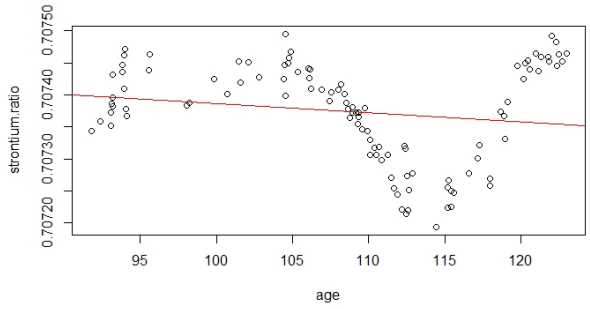
1 Introduction

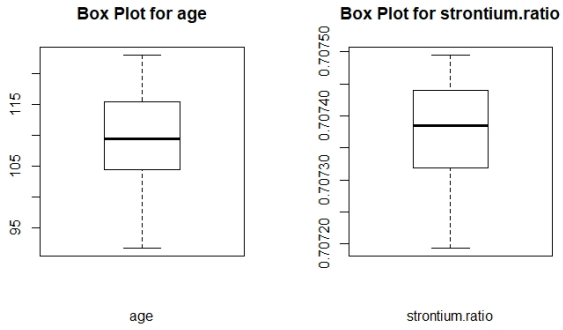
The objective of this study is to find the relationship between the dependent variable (strontium.ratio) and the independent variable (age) which are for investigating dating of a target fossil. In this study, 106 observations are given to analyze in statistical methods including parametric regression model and nonparametric model if it is necessary.

2 Descriptive Statistics

In statistical analysis, scatter plot of a given dataset is a method to approach for analyzing a dataset. Figure 1 represents that the scatter plot of 106 observations of the fossil dataset. As can be seen, the dataset is more suitable to follow a higher-degree polynomial regression model than a linear model. Also, there are no outliers identified from the boxplots in figure 2.



(a) Figure 1: Scatter plot with fitted linear line



(b) Figure 2: Box Plots

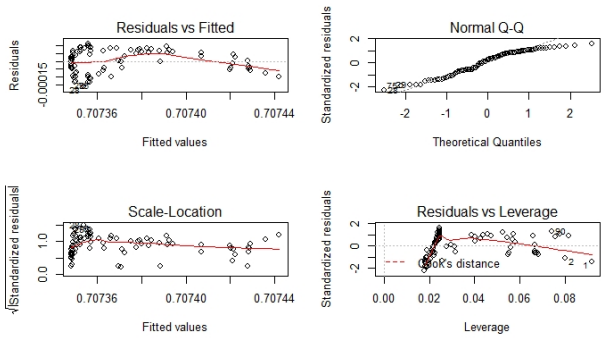
3 Parametric Regression

3.1 Polynomial Regression

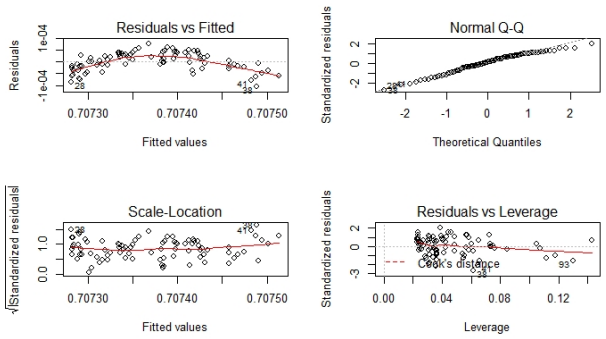
A quadratic regression model and a cubic model use for polynomial regression in this study. In statistical analysis, to confirm a model is a valid regression model, the model assumptions have to be satisfied. Diagnostics plots for the quadratic and the cubic model of the data set can be seen in figure 3 and 4, respectively. In figure 3, Q-Q plot (top-right) for the quadratic model reveals that the regression model tends to derail its straight line at the tails of the plot. In addition, the Scale-Location plot (bottom-left) has evidence that the quadratic regression model is not homoscedasticity because residuals are highly concentrated on the left side of the plot. The Scale-Location plot (bottom-left) for the cubic model in figure 4 shows the residuals are evenly distributed. In addition, for the normality assumption of the cubic model, Q-Q plot (top-right) in figure 4 represents that the regression model provide no evidence to conclude

non-normality. However, residuals vs fitted values plot (top-left) in figure 4 reveals that the curved line indicates that the regression model does not meet the condition for normality.

For a clear decision on the normality condition and homoscedasticity, the Shapiro-Wilk test for normality condition and the Breusch-Pagan test for constant variance use to conclude the model assumptions. The outputs of the tests are in table 1 and table 2. For normality confirmation, as the null hypothesis of the Shapiro Wilk test is a model is distributed normally, the result of the quadratic model rejects the null hypothesis since the p-value is less than α=0.05. On the other hand, the p-value of the cubic model is 0.07982 where greater than 0.05. Therefore, it can be concluded that the null hypothesis of the cubic model is not rejected. Table 2 is the outputs of the Breusch-Pagan test which is able to make a decision for homoscedasticity for each polynomial model. The null hypothesis of the test assumes that a testing model has homoscedasticity. In this case, the null hypothesis of the quadratic model has to be rejected, whereas the test for the cubic model is not rejected. In R software, bptest function in "lmtest" package is able to perform the test.



(c) Figure 3: Quadratic polynomial model



(d) Figure 4: Cubic polynomial model

Table 1: the Shapiro-Wilk test

Model	p-value
quadratic model	0.0006954
cubic model	0.07982

Table 2: the Breusch-Pagan test

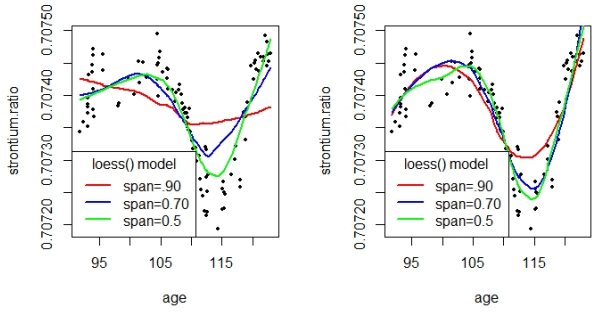
Model	p-value
quadratic model	0.0494
cubic model	0.5389

Based on the diagnostics plots and results from the tests, it can be concluded the cubic regression model is more desirable model than the quadratic model. Therefore, the equation of the cubic model is as below:

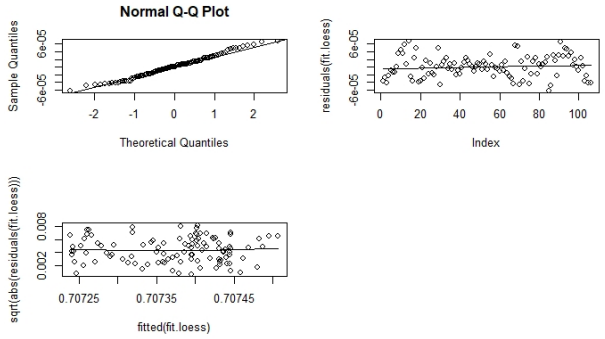
$$strontium.ratio = 5.800e^{-01} + 3.618e^{-03}age - 3.407e^{-05}age^2 + 1.064e^{-07}age^3 \qquad (1)$$

4 Nonparametric regression

A nonparametric method can be considered for this statistical analysis since the linear regression model is not sufficient to interpret the association between the response variables(strontium.ratio) and the independent/predictor variable (age). In this case, we use the method of the local polynomial method which is able to combine robustness and local fitting. In R software, loess function is able to perform the local polynomials method. The first step of the method is to find a proper window span with an optimal polynomial degree for the sake of capturing most information of observations in the dataset. For searching a sufficient nonparametric model, we test linear degree with three different span options; 0.90, 0.70 and 0.50. Same as the linear degree, second degree with the same span values also performs in this study. As can be seen in figure 5, the curved line of the second degree with span=0.5 reveals that most suitable nonparametric model as capturing most observations of the dataset according to the right plot of figure 5. In figure 6, the diagnostic plots of the loess function nonparametric method with the second degree and span=0.5 provide appropriateness of the model. According to the diagnostic plots of the nonparametric model, there is no critical evidence identified that the model is either non-normality or heteroscedasticity. Therefore, the nonparametric model with the second degree and span=0.5 can be determined as a valid nonparametric model. Prediction result with 95% confidence interval and the range from 95 to 125 is in table 3. Also, table 4 provide the model is nonlinear since the p-value of the nonlinearity test is zero.



(e) Figure 5: Left(degree 1) and Right (degree 2)



(f) Figure 6: diagnostic plots of degree 2 and span=0.5

Table 3: Prediction with 95% CI, by=5

Prediction	Lower	Upper
0.7074114	0.7073976	0.7074251
0.7074316	0.7074113	0.7074519
0.7074399	0.7074261	0.7074538
0.7073353	0.7073234	0.7073472
0.7072405	0.7072257	0.7072554
0.7073920	0.7073800	0.7074040

Table 4: Nonlinearity test

	test result
F-value	181.164
F-critical	2.295901
p-value	0

5 conclusion

In this study, we investigate to search a proper regression model to interpret and predict the relationship between the ratios of strontium isotopes and fossil age. Due to the nature of the dataset, higher-degree polynomial regression models use for this statistical analysis and then a cubic regression model has a superior outcome in the model assumptions. In the last section of this study, the quadratic degree with span=0.5 regression model in local polynomial method performs optimal interpretation of this statistical analysis.