CHAPTER 4

DATA ANALYSIS

4.1 Basic Statistical Characteristics

The basic characteristics of the data are described in Table 1, Table 2, Figure 1, Figure 2 and Figure 3. According to the results in Table 1, the mean and median of the age of the child in days are 277.7 and 219.5. The mean is larger than the median because the mean is more affected by extreme values. The gestational age of the child in days has the smallest value, which is 32.00. DiPietro and Allen (1991) said that normally gestational age can be 38 to 42 weeks. When the gestational age is small, the mother can usually supplement some nutrients. The first quartile of the birth weight in grams is 3150, indicating that 25% of the birth weight data is below this point. The third quartile of the weight measurement in kilograms is 10.775, implying that 75% data of weight measurement in kilograms lies below 10.775. The correlation between the interested variables is depicted in Table 2. It can be seen that explanatory and response variables are positively correlated. As observed in Figure 1, the weight measurement in kilograms shows an obvious growth trend over the age of the child in days. There is a strong positive linear correlation. Figure 2 represent the relationship between the gestational age of the child in days and weight measurement in kilograms. Similarly, Figure 3 is the relationship between birth weight in grams and weight measurement in kilograms. There is a weak positive correlation between each pair of variables as Figure 2 and 3.

Table 1: Summary statistics of interested variables

	agedays	ga	bw	wtkg
Min	0.0	32.00	1180	1.180
1st Qu	60.0	39.00	3150	5.170
Median	219.5	40.00	3500	8.305
Mean	277.7	39.76	3496	8.171
3rd Qu	457.0	41.00	3900	10.775
Max	978.0	43.00	5100	16.500

Table 2: Correlation of interested variables

	agedays	ga	bw	wtkg
agedays	1.000	-0.011	-0.008	0.936
ga	-0.011	1.000	0.606	0.027
bw	-0.008	0.606	1.000	0.109
wtkg	0.936	0.027	0.109	1.000

Figure 1: Relationship between age of the child in days and weight measurement in kg

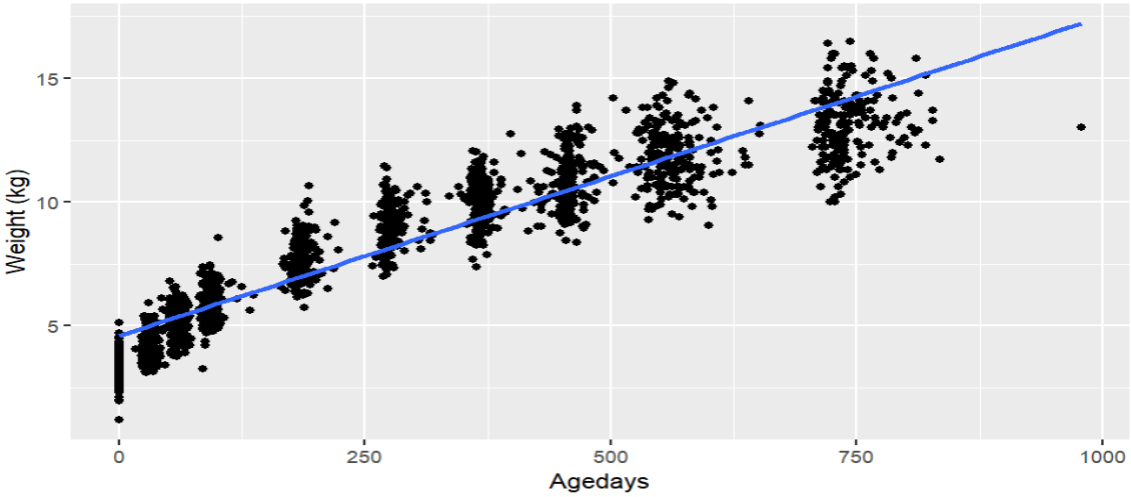


Figure 2: Relationship between gestational age of the child in days and weight measurement in kg

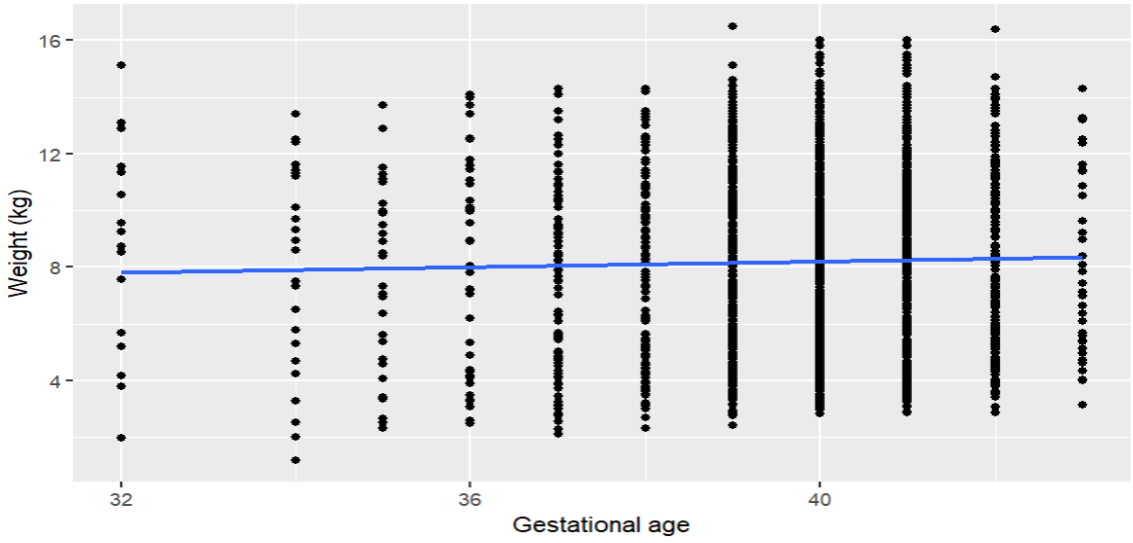
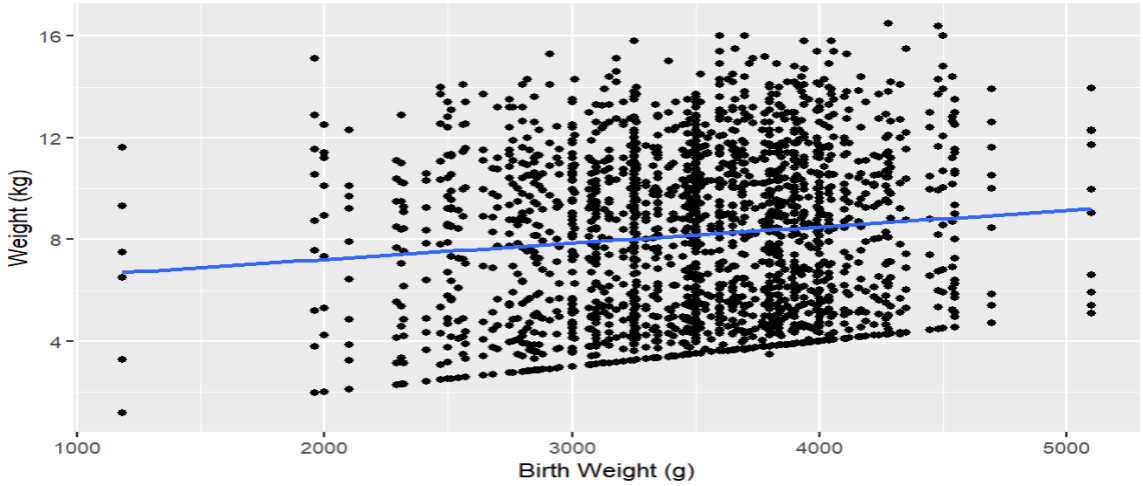


Figure 3: Relationship between birth weight in grams and weight measurement in kg



4.2 Linear Regression Model Results

Table 3 contains the estimates of the parameters from the fitted linear regression model. It also shows the 95% confidence interval and the p-value of the parameters. The parameters of the age of the child in days, the sex of the male child and the birth weight in grams are positive, indicating that the increase in these explanatory variables will contribute to an associated increase in the weight measurement in kilograms. Conversely, the parameter of the gestational age of the child in days is negative, suggesting that the increase in the gestational age will lead to a corresponding decrease in the weight measurement in kilograms. For example, the weight measurement in kilograms will decrease by 0.075 for every one unit increase in the gestational age as noted by the Table 3. Since 0 does not lie within the confidence interval of these explanatory variables, it can be concluded that all explanatory variables are significant. In addition, it can be inferred that there is evidence of the linear relationship between the response variable and explanatory variables. The p-value of all explanatory variables are less than 0.05, so the explanatory variables have the impact on the weight measurement in kilograms. Hence, people should contain all explanatory variables in the future model. After fitting the model, the assumptions are checked from Figure 4 to Figure 6. The scatterplot of the residuals against explanatory variables is presented in Figure 4. It can be seen that the residuals of all explanatory variables have mean zero. Comparing the graphs in Figure 4, it's clear that the residuals of the age of the child in days have non-constant variance. Figure 5 provides the scatterplot of the residuals against fitted values. It depicts the residuals of the age of the child in days and fitted values have mean zero and non-constant variance. As it can be observed in Figure 6, the residuals do not appear to be normally distributed. In total, the linear regression model does not follow the assumptions. People may solve this problem

by taking logarithm to the response variable. The plot of observed and predicted values is presented in Figure 7.

Table 3: Estimates of the parameters from the fitted linear regression model

Term	Estimate	CI Lower Bound	CI Upper Bound	P-Value
intercept	4.623	3.472	5.774	0.000
agedays	0.013	0.013	0.013	0.000
sex: Male	0.355	0.258	0.451	0.000
ga	-0.075	-0.108	-0.041	0.000
bw	0.001	0.001	0.001	0.000

Figure 4: Scatterplot of the residuals against explanatory variables

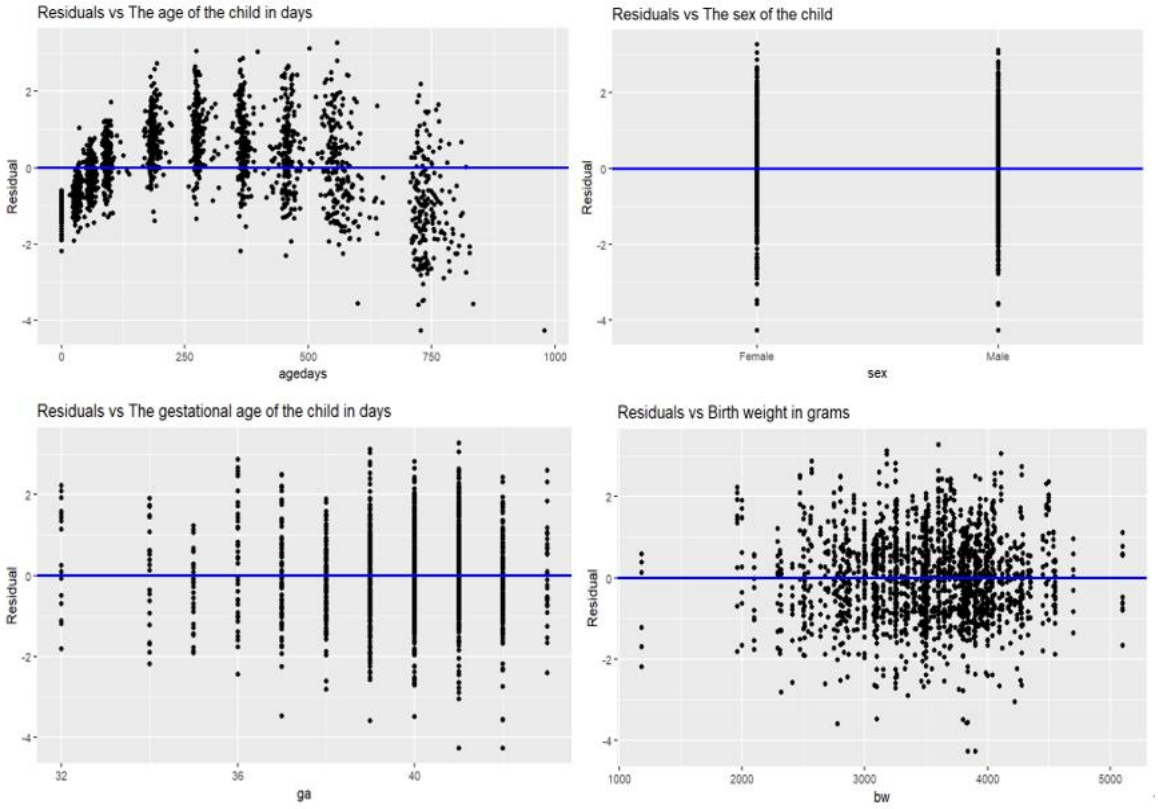


Figure 5: Scatterplot of the residuals against fitted values

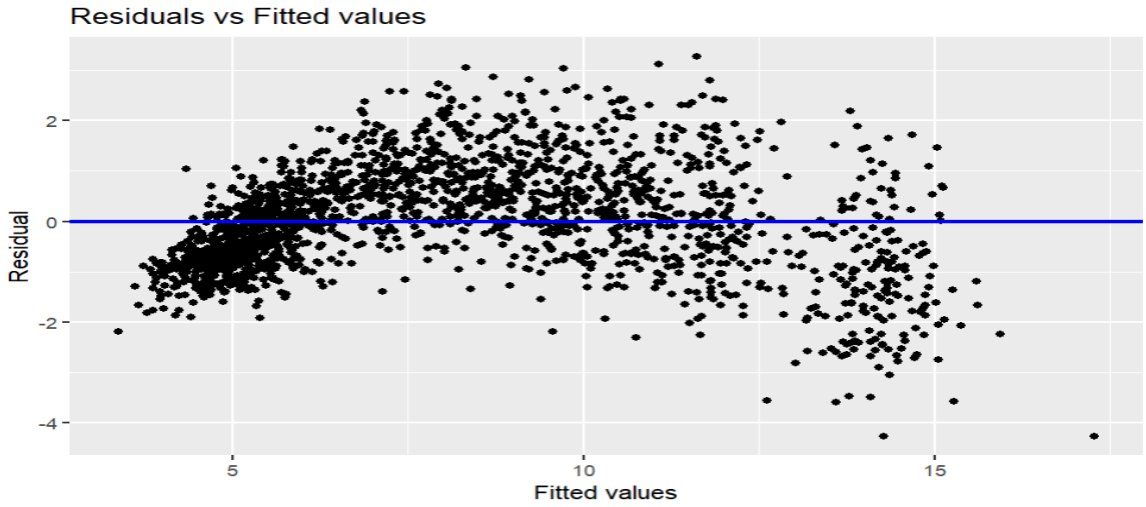


Figure 6: Histogram of residuals

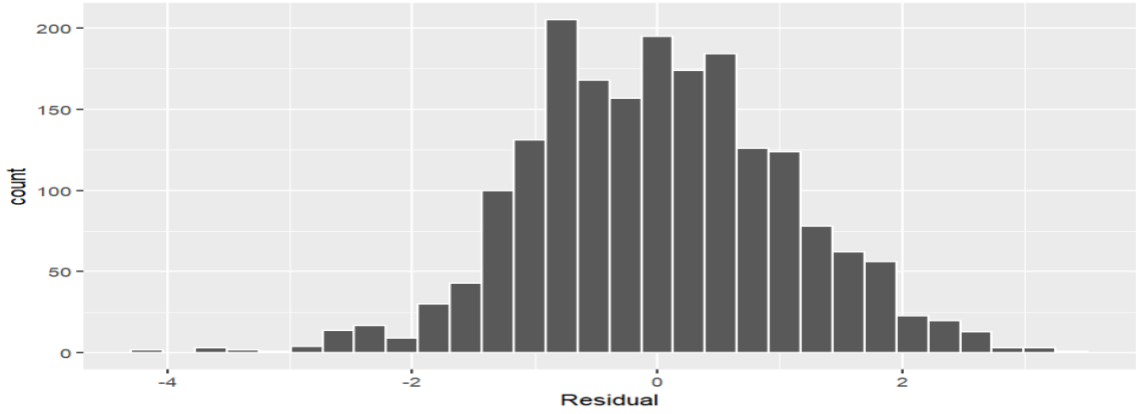
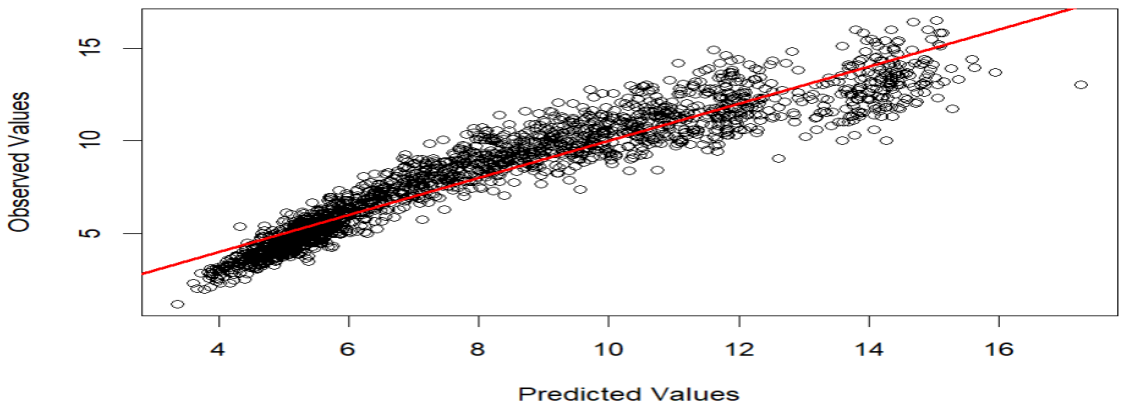


Figure 7: Plot of observed and predicted values



4.3 Generalized Additive Model Results

The generalized additive model of this article assumes a Gaussian or normal distribution of the errors. The parametric terms of this model are based on the factor variable. It shows the coefficients for the linear terms in the model, Table 4 shows the parametric coefficients of the intercept and the sex of the child. It can be observed that the weight measurement in kilograms will increase in 0.372 with every one unit increase in the sex of the male child. The standard error of the sex of the male child is 0.037, which is small. A small standard error indicates that the means are closer together. Thus, it is more likely that the sample mean is an accurate representation of the true population mean. At the same time, people can see that the model intercept and the fixed effect of the sex of the child are significant at the 0.05 level. Table 5 presents the smooth terms of generalized additive model. Smooths coefficients are not printed in this chart because each smooth has several coefficients. The value of effective degrees of freedom represents the complexity of the smooth. Janson, Fithian and Hastie (2015) inferred that the higher effective degrees of freedom can describe more wiggly curves. It is obvious that all explanatory variables in the Table 5 are not straight lines or quadratic curves because all effective degrees of freedom are greater than 2. They are complex and wiggly. People can use the reference degrees of freedom and F column to test overall significance of the smooth. The result of this test can be drawn from the p-value. It is easy to conclude that the smooth terms of all explanatory variables are significant since the p-value are less than 0.05. Partial effect plots in Figure 8 depicts how the response variable changes with the explanatory variables. The Y-axis contains the values of the weight measurement in kilograms and the X-axis contains the values of explanatory variables. From the plot people can know that the weight measurement in kilograms first increases with the age of the child in days then decreases after around 800. For the variable of the birth weight in grams, the weight measurement in kilograms keep increasing normally. It seems that there is a decrease in the weight measurement in kilograms when the gestational age of the child in days increases. The weight measurement in kilograms also increases monotonically for the categorical variable of the sex of child. Hence, the generalized additive model is an effective way of fitting nonlinear functions on several variables. Figure 9 shows diagnostic plots for model checking. The Q-Q plot is on the top-left, which compares the model residuals to a normal distribution. The residuals will be near to a straight line if they come from the well-fit model. As it can be observed that the histogram of residuals has a symmetrical bell shape on top-right plot. The plot of residual values is on the bottom-left. These values are approximately evenly distributed around the zero. Finally, the plot of response against fitted values can be seen from the bottom-right. It clusters around the 1-to-1 line, which means the model fits nearly perfectly.

Table 4: Parametric terms of generalized additive model

	Estimate	Standard Error	P-Value
Intercept	7.988	0.025	0.000
sexMale	0.372	0.037	0.000

Table 5: Smooth terms of generalized additive model

	Effective Degrees of Freedom	Reference Degrees of Freedom	F	P-Value
s(agedays)	8.718	8.974	3527.039	0.000
s(bw)	8.472	8.921	47.605	0.000
s(ga)	8.782	8.981	9.501	0.000

Figure 8: Partial effect plots of generalized additive model

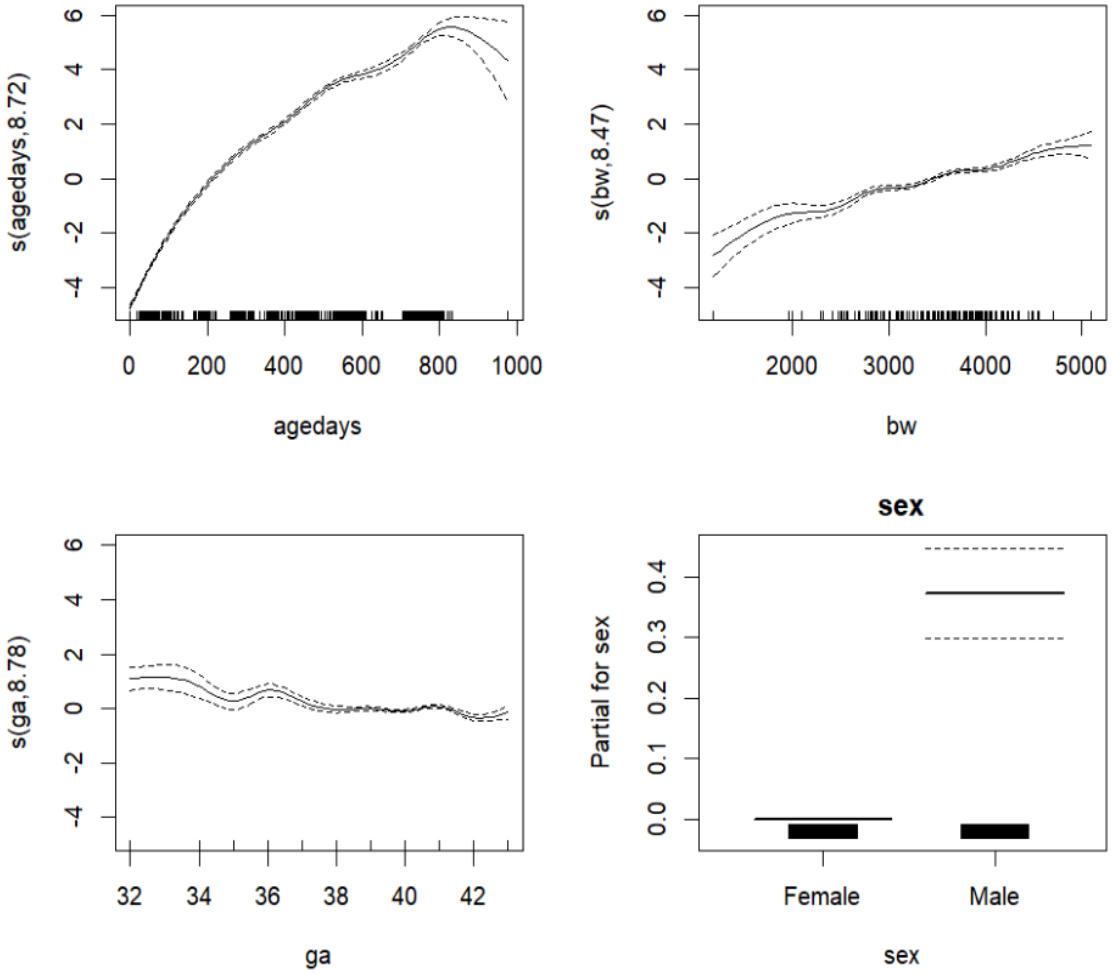
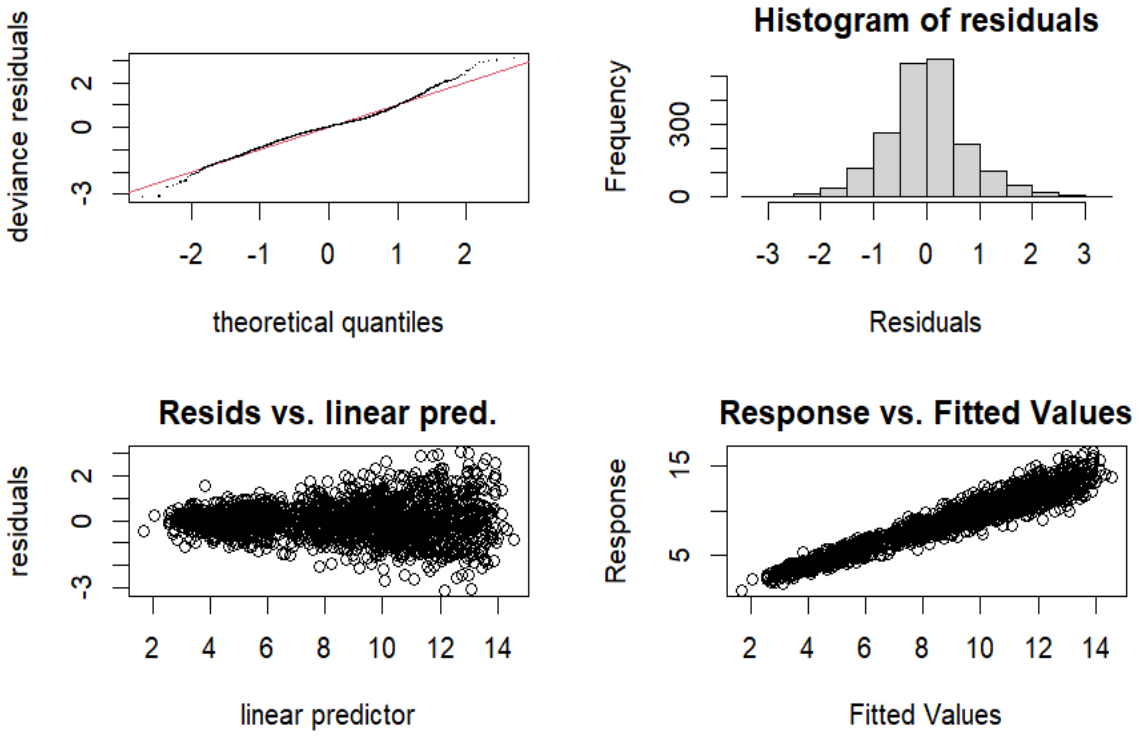


Figure 9: Diagnostic plots of generalized additive model



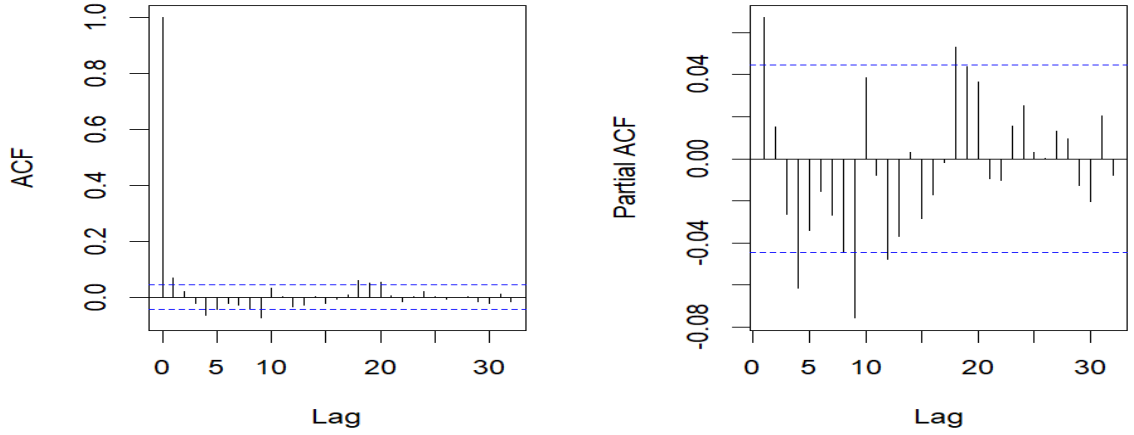
4.4 Generalized Additive Mixed Model Results

Rigby and Stasinopoulos (2005) noted that the generalized additive model has the assumption that residuals are identically and independently distributed. However, the errors sometimes will be correlated in real practice. People call this phenomenon autocorrelation, which implies results from the model may be biased. The problem can be handled by adding autoregressive model for residuals. As it can be observed in Table 6, all explanatory variables are significant due to their p values are less than 0.05. Figure 10 is the plot of autocorrelation function and partial autocorrelation function with AR (1). The ACF drops off from lag 0. The values of partial ACF mostly between dashed blue lines, which can lessen the autocorrelation of residuals.

Table 6: Estimated values of generalized additive mixed model with AR (1)

	Value	Std.Error	DF	t-value	p-value
X(Intercept)	8.100	0.055	1943	146.109	0.000
XsexMale	0.163	0.052	1943	3.157	0.002
Xs(agedays)	3.027	0.336	1943	8.997	0.000
Xs(bw)	0.540	0.034	1943	15.670	0.000
Xs(ga)	-0.919	0.298	1943	-3.086	0.002

Figure 10: ACF and PACF of residuals from generalized additive model with AR (1)



4.5 Model Performance Comparison

The model comparison can be found on the Table 7. Anderson and Burnham (2004) concluded that the absolute values of the AIC and BIC are not important. The best fitting model comes from the minimum AIC and BIC. Based on the AIC and BIC, people can infer that it is better to use Z-scores data than raw data to fit the model. The generalized additive mixed model of Z-scores has the minimum AIC and BIC, which means it is the most suitable model for characterizing growth patterns in children's weight.

Table 7: AIC and BIC of six models

	AIC	BIC
LRM	5748.731	5782.178
GAM	4566.587	4729.117
GAMM	3053.050	3108.796
LRM Z-scores	1141.231	1174.679
GAM Z-scores	-32.750	128.757
GAMM Zscores	-1554.449	-1498.704