Predict Bone Mineral Content and Density Using Multiple Regression Models

by Shuangyan Wu

May 20, 2023

Contents

List of Figures

List of Tables

Summary

Bone health, which is closely related to body bone mineral density (BMD) and bone mineral content (BMC), may be negatively affected by Type 2 diabetes. To assess whether pre-diabetes can significantly alter the bone health of children and adolescents at age 12-20, a study is conducted to investigate the relationship between pre-diabetes and bone health values including the total body BMD (TotalBMD), Total body BMC (TotalBMC), Spine BMD (SpineBMD) and Spine BMC (Spine BMC). Apart from the Pre-diabetes variable, other factors including Age, Race, Gender, Height, Weight, and Lean body mass index (LBMI) may also influence the bone health of adolescents. Therefore, these factors are also included for data analysis. In this report, a multiple linear regression model is fitted for each response variable. The predictor variables are selected based on the model performance (R^2 and $RMSE$). The collinearity problem (correlation between explanatory variables) is solved by checking the variance inflation factor (VIF) and reducing variables. Significant variables are found to be Pre-diabetes, Age, Race, Gender, Height, and LBMI for the bone health outcomes at $\alpha = 0.01$. Race, however, shows no significant effect on the SpineBMC. Weight and LBMI are found to be strongly correlated, so only LBMI is used for the models. The final models fit well with provided data ($R^2 : 0.61 - 0.85$), and the RMSE values are 0.08, 188.90, 0.10, and 8.2 for the TotalBMD, TotalBMC, SpineBMD, and SpineBMC models, respectively. The RMSE values are the smallest among all other corresponding sub-models with no collinearity problem suggesting good model fitting. Equations (1-4) show the final models for all four outcomes. As seen from the model, increasing Age, Height, and LBMI leads to higher predicted TotalBMD, TotalBMC, SpineBMD, and SpineBMC. Black adolescents at age 12-20 have higher predicted TotalBMD, TotalBMC, and SpineBMD. Male adolescents have lower predicted bone health values after separating all effects from Height, Weight, and LBMI. And adolescents with pre-diabetes conditions show lower predicted bone health values for all four outcomes. Along with the significance of the Pre-diabetes variable (P-values < 0.01), there is enough evidence to show that the change of bone health values (TotalBMD, TotalBMC, SpineBMD, and SpineBMC) significantly depends on the pre-diabetes condition for children at age 12-20 at 99% confidence level.

$$\begin{aligned} \widehat{TotalBMD} = &- 0.348 + 0.011 \times Age + 0.006 \times Height + 0.017 \times LBMI + 0.026 \times I_b \\ &- 0.018 \times I_m - 0.019 \times I_p \end{aligned} \tag{1}$$

$$\begin{aligned} \widehat{TotalBMC} = &- 4409.0 + 22.9 \times Age + 28.4 \times Height + 74.2 \times LBMI + 66.1 \times I_b \\ &- 82.6 \times I_m - 49.4 \times I_p \end{aligned} \tag{2}$$

$$\begin{aligned} \widehat{SpineBMD} = &- 0.604 + 0.019 \times Age + 0.006 \times Height + 0.024 \times LBMI + 0.026 \times I_b \\ &- 0.146 \times I_m - 0.020 \times I_p \end{aligned} \tag{3}$$

$$\widehat{SpineBMC} = -131.54 + 1.43 \times Age + 0.88 \times Height + 1.46 \times LBMI - 6.61 \times I_m - 2.65 \times I_p \tag{4}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has a pre-diabetes condition, respectively.

1 Introduction

Type 2 diabetes may negatively affect the bone health of children and adolescents at the age of peak bone mass. Bone health can be reflected by factors such as bone mineral density (BMD), bone mineral content (BMC), Spine bone mineral density (SpineBMD), and Spine bone mineral content (SpineBMC). A higher bone mineral density or bone mineral content normally indicates better bone health. To assess the relationship between pre-diabetes and bone health of children and adolescents at age 12-20, the effects of pre-diabetes condition and other variables (age, race, gender, height, weight, lean body mass index/LBMI) on BMD, BMC, SpineBMD, and SpineBMC of adolescents need to be investigated. Here, the data set is obtained from the national representative National Health and Nutrition Examination Survey 2005, which includes related variables and observations. An overall data summary and data analysis using the multiple linear regression method are presented in this report.

2 Exploratory data analysis

2.1 Data summary

The provided data includes 18 variables (4 categorical and 14 continuous) and 6178 observations. Table 1 shows the notations, descriptions, and types for these variables. The age range of the provided data is 12-85 years old. However, in this study, the researchers are particularly interested in the effects of pre-diabetes condition on the bone health of adolescents at age 12-20. Therefore, the observations at ages 12-20 (2075 observations) are selected. The 4 response variables are TotalBMD, TotalBMC, SpineBMD, and SpineBMC for bone health. There are 9 potential explanatory variables including Age, Race, Gender, Height, Weight, LBMI, FastGluType, Glu2HourType, and HbA1cType. The last three are different assessments for pre-diabetes. The SEQN variable shows sample ID, therefore is not used for data analysis. The Age_mon variable is similar to the Age variable and it has missing data, so only the Age variable is used. BMI variable is not included since it is calculated based on Height and Weight. Similarly, LBMI is related to TotalFat and TotalLean, so only LBMI is included as an overall index. The data summary (statistics or frequency counts) for all interested variables for observations at ages 12-20 is shown in Table 2-3. No extreme input or typos were found in the dataset.

2.2 Data cleaning and visualization

Before visualizing the data, data cleaning is completed. As shown in Table 4, there are 0.4%, 0.5%, 3.4%, 52.4%, 60.7%, and 0.2% of the data missing for the explanatory variables: Height, Weight, LBMI, FastGluType, Glu2HourType, and HbA1cType variables, respectively. There are also small amounts (3.2%- 14.5%) of the data missing for the response variables (TotalBMD, TotalBMC, SpineBMD, and SpineBMD). However, it is not recommended to use explanatory variables to impute response variables and then use imputed data to fit models. Therefore, the missing data in these variables are not imputed. In addition, 0.1%-3.2% of data is missing for the Age_mon, BMI, TotalFat, and TotalLean variables, which are not the interested variables for model building, but these variables can be included for data imputation. More information about the numbers and distribution of missing data in all variables can be found in the Appendix (Figures 3-4).

Table 1: Information for all provided variables

Notation	Description	Type
SEQN	Sample ID	Continuous
Age	Chronological age in years	Continuous
Age_mon	Chronological age in months	Continuous
Gender	Male or female	Categorical
Race	Race/ethnicity: 1 black, 0 nonblack	Categorical
Height	Standing height (cm)	Continuous
Weight	Body weight (kg)	Continuous
BMI	Body mass index	Continuous
TotalFat	Whole body fat mass	Continuous
TotalLean	Whole body fat mass	Continuous
LBMI	Lean body mass index	Continuous
FastGluType	Fasting blood glucose	Categorical
Glu2HourType	Glucose 2 hours after OGTT	Categorical
HbA1cType	Glycated hemoglobin	Categorical
TotalBMD	Whole body (minus head) bone mineral density(BMD)	Continuous
TotalBMC	Whole body (minus head) bone mineral density(BMC)	Continuous
SpineBMD	Lumbar spine bone mineral density (BMD)	Continuous
SpineBMC	Lumbar spine bone mineral content (BMC)	Continuous

Table 2: Summary for interested continuous variables

Variable	Age	Height	Weight	LBMI	TotalBMD	TotalBMC	SpineBMD	SpineBMC
n	2075	2066	2065	2005	2008	2008	1774	1774
Mean	16	165.5	67.4	15.4	1.0	1771.7	0.96	54.2
Median	16	164.8	63.3	15.0	1.0	1707.4	0.96	54.2
Min	12	135.6	28.4	9.4	0.6	635.4	0.48	19.9
Max	20	200.1	215.3	35.7	1.6	5014.4	1.51	109.5
SD	2	10.3	20.4	2.9	0.1	483.8	0.16	14.7

As mentioned, there is a large amount of data missing for both FastGluType (52.4%) and Glu2HourType (60.7%) variables. And 0.2% data is missing for HbA1cType. However, each observation has data for at least one of the three assessments. So a new explanatory variable ”Pre-diabetes” is created by combining all three assessment results. The new variable also has 2 levels: Pre-diabetes and Healthy, it is determined as Healthy only if all non-missing assessments for the observation are Healthy. Otherwise, it is determined as Pre-diabetes. The frequency counts for this new variable are shown in Table 5. Then The missing values for potential explanatory variables (Height, Weight, and LBMI) were imputed using the predictive mean matching (*pmm*) function in the R MICE package, which imputes a missing data point by selecting the original non-missing data point which has the nearest predicted value to the predicted value of the missing one, using other continuous variables (not the response variables) as predictors in this case [1].

Data visualization: After exploring plots with different variables, some interesting trends are found between the continuous variables and TotalBMD values as shown in Figure 1 (left). With increasing age, height, weight, or LBMI, the TotalBMD of sampled children and adolescents shows a roughly linear increasing trend. Besides, each pair of the Height, Weight, and LBMI variables also show linear correlations to different contents. In particular, the Weight and LBMI variables are strongly correlated since their scatter plot shows high linearity. As expected, Age increase leads to higher TotalBMD, Weight, and Height. The right plot in Figure 1 shows the same increasing trend for

Table 3: Frequency count for categorical variables.

Race	Gender	FastGluType	Glu2HourType	HbA1cType
1372 (Nonblack)	1044(Female)	796 (Healthy)	782 (Healthy)	1985(Healthy)
703 (Black)	1031 (Male)	192 (Pre-diabetes)	34 (Pre-diabetes)	85 (Pre-diabetes)

Table 4: Missing data percentages.

Age	Race	Gender	Height	Weight	LBMI
0	0	0	0.4	0.5	3.4
FastGluType	Glu2HourType	HbA1cType	TotalBMD	TotalBMC	SpineBMD
52.4	60.7	0.2	3.2	3.2	14.5
SpineBMC	SEQN	Age_mon	BMI	TotalFat	TotalLean
14.5	0	0.1	0.5	3.2	3.2

TotalBMD with increasing height, but in the overlapped region, female children tend to have slightly higher TotalBMD than male children with the same Height. The boxplots between TotalBMD and the categorical variables (Race, Gender, Pre-diabetes) can be found in the Appendix (Figures 5-6). After replacing the response variable, similar boxplot trends can be found between the continuous variables and TotalBMC, SpineBMD, or SpineBMC (Appendix, Figures 7-9).

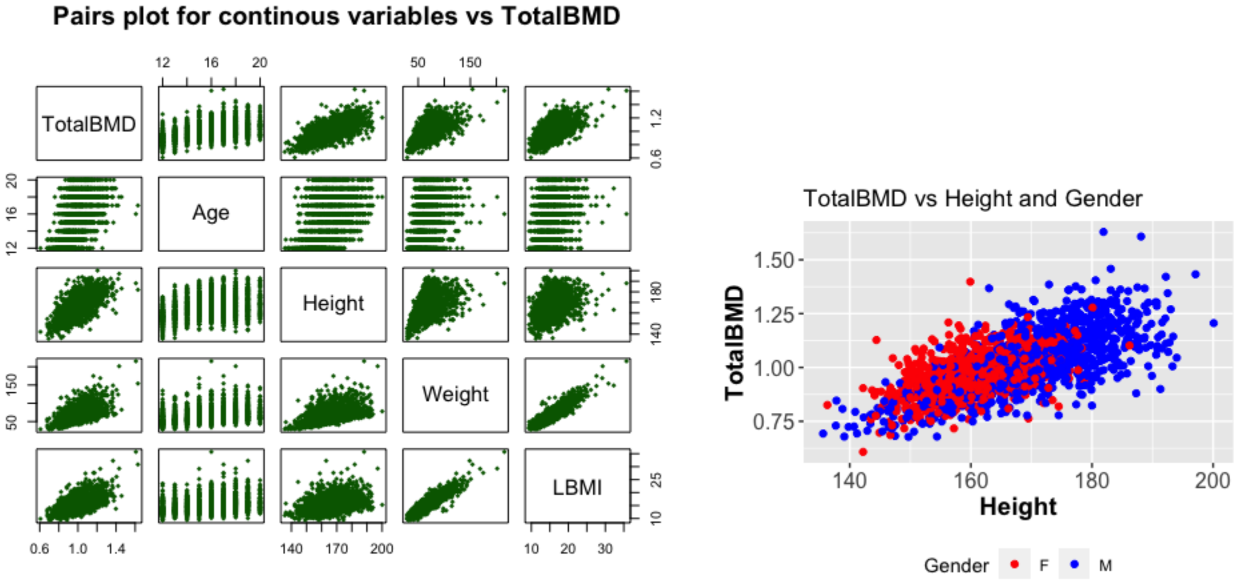

Figure 1: Matrix plots of the continuous variables with TotalBMD (left) and plot of TotalBMD with Height and Gender (right).

3 Statistical methods

To determine whether the pre-diabetes condition can significantly affect bone health (TotalBMD, TotalBMC, SpineBMD, and SpineBMD), all possible confounding variables (Age, Race, Gender, Height, Weight, LBMI) need to be considered for model building. In this analysis, multiple regression methods are used with both continuous and categorical variables. All 7 potential explanatory variables (Pre-diabetes, Age, Race, Gender, Height, Weight, LBMI) are fitted into the model gradually starting with the demographic variables for each response variable. Model performance is evaluated based on

Table 5: Frequency count for the created Pre-diabetes variable.

Healthy	Pre-diabetes
1793	282

R^2 and root mean square error (RMSE). Residual plots are generated to check the model assumptions. Possible collinearity is detected using the variance inflation factor, which measures the correlation and strength of correlation between explanatory variables in regression models [2-3]. R studio software is used to analyze the data and generate plots.

4 Results and discussion

4.1 Model for TotalBMD

4.1.1 Model building

In this section, a multiple regression model is built for TotalBMD by fitting the demographic variables (Age, Race, Gender) first and then adding one more variable each time. Although the Race variable is coded with 0 and 1, it is fitted as a factor. The general form of the full model is shown in equation (5). The hypotheses are listed below. The model fitting R^2, RMSE, and significance of the parameters are collected and compared for different models as shown in Table 6. As seen, the model fits better (larger R^2, smaller RMSE) with more explanatory variables. All parameters are significant at 95% confidence level, and the full model (5) has the highest $R^2 = 0.68$ and the lowest $RMSE = 0.07$, which indicates the best fitting. However, Table 7 shows that the LBMI and Weight variables in the full model are highly correlated with VIF values (8.8 and 9.2) greater than 5. This was also suggested by the matrix plot in section 2. In this case, the parameter estimates and p-values for the model are likely unreliable. So models (4) and (6) with either LBMI or Weight are fitted and compared to avoid collinearity problems. As seen, neither of the two models shows a strong correlation between explanatory variables since their VIF values are all smaller than 2. But model (4) has higher $R^2 = 0.66$ and lower $RMSE = 0.07$ than model (6). So, it is chosen as the final model for predicting TotalBMD. Table 8 shows the model summary and 95% confidence intervals (CIs) for the estimates. The negative parameter estimate (-0.019) for Pre-diabetes and p-value< 0.001 indicate that the pre-diabetes condition does lower the total body bone mineral density at a 99.9% confidence level. Its significance at $\alpha = 0.05$ is also confirmed by the negative range of the confidence interval. Equation (5) shows the final model for TotalBMD prediction ($R^2 = 0.66$, $RMSE = 0.07$). It can be interpreted as the predicted TotalBMD increases by 0.011 with a 1-year increase in age, by 0.006 with a 1cm increase in height, and by 0.017 with a 1 unit of LBMI increase. Compared to non-black adolescents, black adolescents have 0.026 higher predicted TotalBMD. Male adolescents have 0.018 lower predicted TotalBMD than female adolescents. And adolescents with pre-diabetes have 0.019 lower predicted TotalBMD than their counterparts.

$$\begin{aligned} TotalBMD =& \beta_0 + \beta_1 \times Age + \beta_2 \times Height + \beta_3 \times Weight + \beta_4 \times LBMI + \beta_5 \times I_b \\ & + \beta_6 \times I_m + \beta_7 \times I_p + \varepsilon \end{aligned} \tag{5}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has pre-diabetes condition, respectively, and ε follows N(0, σ_ε^2).

Hypotheses:

H_o: all parameters except the intercept β_0 are equal to 0; H_a: Not all of the non-intercept parameters are 0.

Table 6: Model comparison and selection.

Model parameters	significant	R^2	RMSE
(1) Age, Race, Gender	All	0.36	0.1017
(2) Age, Race, Gender, Height	All	0.54	0.0859
(3) Age, Race, Gender, Height, LBMI	All	0.65	0.0748
(4) Age, Race, Gender, Height, LBMI, Pre_diabetes	All	0.66	0.0746
(5) Age, Race, Gender, Height, LBMI, Pre_diabetes, Weight	All	0.68	0.0718
(6) Age, Race, Gender, Height, Weight, Pre_diabetes	All	0.60	0.0800

Table 7: VIF values for the parameters in three models.

Model/Variable	Age	Race	Gender	Height	LBMI	Pre-diabetes	Weight
(4)	1.3	1.0	1.4	1.6	1.4	1.0	-
(5)	1.3	1.1	2.2	2.4	8.8	1.0	9.2
(6)	1.3	1.0	1.4	1.9	-	1.0	1.4

Table 8: Model summary and 95%CI for estimates.

	Estimates	P-value	2.5%	97.5%
Intercept	-0.348	<0.001	-0.406	-0.291
Age	0.011	<0.001	0.009	0.012
Race(black)	0.026	<0.001	0.019	0.033
Gender(male)	-0.018	<0.001	-0.026	-0.010
Height	0.006	<0.001	0.005	0.006
LBMI	0.017	<0.001	0.016	0.018
Pre_diabetes (yes)	-0.019	<0.001	-0.029	-0.009

$$\begin{aligned} Total\hat{B}MD = &-0.348 + 0.011 \times Age + 0.006 \times Height + 0.017 \times LBMI + 0.026 \times I_b \\ &- 0.018 \times I_m - 0.019 \times I_p \end{aligned} \tag{6}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has a pre-diabetes condition, respectively.

4.1.2 Model performance

As obtained in section 4.2.1, the $R^2 = 0.66$, the final model explains 66% of the sample variance indicating a reasonably good fitting for the model. And its $RMSE = 0.075$, it is lower than that of all other sub-models. Figure 2 shows its residual and Q-Q plots, which are used to check the normal distribution and constant variance assumption for the residuals in the regression model. It is seen the data points in the residual vs fitted plot are randomly scattered around mean 0 with no specific pattern, and the data points in the Q-Q plot follow an approximately straight line, therefore both assumptions are met. The model performs well with provided data.

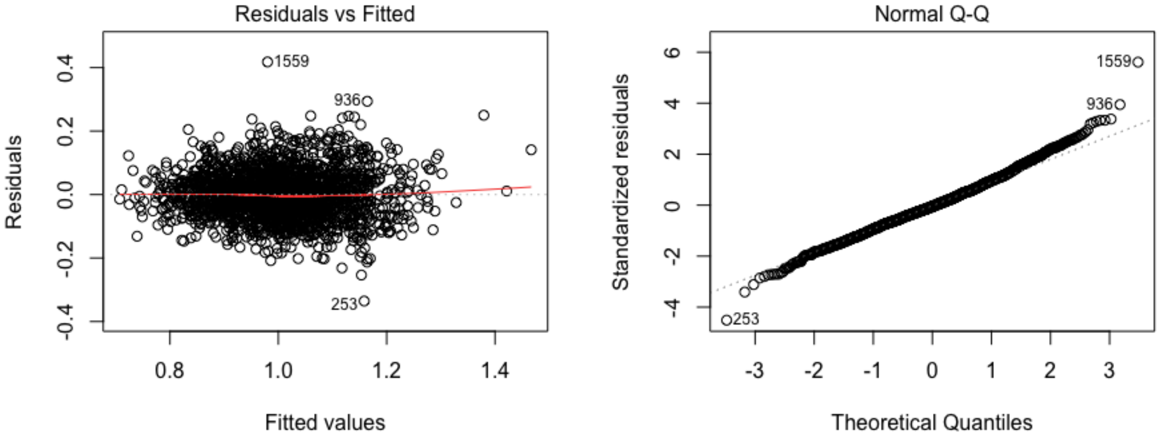

Figure 2: Residual vs fitted (left) and Q-Q (right) plots for the TotalBMD model.

4.2 Model for TotalBMC

The same step-by-step approach is used for selecting models for the TotalBMC response variable and avoiding collinearity. Similar general model forms and hypotheses apply here with only a different response variable, so they are not repeated here. The final model turns out to have the same predictor variables (AGE, Race, Gender, Height, LBMI, and Pre_diabetes) as shown in equation (7). All included parameters are significant at 99.9% confidence level (P-values$<$ 0.001). And all included explanatory variables show a low correlation with VIF values smaller than 2. The model indicates increasing age, height and LBMI leads to higher predicted TotalBMC. Black children (12-20 years old) have 66.1 higher predicted TotalBMC than non-black children and male children have 82.6 lower predicted TotalBMC than female children. And pre-diabetes condition leads to 49.4 lower predicted TotalBMC, which significantly decreases TotalBMC at 99.9% confidence level. For model performance, the model has a high $R^2 = 0.85$ indicating a good fitting. And the RMSE of the model is 188.90, lower than that of all other sub-models with no strong correlation between explanatory variables. The residual plots (Appendix-Figure 10) indicate normality and constant variance assumptions of residuals are reasonably followed.

$$\begin{aligned}Total\hat{B}MC =& -4409.0 + 22.9 \times Age + 28.4 \times Height + 74.2 \times LBMI + 66.1 \times I_b \\ & - 82.6 \times I_m - 49.4 \times I_p\end{aligned} \tag{7}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has a pre-diabetes condition, respectively.

4.3 Model for SpineBMD

Similarly, the final model for predicting SpineBMD is selected as shown in equation (8). All included parameters are significant at $\alpha = 0.01$ (P-values$<$ 0.01) and the coefficient estimate for Pre-diabetes is negative (-0.020). These indicate the pre-diabetes condition significantly decreases SpineBMD (spine bone mineral density) in children at age 12-20. The explanatory variables show a low correlation with $VIF < 2$ ensuring the model results are reliable. And the model fits reasonably well with $R^2 = 0.61$ and $RMSE = 0.10$ (lower than that of all other sub-models which show no strong correlation between explanatory variables). Both residual plots (Appendix-Figure 11) confirm that the model assumptions are met.

$$\begin{aligned} Spine\hat{B}MD = & -0.604 + 0.019 \times Age + 0.006 \times Height + 0.024 \times LBMI + 0.026 \times I_b \\ & -0.146 \times I_m - 0.020 \times I_p \end{aligned} \tag{8}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has a pre-diabetes condition, respectively.

4.4 Model for SpineBMC

At last, a model for predicting SpineBMC is selected using the same procedure. Equation (9) shows the final model. Unlike previously selected models, this model does not include the Race variable since it is not significant at 95% confidence level (P-value$>$ 0.05). All P-values for the estimates of included variables are smaller than 0.001. The effect (increase/decrease) of each variable on the SpineBMC is similar to it on other response variables. Pre-diabetes condition significantly decreases the SpineBMC, spine bone mineral content, of children at age 12-20 at 99.9% confidence level. The predicted SpineBMC is 2.65 lower for children with pre-diabetes than healthy ones. The model fits reasonably well with $R^2 = 0.69$. The $RMSE$ is 8.26 (lower than that of other sub-models in which the predictor variables are not strongly correlated). The model assumptions are met by checking the residual plots (Appendix-Figure 12).

$$Spine\hat{B}MC = -131.54 + 1.43 \times Age + 0.88 \times Height + 1.46 \times LBMI - 6.61 \times I_m - 2.65 \times I_p \tag{9}$$

Where I_b, I_m, and I_p are the indicators that the individual is black, male, and has a pre-diabetes condition, respectively.

5 Conclusion

In this report, multiple linear regression was used to build prediction models for bone health-related response variables: TotalBMD, TotalBMC, SpineBMD, and SpineBMC. It was found the explanatory variables - Age, Race (except for SpineBMC), Gender, Height, LBMI, and Pre-diabetes all significantly affect the TotalBMD, TotalBMC, SpineBMD, and SpineBMC at a minimal 99% confidence level. Adolescents (age 12-20) with increasing age, height, and LBMI have higher predicted bone health values (all four response variables). Male adolescents have lower predicted bone health values than female adolescents without considering related factors such as height, weight, and LBMI since male adolescents tend to have higher LBMI, weight, and height. And adolescents with pre-diabetes have lower predicted bone health values, so the pre-diabetes condition indeed negatively affects children's bone health at $\alpha = 0.01$ (assuming higher values represent better bone health). While black adolescents show higher predicted TotalBMD, TotalBMC, and SpineBMD than non-black adolescents, race does not show a significant effect on SpineBMC at $\alpha = 0.05$. Besides, Weight and LBMI were found to have a strong correlation, therefore are not included in the model spontaneously. The final models show good fitting with R^2 in the range of 0.61-0.85. The RMSE values are 0.08, 188.90, 0.10, and 8.26 for TotalBMD, TotalBMC, SpineBMD, and SpineBMC models, respectively. Overall, there is enough evidence to state the change of TotalBMD, TotalBMC, SpineBMD, and SpineBMC in adolescents at age 12-20 depends on the pre-diabetes status at 99% confidence level.

Reference

1. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.
2. https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.
3. https://www.statology.org/variance-inflation-factor-r/.

Appendix 1-Figures

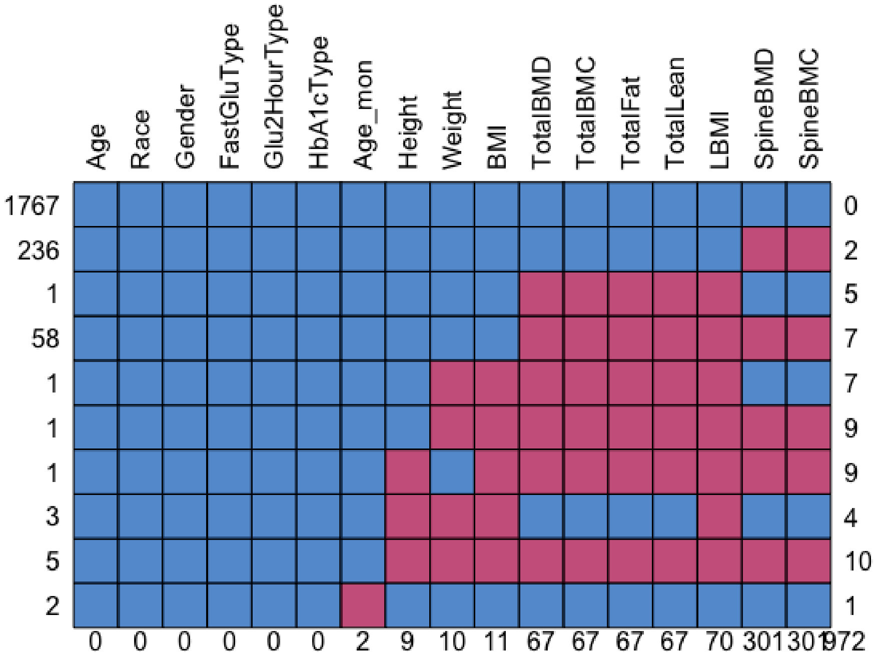

Figure 3: Numbers and distribution of missing data.

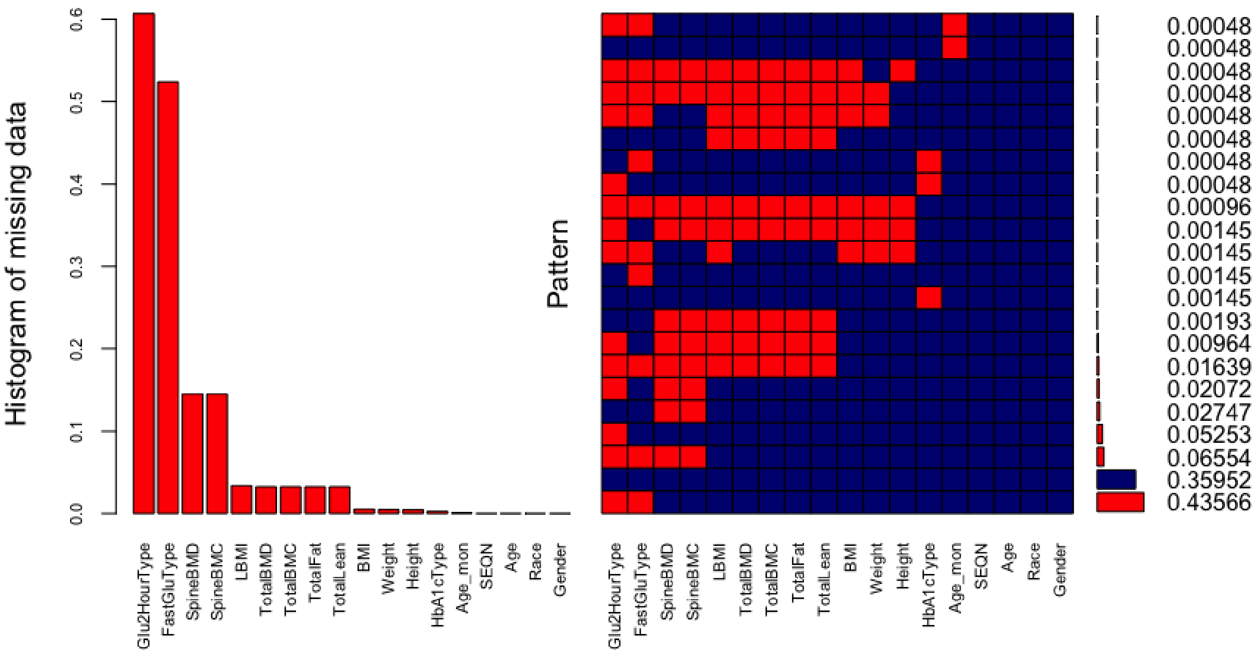

Figure 4: Percentages and distribution of missing data.

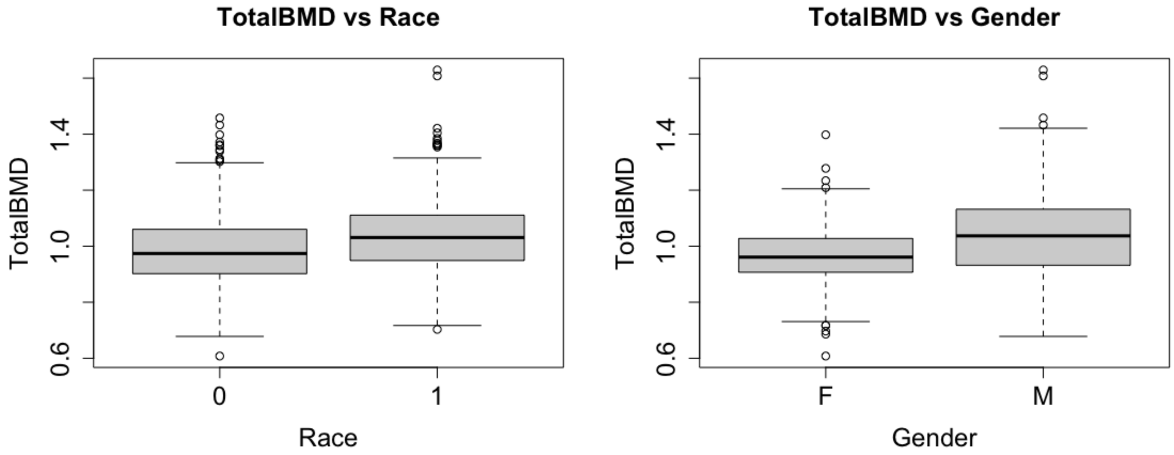

Figure 5: Box-plots of TotalBMD vs Race(left) and Gender (right).

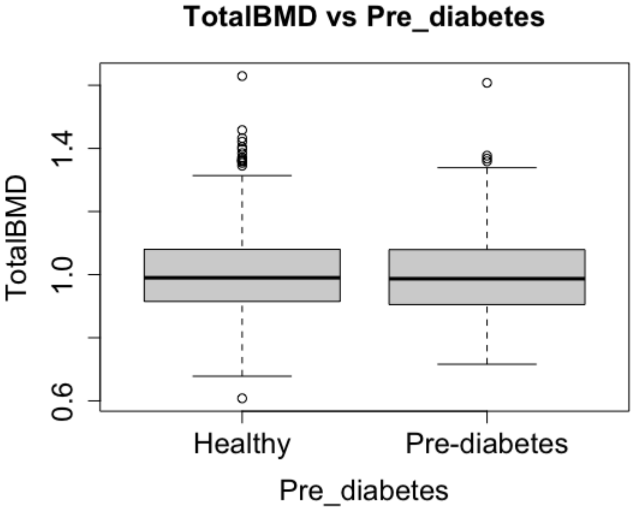

Figure 6: Box-plots of TotalBMD vs Pre_diabetes.

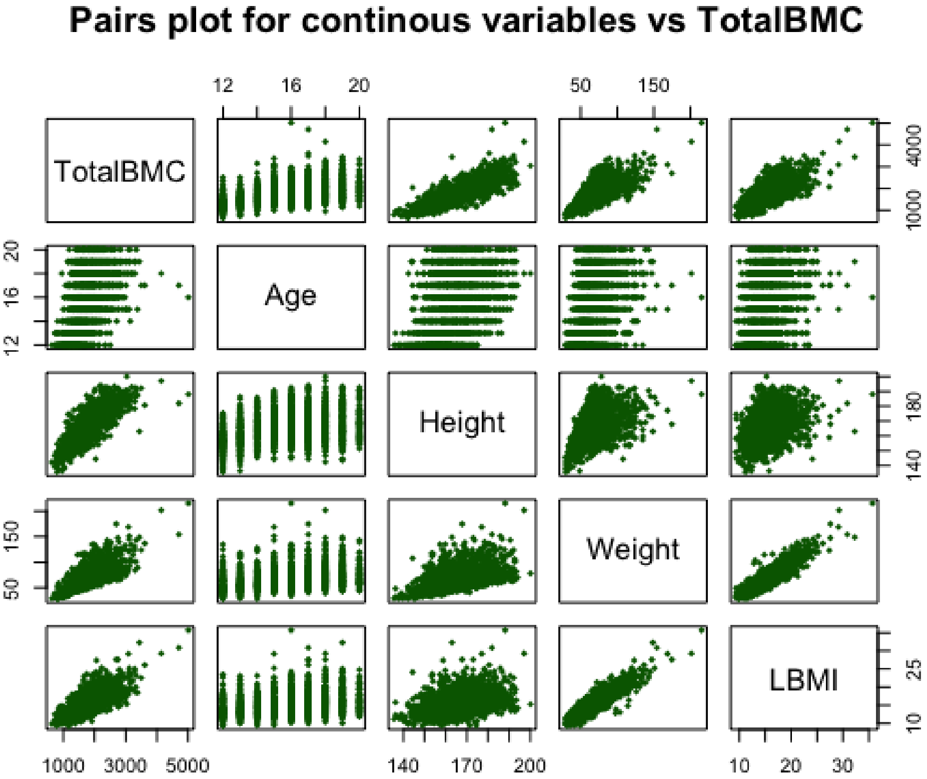

Figure 7: Matrix plots of the continuous variables with TotalBMC.

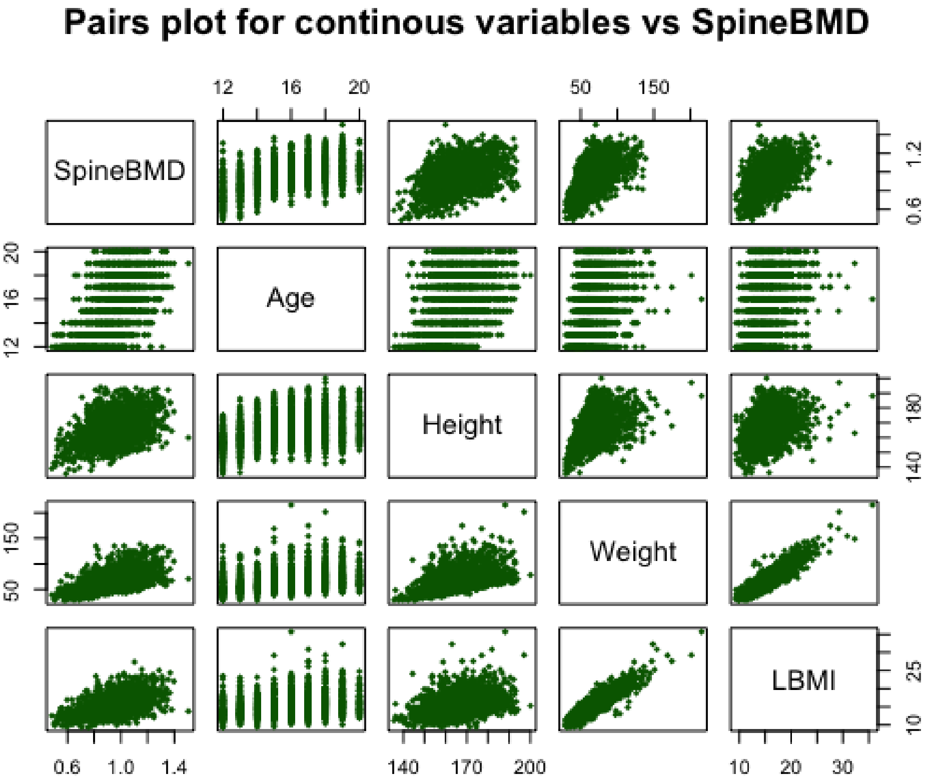

Figure 8: Matrix plots of the continuous variables with SpineBMD.

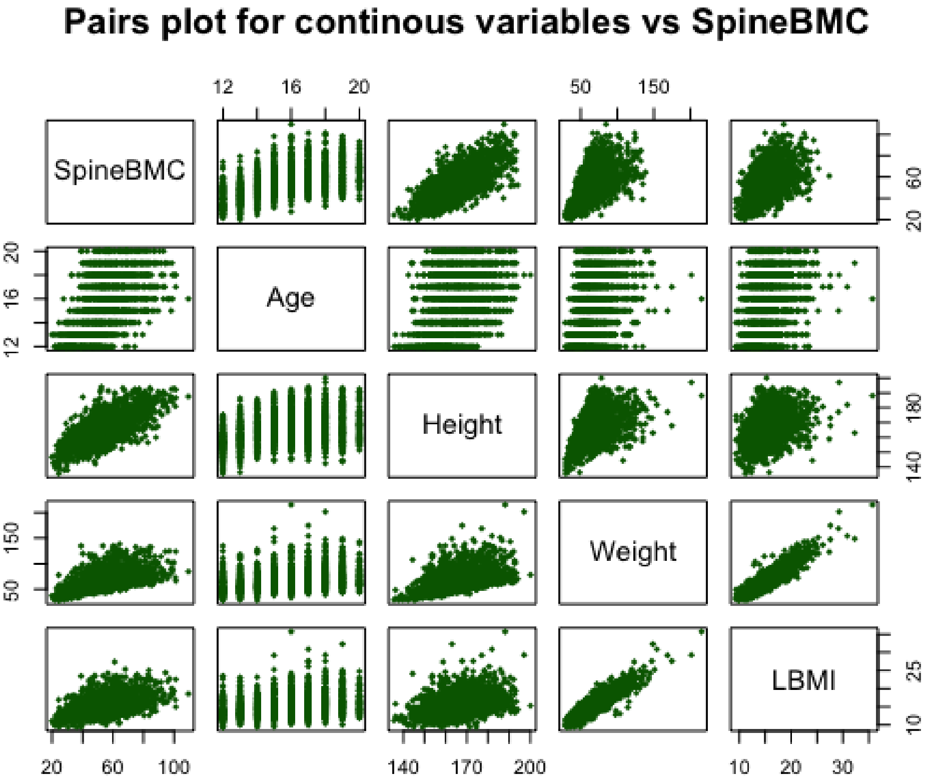

Figure 9: Matrix plots of the continuous variables with SpineBMC.

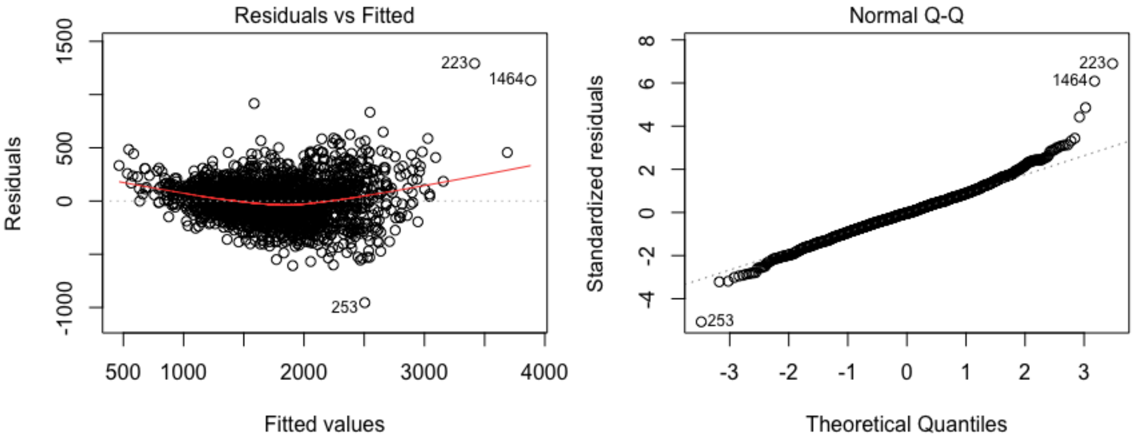

Figure 10: Residual vs fitted (left) and Q-Q (right) plots for the TotalBMC model.

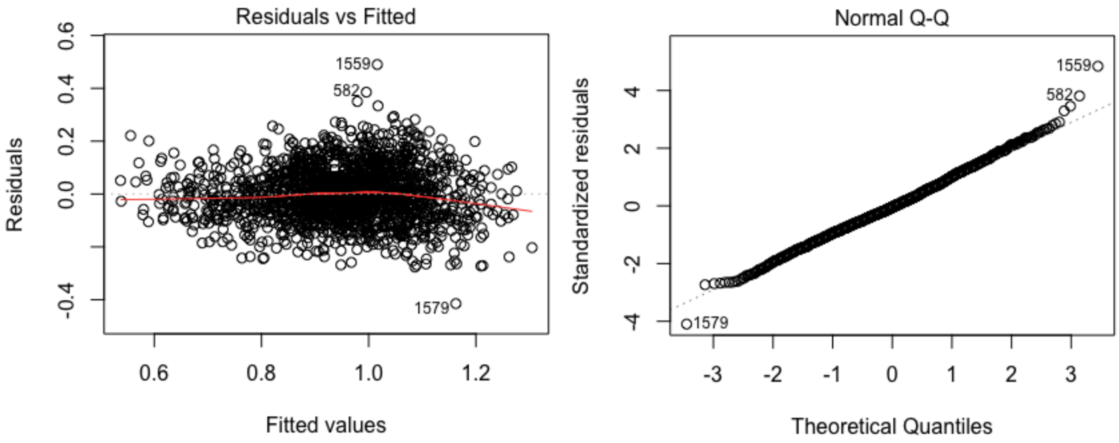

Figure 11: Residual vs fitted (left) and Q-Q (right) plots for the SpineBMD model.

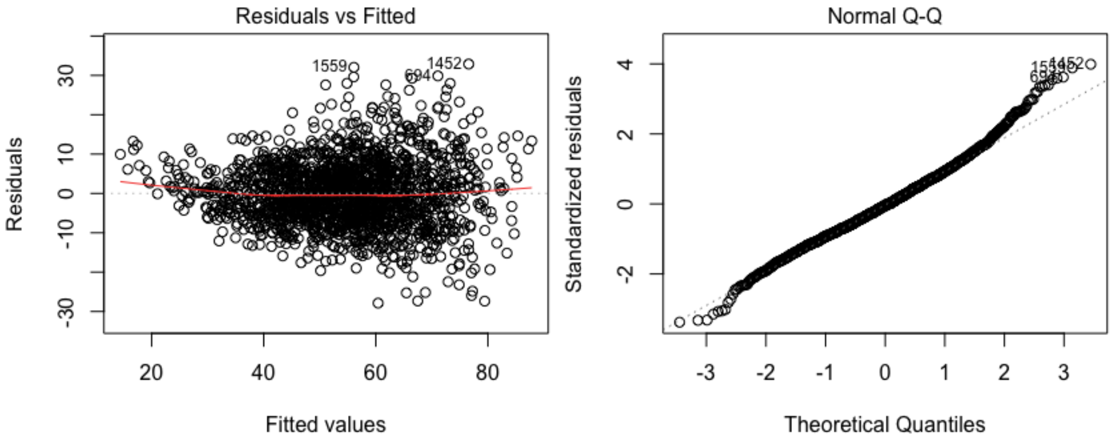

Figure 12: Residual vs fitted (left) and Q-Q (right) plots for the SpineBMC model.

Appendix 2 - R Codes

```
library(tidyverse)
library(readxl)
library(car)
library(mice)
library(VIM)

#######################################################################
#                                                                     #
#                              Load data                              #
#                                                                     #
#######################################################################
data <- read_excel("BoneDensity.xls", col_names = TRUE, na = c("","NA"))

# slicing age
df <- data [data$Age >= 12 & data$Age <= 20, ]
df
#######################################################################
#                                                                     #
#                        Exploratory Data Analysis                    #
#                                                                     #
#######################################################################

#############################################
#                                           #
#                Data Summary               #
#                                           #
#############################################

# Continuous:
summary(df[, c(2,6, 7, 15, 9, 10, 13,14)])
sd(df$Age,na.rm = TRUE)
sd(df$Height,na.rm = TRUE)
sd(df$Weight,na.rm = TRUE)
sd(df$LBMI,na.rm = TRUE)
sd(df$TotalBMD,na.rm = TRUE)
sd(df$TotalBMC,na.rm = TRUE)
sd(df$SpineBMD,na.rm = TRUE)
sd(df$SpineBMC,na.rm = TRUE)

# Categorical
table(df$Race)
table(df$Gender)
table(df$FastGluType)
table(df$Glu2HourType)
table(df$HbA1cType)

##############################################
#                                            #
#     Data Cleaning and Preprocessing        #
#                                            #
##############################################
# Missing data
pMiss <- function(x){sum(is.na(x))/length(x)*100}
apply(df,2,pMiss)  # missing %
md.pattern(df[,-1],rotate.names=TRUE)
aggr(df, col=c('navyblue','red'), numbers=TRUE, sortVars=TRUE,
     labels=names(df), cex.axis = 0.7, gap = 0.5, ylab = c("Histogram of missing data","Patte

# Combining Three Pre-diabetes Test Results into One
dim (df[is.na(df$FastGluType)&is.na(df$Glu2HourType)&is.na(df$HbA1cType), ])[1]
# no rows missing all three tests

df$FastGluType[is.na(df$FastGluType)] <- "miss"
df$Glu2HourType[is.na(df$Glu2HourType)] <- "miss"
df$HbA1cType[is.na(df$HbA1cType)] <- "miss"

test <- df$FastGluType == "Pre-diabetes"|df$Glu2HourType == "Pre-diabetes"|df$HbA1cType == "P
Prediabetes <- replace(test, test == TRUE,"Pre-diabetes")
Prediabetes2 <- replace(Prediabetes, Prediabetes == FALSE,"Healthy")

# Created Pre_diabetes column
df2 <- mutate(df, Pre_diabetes = Prediabetes2)
table(df2[, 19])
df2[,c(1, 9, 10,13,14)]

# X data
df2_x <- df2[, - c(1, 9, 10,13,14)]
df2_x

# Impute missing data using MICE pmm
tempData <- mice(df2_x, m=5, maxit=50, meth='pmm', seed=500)
summary(tempData)
completedData <- complete(tempData, 1)
df3 <- as_tibble(cbind(completedData, df2[, c(1, 9, 10,13,14)]))

##############################################
#                                            #
#              Data Visualization            #
#                                           #
#############################################
TotalBMD <- df3$TotalBMD
TotalBMC <- df3$TotalBMC
SpineBMD <- df3$SpineBMD
SpineBMC <- df3$SpineBMC

Age <- df3$Age;
Height <- df3$Height
Weight <- df3$Weight
LBMI <- df3$LBMI
Race <- df3$Race
Gender <- df3$Gender
Pre_diabetes<-df3$Pre_diabetes

# Matrix Plots
x <- cbind(TotalBMD, Age, Height, Weight, LBMI)
pairs(x, cex = 0.8, col = "darkgreen", pch = 18,
      main = " Pairs plot for continous variables vs TotalBMD")

x2 <- cbind(TotalBMC, Age, Height, Weight, LBMI)
pairs(x2, cex = 0.8, col = "darkgreen", pch = 18,
      main = " Pairs plot for continous variables vs TotalBMC")

x3 <- cbind(SpineBMD, Age, Height, Weight, LBMI)
pairs(x3, cex = 0.8, col = "darkgreen", pch = 18,
      main = " Pairs plot for continous variables vs SpineBMD")

x4 <- cbind(SpineBMC, Age, Height, Weight, LBMI)
pairs(x4, cex = 0.8, col = "darkgreen", pch = 18,
      main = " Pairs plot for continous variables vs SpineBMC")

# Plot with TotalBMD, Height and Gender
df3$Gender <- factor(df3$Gender)
sp <- ggplot(data=df3) +
  geom_point(aes(x = Height, y = TotalBMD, group = 1,color = Gender)) +
  ylab("TotalBMD") + xlab("Height") + ggtitle("TotalBMD vs Height and Gender")

sp + scale_color_manual(values = c("red", "blue")) +
  theme(axis.text = element_text(size = 12),
        axis.title = element_text(size = 14,face = "bold"),
        legend.position = "bottom")

# Boxplots for TotalBMD vs Race, Gender, and Pre_diabetes
plot(TotalBMD ~ factor(Race), ylab="TotalBMD", xlab="Race",
     cex.axis=1.5, cex.lab=1.5, cex.main=1.5, main="TotalBMD vs Race")
plot(TotalBMD ~ factor(Gender), ylab="TotalBMD", xlab="Gender",
     cex.axis=1.5, cex.lab=1.5, cex.main=1.5, main="TotalBMD vs Gender")
plot(TotalBMD ~ factor(Pre_diabetes), ylab="TotalBMD", xlab="Pre_diabetes",
     cex.axis=1.5, cex.lab=1.5, cex.main=1.5, main="TotalBMD vs Pre_diabetes")

########################################################################
#                                                                      #
#                       Model Building - TotalBMD                      #
#                                                                      #
########################################################################
# initial model with demographic variables
df3$Race <- factor(df3$Race)
g11 <- lm (TotalBMD ~ Age + Race + Gender, data = df3)
summary(g11)
anova(g11)
vif(g11)

# Add more predictors
g12 <- lm (TotalBMD ~ Age + Race + Gender + Height, data = df3)
summary(g12)
anova(g12)
vif(g12)

g13 <- lm (TotalBMD ~ Age + Race + Gender + Height + LBMI, data = df3)
summary(g13)
anova(g13)
vif(g13)

# Final model
g14 <- lm (TotalBMD ~ Age + Race + Gender + Height + LBMI + Pre_diabetes, data = df3)
summary(g14)
anova(g14)
vif(g14)

# 95% Confidence Interval for Estimates
confint(g14)

# Residual Plots
plot(g14, 1:2)

g15 <- lm (TotalBMD ~ Age + Race + Gender + Height + LBMI + Pre_diabetes + Weight, data = df3
summary(g15)
anova(g15)
# LBMI and Weight highly correlated
```

```
vif(g15)

g16 <- lm (TotalBMD ~ Age + Race + Gender + Height + Pre_diabetes + Weight, data = df3)
summary(g16)
anova(g16)
vif(g16)

########################################################################
#                                                                      #
#                        Model Building - TotalBMC                     #
#                                                                      #
########################################################################
# Initial model with demographic variables
g21 <- lm (TotalBMC ~ Age + Race + Gender, data = df3)
summary(g21)
anova(g21)
vif(g21)

# Add more predictors
g22 <- lm (TotalBMC ~ Age + Race + Gender + Height, data = df3)
summary(g22)
anova(g22)
vif(g22)

g23 <- lm (TotalBMC ~ Age + Race + Gender + Height + LBMI, data = df3)
summary(g23)
anova(g23)
vif(g23)

# Final model
g24 <- lm (TotalBMC ~ Age + Race + Gender + Height + LBMI + Pre_diabetes, data = df3)
summary(g24)
anova(g24)
vif(g24)
# CI for Estimates
confint(g24)

# Residual Plots
plot(g24, 1:2)

g25 <- lm (TotalBMC ~ Age + Race + Gender + Height + LBMI + Pre_diabetes + Weight, data = df3
summary(g25)
anova(g25)
vif(g25)
```

```
g26 <- lm (TotalBMC ~ Age + Race + Gender + Height + Pre_diabetes + Weight, data = df3)
summary(g26)
anova(g26)
vif(g26)

########################################################################
#                                                                      #
#                       Model Building - SpineBMD                      #
#                                                                      #
########################################################################
# Initial model with demographic variables
g31 <- lm (SpineBMD ~ Age + Race + Gender, data = df3)
summary(g31)
anova(g31)
vif(g31)

g32 <- lm (SpineBMD ~ Age + Race + Gender + Height, data = df3)
summary(g32)
anova(g32)
vif(g32)

g33 <- lm (SpineBMD ~ Age + Race + Gender + Height + LBMI, data = df3)
summary(g33)
anova(g33)
vif(g33)

# Final model
g34 <- lm (SpineBMD ~ Age + Race + Gender + Height + LBMI + Pre_diabetes, data = df3)
summary(g34)
anova(g34)
vif(g34)

# Confidence Interval for Estimates
confint(g34)

# Residual Plots
plot(g34, 1:2)

g35 <- lm (SpineBMD ~ Age + Race + Gender + Height + LBMI + Pre_diabetes + Weight, data = df3
summary(g35)
anova(g35)
# LBMI and Weight highly correlated
vif(g35)

g36 <- lm (SpineBMD ~ Age + Race + Gender + Height + Pre_diabetes + Weight, data = df3)
```

```r
summary(g36)
anova(g36)
vif(g36)

######################################################################
#                                                                    #
#                       Model Building - SpineBMC                    #
#                                                                    #
######################################################################
# Initial model with demographic variables
g41 <- lm (SpineBMC ~ Age + Race + Gender, data = df3)
summary(g41)
anova(g41)
vif(g41)

g42 <- lm (SpineBMC ~ Age + Race + Gender + Height, data = df3)
summary(g42)
anova(g42)
vif(g42)

g43 <- lm (SpineBMC ~ Age + Race + Gender + Height + LBMI, data = df3)
summary(g43)
anova(g43)
vif(g43)

g44 <- lm (SpineBMC ~ Age + Race + Gender + Height + LBMI + Pre_diabetes, data = df3)
summary(g44)
Anova(g44, type = 3)
vif(g44)

# LBMI and Weight highly correlated
g45 <- lm (SpineBMC ~ Age + Race + Gender + Height + LBMI + Pre_diabetes + Weight, data = df3
summary(g45)
anova(g45)
vif(g45)

g46 <- lm (SpineBMC ~ Age + Race + Gender + Height + Pre_diabetes + Weight, data = df3)
summary(g46)
anova(g46)
vif(g46)

# From g44: Race not significant, remove Race, final model
g47 <- lm (SpineBMC ~ Age + Gender + Height + LBMI + Pre_diabetes, data = df3)
summary(g47)
Anova(g47, type = 3)
```

```
vif(g47)

# CI for Estimates
confint(g47)

# Residual Plots
plot(g47, 1:2)
```