Modelling Stillbirths with Natural Cubic Splines

by

Jie Yao

BMed, Zhejiang University, China, 2016

Submitted to the Graduate Faculty of

Department of Biostatistics

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Jie Yao

It was defended on

May 30, 2019

and approved by

Hong Wang, PhD, Research Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Tianjiao Chu, PhD, Associate Professor,
Department of Obstetrics, Gynecology & Reproductive Sciences
School of Medicine, University of Pittsburgh

Thesis Advisor: Gong Tang, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Gong Tang, PhD

Modelling Stillbirths with Natural Cubic Splines

Jie Yao, MS

University of Pittsburgh, 2019

Abstract

Fetal deaths such as spontaneous abortion and stillbirth are the most common adverse events during the process of pregnancy. According to the Center for Disease Control and Prevention, in 2016, about 1 in 10 babies failed to reach full term birth in the U.S. Spontaneous abortion occurs when the fetus dies in the uterus prior to 20 weeks of gestational age and life birth is not viable then. After 21 weeks, stillbirth may occur when the baby dies during delivery and the risk of stillbirth changes dramatically over the late course of pregnancy. Current studies focus on assessing the overall risk of stillbirth and discovery of preconception markers for the overall risk. In this thesis, we propose a flexible logistic regression model with time-varying, modelled by natural cubic splines, to study the temporal pattern in the risk of stillbirth over time and characterize the temporal association between preconception markers and the risk of stillbirth accordingly. The proposed method was illustrated via analysis of pregnancy outcome data from the Longitudinal Indian Family HEalth (LIFE) study and simulation studies.

Public health importance: Application of the proposed method provides accurate assessment of the risk of stillbirth over the course of pregnancy and identify potential preconception makers which leads to the development of novel intervention strategies and treatments and improvement in the well-being of both mothers and babies.

Table of Contents

List of Tables

List of Figures

Preface

I would first like to thank my thesis advisor Dr. Gong Tang for his excellent instruction during the process. The suggestions and guidance really help me a lot.

I would also thank Dr. Hong Wang and Dr. Tianjiao Chu for serving on my thesis committee.

1.0 Introduction

Pregnancy, also known as gestation, is the process that maintains the continuation of the human race. Since pregnancy is an important part of human reproduction, research on pregnancy is of great interest to many investigators. Human pregnancy typically lasts 40 weeks or about 9 months, and is divided into three stages: first trimester, second trimester and third trimester [1]. The first trimester is defined as 1 to 12 weeks of gestation. The zygote divides rapidly and important organs such as the heart, lungs, and brain start to develop during this period. For the second trimester, which is from 13 to 27 weeks of gestation, the function of each organ tends to perfect, but some organs are inactive despite being fully formed. In the third trimester (28-40 weeks of gestation), fetal skeletal development is essentially complete. The fetus has reached the stage of complete development, moves down to its underbelly and turns, ready to be born after 37 weeks of pregnancy.

During pregnancy, many adverse events could occur and cause the death of the infants. Spontaneous abortion (SAB), stillbirth, and preterm birth (PTB) are common abnormal outcomes of pregnancy. Spontaneous abortion refers to the natural death of an embryo or fetus before the 20 weeks of gestation, before a baby can survive independently [2]. In other words, a baby would have a chance of surviving only if he or she is delivered after 20 weeks of gestation and that chance of survival increases over time. According to MedlinePlus, 10-25% of pregnant women lose their babies due to miscarriage, and most cases of miscarriage happen in first trimester [3]. In contrast to miscarriage, a stillbirth refers to the loss of a baby after 20 weeks of pregnancy [4]. Stillbirths babies do not have any vital signs when born. Statistics from the National Center for Health Statistics reveal that approximately 24,000 stillbirths occurred in the U.S. in 2014 [5]. Depending

on the time of infant death, a stillbirth can be classified into three categories. If a baby dies between 20 and 27 weeks of gestation, the death would be an “early stillbirth”. If the death occurs between 28 and 36 weeks of pregnancy, then it is a “late stillbirth”. A term “stillbirth” is the death of a baby after 36 weeks of pregnancy [4]. Preterm birth is defined as a baby who is delivered before the 37^{th} week of gestation [6]. Based on data from the National Center for Health Statistics, 9.85% of babies born in the U.S. in 2016 were preterm births [7]. Due to their immature organs, these preterm babies have a high risk of diseases, such as pulmonary disease, brain development problems, and vision or hearing problems. In addition to the health problems of preterm babies, the ethical problem is also difficult to solve. For instance, if a mother needs to deliver her baby between 21 and 37 weeks of gestation, there is the problem of deciding whether to provide professional life support for this premature infant. This is especially true if the baby is born closer 21 weeks of gestation, as there is no clear guide instructing doctors on when to implement intervention for higher probability of saving infants.

Recent studies have focused on the determining the threshold of viability. Fetal viability describes the potential of a baby outside the uterus. The threshold of viability changed to 22 weeks of gestation in 1990. These studies highlight the theoretical limit of time when infants can be saved after intervention and suggest intervention should be given as early as possible. However, there is no related study to model the survival rate of infants born at 21 weeks, though this could provide a comprehensive and clear understanding of infant survival after 21 weeks of gestation. Hence, it is unknown whether providing life support as early as possible is beneficial in practice. Under this scenario, having a model for preterm live birth during this critical period will facilitate an understanding of the survival of preterm birth babies over time, as well as which markers are informative for survival and influence patient care and decision making during this important

process. With an estimate of how the probability of survival evolves during this critical period, clinicians will have a better idea of how and when to take necessary steps to prevent mothers from losing their babies. For example, doctors can determine what time to give intervention is more beneficial because they know the probability of saving babies at this time point or they may try to control some maternal health indicators associated with the survival rate of stillbirths to improve the chances of a fetus being born alive. Hence, it is necessary to develop an accurate model for stillbirth to reflect the temporal nature and guide the decision making and consideration of necessary interventions in those pregnancies.

As our outcome variable is fetal death, which is a binary variable, we need to find a model to fit the binary data. Logistic regression is one of the most commonly used methods to study the association between covariates and binary outcomes. However, we cannot directly use a simple logistic regression model for this process that assumes that the relationship between independent variables and the associated odds ratios is linear. According to the previous study [8], the viability of a baby increases by 3-4% per day between 23 and 24 weeks of gestation, so there is an association between the survival rate and time, but the association may not follow a linear function. In this case, we would resort to a more flexible logistic regression model with time-varying coefficients. Natural cubic spline is used to model the time-varying coefficients and provide different functional forms for different time intervals, in other words, the model fits a curve that smoothly passes through defined data points which is also called knots, so the fitted curve could be flexible. As natural cubic spline is continuous at every knot and piecewise-linear beyond the boundary knots, this method would provide a robust characterization of the association between time and the likelihood of a preterm live birth.

In this study, we aimed to access the relationship between gestational age of mothers and the survival rate of babies after delivery, as well as explore factors other than gestational age that influence fetal death. Using natural cubic splines, the time of viability could be separated into several intervals to cope with the situation where fitting just a logistic model is not proper. We expected to discover how the probability changes with gestational age and obtain enough information to determine whether to implement or decline life support for preterm babies at different weeks of gestations. We additionally investigated whether there are other factors that influence the chance of survival after 21 weeks' gestation, such as AHDL or LDL.

2.0 Natural Cubic Splines

Logistic regression provides a powerful tool to assess the association between predictors and a binary outcome variable [9]. In a typical logistic regression model, the logarithm of odds is modeled as a linear function of the predictors and the association is reflected via a multiplication factor to the odds, corresponding to a one-unit change in the predictor of interest. Logistic regression models with polynomial components or spline components have also been used to model the nonlinear association between predictors and the log-odds [10]. The predictors and the binary outcome could be cross-sectional or follow a temporal sequence, for example, the predictors are baseline preconception maternal characteristics or markers and the outcome is stillbirth in a pregnancy. Although a logistic regression model could help identify potential predictors for the chance that a pregnancy ends in stillbirth and provide the magnitude of the corresponding association in odds ratios, it would not provide a satisfactory description of the temporal nature of such an association that the risk of a stillbirth decreases over time after gestational age of 21 weeks. A logistic regression model with time-varying coefficients is necessary to model this time-varying association [11, 12]. Within that general framework, we introduced a natural cubic spline-based method to assess the time-varying relationship between preconception markers and the risk of stillbirth over time.

In linear regression models, the mean response is modeled as a linear function of predictors. Smoothing splines are sometimes incorporated to model the non-linear association between the mean response and some predictors. Compared to linear regression, which is fitted by least squares, spline functions can provide a more flexible model of the association, and one can control the over-fitting through a roughness penalty approach [13]. We know that second derivatives are usually

used to evaluate the shape of a curve. If a curve is twisted somewhere, the second derivative is high at this point. However, we cannot just integrate second derivatives together as positive and negative second derivatives can be integrated to a low value when the shape of a curve is extremely fluctuant. In general, the way to quantify the roughness of a curve is to measure the integration of the squared second derivatives $\int_a^b \{g''(t)\}^2 dt$ over the interval [a, b], where $g(t)$ represents the curve.

Natural cubic splines are the smoothest splines to interpolate a set of data points. For selected knots $a<t_1<t_2<t_3<\ldots t_n<b$ in an interval [a, b], a function g(t) is said to be a natural cubic spline when the following three conditions are met [13].

(i) There is a function of cubic polynomial in each interval (a, t_1), (t_1, t_2) …, (t_n, b).

(ii) To pass smoothly at each time point, the first and second derivatives should be continuous at each time point t_i.

(iii) The value of the second and third derivatives at boundary knots t_1 and t_n are zero, in other words, they are linear functions beyond t_1 and t_n.

The reason why natural cubic splines are said to be the smoothest interpolators among all of functions interpolating data points on [a, b] are based on the previous mentioned criteria: $J(g) = \int_a^b \{g''(t)\}^2 dt$. That is to say, the penalty $J(g)$ is minimized for natural cubic splines when compared to any other functions that interpolate the same data points.

Green and Silverman (1994) illustrated a proof for this [13]. Let $\hat{g}(t)$ be an interpolator of data over [a, b] where $g(t)$ is a natural cubic spline interpolant of the same data. Let $h(t) = \hat{g}(t) - g(t)$. We need to calculate the penalty term

$$J(\hat{g}) = \int_a^b (g''(t) + h''(t))^2 dt = \int_a^b g''(t)^2 dt + 2\int_a^b g''(t)h''(t)dt + \int_a^b h''(t)^2 dt$$

H(t) should be equal to 0 at $t = t_i$, for i =1, …, n. Indeed, $g''(t)$ is equal to zero at boundary points a and b and $g'''(t)$ is constant over each interval (t_j, t_{j+1}), the value of which is $g'''(t_j^+)$. Via integration by part,

$$\int_a^b g(t)''h(t)''dt = g''(b)h'(b) - g''(a)\,h'(a) - \int_a^b g'''(t)h'(t)dt = -\int_a^b g'''(t)h'(t)dt$$

$$= -\sum_{j=1}^{n-1} g'''(t_j^+)\int_{t_j}^{t_{j+1}} h'(t)\,dt = -\sum_{j=1}^{n-1} g'''(t_j^+)\{h(t_{j+1}) - h(t_j)\} = 0$$

Then we have

$$J(\hat{g}) = \int_a^b g''(t)^2dt + \int_a^b h''(t)^2dt \geq \int_a^b g''(t)^2dt$$

This means that if and only if $\hat{g}(t)$ and $g(t)$ are identical, can we obtain the optimal interpolation.

Here we introduce another concept called smoothing splines. To estimate the function of smoothing splines, a concept called the penalized least square must be defined:

$$S(g) = \sum_{i=1}^{n}\{Y_i - g(t_i)\}^2 + \lambda\int_a^b\{g''(x)\}^2dx \ .$$

As $\sum_{i=1}^{n}\{Y_i - g(t_i)\}^2$ is the residual sum of squares which represents how good a curve fit to the data and $\lambda\int_a^b\{g''(x)\}^2dx$ represents the roughness of a curve, minimizing $S(g)$ means we strike a balance between smooth and goodness of fit when the parameter λ is set, where integration is over the range of x and λ is a tuning parameter. As λ goes to zero, there is no penalty term in penalized least square, then the fitted curve would be quite overfitting, since it passes through every data point. As λ goes to infinity, the penalty is the main part of penalized least square and the curve would become a linear regression fit, which provides maximum smoothness (the second derivative is always 0), but there may be a poor fit. Interestingly, it can be demonstrated that minimizing the penalized least square for a fixed λ over the space of all continuous differentiable

functions leads to a unique solution. This solution is a natural cubic spline with knots at the data points since the minimizer of roughness penalties term would be obtained only if the function is a natural cubic spline as we proved above.

However, interpolation splines are used in our study, rather than smoothing spines, as smoothing splines have a flaw for study. As mentioned previously, though smoothing splines do not need to choose knots, they treat every data point as a knot, and the number of parameters of smoothing splines is the same as the number of data points. Few parameters among them are useful, so most parameters contribute a little to interpolation. The effect of λ is just to build a spline that is smoother than a spline with the same degrees of freedom would imply. In practice, fitting a model like this is extravagant. Hence, we chose to do interpolation splines with selected knots, and we just need to estimate as many parameters as the number of knots.

It is now appropriate to show the algorithm behind the interpolation of natural cubic splines [13]. Suppose g(t) is a natural cubic spline that has n fixed knots $a<t_1<t_2<\ldots<t_n<b$ in time interval [a, b], we define $g_i = g(t_i)$ and $\gamma_i = g''(t_i)$. Based on the requirements of a natural cubic spline, $\gamma_1 = \gamma_n = 0$. Let G = $\begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}$ and γ = $\begin{bmatrix} \gamma_2 \\ \vdots \\ \gamma_{n-1} \end{bmatrix}$.

We also construct the n * (n-2) matrix Q and the (n-2) * (n-2) matrix R for natural cubic spline representation. Before the construction of Q and R, we need to define $h_i = t_{i+1}-t_i$, for i = 1, …, n. Now, we have the elements for building Q and matrix R. Q is a band matrix and has the components $q_{i,j}$, for i = 1, …, n and j = 2, …, n-1, which are constructed by

$$q_{j-1,j}=h_{j-1}^{-1},\ q_{j,j}=-h_{j-1}^{-1}-h_j^{-1},\ q_{j+1,j}=h_j^{-1} \text{ for } j=2, \ldots, n-1,$$

and $q_{ij} = 0$ for $|i - j| \geq 2$.

The symmetric matrix R consists of components $r_{i,j}$, for i = 2, …, n-1 and j = 2, …, n-1, which gives

$$r_{i,i}=(h_i^{-1} -h_i)/3 \text{ for } i = 2, \ldots, n-1,$$

$$r_{i,i+1}= r_{i+1,i} =h_i/6 \text{ for } i = 2, \ldots, n-1,$$

and $r_{ij} = 0$ for $|i - j| \geq 2$.

If and only if the equation $Q^TG=R\gamma$ is satisfied, can we specify a natural cubic spline function. As the matrix G is known, we could obtain γ through the equation $\gamma = R^{-1}QG$. Therefore, for the interval $[t_L, t_R]$, if we define $g(t_L) = g_L$, $g(t_R) = g_R$, $g''(t_L) = \gamma_L$ and $g''(t_R) = \gamma_R$ and $h = t_R - t_L$.To express the function of natural cubic splines, we first need to get the expression of $g''(t)$. As the function of $g(t)$ is a cubic polynomial, $g''(t)$ is linear over $[t_L, t_R]$ with

$$g'''(t) = \frac{\gamma_R-\gamma_L}{h}.$$

Then we express g'' as

$$g''(t) = \frac{(t-t_L)\gamma_R+(t_R-t)\gamma_L}{h}.$$

By integrating twice, we can get

$$g(t) = \frac{g_R}{6h}(t - t_L)^3 + \frac{g_L}{6h}(t_R - t)^3 + C(t - t_L) + D(t_R - t).$$

After plugging $g(t_L) = g_L$ and $g(t_R) = g_R$ into this formula, the expression of g is shown as

$$g(t)=\frac{(t-t_L)g_R+(t_R-t)g_L}{h}+\frac{1}{6}(t - t_L)(t_R - t)\left\{(1 + \frac{t-t_L}{h})\gamma_R + (1 + \frac{t_R-t}{h})\gamma_L\right\}.$$

The above theorem is based on the properties of natural cubic splines. As we have defined $\gamma_i = g''(t_i)$ and the second derivatives are continuous over the interval, we get to know,

$$g''(t_j^+) = g''(t_j^-) = \gamma_j, \text{ for } j=2, \ldots, (n-1).$$

In addition to the second derivatives, the first derivatives are continuous at each knot as well. Through taking the obtained derivative of $g(t)$ we got above, for j = 2, …, (n-1), we can express $g'(t_i)$ as

$$g'(t_j^-) = \frac{g_j - g_{j-1}}{h_{j-1}} + \frac{1}{6} h_{j-1}(\gamma_{j-1} + 2\gamma_j)$$

and

$$g'(t_j^+) = \frac{g_{j+1} - g_j}{h_j} - \frac{1}{6} h_j(2\gamma_j + \gamma_{j+1}),$$

Therefore,

$$\frac{g_{j+1} - g_j}{h_j} - \frac{g_j - g_{j-1}}{h_{j-1}} = \frac{1}{6} h_{j-1}\gamma_{j-1} + \frac{1}{3}(h_{j-1} + h_j)\gamma_j + \frac{1}{6} h_j \gamma_{j+1},$$

This is the proof of the theorem regarding why $Q^T G = RY$ can specify a natural cubic spline.

3.0 A Logistic Regression Model with Time-Varying Coefficients

In application, we would introduce a concept called varying-coefficient models to apply natural cubic splines into a logistic regression. According Colin Wu [7], a linear time-varying coefficient model can have the form:

$$Y(t) = X^T(t)\beta(t) + \varepsilon(t),$$

where covariate vectors $X^T(t) = (1, X_1(t), \dots, X_k(t))^T$, $\beta^T(t) = (\beta_0(t), \beta_1(t), \dots, \beta_k(t))^T$, and $\varepsilon(t)$ is a mean 0 stochastic process with variance $\sigma^2(t)$ and $X^T(t)$ and $\varepsilon(t)$ are independent.

Through the application of generalized linear models, we can extend above model to logistic regression. The model would have a link function $g(\mu) = \log(\frac{\mu}{1-\mu})$ and $Y(t)$ is a binomial variate.

As the coefficients of a variable could change smoothly with time in time-varying coefficient models, time-varying coefficient models is better than linear regression models in modeling the interaction between time and other factors. The function of coefficients can be different forms. For example, if $\beta(t) = kt$, the function of t is a linear form so the model will have an interaction term kXt. It can also be flexible parametric representations such as Fourier series and smoothly nonparametric functions.

We used the function "ns()" of the R package "splines" to get the basic matrix of natural cubic splines. Through this function, for K knots, we can get K natural cubic splines basic functions as [14]:

$$N_1(T) \equiv 1,\ N_2(T) = t,\ N_{k+1}(T) = d_k(T) - d_{K-1}(T),$$

where

$$d_k(T) = \frac{(T-\xi_k)_+^3 - (T-\xi_K)_+^3}{\xi_K - \xi_k},$$

and ξ_k represents the value at k knots. Then we get $\beta(t) = \sum_{j=1}^{k} \beta_j N_j(t)$, where $N_j(t)$ is the basic matrix of natural cubic splines.

Since preterm births are those babies delivered before 37 weeks of gestation, and only those infants born after 21 weeks of gestation have a probability of surviving, we set 22 and 36 weeks as our boundary knots for the natural cubic spline. We chose 29 weeks of gestation as the interior knots to fit the model. As the interior knots are commonly set according to the quantile, we chose 29 weeks of gestation which is the middle of 22 and 36 weeks as the interior knot to fit the model.

4.0 Data Analysis

4.1 Description of the Data

Our motivating dataset are from the Longitudinal Indian Family Health (LIFE) study. LIFE is a prospective study which is conducted on the edge of the Hyderabad city in India. This study recruited 1,227 women between October 2009 and August 2011, among whom 80% were in the preconception stage and 20% were in their first trimester. In the first and third trimester, the health condition of the women was accessed, and the results were recorded along with the details of delivery. We included 330 women who had different pregnancy outcomes during weeks 21 through 38 of gestation. Among those women, 137 subjects had pregnancy outcomes during 21 to 36 weeks of gestation. Most of loss of pregnancies occurred before 28 weeks of gestation. We present the pregnancy outcomes from 21 to 36 weeks in table 1.

Table 1. Descriptive data for pregnancy outcomes from 21 to 36 weeks

	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Loss	2	2	2	6	1	2	3	3	0	0	0	1	0	1	0	1
PTB	0	0	0	0	0	1	1	0	3	1	8	6	10	12	25	46

Loss: Pregnancy loss
PTB: Preterm birth

Of these 137 subjects, 86 subjects have the records for covariates cholesterol, triglycerides, AHDL, LDL and VLDL. Table 2 summarizes the descriptive statistics for these five covariates.

Table 2. Descriptive data for covariates

Covariate	N	Mean(±SD)	Median (IQR)
cholesterol	86	4.99±0.17	4.98 (4.85, 5.1)
triglycerides	86	1.42±0.13	1.41 (1.32, 1.5)
AHDL	86	46.27±10.81	44.5 (38, 53.8)
LDL	86	87.18±23.4	85.1 (71.85, 98)
VLDL	86	15.61±11.05	12 (8.25, 17)

4.2 Application of the Logistic Regression Model with Natural Cubic Splines

Because our data included the loss of pregnancy, we set fetal death as the outcome variable. Additionally, since we do not know whether the covariate X is influential or whether there is an interaction between gestational age and X, we fit three models to explore, step by step, the relationship between gestational age and the probability of fetal death. We denote $logit(p) = \log(\frac{p}{1-p})$ in these models.

Model 1:

$$\text{Logit}\{\text{pr}(Y(t) = 1| t)\} = \beta_0(t)$$

In this model, Y(t) denotes the indicator for fetal death at t. For all $t \in [a, b]$, $\beta_0(t)$ is the natural cubic spline function of t. This model considers gestational age as the only variable that influences the probability of a baby's survival.

Model 2:

$$\text{Logit}\{\text{pr}(Y(t) = 1| x, t)\} = \beta_0(t) + \beta_1 x$$

In Model 2, Y(t) denotes the indicator for fetal death at t. X denotes the covariate. For all $t \in [a, b]$, $\beta_0(t)$ is the natural cubic spline function of t. β_1 is the constant coefficient for X. In this model, we add the covariate X to explore whether there are other factors that impact the change in the probability of fetal death.

Model 3:

$$\text{Logit}\{\text{pr}(Y(t) = 1 | x, t)\} = \beta_0(t) + \beta_1(t)x$$

Y(t) denotes the indicator for fetal death at t. X denotes the covariate. For all $t \in [a, b]$, $\beta_0(t)$ and $\beta_1(t)$ are natural cubic spline functions of t. In Model 3, we treat the coefficient of X as time-varying, so that the coefficients for covariate may differ in different time intervals.

As the fitting of Model 3 did not converge, we would just have the results for Model 1 and 2. Table 3 displays the coefficients and p-values for each natural cubic spline basis functions and different covariates. From this table, as we know that $\beta_0(t) = \sum_{j=1}^{3} \beta_{0j} N_j(t)$, where $N_j(t)$ is the basic matrix of natural cubic splines with knots at 22, 29 and 36 weeks of gestation, we discovered that the natural cubic spline in Model 1 is statistically significant; thus, the probability of fetal deaths decreases over gestational age. However, this table also illustrates that none of the covariates we selected are statistically significant, so we cannot conclude that there is a marker influencing the survival rate of infants.

Table 3. Coefficients of Model 1 and 2

Model	β_{01}		β_{02}		β_{03}		Marker		AIC
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value	
Model 1	24.8	0.04	-50.3	0.03	-15.6	0.01			38.696
Model2-cholesterol	24.8	0.04	-50	0.03	-15.5	0.01	-0.0015	0.94	40.689
Model2-triglycerides	24.8	0.04	-50.6	0.03	-15.7	0.01	0.0015	0.9	40.682
Model2-AHDL	24.8	0.05	-50.3	0.03	-15.6	0.01	0.00011	0.99	40.696
Model2-LDL	24.8	0.04	-49.8	0.03	-15.5	0.01	-0.0032	0.89	40.675
Model2-VLDL	24.9	0.04	-50.7	0.03	-15.7	0.01	0.0083	0.89	40.679

Figure 1 displays the predicted probability of fetal death over gestational age, along with its 95% confidence band, based on Model 1.

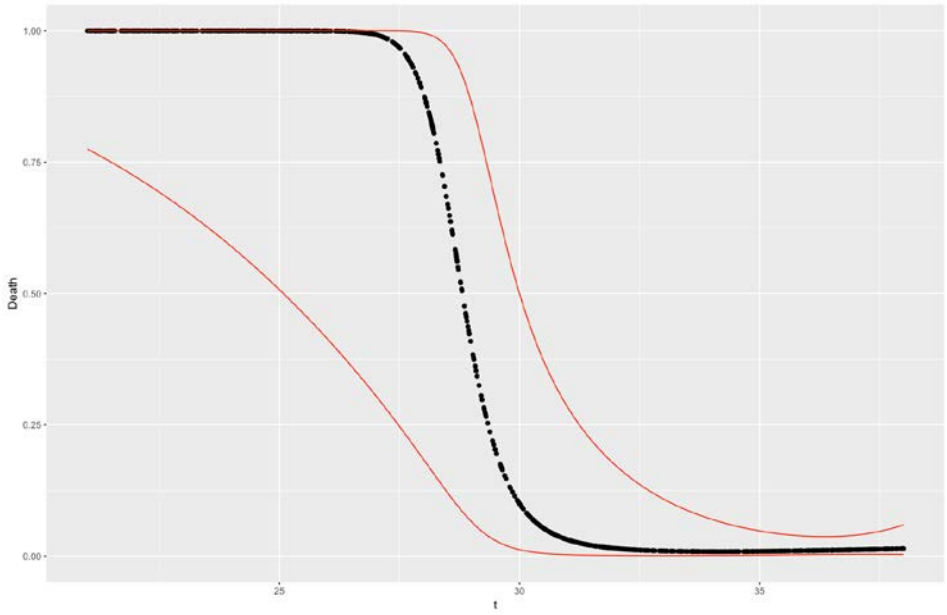

Figure 1. Probability of fetal death over gestational age

This clearly illustrates that the chance for stillbirth is almost 100% by week 26 and drops to about 50% around week 29. A live birth is almost certain when the gestational age reaches week 33.

We were not able to identify statistically significant preconception predictors in this analysis, which may due to the limited sample size.

5.0 Simulation

5.1 Simulation Settings

We conducted two simulation studies to evaluate how the natural cubic splines modeling the data perform for a similar process in which the odds of an event changes over gestational age. By comparing the parameters from the simulation settings to the parameters from natural cubic splines model, we could explore whether natural cubic splines fit the simulated data well. Two simulation settings are towards Model 2 and 3. For Model 2, as the parameter of covariate is a constant, we used empirical bias and empirical standard deviation to evaluate the difference between the fitted parameters and the true values of parameter. For Model 3, as the parameter of the covariate is a function of t, we would plot confidence bands of fitted parameters for each t = 22, 23, …, 38, so we can see whether the true parameters for each t can fall into this band.

5.1.1 Setting one

The sample size for each simulation study was n=1000. For each subject of a sample, we first simulated the time to delivery variable T=38-Z, where Z followed an exponential distribution with parameter λ and Z was truncated by 16. According to our data, there were 22 loss of pregnancies among 328 pregnancies, so the survival rate between 22 to 38 weeks of gestation was 0.93(306/328). We used $\lambda = -\frac{\log(0.93)}{16} = 0.002$ because the survival function for an exponential distribution is $S(z) = \Pr(Z \geq z) = e^{-\lambda z} = 0.93$, where z = 16. In the next, we simulated a covariate X randomly from the uniform distribution unif (0, 1). Denoting Y(t) as the indicator of

stillbirth when the fetus is delivery at time t, we modeled the probability of fetal death through the function

$$\Pr(Y(t) = 1|X) = logit^{-1}(\beta_0(t) + \beta_1 X)$$

with

$$\beta_0(t) = 1 - 0.01 * (t - 22) - 0.0006 * (t - 22)^3$$

and

$$\beta_1 \equiv 4.04.$$

Subsequently the stillbirth status Y(t) was simulated via a Bernoulli distribution with success probability $\Pr(Y(t) = 1|X)$ for each subject. Then we used the Model 2 to fit the simulated data and obtain the fitted parameter for X, which is defined as $\hat{\beta}_1$ and we calculated the $bias = \hat{\beta}_1 - \beta_1$ for each dataset. The above setting would be run for D=1000 times, so we can get the

$$\text{empirical bias} = \frac{1}{1000}\sum_{d=1}^{1000}(\hat{\beta}_1^{(d)} - \beta_1),$$

and

$$\text{empirical standard deviation} = \sqrt{\frac{1}{1000-1}\sum_{d=1}^{1000}(\hat{\beta}_1^{(d)} - \bar{\hat{\beta}}_1)^2},$$

where $\bar{\hat{\beta}}_1 = \frac{1}{1000}\sum_{d=1}^{1000}\hat{\beta}_1^{(d)}$.

Additionally, we use the Model 3 to fit the simulated data of setting one. For each sample D^d, we can calculate the $\hat{\beta}_1^{(d)}(t)$, where t = 22, 23, …, 38 and d = 1, …, 1000. For each t, we constructed the confidence band based on 2.5th and 97.5th percentiles. In the end, we plot $\{\hat{\beta}_{1ub}(t), \hat{\beta}_{1lb}(t), \beta_1\}$ together for t=22, 23, …, 38, so we could see whether the true value of β_1 would fall into this band.

5.1.2 Setting two

In setting two, the sample size for each simulation study was also n=1000, and the process to simulate the time to delivery and covariate X was the same as setting one. Denoting Y(t) as the indicator of stillbirth when the fetus is delivered at time t, we modeled the probability of fetal death through the function

$$\Pr(Y(t) = 1|X) = logit^{-1}(\beta_0(t) + \beta_1(t)X),$$

where

$$\beta_0(t) = 1 + 0.01 * (t - 22) - 0.02 * (t - 22)^3$$

and

$$\beta_1(t) = 16 + 0.2 * (t - 22) + 0.272 * (t - 22)^2 - 0.015 * (t - 22)^3.$$

In the same way, we ran the simulate D=1000 times. For each sample D^d, we could calculate the $\hat{\beta}_1^{(d)}(t)$, where t = 22, 23, …, 38 and d = 1, …, 1000. For each t, we constructed the confidence band based on 2.5^{th} and 97.5^{th} percentiles. For example, we calculate the upper value $\hat{\beta}_{1ub}(t_{22})$, where this value was the 97.5^{th} out of 1000 $\hat{\beta}_1(t_{22})$ and similarly calculated the lower value $\hat{\beta}_{1lb}(t_{22})$ based on the 2.5^{th} out of 1000 $\hat{\beta}_1(t_{22})$. In the end, we plotted $\{\hat{\beta}_{1ub}(t), \hat{\beta}_{1lb}(t), \beta_1(t)\}$ together for t=22, 23, …, 38, so we could see whether the true value of $\beta_1(t)$ would fall into this band.

5.2 Simulation Results

5.2.1 Setting one

As we set the $\beta_1 = 4.04$, after repeating D=1000 times, we can get the empirical bias= $\frac{1}{1000}\sum_{d=1}^{1000}(\hat{\beta}_1^{(d)} - \beta_1) = 0.042$ and empirical standard deviation = $\sqrt{\frac{1}{1000-1}\sum_{d=1}^{1000}(\hat{\beta}_1^{(d)} - \bar{\hat{\beta}}_1)^2} = 0.44$. The empirical bias showed that there was negligible bias in the estimate of β_1.

This Figure 2 is the result of simulated data of setting one modelled by Model 3. The red line represents the upper band and the seagreen line represents the lower band. The black line represents the constant β_1 which equals to 4.04. It showed that true β_1 was in the confident band of $\hat{\beta}_1^{(d)}(t)$.

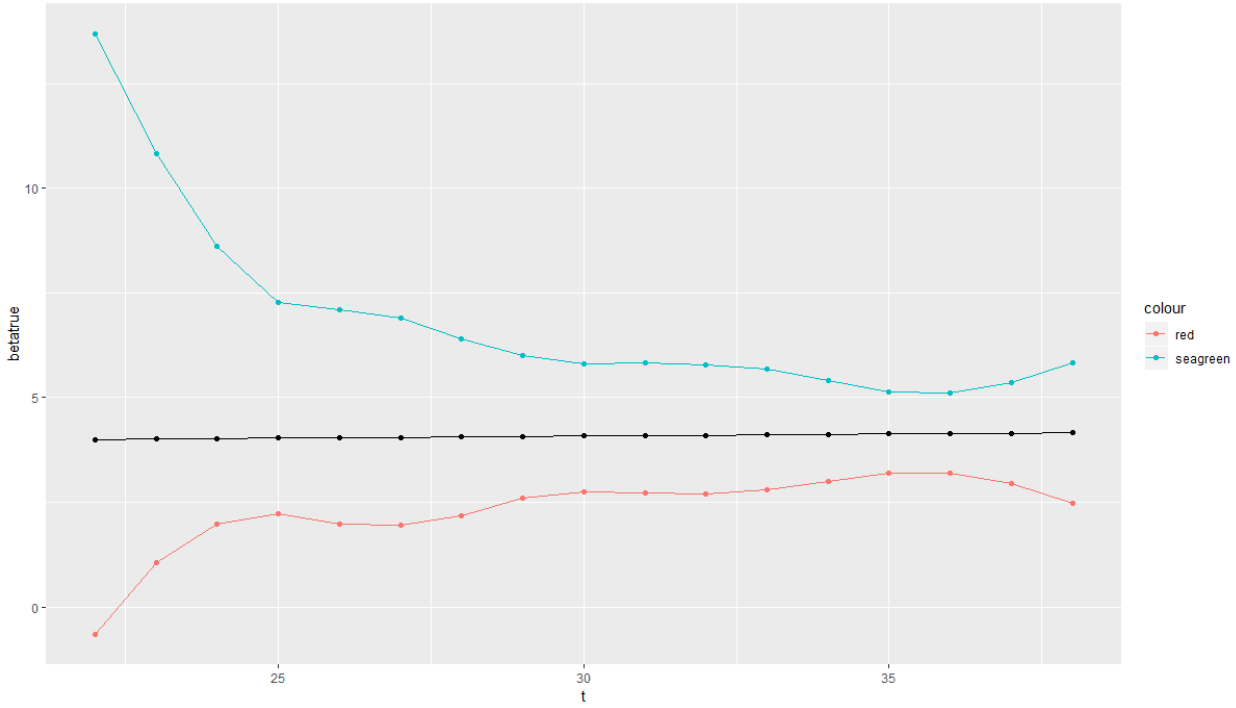

Figure 2. True β_1 and confidence band of $\widehat{\beta}_1$

5.2.2 Setting two

In Figure 3, the red points represent the true $\beta_1(t)$ for t=22, 23, …, 38, green points represent the upper band of $\hat{\beta}_1(t)$ and the blue points represented the lower band of $\hat{\beta}_1(t)$. From this figure, it is clear the true $\beta_1(t)$ all fall between the upper value $\hat{\beta}_{1ub}(t)$ and the lower value $\hat{\beta}_{1lb}(t)$.

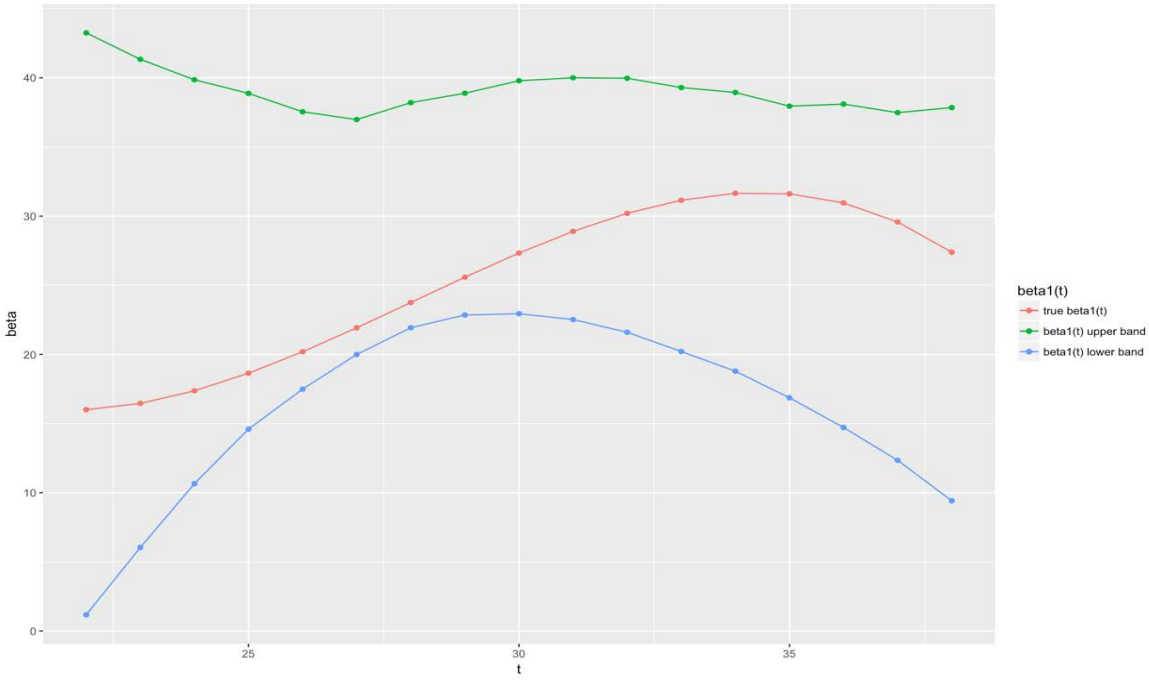

Figure 3. True $\boldsymbol{\beta_1(t)}$ and confidence band of $\boldsymbol{\hat{\beta}_1}$

6.0 Discussion

According to the results of the data analysis, Model 1 performed well in modeling the probability of fetal death over gestational age. However, there was no makers' coefficient being statistically significant in Model 2, and Model 3 did not converge.

From the simulation, we observed that Model 2 and 3 performed well in modeling the simulated data. For simulation setting one, the empirical bias and standard deviation were small, which indicates that the fitted coefficients were not far from the true coefficient and the coefficients were not widely spread. For simulation setting two, as the true coefficient could fall into the confidence band of fitted coefficients, we can say that natural cubic splines can model the data with time-varying coefficients.

Based on the result of our study, natural cubic splines can help us model a non-linear relationship. One advantage of the logistic regression models with time-varying coefficients, modeled by natural cubic splines is that it provides flexibility to allow the coefficients of a covariate to vary with other factors, such as time.

One limitation of our study is the small sample size. Although we find that a natural cubic spline can model a non-linear relationship well in simulation, we need to treat the association carefully. Future research should include more data into our study and try to find other markers that may influence the stillbirth, so that we can study the model of stillbirth more in depth.

Appendix: R sample codes

```
MyData <- read.csv(file="SAB_Yao.csv", header=TRUE, sep=",")
install.packages("RGeode")
install.packages("Rlab")
install.packages("gmodels")
install.packages("pastecs")
library(pastecs)
library(Rlab)
library(RGeode)
library(splines)
library(gmodels)
library(ggplot2)
MyData1<-data.frame(MyData)
MyData2<-na.omit(MyData1)
MyData3<-subset(MyData2,GA_ALL_CLEAN2016<=36)
#describe the data
describe(MyData2)
z<-ns(MyData1$GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36))
#fit model logit(pr(live))=beta0(t)
mylogit<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) ), data=MyData1, fam
ily = binomial(link="logit") )
summary(mylogit5)
#fit model logit(pr(live|x))=beta0(t)+X
mylogit1<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+CHOLESTEROL_REG ,
data=MyData2, family = binomial(link="logit"))
mylogit2<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+TRIGLYCERIDES_REG
 , data=MyData2, family = binomial(link="logit"))
mylogit3<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+AHDL_REG , data=My
Data2, family = binomial(link="logit"))
mylogit4<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+LDL_REG , data=MyD
ata2, family = binomial(link="logit"))
mylogit5<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+VLDL_REG , data=My
Data2, family = binomial(link="logit"))
#fit model logit(pr(live|x))=beta0(t)+beta1(t)X
mylogit6<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+ns(GA_ALL_CLEAN201
6, knots=c(29), Boundary.knots =c(22,36) )*CHOLESTEROL_REG  , data=MyData2, family = bi
nomial(link="logit"))
mylogit7<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+ns(GA_ALL_CLEAN201
6, knots=c(29), Boundary.knots =c(22,36) )*TRIGLYCERIDES_REG  , data=MyData2, family =
binomial(link="logit"))
mylogit8<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+ns(GA_ALL_CLEAN201
6, knots=c(29), Boundary.knots =c(22,36) )*AHDL_REG , data=MyData2, family = binomial(l
ink="logit"))
mylogit9<-
glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+ns(GA_ALL_CLEAN201
```

```
   6, knots=c(29), Boundary.knots =c(22,36) )*LDL_REG , data=MyData2, family = binomial(li
   nk="logit"))
32. mylogit10<-
   glm(SAB~ns(GA_ALL_CLEAN2016, knots=c(29), Boundary.knots =c(22,36) )+ns(GA_ALL_CLEAN201
   6, knots=c(29), Boundary.knots =c(22,36) )*VLDL_REG , data=MyData2, family = binomial(l
   ink="logit"))
33.
34. #plot the probability of death over time for model logit(pr(live))=beta0(t)
35. t<-runif(1000,min=21,max=38)
36. aa<-data.frame(GA_ALL_CLEAN2016=t)
37. beta0<-predict(mylogit,newdata = aa,type = "link", se=TRUE)
38. Death<-plogis(beta0$fit)
39. #calculate upper band and lower band
40. LL <- plogis(beta0$fit - (1.96 * beta0$se.fit))
41. UL <- plogis(beta0$fit + (1.96 * beta0$se.fit))
42. aa$lwr <- LL
43. aa$upr <- UL
44. #plot the probability of death over time and confident band
45. ggplot(data=aa, mapping=aes(x=t, y=Death)) + geom_point() +
46.   geom_line(data=aa, mapping=aes(x=t, y=upr), col="red") +
47.   geom_line(data=aa, mapping=aes(x=t, y=lwr), col="red")
48.
49.
50. #Setting one: simulation for model logit(pr(live|x))=beta0(t)+x
51. set.seed(11111)
52. sum<-0
53. dhat<-rep(NA,1000)
54. bhat<-rep(NA,1000)
55. #simulate 1000 datasets
56. for(s in 1:1000){
57. #generate time to delivery
58.   t2<-rexptr(n =1000, lambda =0.002 ,range=c(0,16))
59.  t3<-38-t2
60. #generate covarate based on uniform distribution
61. x1<-runif(1000, min=0, max=1)
62. b00<-1-0.01*(t3-22)-0.0006*(t3-22)^3
63. b11<-4.04
64. pr1<-plogis(b00+b11*x1)
65. z<-0
66. for(i in 1:1000){z[i]<-rbern(n=1,prob=pr1[i])}
67. newdata4<-data.frame(z,t3,x1)
68. #fit the function pr(y(t)=1|x)=logit^-1(beta0(t)+beta1*x)
69. mylogit18<-
   glm(z~ns(t3, knots=c(26,31), Boundary.knots =c(22,36) )+x1, data=newdata4, family = bin
   omial(link="logit") )
70. #calculate beta1 hat for each dataset
71. bhat[s]<-coef(mylogit18)[5]
72. #calculate absolute value of the difference between beta1 hat and true beta1
73. dhat[s]<-bhat[s]-b11
74. sum<-sum+dhat[s]
75. }
76. #calculate empirical bias
77. bias<-sum/1000
78. #calculate the empirical standard deviation
79. sd<-sd(bhat)
80.
81. #Setting two: simulation for model logit(pr(live|x))=beta0(t)+beta1(t)x
82. set.seed(12345)
83.  beta1<-matrix(NA,nrow=1000,ncol=17)
84.  for(s in 1:1000){
85.  t<-rexptr(n =1000, lambda =0.002 ,range=c(0,16))
```

```
  t1<-38-t
  x<-runif(1000, min=0, max=1)
 b0<-1+0.01*(t1-22)-0.02*(t1-22)^3
  b1<-16+0.2*(t1-22)+0.272*(t1-22)^2-0.015*(t1-22)^3

  pr<-plogis(b0+b1*x)
  for(i in 1:1000){y[i]<-rbern(n=1,prprob=pr[i])}
  newdata3<-data.frame(y,t1,x)
  newdata31<-data.frame(t1=c(22:38),x=0)
  newdata32<-data.frame(t1=c(22:38),x=1)
  #fit the function pr(y(t)=1|x)=logit^-1(beta0(t)+beta1(t)*x)
  mylogit16<-
glm(y~ns(t1, knots=c(29), Boundary.knots =c(22,36) )+ns(t1, knots=c(29), Boundary.knots
 =c(22,36) )*x, data=newdata3, family = binomial(link="logit") )
  newfit<-predict(mylogit16,newdata = newdata31)
  newfit1<-predict(mylogit16,newdata = newdata32)
   # calculate the beta1(t) hat for t=22,23, ...., 38
    beta1[s,]<-newfit1-newfit}
 betaub<-rep(NA,17)
 betalb<-rep(NA,17)
 betatrue<-rep(NA,17)
 #calculater the upper band and lower band for t=22,23, ...., 38 based on precentile
for(g in 1:17){betaub[g]=quantile(beta1[,g],c(0.975))
betalb[g]=quantile(beta1[,g],c(0.025))
betatrue[g]=16+0.2*(g-1)+0.272*(g-1)^2-0.015*(g-1)^3}
 betanew<-data.frame(t=c(22:38),betaub,betalb,betatrue,beta=0)
#plot beta1(t), upper band and lower band of beta1(t) hat for t=22,23, ...., 38
 ggplot(data=betanew, aes(x=t, y=beta)) +
  geom_point(data=betanew, aes(x=t, y=betatrue,col="black"))+
  geom_line(data=betanew, aes(x=t, y=betatrue,col="black"))+
  geom_point(data=betanew, aes(x=t,y=betaub,col="red"))+
  geom_line(data=betanew, aes(x=t, y=betaub,col="red"))+
 geom_point(data=betanew, aes(x=t,y=betalb,col="seagreen"))+
  geom_line(data=betanew, aes(x=t, y=betalb,col="seagreen"))+
  scale_color_discrete(name = "beta1(t)", labels = c("true beta1(t)","beta1(t) upper
band", "beta1(t) lower band"))


```

Bibliography

1. U.S. Department of Health & Human Service. (2019). *Stages of pregnancy*. Retrieved from https://www.womenshealth.gov/pregnancy/youre-pregnant-now-what/stages-pregnancy

2. Jessica L. Bienstock, Harold E. Fox, Edward E. Wallach, Clark T. Johnson, and Jennifer L. Hallock. (Eds.). (2012). *The Johns Hopkins Manual of Gynecology and Obstetrics (5th edition*). Philadelphia, PA. Wolters Kluwer Health.

3. Medicine. (2014). *Miscarriage*. Retrieved from https://medlineplus.gov/ency/article/001488.htm.

4. Centers for Disease Control and Prevention. (2019). *Stillbirth*. Retrieved from https://www.cdc.gov/ncbddd/stillbirth/facts.html

5. National Center for Health Statistics. (2016). *Cause of Fetal Death: Data From the Fetal Death Report, 2014*. Retrieved from https://www.cdc.gov/nchs/data/nvsr/nvsr65/nvsr65_07.pdf.

6. Centers for Disease Control and Prevention. (2014). *Preterm Birth*. Retrieved from https://www.cdc.gov/reproductivehealth/maternalinfanthealth/pretermbirth.htm

7. National Center for Health Statistics. (2018). *Describing the Increase in Preterm Births in the United States, 2014–2016*. Retrieved from https://www.cdc.gov/nchs/data/databriefs/db312.pdf.

8. Marilee C. Allen, Pamela K. Donohue, and Amy E. Dusman. (1993). The Limit of Viability -- Neonatal Outcome of Infants Born at 22 to 25 Weeks' Gestation. New England Journal of Medicine. 329 (22): 1597–1601. doi:10.1056/NEJM199311253292201

9. Walker, SH; Duncan, DB (1967). *Estimation of the probability of an event as a function of several independent variable*s. Biometrika. 54 (1/2): 167–178. doi:10.2307/2333860. JSTOR 2333860.

10. Hastie, Trevor and Tibshirani, (1986), Robert. *Generalized Additive Models*. Statist. Sci. 1(3), 297--310. doi:10.1214/ss/1177013604. https://projecteuclid.org/euclid.ss/1177013604.

11. Ramsay, J. O., and Dalzell, C. J. (1991), *Some Tools for Functional Data Analysis*, Journal of the Royal Statistical Society, Ser. B, 53, 539-572.

12. Wu, Colin, Chiang, Chin-Tsang and R. Hoover, Donald. (1998). *Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data*. Journal of The American Statistical Association 93. 1388-1402.

13. Green, P. J and. Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. London, UK. Chapman & Hall.

14. Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. New York, NY. Springer-Verlag New York Inc.