mass, as shown in sections 2.3.1, 2.3.5 and 5, individual models were fit for both fat percentages and lean percentages throughout the body using the same data imputation methods and model search algorithms as those employed in section 5.

6.2 Models for Regional Fat Mass Percentage Distributions

For the regional fat percentage models, the imputation method didn't make too much of a difference to the models found (table 6.1).

Imputation	Algorithm	AIC	BIC
Mean	BIC	530.29	904.47
LOCF	BIC	530.29	904.47
Mean	CV	518.25	936.48
LOCF	CV	519.03	928.65

Table 6.1: AIC and BIC values for each of the four regional fat percentage models

Using the BIC grid search algorithm, the models retain only `age`, with all other variable coefficients shrinking to 0, as seen in table A.2 in Appendix A.3, and so it could be argued that these models are being over penalised. The cross validation model searches, however, provide significantly different models, to those of the BIC search, which are shown in tables A.3 and A.4 in Appendix A.3. These latter models do retain variables in the model which are non-significant ($p > 0.05$) though and so one could argue that these models are under penalised.

It was mentioned in section 4.3.3 that there were issues standardising the variables. The problem that this causes is emphasised in this model output. The variable `TPA` (total physical activity) has a coefficient value of -0.0000015; which could potentially mean one of two things. The first is that the model algorithm will remove this parameter for a more accurate value of λ as it will be

further shrunk towards zero and thus removed from the model. This is a limitation of using grid search methods; we cannot possibly search over all possible values of the penalization parameter, λ. The second, more likely, reason is that TPA data values are much larger (mean TPA is 441842.5) than, say, fat mass percentages (mean trunk fat percentage is 25.46) and so we would expect the coefficient value to be smaller, otherwise this variable would dominate the model. Of course this issue would be alleviated were we able to standardise the variables but this was not possible for reasons described in section 4.3.3.

The amount of insight that can be gained from the regional fat model therefore is limited. What it does show, however, is that there is some statistical evidence to suggest that the fat percentage in the trunk is indicative of impaired fasting glucose in children ($p = 0.042$; table A.3) as well as its interaction with gender ($p = 0.048$; table A.3) where females with a larger fat mass have higher log-odds of being pre-diabetic than males; though gender itself is not statistically significant it does have a large coefficient indicating that it does have some effect on the log-odds of being pre-diabetic. However sex has a very large standard error, 3.147. Interestingly, the models shown in tables A.3 and A.4 do not classify age as significant as was found in section 5 but again have substantially sized coefficient values for age relative to the other coefficients remaining in the model.

Furthermore the model shown in table A.3 has a larger coefficient for the fat percentage in the arms of the children compared with the model results shown in A.4 which places more emphasis on the fat percentage in the legs.

Model diagnostics (figure A.2; Appendix A.3) showed similar issues to those seen in the total mass model; there are a few outlying values which are not dealt with in this report.

6.3 Models for Regional Lean Mass Percentage Distributions

Imputation	Algorithm	AIC	BIC
Mean	BIC	528.15	904.60
LOCF	BIC	528.09	904.46
Mean	CV	526.87	930.40
LOCF	CV	526.14	905.65

Table 6.2: AIC and BIC values for each of the four regional lean percentage models

AIC and BIC values for the lean regional percentage models are similar for each of the models expect for the model found with mean data imputation by the cross validation method (table 6.2). For the three other models, they retain only age and the interactions between age and lean mass percentage in the arms and legs (tables A.5, A.4 and A.7; Appendix A.4), though the model found

using the LOCF data imputation method and CV grid search also retains the interaction between age and trunk lean mass percentage. For each of these models, the coefficient values and standard errors are very small so we shouldn't read into their results too much, particularly as all parameters are non-significant ($p > 0.05$).

The model that stands out therefore is the one found by the cross validation algorithm with mean data imputation (table A.8). It should be noted, however, that this model has the largest BIC value overall for each of the lean percentage models as seen in table 6.2. The model does appear to be a comparison between arm, leg and trunk lean mass percentages, with parameter estimates of 0.648, 0.119 and -0.66. Only the trunk lean mass percentage variable shows some evidence of being statistically significant, however ($p = 0.043$). This corroborates with the evidence shown in the PCA in section 2.3.5 in which one principal component was always a measure of overall size. Again, age has a large coefficient value, 0.59, but in this case is not statistically significant, however the interaction between age and trunk lean mass percentage shows weak evidence of being statistically significant ($p = 0.085$). We could therefore say that impaired fasting glucose is in fact influenced by the level of lean mass in the whole body of children, as suggested by the model found in section 5, though there is some evidence to suggest that the level of lean mass in the trunk plays a larger role than that of the arms and legs.

Model diagnostics (figure A.3; Appendix A.4) showed similar issues to those seen in the total mass model; there are a few outlying values which are not dealt with in this report.

6.4 Summary

The grid search algorithms do not perform as well for the regional models as they did for the total models; there are discrepancies between the cross validation and BIC algorithms employed for this report.

The regional models using cross validation discussed in sections 6.2 and 6.3 both seem to imply that the risk of developing impaired fasting glucose in children is linked to the level of mass within the trunk since there is some statistical evidence to suggest that this variable is statistically significant in both models ($p = 0.042$ and $p = 0.043$). The regional fat model differs from the regional lean model however as it suggests that that the interaction between trunk fat percentage and gender is statistically significant ($p = 0.048$) whereas the regional lean model suggests that the interaction between age and trunk lean percentage is (weakly) statistically significant for the model using mean data imputation ($p = 0.085$). It would appear then that the trunk is possibly the area of interest in which further studies should be focused. The fact that the lean and fat mass percentage models differ in the importance of age and gender suggests that the interaction between these parameters and the mass parameters is complex, which supports hypothesis 3.

Both models would appear to suggest that arms and legs also contribute towards the risk of impaired fasting glucose in children, though their contribution is not statistically significant ($p > 0.05$ in all cases). This corroborates what the principal component analysis suggested, whereby the overall size is of primary importance.

7 Investigation of Trunk to Leg Volume Ratio as a Predictor of Pre-Diabetes

7.1 Introduction

In order to test Hypothesis 2, that trunk to leg volume ratio is independent of total and regional fat distributions and predictive of impaired fasting glucose, a little more background is needed. The total body volume is an important health metric used to measure body density, shape and multi-compartmental body composition however it is currently only available through underwater weighing or air displacement plethysmography (ADP). Wilson et al. [4] performed a study to investigate whether an accurate measure of body volume could be derived from DEXA reported measures. Using a volunteer group of 25 patients, whole body DEXA scans and ADP measurements were taken at baseline and again at six months (only 22 patients were included at six months). The trunk to leg volume formula is calculated as such [4]

$$\text{DEXA}_{\text{volume}} = \nu_{\text{Fat}} \times \text{Fat} + \nu_{\text{Lean}} \times \text{Lean} + \nu_{\text{BMC}} \times \text{BMC} + \nu_{residual}$$

where fat, lean and BMC were the mass measures reported by the DEXA software. $\nu_{residual}$ was the residual volume not explained by the lean, fat and bone mineral content (BMC) components, and ν_{Fat}, ν_{Lean} and ν_{BMC} were the coefficients for each respective compartment and correspond to the inverse densities of each of the DEXA components [4]. The authors used linear regression, Student's t-test and Bland-Altman analyses to compare the change in DEXA-volume to change in ADP-volume between baseline and 6 month measures, eventually coming up with the following formula, using coefficients from the linear regression model.

$$\text{DEXA}_{\text{volume}} = \frac{\text{Fat}}{0.88} + \frac{\text{Lean}}{1.05} + \frac{\text{BMC}}{4.85} + 0.01$$

The baseline had no statistically significant differences in characteristics from the six month follow-up measures by the Student's t-test. DEXA and ADP volumes were highly correlated ($R^2 = 0.99$). The precision of the formula was measured using a second sample of self-selected patients from the NHANES study[2]. Using this data ($n = 385$) it was determined that there was no significant change

[2]The National Health and Nutrition Examination Surveys (NHANES) is a program of studies that began in the 1960s designed to assess the health and nutritional status of adults and children in the United States [5]. Each year approximately 5000 persons, located in counties across the US, are recruited to take part in the NHANES survey as well as undergo a laboratory examination. The aim is to collect demographic, socio-economic, dietary and health-related data using the survey and then in the laboratory examination collect medical, dental and physiological measurements such as anthropometric measurements and full body DEXA scans. This data is very similar to that collected during the EarlyBird study, though demographics will of course differ.

in the coefficients for the model and the precision was in fact better for volume ratios compared with their equivalent fat mass ratios (e.g. trunk to leg volume compared to trunk to leg fat mass).

7.1.1 Limitations of the Study

The study was limited by its small sample size ($n = 47$), which the authors confess to. Whilst the authors state the brand of DEXA scanner they used for the study, what they fail to mention is the known differences between DEXA scanners [15,16]. These two facts directly impact the confidence in the coefficient values used for the model since they could plausibly be different given another sample population or another brand of DEXA scanner, hence this formula may not be applicable to all data sets. For the second sample, no statistical sampling was applied and participants were self-selected volunteers from NHANES, thus again, a properly powered study would be required to use the model formula in other studies with any real confidence; not to mention the potential for selection bias. Finally as this model is based on adult patients, it may not be applicable for children as the human body undergoes large changes during adolescence and the model coefficients may well vary for different genders and body transitions.

7.1.2 Comparisons with EarlyBird

The results of the original paper [4] that calibrated the trunk to leg volume ratio variable relied upon air displacement plethysmography data. Unfortunately, no air displacement plethysmography data are available for the EarlyBird study as this method of body measurement was never a part of the study. This means that we cannot calibrate a similar formula to that calculated in [4], which would have potentially been more applicable to children. Therefore in order to investigate the trunk to leg volume ratio using the EarlyBird data, we must use the same calibrations from the Wilson et. al paper [4].

7.2 Exploration

The trunk to leg volume ratio variable is highly correlated with the trunk to leg fat mass variable for both genders and all age groups, as seen in figure 7.1. Given the limitation issues discussed in section 7.1.1, this begs the question of why we should go through the trouble of using a transformation of the data, which was calibrated using only a small subset of volunteers from the population, to analyse our data?

However as Wilson et al. note, trunk to leg volume ratio does appear to be independent of other DEXA related variables as they are not highly correlated; see figure 2.3. This would appear to

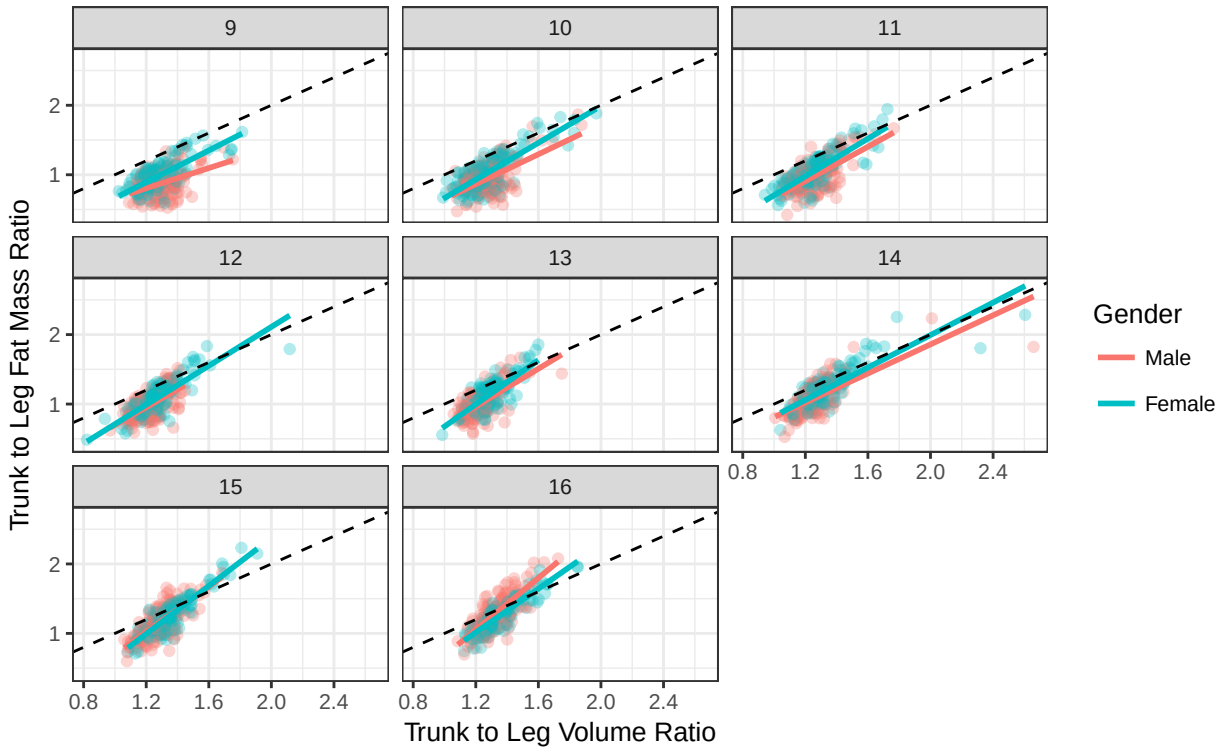

Figure 7.1: Correlation between trunk to leg volume ratio and trunk to leg fat mass ratio

satisfy part of hypothesis 2, which states that trunk to leg volume ratio is independent of other total and regional fat distributions, but we need to test whether trunk to leg volume ratio is predictive of impaired fasting glucose.

7.3 Modelling

When modelling the regional fat models using the BIC grid search method to identify the optimal penalisation parameter, λ, the parameter `age` was the only variable left in three of the four models. The same occurs for the trunk to leg volume ratio models, however again similar to the regional fat models, the cross validation models are not over penalised and do give more interesting results. These models return the same parameters, though with slightly different coefficients and can be seen in tables A.9 and A.10 in Appendix A.5. Both model searches leave age, the interaction between age and trunk to leg volume ratio and the interaction between age and arm lean mass percentage as variables in the model where the latter is the only parameter with a statistically significant p-value, though this is only weak evidence ($p = 0.092$). No issues were detected in the residual plot except for the outliers detected in other models (figure A.4). Therefore given the lack of statistically significant parameters, and the removal of the main effects from the model, it would appear that trunk to leg volume ratio is not overly indicative of impaired fasting glucose in children.

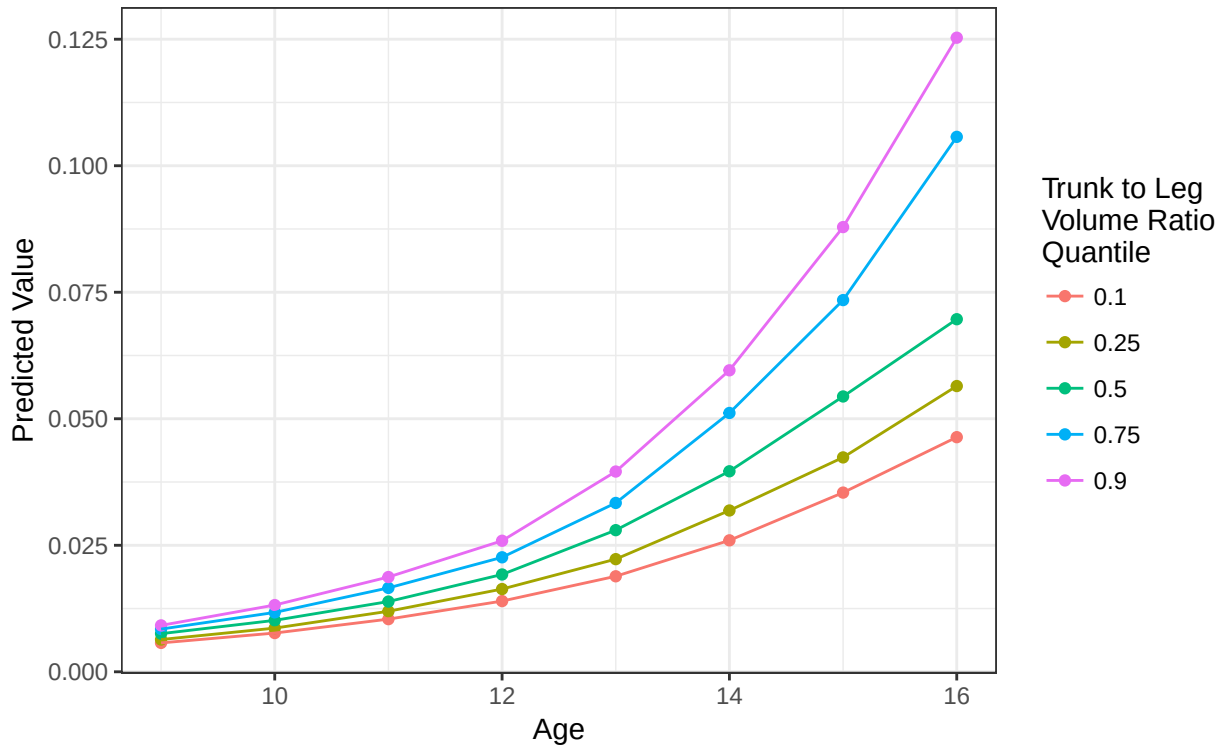

Figure 7.2: Predicted probabilities of pre-diabetes in children for varying ages at varying quantiles of TLVR

7.4 Prediction

Interestingly this model gives a different prediction to that of the Total Mass Index model as the model suggests that as children age, their odds of pre-diabetes increase and are more extreme for those children who are larger. This does make sense and supports findings in previous studies - that overall size and overweight are indicative of diabetes and pre-diabetes.

7.5 Summary

Trunk to leg volume ratio is a novel measure designed by Wilson et. al. This measure was calibrated using a small sample of American adults and so the formula defined to calculate it is probably not applicable to English children. The regional models calculated in this section have shown that trunk to leg volume ratio is not indicative of impaired fasting glucose and therefore does not satisfy Hypothesis 2. However further testing is needed to see if a similar formula can be calibrated using a more representative and larger sample. The probability of developing pre-diabetes increases as children age and is increased further for those children with a larger trunk to leg volume ratio.

8 Sensitivity Analysis

8.1 Introduction

Given we have imputed a large amount of the data, we should really test whether imputing more data will impact the results of our model fits. This will give us an idea of how robust our imputation methods are. The total mass indices models were therefore refit for children with up to 3 missing time points for any given variable, whereas in the previous sections only children with up to 2 missing time points were considered.

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-42.454	0.226	-187.616	< 0.001
age	2.480	0.664	3.734	< 0.001
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
lmi_total	2.606	0.690	3.778	< 0.001
fmi_total	0.000			
age:sexFemale	0.000			
age:lmi_total	-0.163	0.047	-3.449	< 0.001
sexFemale:lmi_total	0.000			
age:fmi_total	0.000			
sexFemale:fmi_total	0.000			

Table 8.1: Fixed effect model results from the total mass index model with mean data imputation and BIC grid search

8.2 The Results

If we compare the results shown in table 8.1 with those in table 5.1, we can see that there isn't a great deal of change with the parameter values. The same can be said for the model generated with mean data imputation and the BIC grid search, however the model generated with LOCF imputation and cross validation grid search contains both total physical activity and the interaction between age and total fat mass index. The main concern, however, is with the model that uses mean data imputation and cross validation grid search which leaves many additional, albeit non-significant, parameters in the model (table 8.2). These parameters have a mixture of parameter values, some are larger, for example age at peak height velocity is -0.067, but some are small for example total physical activity is -0.0000014, though this is expected due to the parameters not being scaled.

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-46.338	0.260	-178.516	< 0.001
age	2.967	0.756	3.925	< 0.001
sexFemale	0.000			
imd2004	0.011	0.013	0.793	0.4278
APHV	-0.067	0.221	-0.303	0.7616
TPA	-0.000	0.000	-1.118	0.2638
lmi_total	2.926	0.778	3.760	< 0.001
fmi_total	0.000			
age:sexFemale	-0.052	0.047	-1.106	0.2687
age:lmi_total	-0.192	0.053	-3.603	< 0.001
sexFemale:lmi_total	0.000			
age:fmi_total	-0.003	0.005	-0.687	0.4923
sexFemale:fmi_total	0.000			

Table 8.2: Fixed effect model results from the total mass index model with mean data imputation and CV grid search

8.3 Summary

To conclude, our model seems to be relatively robust to the data imputation methods we have used, though there is some concern over the cross validation grid search method which has kept additional parameters in the model. However for the LOCF data imputation method, these were insignificant both statistically and how they would affect the predictions.

9 Conclusions

9.1 Final Summary

The main aims of this study were to test the following hypotheses

Hypothesis 1 *DEXA related weight variables are predictive of impaired fasting glucose.*

Hypothesis 2 *Trunk to leg volume ratio is independent of total and regional fat distributions and predictive of impaired fasting glucose.*

Hypothesis 3 *Growth and development of children as well as the outcome of impaired fasting glucose in adolescence are not mutually exclusive issues. Body composition changes over time and so too does the risk of impaired fasting glucose.*

For the first hypothesis it would appear that we can conclude that there is evidence to suggest that DEXA related variables are predictive of impaired fasting glucose. It seems though that this is best described with an overall measure of size. We saw in section 2.3.5 that the variables reduced to two main components for each year group tested; an overall measure of size and a comparison between fat mass and lean mass. This outcome was replicated when modelling pre-diabetes as in section 5, the total lean mass index is overwhelmingly significant. In the regional models described in sections 6 the models were much less clear. For the models that didn't shrink all coefficients to zero, the arm, leg and trunk mass percentage variables remained in the model but only the trunk showed some statistical evidence of being significant.

From the results in section 7.3, it is evident that we can reject Hypothesis 2 and state that trunk to leg volume ratio is not predictive of pre-diabetes. However as has been discussed, this should not be the end of the study into this novel measure as we were unable to calibrate a similar formula to that described by Wilson et. al [4] since we do not have the air displacement plethysmography data required to do so. It would appear as though the trunk to leg volume ratio measure is in fact independent of other DEXA related measures since it is so uncorrelated with them (section 2.3.3).

The prevalence of impaired fasting glucose would also appear to be dependent on age suggesting that we cannot reject Hypothesis 3. Age was overwhelmingly significant in the model presented in section 5 and the number of children with impaired fasting glucose was seen to increase over time (section 2.2). The interaction between age and total lean mass index was also overwhelmingly significant in the section 5 model, but also complex, meaning we can conclude that prevalence of impaired fasting glucose and body composition changes over time are not mutually exclusive events. Similar results were seen in the regional body models, though not for all models. In addition, the interaction between age and mass percentage variables was not always significant.

9.2 Further Considerations

The analysis and conclusions presented in this report are by no means exhaustive and there are a number of other hypotheses we could test, and statistical methods we could try given the richness of this data set; particularly if the results presented here were to go to publication. The models produced in this report provide an interesting insight into impaired fasting glucose during adolescence but do not paint a full picture. Children's bodies go through large changes as they grow and the models in this report would appear to suggest that these changes impact on levels of glucose, and therefore the risk of pre-diabetes. However children grow at individual rates and their levels of glucose equally change over time at different rates; therefore it would be useful to account for these individual effects by way of a random slope effect - yet the computing power needed for the calculations to do this was not available at the time of writing. Therefore it would be useful in further work to do two things. The first would be to review the code used in the **glmmLasso** package and optimise the code where possible. Secondly, if the code is fully optimised and still the computer runs out of memory, access to a more powerful machine would be required. Another consideration would be to model glucose which is a continuous variable, rather than a cut off for impaired fasting glucose, since impaired fasting glucose is more interesting from a clinical point of view but continuous variables are easier to model.

Pros and cons were given for various data imputation methods but only two were applied in the modelling process. Of course, the data imputation methods discussed were relatively simple and more complex methods such as generalised estimating equations (GEE) could be applied to the data. Furthermore, no work was undertaken to test for completely random dropouts, that is that the probability a child drops out at time t_j is independent of the observed sequence of measurements on that child at times $t_1, ..., t_{j-1}$. This is then viewed as a screening device to avoid any parametric assumptions about the process which imputes the data [20].

Model residual checks showed several outlying values which would need to be dealt with in a more sophisticated manner were the results to be published. Some simple approaches could be to impute new values or we could just remove them all together.

Producing confidence intervals for the parameter estimates in a frequentist framework is not simple as one must use bootstrap techniques to acquire them. Confidence intervals around parameter estimates can provide insight into how plausible it is for these parameters to be equal to 0. A better approach to this problem, and possibly to this analysis would be to use Bayesian methods. Bayesian approaches to GLMM inference offers several advantages over frequentist and information-theoretic methods [28]. MCMC naturally provides confidence intervals on the parameter estimates which naturally averages over the uncertainty in both the fixed and random effect parameters; thus avoiding approximations made using frequentist hypothesis testing [29].

References

1 Wilson J, Kanaya A, Fan B *et al.* Ratio of trunk to leg volume as a new body shape metric for diabetes and mortality. 2013;**8**:e68716. doi:10.1371/journal.pone.0068716

2 Gómez-Ambrosi J, Silva C, Galofré JC *et al.* Body adiposity and type 2 diabetes: Increased risk with a high body fat percentage even having a normal bmi. *Obesity (Silver Spring)* 2011;**19**:1439–44. doi:10.1038/oby.2011.36

3 Weiss R, Caprio S. The metabolic consequences of childhood obesity. *Best Pract Res Clin Endocrinol Metab* 2005;**19**:405–19. doi:10.1016/j.beem.2005.04.009

4 Wilson J, Fan B, Shepherd J. Total and regional body volumes derived from dual-energy x-ray absorptiometry output. 2013;**16**:368–73. doi:10.1016/j.jocd.2012.11.001

5 *NHANES - about the national health and nutrition examination survey.* http://www.cdc.gov/nchs/nhanes/about_nhanes.htm 2016.

6 Groll A, Tutz G. Variable selection for generalized linear mixed models by L_1-penalized estimation. *Statistics and Computing* 2012;**24**:137–54. doi:10.1007/s11222-012-9359-z

7 Schelldorfer J, Meier L, Buhlmann P. GLMMLasso: An algorithm for high-dimensional generalized linear mixed models using ℓ_1-penalization. *Journal of Computational and Graphical Statistics* 2014;**23**:460–77. doi:10.1080/10618600.2013.773239

8 Goran M, Treuth M. Energy expenditure, physical activity, and obesity in children. *Pediatr Clin North Am* 2001;**48**:931–53. doi:10.1016/S0031-3955(05)70349-7

9 *Facts and stats.* https://www.diabetes.org.uk/Documents/Position%20statements/Diabetes%20UK%20Facts%20and%20Stats_Dec%202015.pdf 2016.

10 *National paediatric diabetes audit report 2013-14 part 1: Care processes and outcomes.* http://www.rcpch.ac.uk/system/files/protected/page/Revised%20Sept%202014%20NPDA%20Report%201%20FINAL.pdf 2016.

11 *How a dexa scan is performed.* http://www.nhs.uk/Conditions/DEXA-scan/Pages/How-is-it-performed.aspx 2014.

12 El Maghraoui A, Roux C. DXA scanning in clinical practice. 2008;**101**:605–17. doi:10.1093/qjmed/hcn022

13 Wilson K. Practical considerations when replacing a dxa system. http://www.hologic.com/sites/default/files/white-papers/WP-00054_DXA%20Migration_WhitePaper_10-11.pdf 2011.

14 Shepherd J, Ying L, Wilson K *et al.* Cross-calibration and minimum precision standards for dual-energy x-ray absorptiometry: The 2005 iscd official positions. *Journal of Clinical Densitometry* 2006;**9**:31–6. doi:10.1016/j.jocd.2006.05.005

15 Hangartner T. A study of the long-term precision of dual-energy x-ray absorptiometry bone densitometers and implications for the validity of the least-significant-change calculation. *Osteoporos Int* 2007;**18**:513–23. doi:10.1007/s00198-006-0280-1

16 Frost S, Nguyen N, Center J *et al.* Discordance of longitudinal changes in bone density between

densitometers. *Bone* 2007;**41**:690–7. doi:10.1016/j.bone.2007.07.002

17 Ho-Pham LT, Nguyen UDT, Nguyen TV. Association between lean mass, fat mass, and bone mineral density: A meta-analysis. 2014;**99**:30–8. doi:10.1210/jc.2013-3190

18 Wells JCK, Cole TJ, team A study. Adjustment of fat-free mass and fat mass for height in children aged 8 y. *International Journal of Obesity* 2002;**26**:947–52. doi:10.1038/sj.ijo.0802027

19 Vangipurapu J, Stancáková A, Jauhiainen R *et al.* Short adult stature predicts impaired β-cell function, insulin resistance, glycemia, and type 2 diabetes in finnish men. *The Journal of Clinical Endocrinology & Metabolism* 2017;**102**:443–50. doi:10.1210/jc.2016-2933

20 Diggle PJ, Heagerty P, Liang K-Y *et al. Analysis of longitudinal data*. 2nd ed. Great Clarendon Street, Oxford, OX2 6DP, United Kingdom:: Oxford University Press 2012.

21 Siddique J, Brown C, Hedeker D *et al.* Missing data in longitudinal trials. part b. analytic issues. *Psychiatr Ann* 2008;**38**:793–801.

22 Molnar F, Hutton B, Fergusson D. Does analysis using 'last observation carried forward' introduce bias in dementia research? *CMAJ* 2008;**179**:751–3.

23 Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 1994;**58**:267–88. doi:10.2307/41262671

24 Groll A. *GlmmLasso: Variable selection for generalized linear mixed models by l1-penalized estimation*. 2016. https://CRAN.R-project.org/package=glmmLasso

25 Box GEP. Robustness in the strategy of scientific model building. *Robustness in Statistics, Academic Press* 1979;201–36.

26 Wickham H. *Advanced R (Chapman & Hall/CRC The R Series)*. 1st ed. Paperback; Chapman; Hall/CRC 2014.

27 Group USC. Introduction to generalized linear mixed models - idre stats.

28 Gelman A, Hill J. *Data analysis using regression and multilevel/hierarchical models*. 1st ed. Cambridge University Press 2007.

29 Bolker BM, Brooks ME, Clark CJ *et al.* Generalized linear mixed models: A practical guide for ecology and evolution. 2009;**24**:127–35.

30 Ushey K, McPherson J, Cheng J *et al. Packrat: A dependency management system for projects and their r package dependencies*. 2016. https://CRAN.R-project.org/package=packrat

A Appendices

A.1 Screeplots for Princinpal Component Analyses

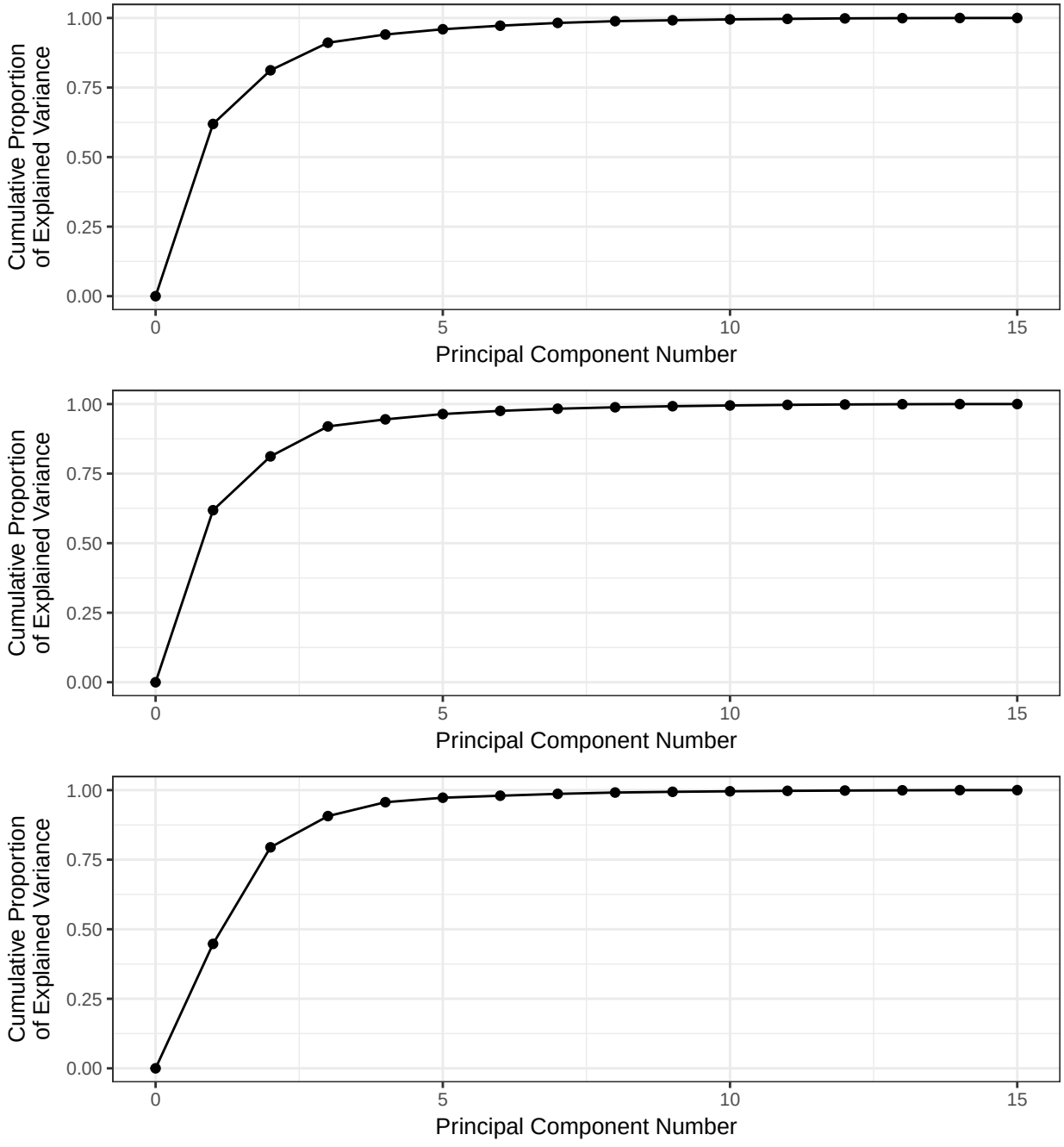

Figure A.1: Screeplots for the princinpal components analysis on all children for ages 9 (top), 12 (middle) and 16 (bottom)

A.2 Fat Mass Index Only Model

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-6.668	0.230	-28.967	< 0.001
age	0.315	0.086	3.666	< 0.001
sexFemale	-2.792	2.085	-1.339	$= 0.1806$
imd2004	0.000			
APHV	0.000			
TPA	-0.000	0.000	-0.826	$= 0.409$
fmi_total	0.000			
age:sexFemale	0.146	0.148	0.984	$= 0.3251$
age:fmi_total	-0.003	0.006	-0.431	$= 0.6667$
sexFemale:fmi_total	-0.000	0.126	-0.000	$= 0.9999$

Table A.1: Model results from the fat mass index only model with LOCF data imputation and CV grid search

A.3 Regional Fat Percentage Models

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.262	0.211	-39.237	< 0.001
age	0.363	0.065	5.562	< 0.001
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
arms_fat_perc	0.000			
legs_fat_perc	0.000			
trunk_fat_perc	0.000			
age:sexFemale	0.000			
age:arms_fat_perc	0.000			
sexFemale:arms_fat_perc	0.000			
age:legs_fat_perc	0.000			
sexFemale:legs_fat_perc	0.000			
age:trunk_fat_perc	0.000			
sexFemale:trunk_fat_perc	0.000			

Table A.2: Model results from the regional fat percentage model with mean data imputation and BIC grid search

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-6.076	0.300	-20.284	< 0.001
age	0.318	0.224	1.417	0.1565
sexFemale	-0.456	3.147	-0.145	0.8848
imd2004	0.000			
APHV	0.000			
TPA	-0.000	0.000	-1.155	0.2479
arms_fat_perc	-0.583	0.437	-1.335	0.182
legs_fat_perc	-0.141	0.369	-0.381	0.7033
trunk_fat_perc	0.702	0.345	2.036	0.0418
age:sexFemale	0.155	0.206	0.753	0.4512
age:arms_fat_perc	0.043	0.032	1.351	0.1766
sexFemale:arms_fat_perc	-0.227	0.152	-1.491	0.1361
age:legs_fat_perc	0.007	0.027	0.272	0.7854
sexFemale:legs_fat_perc	-0.129	0.145	-0.895	0.3706
age:trunk_fat_perc	-0.050	0.025	-2.013	0.0441
sexFemale:trunk_fat_perc	0.259	0.131	1.979	0.0478

Table A.3: Model results from the regional fat percentage model with mean data imputation and CV grid search

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-4.843	0.286	-16.913	< 0.001
age	0.205	0.205	0.999	0.3176
sexFemale	-0.462	3.134	-0.147	0.8829
imd2004	0.000			
APHV	0.000			
TPA	-0.000	0.000	-0.975	0.3297
arms_fat_perc	-0.019	0.101	-0.186	0.8521
legs_fat_perc	-0.438	0.303	-1.446	0.1482
trunk_fat_perc	0.486	0.287	1.690	0.091
age:sexFemale	0.186	0.204	0.914	0.3609
age:arms_fat_perc	0.000			
sexFemale:arms_fat_perc	-0.171	0.147	-1.162	0.2454
age:legs_fat_perc	0.030	0.021	1.431	0.1525
sexFemale:legs_fat_perc	-0.168	0.141	-1.192	0.2334
age:trunk_fat_perc	-0.033	0.021	-1.623	0.1046
sexFemale:trunk_fat_perc	0.240	0.128	1.874	0.0609

Table A.4: Model results from the regional fat percentage model with LOCF data imputation and CV grid search

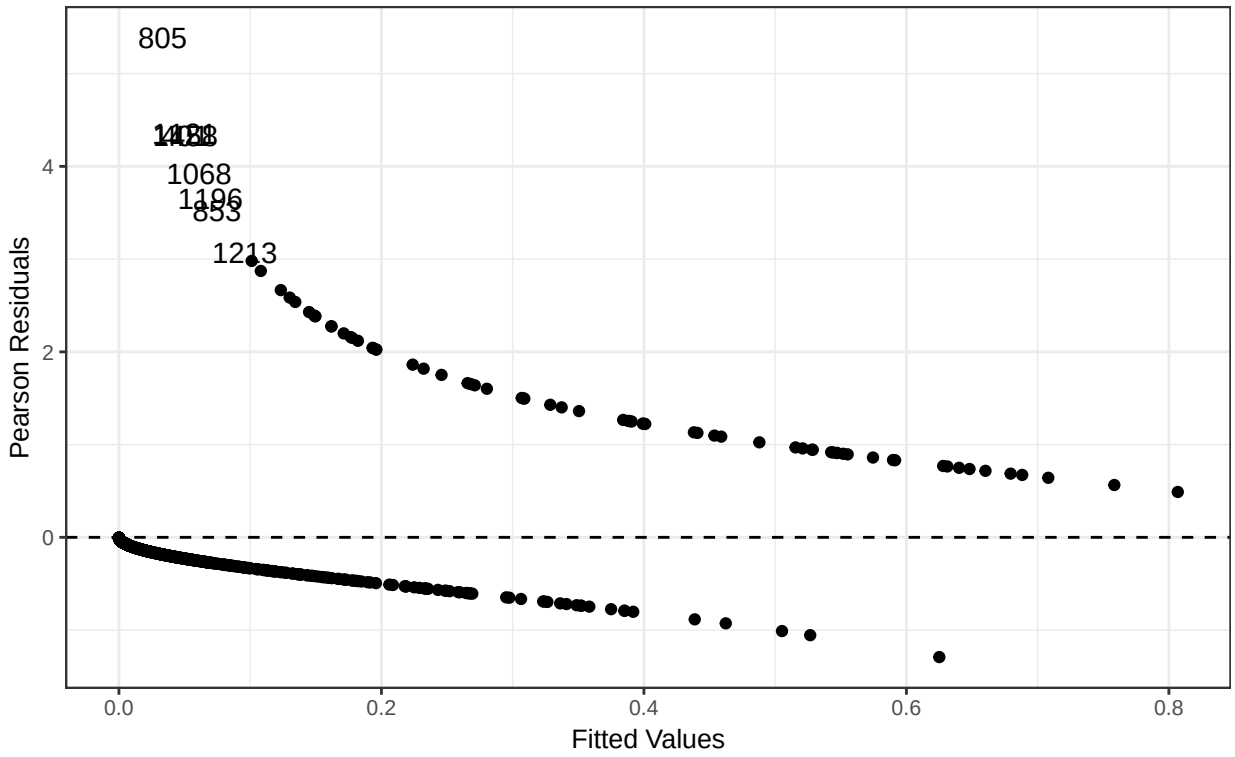

Figure A.2: Residuals vs. fitted values for the Fat Mass Percentage model with mean data imputation and CV algorithm

A.4 Regional Lean Percentage Models

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.166	0.213	-38.427	< 0.001
age	0.217	0.113	1.926	0.0541
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
arms_lean_perc	0.000			
legs_lean_perc	0.000			
trunk_lean_perc	0.000			
age:sexFemale	0.000			
age:arms_lean_perc	0.001	0.004	0.164	0.8696
sexFemale:arms_lean_perc	0.000			
age:legs_lean_perc	0.001	0.004	0.292	0.7699
sexFemale:legs_lean_perc	0.000			
age:trunk_lean_perc	0.000			
sexFemale:trunk_lean_perc	0.000			

Table A.5: Model results from the regional lean percentage model with LOCF data imputation and BIC grid search

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-7.381	0.214	-34.490	< 0.001
age	0.150	0.122	1.231	0.2182
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
arms_lean_perc	0.000			
legs_lean_perc	0.000			
trunk_lean_perc	0.000			
age:sexFemale	0.000			
age:arms_lean_perc	0.004	0.005	0.846	0.3976
sexFemale:arms_lean_perc	0.000			
age:legs_lean_perc	0.004	0.005	0.909	0.3634
sexFemale:legs_lean_perc	0.000			
age:trunk_lean_perc	-0.006	0.004	-1.586	0.1127
sexFemale:trunk_lean_perc	0.000			

Table A.6: Model results from the regional lean percentage model with LOCF data imputation and CV grid search

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.167	0.212	-38.436	< 0.001
age	0.217	0.113	1.930	0.0537
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
arms_lean_perc	0.000			
legs_lean_perc	0.000			
trunk_lean_perc	0.000			
age:sexFemale	0.000			
age:arms_lean_perc	0.001	0.004	0.188	0.8506
sexFemale:arms_lean_perc	0.000			
age:legs_lean_perc	0.001	0.004	0.265	0.7907
sexFemale:legs_lean_perc	0.000			
age:trunk_lean_perc	0.000			
sexFemale:trunk_lean_perc	0.000			

Table A.7: Model results from the regional lean percentage model with mean data imputation and BIC grid search

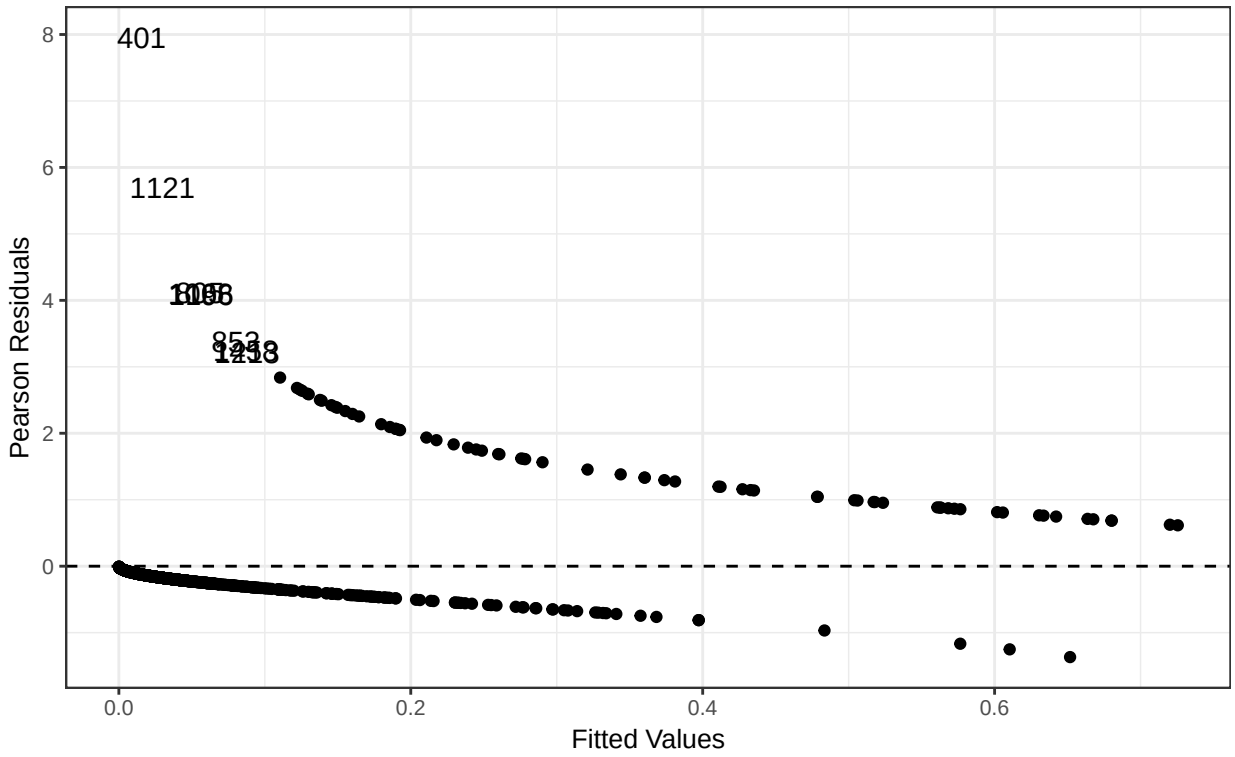

Figure A.3: Residuals vs. fitted values for the Lean Mass Percentage model with mean data imputation and CV algorithm

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-13.860	0.240	-57.710	< 0.001
age	0.590	0.555	1.065	0.2871
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	-0.000	0.000	-1.024	0.3059
arms_lean_perc	0.648	0.417	1.553	0.1204
legs_lean_perc	0.119	0.360	0.330	0.7411
trunk_lean_perc	-0.660	0.326	-2.027	0.0427
age:sexFemale	0.000			
age:arms_lean_perc	-0.041	0.030	-1.384	0.1663
sexFemale:arms_lean_perc	0.000			
age:legs_lean_perc	-0.003	0.026	-0.135	0.8927
sexFemale:legs_lean_perc	0.000			
age:trunk_lean_perc	0.040	0.023	1.721	0.0852
sexFemale:trunk_lean_perc	0.000			

Table A.8: Model results from the regional lean percentage model with mean data imputation and CV grid search

A.5 Trunk to Leg Volume Models

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.087	0.209	-38.634	< 0.001
age	0.201	0.175	1.148	0.2508
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
trunk_leg_vol	0.000			
arms_fat_perc	0.000			
arms_lean_perc	0.000			
age:sexFemale	0.000			
age:trunk_leg_vol	0.005	0.089	0.059	0.9531
sexFemale:trunk_leg_vol	0.000			
age:arms_fat_perc	0.000			
age:arms_lean_perc	0.002	0.001	1.673	0.0944
sexFemale:arms_fat_perc	0.000			
sexFemale:arms_lean_perc	0.000			

Table A.9: Model results from the trunk to leg volume model with LOCF data imputation and CV grid search

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.063	0.209	-38.546	< 0.001
age	0.185	0.175	1.058	0.2902
sexFemale	0.000			
imd2004	0.000			
APHV	0.000			
TPA	0.000			
trunk_leg_vol	0.000			
arms_fat_perc	0.000			
arms_lean_perc	0.000			
age:sexFemale	0.000			
age:trunk_leg_vol	0.016	0.088	0.180	0.8571
sexFemale:trunk_leg_vol	0.000			
age:arms_fat_perc	0.000			
age:arms_lean_perc	0.002	0.001	1.684	0.0923
sexFemale:arms_fat_perc	0.000			
sexFemale:arms_lean_perc	0.000			

Table A.10: Model results from the trunk to leg volume model with mean data imputation and CV grid search

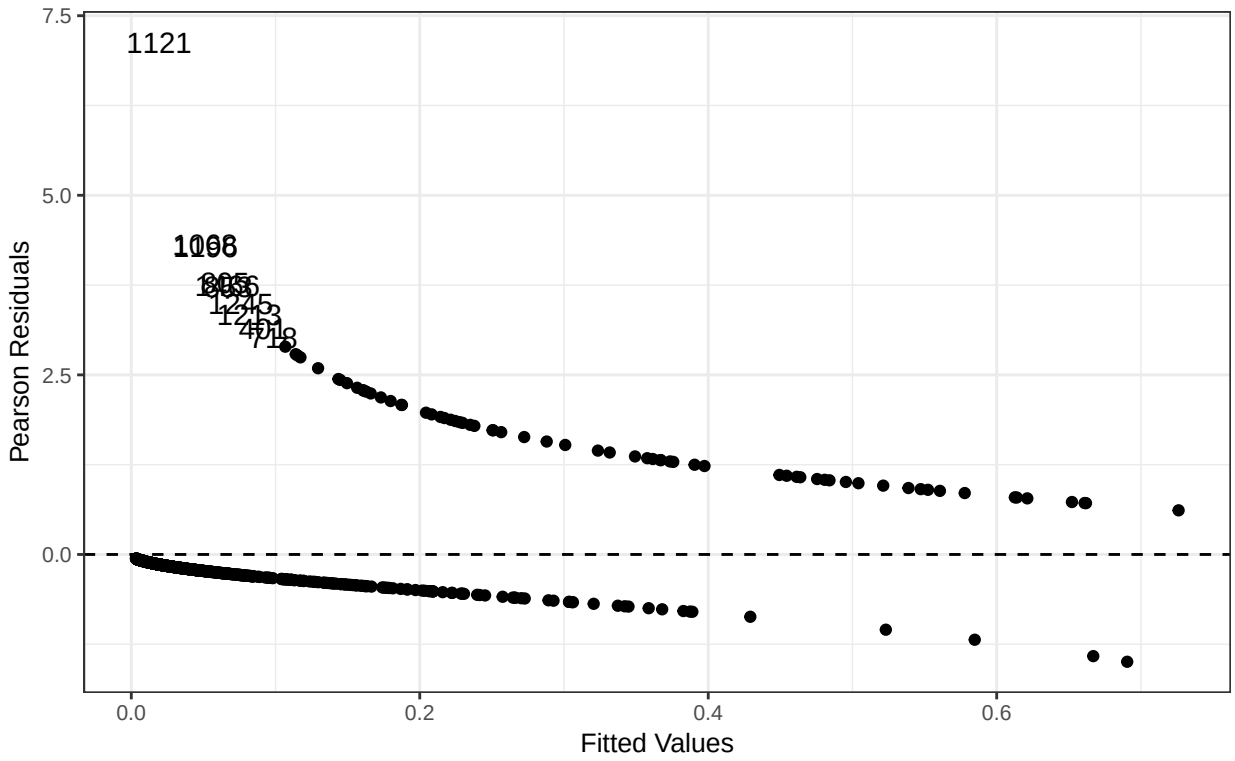

Figure A.4: Residuals vs. fitted values for the Trunk to Leg Volume model with mean data imputation and CV algorithm

A.6 Reproducibility

This report has been produced using R version 3.4.1 (2017-06-30) on a macOS Sierra 10.12.6 operating system (x86_64-apple-darwin16.6.0 (64-bit) platform). All reports, analyses, data and functionality used or created for this report are held within a self-contained R package. The R dependency management was handled by `packrat` [30] which recorded the exact package versions the EarlyBird package depends on; meaning that installing the package on different computers - even across platforms - ensures the same package versions are installed and therefore the results and conclusions made in this report should be reproducible. The packages used to generate this report were: **broom**: 0.4.1, **corrr**: 0.2.0.9000, **dplyr**: 0.5.0, **GGally**: 1.2.9.9999, **ggfortify**: 0.4.1, **gghalfnorm**: 1.1.2, **ggplot2**: 2.2.1.9000, **ggrepel**: 0.6.5, **gridExtra**: 2.2.1, **mice**: 2.25, **moments**: 0.14, **purrr**: 0.2.2, **Rcpp**: 0.12.8, **reshape2**: 1.4.2, **tidyr**: 0.6.0, **xtable**: 1.8-2, **zoo**: 1.7-14.

Exploring predictors of impaired fasting glucose during adolescence - a longitudinal study

Nathan Eastwood

@nathaneastwood_

02/08/2017

School of Mathematics and Statistics, University of Sheffield, Sheffield, U.K.

Overview

Aim

- To see whether the novel measure Trunk to Leg Volume Ratio is predictive of pre-diabetes in children

Motivations

- The prevalence of childhood overweight and obesity are rising and have multiple causes; particularly under-activity and over-nutrition are both believed to contribute
- Obesity is of concern since it is thought to cause the insulin resistance that underlies diabetes and cardiovascular disease

Hypotheses

Trunk to leg volume ratio is...

- *independent of total and regional fat distributions*
- *predictive of impaired fasting glucose in children*

The EarlyBird Study and DEXA data

EarlyBird

- EarlyBird was a unique 12-year non-intervention prospective cohort study observing the health and lifestyle of 347 normal healthy children
- The children were recruited at age 5y and followed up annually until they were 16y
- Data collected include demographics; various anthropometric measurements; and whole body dual energy X-ray absorptiometry (DEXA) scans

DEXA

- Simple measures of adiposity such as BMI are limited due to their inability to distinguish fat from lean mass
- DEXA scans can distinguish between fat mass, lean mass, bone mineral content (BMC) and bone mineral density (BMD)
- This makes DEXA scans the gold standard for body composition assessment

Trunk to Leg Volume Ratio

Development of the Measure

- Total body volume is an important health metric used to measure body density, shape and composition
- However it is only available through underwater weighing or air displacement plethysmography (ADP)
- Wilson et. al [1] derived a novel measure, Trunk to Leg Volume Ratio (TLVR), from DEXA reported measures
- They tested this measure and showed it to be independent of other weight related variables and predictive of diabetes [2]
- This measure has not been tested in children

DEXA Based Formula

$$\text{TLVR} = \frac{\text{Fat Mass (kg)}}{0.88} + \frac{\text{Lean Mass (kg)}}{1.05} + \frac{\text{BMC}}{4.85} + 0.01$$

Independence

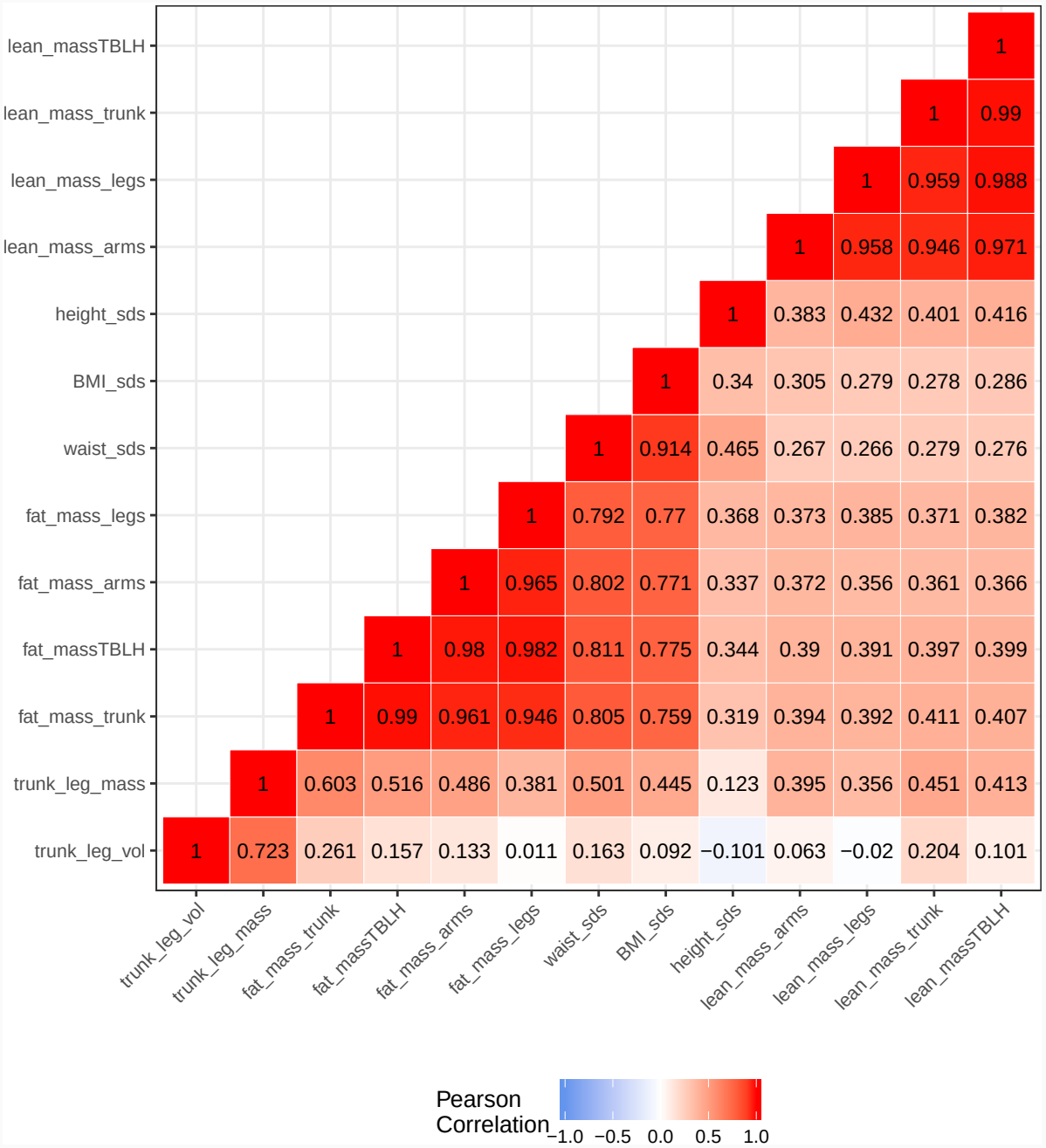

Figure 1: Trunk to Leg Volume Ratio Appears to be Independent of Other Weight Related Variables

Comparison with Trunk to Leg Fat Mass Ratio

Exploration

- Figure 1 shows that TLVR is indeed independent of other anthropometric measures
- However, Figure 2 shows that the TLVR variable is highly correlated with the trunk to leg fat mass variable - so is there a need for a more complicated measure?

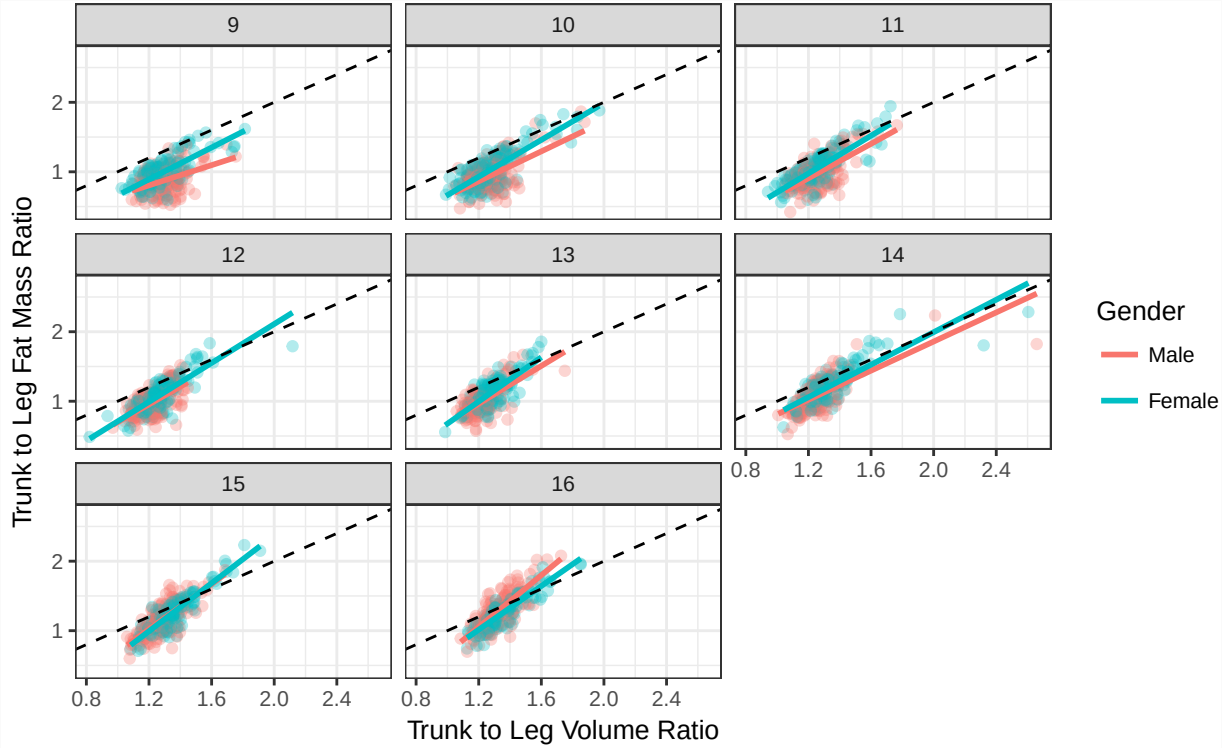

Figure 2: Correlation between TLVR and Trunk to Leg Fat Mass Ratio for each Age

Modelling

Modelling Techniques

- Pre-diabetes (a binary variable) is modelled using a generalised logistic mixed effect model (GLMM)
- Mixed effects are used to allow for within-subject correlation as the effects of individual children are not of interest
- The Lasso is used as a variable selection and shrinkage tool to find variables which are predictive of pre-diabetes
- The **glmmLasso** package is used for the modelling

Limitations of the glmmLasso package

- **glmmLasso** is computationally intensive
- Assuming each child grows at a different rate and has their own trend, i.e. `age | ID`, or standardising variables causes R to crash
- **glmmLasso** will not allow the user to keep main effects when interaction effects are left in the model

Results

Parameter	Estimate	StdErr	z-value	p-value
(Intercept)	-8.063	0.209	-38.546	< 0.001
age	0.185	0.175	1.058	0.2902
sexFemale	-0.000			
imd2004	-0.000			
APHV	-0.000			
TPA	0.000			
trunk_leg_vol	0.000			
arms_fat_perc	0.000			
arms_lean_perc	-0.000			
age:sexFemale	-0.000			
age:trunk_leg_vol	0.016	0.088	0.180	0.8571
sexFemale:trunk_leg_vol	-0.000			
age:arms_fat_perc	0.000			
age:arms_lean_perc	0.002	0.001	1.684	0.0923
sexFemale:arms_fat_perc	-0.000			
sexFemale:arms_lean_perc	-0.000			

Table 1: Model Results from the Trunk to Leg Volume Model

Predictions

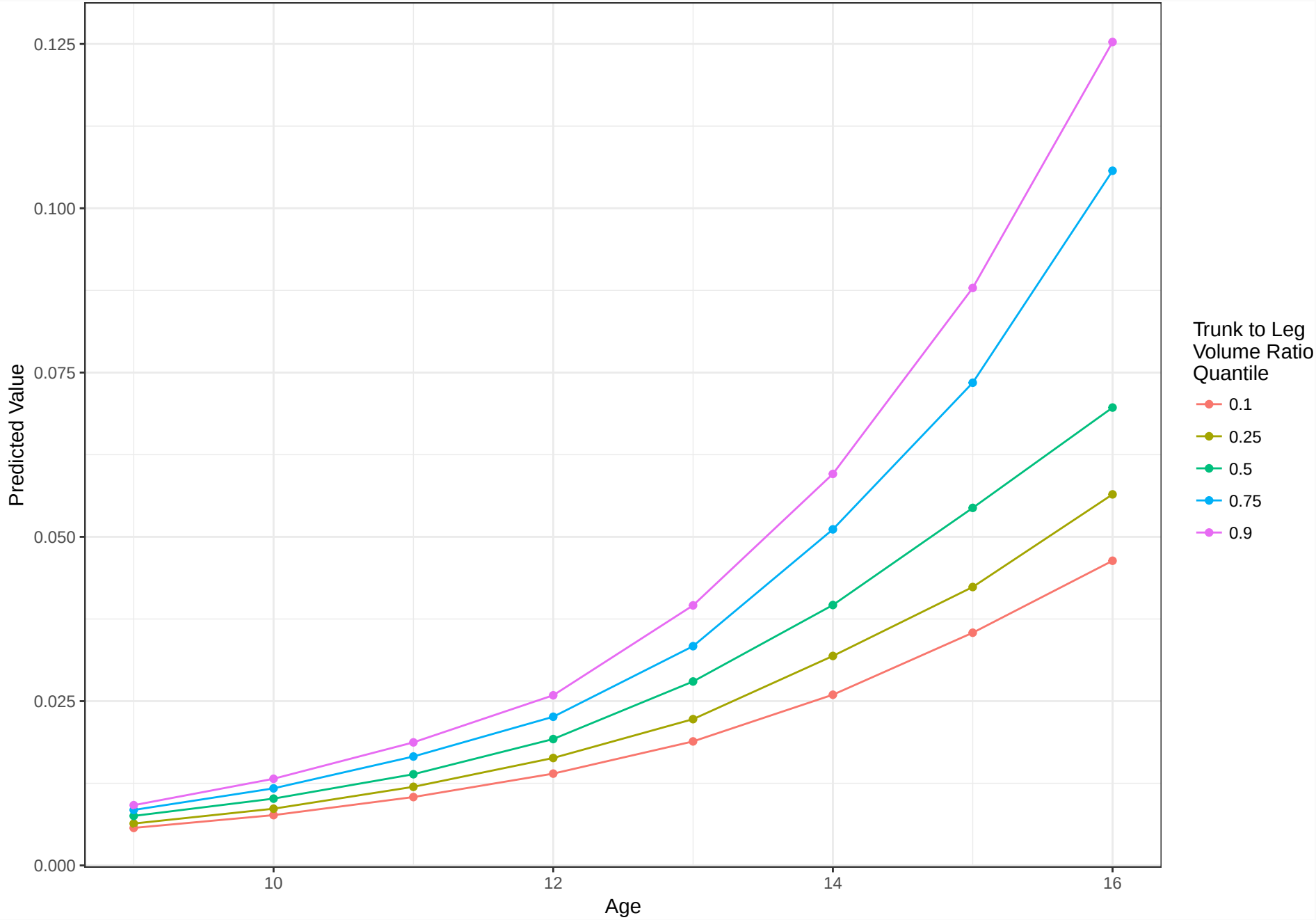

Figure 3: Predicted probabilities of pre-diabetes in children for varying ages at varying quantiles of TLVR

Limitations

Limitations of the Wilson et. al Study

- Wilson et. al used a small sample size ($n = 47$) for their measure calibration
- Their sample contained American adults only whereas EarlyBird contains British children
- Different DEXA scanners are known to produce subtly different results meaning the calibrated model is local to the data used in the study

Limitations of the EarlyBird Study

- The measure was calibrated against ADP data which EarlyBird does not have and so a TLVR formula for children cannot be calibrated

Conclusions

Model Results

- All but age; the interaction between age and TLVR; the interaction between age and arm lean mass percentage, are removed from the model
- The probability of developing pre-diabetes increases as children age and is increased further for those children with a larger trunk to leg volume ratio

Remarks

- TLVR is removed from the model and only appears as an interaction effect with age
- This suggests TLVR is not itself predictive of pre-diabetes in children

Future Work

- Further testing is needed to see if a similar formula can be calibrated using a more representative and larger sample

References

1 Wilson J, Kanaya A, Fan B *et al.* Ratio of trunk to leg volume as a new body shape metric for diabetes and mortality. 2013;**8**:e68716. doi:10.1371/journal.pone.0068716

2 Wilson J, Fan B, Shepherd J. Total and regional body volumes derived from dual-energy x-ray absorptiometry output. 2013;**16**:368–73. doi:10.1016/j.jocd.2012.11.001