MODELLING THE GROWTH OF YOUNG CHILDREN

Project 1

by

Li Lixia
2647470

A thesis submitted to the faculty of
The School of Mathematics and Statistics
at University of Glasgow
in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

School of Mathematics and Statistics

University of Glasgow

September 2022

ABSTRACT

This work takes longitudinal weight measurements for 1933 children from across the Netherlands as the research object and uses linear regression model, generalized additive model and generalized additive mixed model as the research methods. Firstly, the relevant statistical characteristics of infant weight are analyzed. Secondly, a simple linear regression model is constructed. Thirdly, using a generalized additive model to obtain non-linear details. Fourthly, aiming at removing autocorrelation of residuals, this article has developed a generalized additive mixed model that adds new constraints to error terms. Finally, the fitting performance of different models can be compared by the Akaike information criterion and Bayesian information criterion. The result shows that choosing the generalized additive mixed model with Z-scores is helpful for people to reasonably detect significant variables, so as to achieve the goal of predicting children's weight.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Research Background

The growth of children is a concern in the entire world. There are various tools and means to detect factors that impact on the development of infants. Choosing the most suitable model can improve immunity as much as possible and control the disease effectively. This subject combines the theory of some prediction models with the real data of the infant growth, which has analytical and research significance.

This study will take the age of the child in days, the sex of the child, the gestational age of the child in days and the birth weight in grams as the explanatory variables. The response variable will focus on the weight measurement in kilogram. Three modelling approaches can be applied to these variables. The optimal model can be obtained by comparing some criteria.

1.2 Research Question

How to choose the most suitable model for characterizing growth patterns in children's weight?

1.3 Thesis Structure

This article uses longitudinal weight data and focuses on quantitative analysis. The paper applies the theoretical methods of linear regression model, generalized additive model and generalized additive mixed model to1933 children from across the Netherlands. Based on minimum the Akaike information criterion and Bayesian information criterion, the best fitting model can be derived. The structure of this article is mainly divided into introduction, literature review, methodology, data analysis and conclusion.

CHAPTER 2

LITERATURE REVIEW

Diskin (1970) illustrated that linear regression is an analytical method that uses regression equations to model the relationship between one or more independent variables and dependent variables. Hanley (2016) found out that simple linear regression includes only one independent variable, inferring that the condition of more than one independent variable is called multiple linear regression. Linear regression model is a basic model in machine learning. Poole and O'Farrell (1971) claimed that linear regression has five hypotheses. First, there exists a linear relationship between the response and the independent variables. Then, the residuals are independent. Next, there is no multicollinearity between the explanatory variables. In addition, the error terms should have homoscedasticity. Finally, the residuals are normally distributed. If these assumptions are violated, it is suggested that apply non-linear transformation in the form of log or square root to the variables (Benoit, 2011). Shafi and Rusiman (2015) presented that doctors can use linear regression model to predict the probability that a certain disease will occur. This paper mainly uses linear regression model to track children's weight gain.

When the data shows non-linear effects, linear regression model will fail. According to Marra and Wood (2011), generalized additive model can replace the linear component with some smooth functions. It implies that this model can capture nonlinear relationships. It also means that the model will become extremely flexible because there are many different smooth functions. Then, Yu, Park and Mammen (2008) proved that the generalized additive model estimates smooth components through the backfitting algorithm. Therefore, the model can get more features about the curve. Horowitz (2001) observed that there are also some assumptions of the generalized additive model. It can be inferred that the checking procedure is similar to the linear regression model. As found by Jbilou and Adlouni (2012), the principal advantage of the generalized additive model is its ability to model highly complex nonlinear relationships when the number of potential predictors is large. However, the model might miss some non-linear interactions among the predictors. Hastie and Tibshirani (1995) recommend that the generalized additive model can be used to detect the influence of potential prognostic factors on the disease endpoints.

When the assumption that the residuals are not independent is violated in the above models, Fahrmeir and Lang (2001) wrote that generalized additive mixed model can address this problem. It deals with the correlated errors via adding random effects. As noted by Baayen et al (2018), autoregressive model can be used to the residuals that makes the random effects. Wood (2006) demonstrated that the main advantage of the model is that it allows heteroscedasticity, which makes the fitting performance better. However, the shortcoming of the model is also obvious. Wieling (2018) indicated that the computer

calculates the coefficients of this model very slowly. Shadish, Zuur and Sullivan (2014) concluded that people can make the generalized additive mixed model to analyze single case designs.

What types of models are most suitable for describing growth patterns in children's weight is an important question. The purpose of this literature review is to use linear regression model, generalized additive model and generalized additive mixed model to capture the change characteristics of infant weight. In the principle of minimum Akaike information criterion and Bayesian information criterion, the fitting performance of these models can be compared (Regnault et al., 2014). The significance of this literature review is to enable people to choose appropriate models to observe the factors affecting infant weight.

CHAPTER 3

METHODOLOGY

3.1 Linear Regression Model

Pomerance and Krall (1981) proposed that the multiple linear regression can draw the longitudinal growth curve. Their approach enables the child to have the slope and random intercept through the following formula

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon_i \ \ where\ \varepsilon_i \sim N\ (0, \sigma^2) \qquad (1)$$

Here, y_i is the response variable of the i^{th} observation. The weight measurement in kg is the dependent variable in this article. x_i is the explanatory variable of the i^{th} observation. The independent variables include the age of the child in days, the sex of the child, the gestational age of the child in days and the birth weight in grams. β_i represent the regression coefficient that should be estimated. It can be calculated by the least squares method. The ε_i is the residuals. There are some assumptions of the linear regression model. The residuals are required to be independent and normally distributed. In addition, it is perfect that the residuals have constant variance and mean zero. People should try to record the values of the explanatory variables without error. The multiple linear regression was fitted to predict children's weight gain through the data.

3.2 Generalized Additive Model

The Linear model is easy to infer and interpret. However, there are some complex phenomena that cannot be represented by simple linear relationships. If people still choose to fit a linear model to the data in these cases, it won't get a good result. The linear model will not capture key aspects of the data. Hence, Hastie and Tibshirani (1987) demonstrated that people can use the generalized additive model (GAM) to fit data which have complex and nonlinear relationships. It will make good predictions in these situations because it fits data with flexible smooths or splines. The GAM is obtained as follows:

$$y_i = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_i(x_i) + \varepsilon_i \ \ where\ \varepsilon_i \sim N\ (0, \sigma^2) \qquad (2)$$

y_i is a response variable and $x_1, \ldots, x_i$ are independent variables. β_0 is an intercept. $f_1, \ldots, f_i$ are unknown smooth functions. The independent identically distributed random error is ε_i.The flexible smooths in GAM are constructed by some small functions.

These functions are called basis functions. People can fit data with a wide variety of shapes by using the smooth functions in the generalized additive model.

3.3 Generalized Additive Mixed Model

The generalized additive mixed model (GAMM) is a continuation of generalized additive model. When residuals do not follow assumptions, the generalized additive model will fail. This issue can be solved by including autoregressive model to residuals. Groll and Tutz (2012) observed that GAMM can be widely used in the correlated and clustered response variable. For example, the GAMM can capture the dependence structure within the longitudinal data. This model is defined as follows:

$$y_i = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_i(x_i) + e_i \text{ where } e_i = \emptyset e_{i-1} + \varepsilon_i \text{ , } \varepsilon_i \sim N\ (0, \sigma^2) \quad (3)$$

where $e_i = \emptyset e_{i-1} + \varepsilon_i$ is an AR (1) process. $\emptyset$ is an unknown autoregressive coefficient, which should be estimated when fitting GAMM. The autocorrelation of residuals may be eliminated by using GAMM step by step.

3.4 Z-scores and AIC/BIC

Z-scores describe the position of a raw score in terms of its distance from the mean, when measured in standard deviation units (Curtis et al., 2016). It can compare scores on different kinds of scales by standardizing the distribution. A positive Z-score indicates the raw score is higher than the mean. It is possible to transform the normal random variable X into Z-scores by using the following formula

$$z = \frac{x - \mu}{\sigma}$$

where x is the normal random variable, μ is the mean of x, and σ is the standard deviation of x.

People should choose the best model to capture key details of data. Akaike information criterion (AIC) and Bayesian information criterion (BIC) can measure the goodness of fit of statistical models. The smallest AIC and BIC will get the most accurate model. They are defined as follows:

$$AIC = 2k - 2\ln(L)$$

$$BIC = k\ln(n) - 2\ln(L)$$

where k is the number of parameters, which mainly captures complexity of the model. ln(L) is the log-likelihood function of the model, which mostly measures the goodness of fit. Here n is the number of data points. The main difference between AIC and BIC is the

weight of the penalty. The BIC will have more penalties for the model's complexity with the number of data points increases.

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

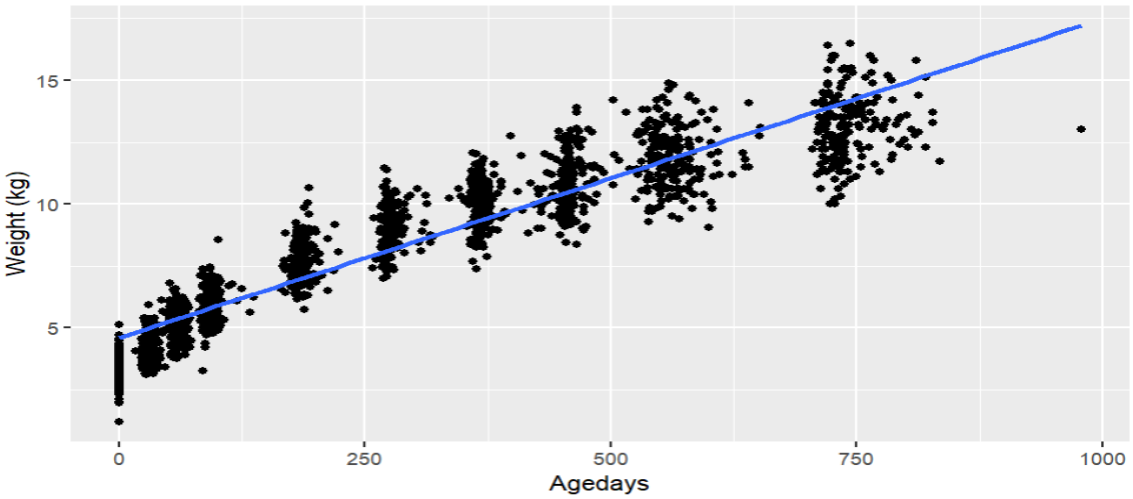


Figure 2: Relationship between gestational age of the child in days and weight measurement in kg

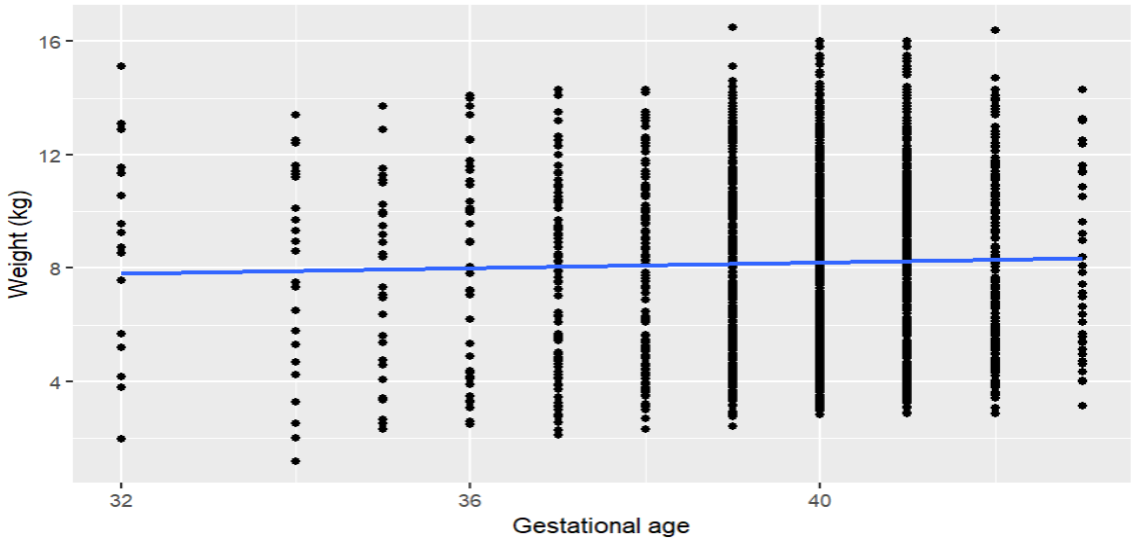
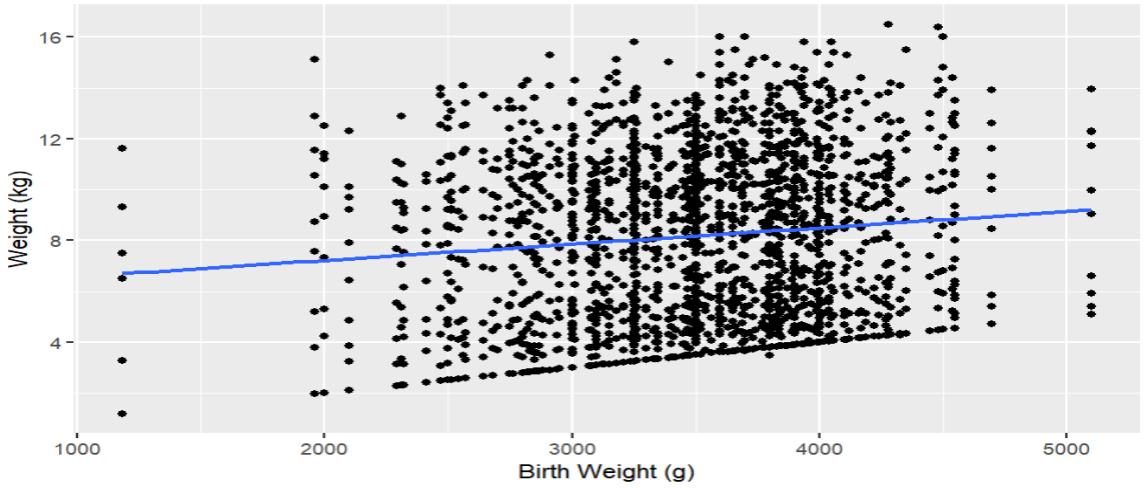


Figure 3: Relationship between birth weight in grams and weight measurement in kg



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

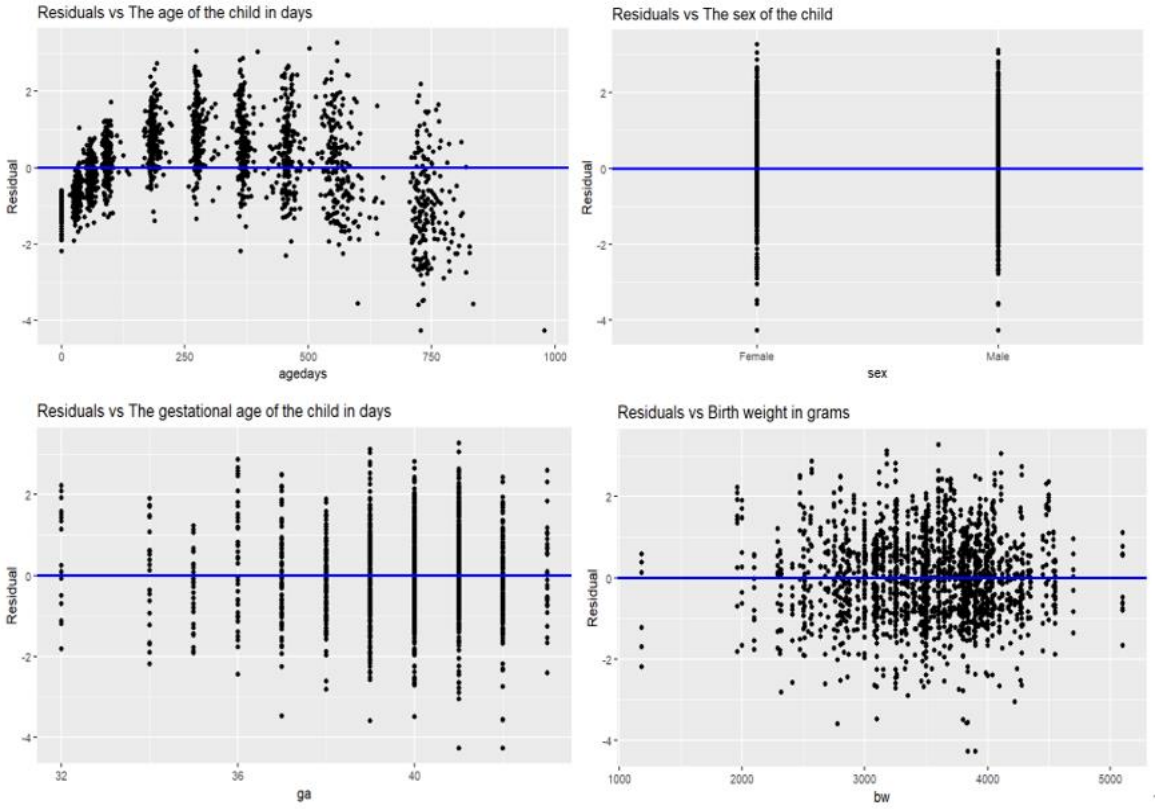


Figure 5: Scatterplot of the residuals against fitted values

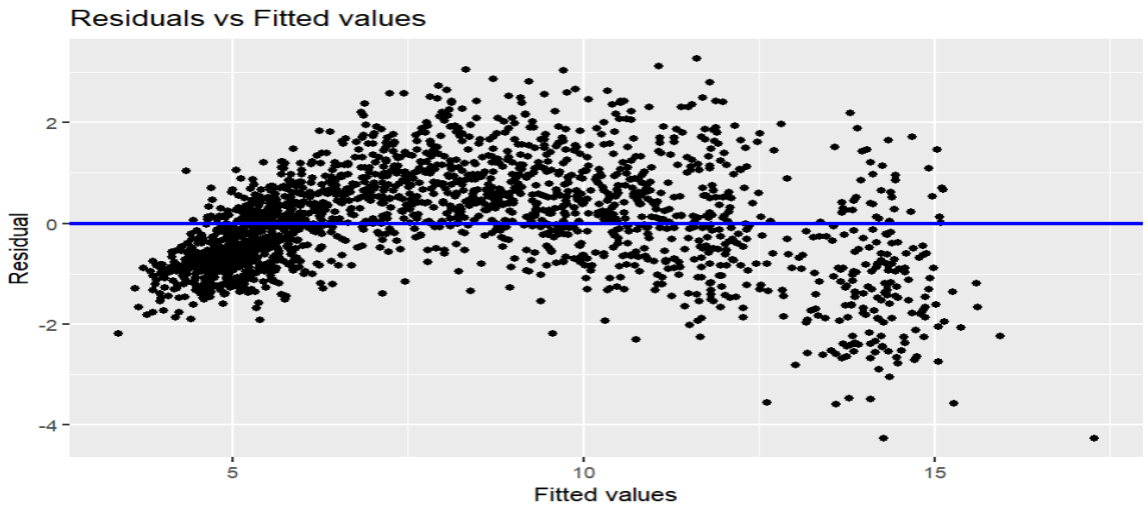


Figure 6: Histogram of residuals

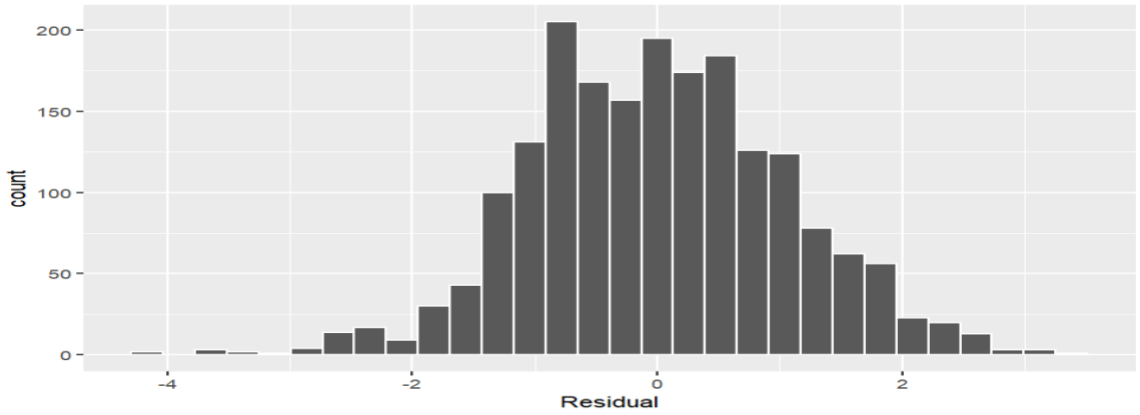
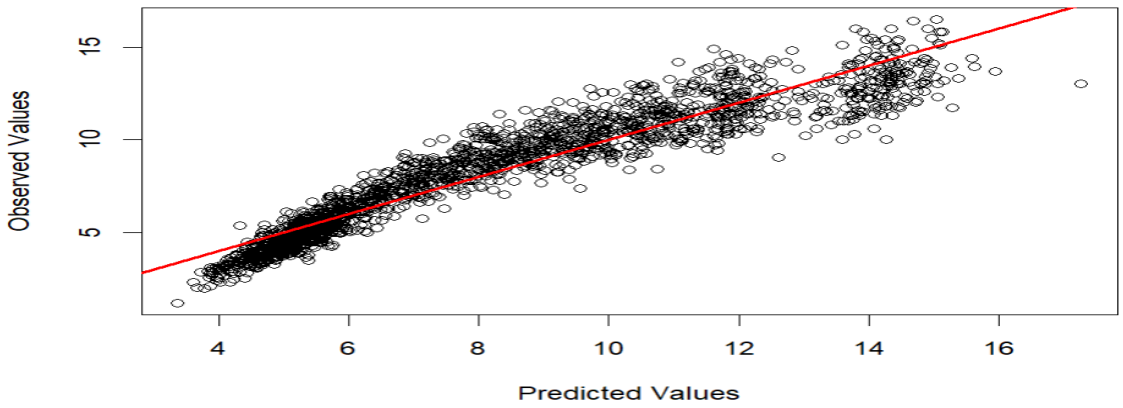


Figure 7: Plot of observed and predicted values



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

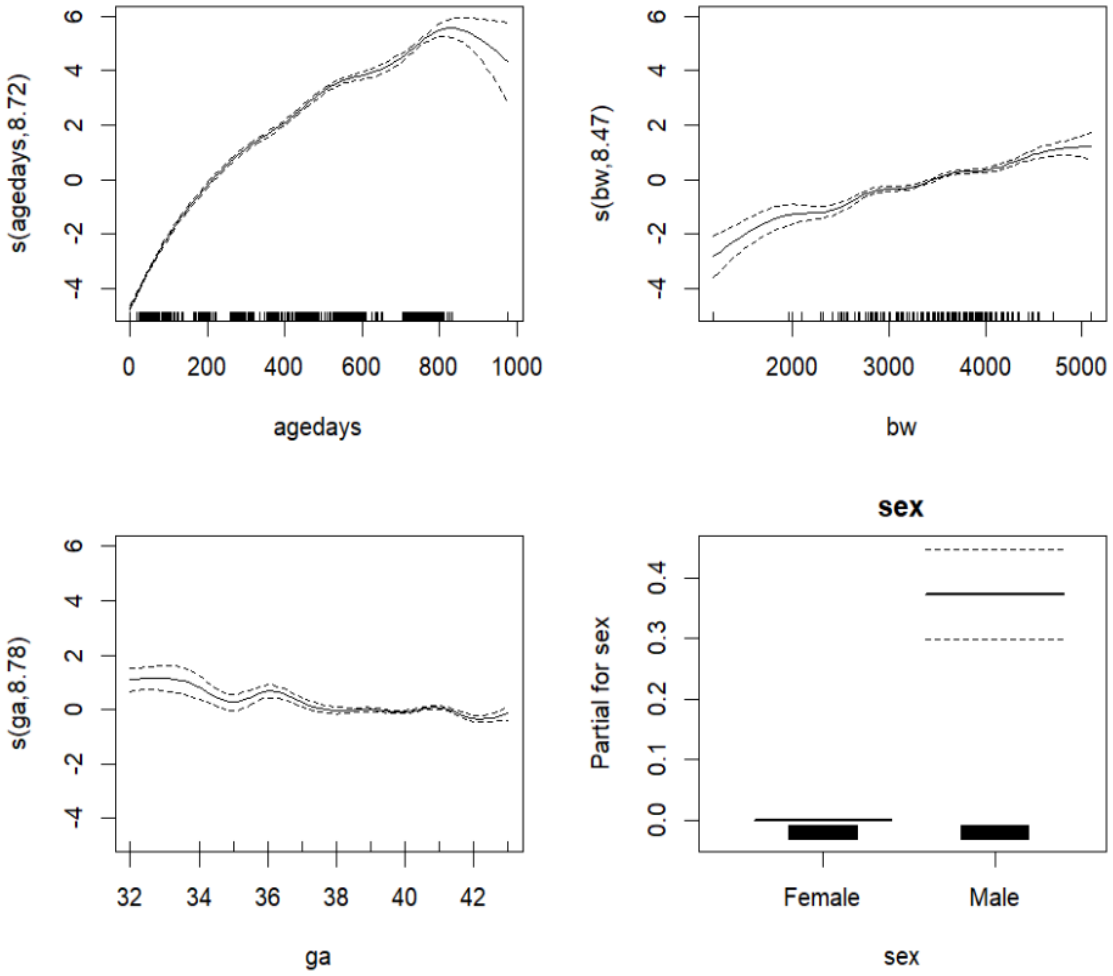
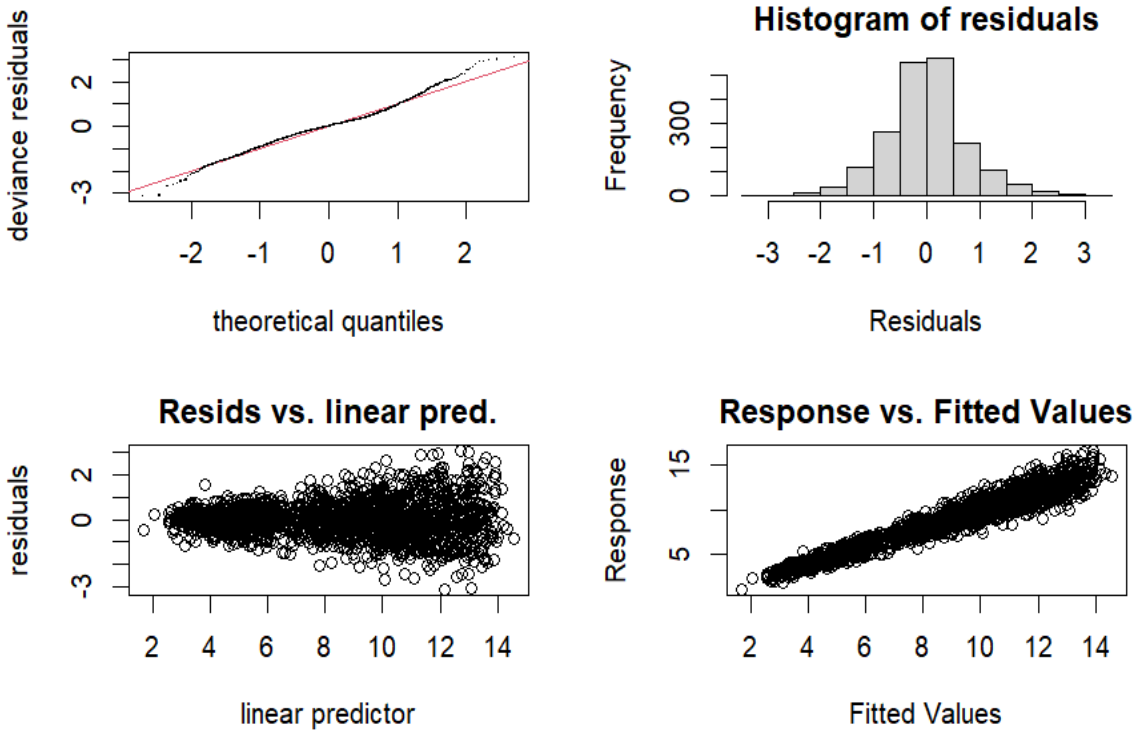


Figure 9: Diagnostic plots of generalized additive model



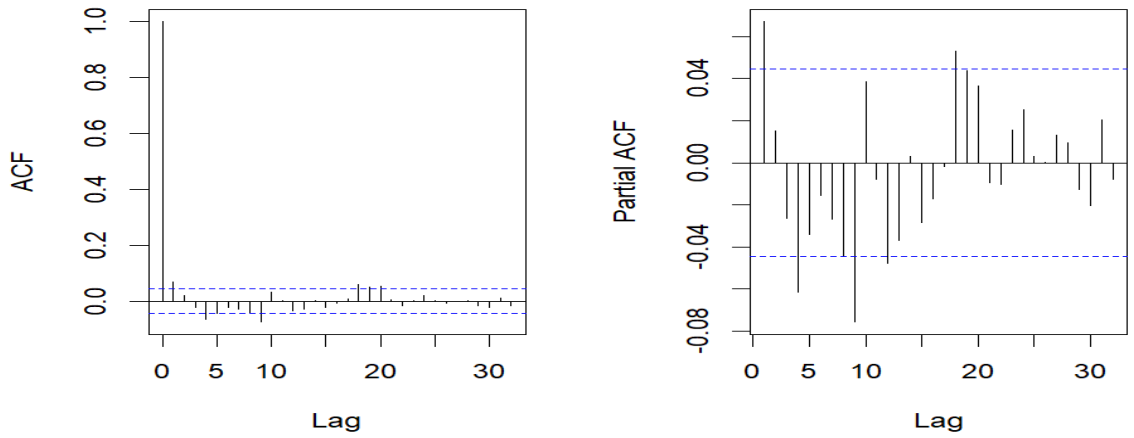
4.4 Generalized Additive Mixed Model Results

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

CONCLUSION

The major achievement of this study was how to choose the most suitable model to predict weight gain in children. In order to achieve this target, linear regression model, generalized additive model and generalized additive mixed model were developed. Then this article constructed Z-scores data to substitute raw data. By comparing with the AIC and BIC, the generalized additive mixed model with Z-scores is the best fit model. In addition, this article also found that all explanatory variables have the significant impact on the subsequent growth in infants.

The models proposed in this article can help people follow the growth trajectory of their children. However, the analysis and research work of this article still has some shortcomings. For example, these above models have the propensity to overfit. They also have no biological explanation. In addition, there is still a lot of work to be completed regarding the application of these models in real practice. In order to solve these limitations, machine learning should be more fully applied in the future.

REFERENCES

Anderson, D., & Burnham, K. (2004). Model selection and multi-model inference. *NY: Springer-Verlag*, *63*(2020), 10.

Baayen, R. H., van Rij, J., De Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by Generalized Additive Mixed Models. *Mixed-effects regression models in linguistics* (pp. 49-69).

Benoit, K. (2011). Linear regression models with logarithmic transformations. *London School of Economics, London*, *22*(1), 23-36.

Curtis, A. E., Smith, T. A., Ziganshin, B. A., & Elefteriades, J. A. (2016). The mystery of the Z-score. *aorta*, *4*(04), 124-130.

DiPietro, J. A., & Allen, M. C. (1991). Estimation of Gestational Age: Implications for Developmental Research. *Child Development*, *62*(5), 1184–1199.

Diskin, M. H. (1970). Definition and uses of the linear regression model. *Water Resources Research*, *6*(6), 1668-1673.

Fahrmeir, L., & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*(2), 201-220.

Groll, A., & Tutz, G. (2012). Regularization for generalized additive mixed models by likelihood-based boosting. *Methods of information in medicine*, *51*(2), 168–177.

Hanley, J. A. (2016). Simple and multiple linear regression: sample size considerations. *Journal of clinical epidemiology*, *79*, 112-119.

Hastie, T., & Tibshirani, R. (1995). Generalized additive models for medical research. *Statistical methods in medical research*, *4*(3), 187-196.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371-386.

Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, *69*(2), 499-513.

Janson, L., Fithian, W., & Hastie, T. J. (2015). Effective degrees of freedom: a flawed metaphor. *Biometrika*, *102*(2), 479–485.

Jbilou, J., & El Adlouni, S. (2012). Generalized additive models in environmental health: a literature review. *Novel Approaches and Their Applications in Risk Assessment*, *120*, 2014-2016.

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*(7), 2372-2387.

Pomerance, H. H., & Krall, J. M. (1981). Linear regression to approximate longitudinal growth curves: revised standards for velocity of weight and length in infants. *Pediatric research, 15*(10), 1390-1395.

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 145-158.

Regnault, N., Gillman, M. W., Kleinman, K., Rifas-Shiman, S., & Botton, J. (2014). Comparative study of four growth models applied to weight and height growth data in a cohort of US children from birth to 9 years. *Annals of Nutrition and Metabolism*, *65*(2-3), 167-174.

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *54*(3), 507–554.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of school psychology*, *52*(2), 149-178.

Shafi, M. A., & Rusiman, M. S. (2015). The use of fuzzy linear regression models for tumor size in colorectal cancer in hospital of Malaysia. *Applied Mathematical Sciences*, *9*(56), 2749-2759.

Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English. *Journal of Phonetics*, *70*, 86-116.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, *62*(4), 1025-1036.

Yu, K., Park, B. U., & Mammen, E. (2008). Smooth backfitting in generalized additive models. *The Annals of Statistics*, *36*(1), 228-260.