Analyses of Repeatedly Measured Blood Pressure Data from A Cohort Study with Splines

by

Jinhong Li

BMed, Huazhong University of Science and Technology, China,

2018

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

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Jinhong Li

It was defended on

April 17, 2020

and approved by

Janet M. Catov, PhD, MS, Associate Professor, Department of Obstetrics, Gynecology & Reproductive Sciences and the Department of Epidemiology, University of Pittsburgh

Yan Lin, PhD, Research Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

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

Analyses of Repeatedly Measured Blood Pressure Data from A Cohort Study with Splines

Jinhong Li, MS

University of Pittsburgh, 2020

Abstract

Cardiovascular disease (CVD) is a group of disease that involve the cardiovascular system. According to World Health Organization, it is the leading cause of death worldwide and hypertension is a major risk factor of CVD. Pregnancy is a special "window" for women and the physiological change could reflect enduring risk for CVD. In a recent study, researchers were interested in whether placental malperfusion could predict risk of CVD 8-10 years after delivery, and blood pressure (BP) is one of the endpoints of interest. In this thesis, we studied whether placental malperfusion is predictive for elevated BP. BP was repeatedly measured for three times during the office visit, thus the data had a longitudinal structure. One challenge is that BP fluctuates with regards to time during a day, and it is essential to adjust for the potential confounding effect of time in regression analyses. Splines provide a powerful tool to adjust for such relationship with abundant flexibility. In this thesis, natural cubic splines (NCS) and smoothing splines (SS) were considered and compared. As a consequence, application of the splines could identify significant predictive biomarkers, with flexible adjustment of the time effect. NCS is easier to use while SS is more flexibile.

Public health importance: Limited number of research have been done about the prognostic utility of placental malperfusion on risk of hypertension and CVD. Splines provide a powerful and flexible tool to characterize such a relationship.

Table of Contents

List of Tables

List of Figures

Preface

I am heartily thankful to my thesis and academic advisor, Dr. Gong Tang, for all the encouragement, guidance and support. Dr. Tang has been a wonderful advisor through my master's years at University of Pittsburgh, and this thesis could not be finished without him.

I also owe a deep sense of gratitude to my committee members, Dr. Yan Lin and Dr. Janet M. Catov, for agreeing to serve on my thesis committee and for their generous help and support on this thesis. I would like to thank Dr. Catov again for the permission to use the dataset for this thesis.

I would also like to express my gratitude to all faculty members, staff and students from the Department of Biostatistics of Pitt Public Health, who gave me tremendous help and support along my way of academic work and seeking further studies. I am thankful to Pitt for giving me this opportunity to join this warm family and spending a wonderful time.

1.0 Introduction

Cardiovascular disease is a life-course disease which involves the heart and blood vessels in human bodies, and the common types of CVD includes coronary heart disease, cerebrovascular disease, peripheral arterial disease and rheumatic heart disease.[1] CVD is the major cause of death in the United States, accounting for more than 28.8 percent of the total death in 2017.[2] The risk of CVD starts accumulating since the early stage of adulthood, and often relates to the onset of CVD later in life.[3] For women, pregnancy is a unique "window" to go through a 'stress test' of the cardiovascular system and has the potential CVD risk factors manifest in their early stage of life.[4]

Hypertension is a common and serious medical condition characterized by elevated blood pressure (BP) measurements.[1] According to the *2017 Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults* by American College of Cardiology, hypertension is diagnosed by the systolic blood pressure (SBP) measurement being ≥ 130 mmHg or the diastolic blood pressure (DBP) measurement being ≥ 80 mmHg.[5] Hypertension is one of the most important risk factors for CVD, and approximately half of the coronary heart disease and stroke cases are related to high BP.[6] Because BP is an important factor affecting the health of one's cardiovascular system, it is critical to detect hypertension early and treat accordingly. However, as much as 38% of hypertension cases were undetected before the age of 40.[7]

In a recent window study initiated in Magee-Womens Hospital in Pittsburgh, PA, 499 women were selected from an retrospective cohort of pregnant women with deliveries at the Magee-Womens Hospital in 2008 or 2009. The primary objective of this study is to identify clinical biomarkers that predict the risk of CVD. A primary endpoint is the BP measured at clinic visits, and the primary biomarker of interest is placental malperfusion. The research subjects' BP were measured at their clinical office visit about 8-10 years after their index delivery, with a standardized manner by trained research staff. The BP were measured after a five-minute rest, and were measured for three times on the non-dominant arm using an appropriately sized cuff with one-minute intervals between each adjacent measurements.

The research subjects' placental pathology were extracted from an existing dataset, as well as other clinical features such as BMI at pregnancy, preeclampsia, age and family history of heart disease. Placental malperfusion, referred to poor placental perfusion, has the main pathological characteristics as decidual vasculopathy, infarcts, abruption and advanced villous maturation.[8] It is known that placental malperfusion has well-established associations with adverse pregnancy outcomes (APO) such as preeclampsia, growth restriction and preterm delivery, and APO is related to increased risk for future CVD in women.[9, 10, 11, 12, 13] In this study, it was proposed that the pathological impairments converge in the placenta to make the placental malperfusion a common pathway to multiple APOs, and thus a consequential, unifying feature that links APOs to maternal cardiometabolic and microvascular risk after delivery.

The placenta malperfusion lesions were determined by both gross and microscopic placental pathology findings following a uniform consensus criteria described in the Amsterdam Placental Workshop Group Consensus Statement.[14] The gross findings are placental hypoplasia, infarction, and retroplacental hemorrhage and the microscopic findings are abnormalities of villous development, which includes distal villous hypoplasia and accelerated villous maturation. In this research, the evidence of malperfusion lesions were grouped in five domains including vasculaopathy, advanced villious maturation, infarction, fibrin deposition and perivillous fibrin, and marked as one union factor called "malperfusion".

One of the challenges in the study of BP is that BP fluctuates during the day and has a strong circadian pattern.[15] For each person, BP measurements are not constant during the day, thus the measurement time should be adjusted to account for potential confoundings. Splines is a powerful tool to provide a flexible adjustment for the fluctuating pattern. It fits multiple functional forms for each specified or default time intervals. Therefore it could fit very flexible curves according to the time range.

Natural cubic spline (NCS) is a widely used tool to provide a flexible fit between a continuous predictor and the continuous outcome. It is specially useful when such a relationship is not the primary concern of the study, but it is necessary to be adjusted for the data analysis. To implement NCS, polynomial curves which are continuous up to the second derivatives are fitted in each pre-defined time intervals. Basically, it is a piecewise smooth curve with

different functional forms on each interval and continuous at the knots (boundary of the intervals), and will provide a more flexible fit than a regular polynomial regression.

Smoothing spline (SS) is a more general spline method than NCS. Instead of fitting the relationship in regard to specific knots, it will fit the curves according to all time points. It is more flexible, and penalizes both the bias and roughness of the fitted curve. However, it is more difficult to program and implement in practice.

The objective of this thesis is to describe the difference of the 8-10 years postpartum BP measurements between woman with placental evidence of maternal malperfusion and those without such lesions during pregnancy and determine whether women with those lesions have excess risk of hypertension.

There are two challenges in the analysis of BP data in the window study. First, each research subject had three repeated measurements, thus the data structure is longitudinal. Second, there is wide variation in time when those measurements were taken. In our data, the office visit time ranged from to 06:53am to 15:34pm, and the comparison of BP should adjust for the measurement time. Based on the data structure and research question, we incorporated NCS and SS in the mixed-effect models separately and compared the results.

In Chapter 2, we will give a more detailed review of the two spline methods. In Chapter 3 we will describe the dataset and apply spline-based methods. We conclude with discussion in Chapter 4.

2.0 Methods

2.1 Non-Parametric Regression Models

Linear regression is one of the most commonly used techniques to model the relationship between the outcome variable y and the predictors $\{x_1, x_2, ..., x_p\}$.[16] The model is characterized by

$$y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i} + \epsilon_i, \ i = 1, 2, ..., n$$

where β_0 is the intercept, $\{\beta_1, ..., \beta_n\}$ are regression coefficients corresponding to predictors $\{x_1, ..., x_p\}$ and ϵ_i is the residual term. Given the observed data, one could fit the linear regression model and use the model to predict outcomes for new observations. Typically, linear regression has the assumption of linear relationship, independent of residuals and equal variance of residuals (homoscedasticity). The model is fitted with least square, which estimates by minimizing residual sum of square $\sum_{i=1}^{n}\{y_i - \beta_0 - \beta_1 x_{1,i} - ... - \beta_p x_{p,i}\}^2$. Linear regression is easy to fit and interpret, in average, y would increase by β_1 with each unit increase in x_1 and other covariates fixed. However, linear regression is not very flexible and could only model linear relationship.

In reality, there are many occasions that the relationship between the outcome variable and the predictors are obviously non-linear. Then, polynomial regression model is more appropriate. Given data points $\{y_i, t_i, i = 1, 2, ..., n\}$, the model has the form

$$y_i = g(t_i) + \epsilon, i = 1, 2, ..., n$$

where $g(t)$ is a polynomial function of t. In practice, it is essential to consider both the goodness-of-fit and the roughness of the curve. The goodness-of-fit of a curve is measured by the residual sum of square $\sum_{i=1}^{n}\{Y_i - g(t_i)\}^2$. The roughness of a curve defined on an interval $[a, b]$ can be measured by the second derivative $g''(t)$ if the curve is twice differentiable, and the absolute value of g'' indicates the degree of fluctuation. Therefore, the integrated squared second derivative $\int_a^b \{g''(t)\}^2 \, dt$ is a global measure of roughness, which has computational advantages compared to $\int_a^b |g''(t)| \, dt$. In order to balance between the residual sum of squares

and the roughness of the fitted curve, a roughness penalty approach is introduced. Given a smoothing parameter $\alpha > 0$, the penalized sum of squares is defined as

$$S(g) = \sum_{i=1}^{n} \{Y_i - g(t_i)\}^2 + \lambda \int_a^b \{g''(t)\}^2 \, dt.$$

where λ is the smoothing parameter. The penalized least squares estimator is the curve function $\hat{g}$ that minimizes the penalized sum of squares function $S(g)$ for a fixed λ. Compared to least square estimator, the penalized least squares estimator takes both the goodness-of-fit of the data and the roughness of the curve into consideration. As λ goes to zero, the penalty term in penalized least square goes to zero, thus the fitted curve would be interpolating every data point and could overfit a new dataset. As λ goes to infinity, the penalty term is the dominating part of the penalized least squares and it would force the roughness term to be zero, thus the curve would become a linear regression fit but there would be a poor fit for the data. Such a nonparametric regression model could provide a flexible description on the association between the outcome and predictors.

2.2 Natural Cubic Splines

Compared to regression models, spline functions are more flexible to model this relationship. In this section, we will describe natural cubic splines with more details.

Given knots $t_1, t_2, ..., t_n$ on an interval [a,b], such that $a < t_1 < t_2 < ... < t_n < b$, a curve function $g(t)$ defined on $[a, b]$ is a cubic spline if it satisfies two conditions.

(i) $g(t)$ is a cubic polynomial on each interval $(a, t_1), (t_1, t_2), ..., (t_n, b)$.

(ii) The function and its first and second derivatives are continuous at each knot t_i, so that the polynomial pieces of $g(t)$ fit together at the knots t_i.

A cubic spline defined on $[a, b]$ is said to be a natural cubic spline if its second and third derivatives at a and b are zero. This means that $g(t)$ is linear on the intervals $[a, t_1]$ and $[t_n, b]$.

It is natural to express the natural cubic spline defined on $[a, b]$ as

$$g(t) = d_i(t - t_i)^3 + c_i(t - t_i)^2 + b_i(t - t_i) + a_i \ for \ t_i \leq t \leq t_{i+1}, \ i = 0, 1, ..., n \qquad (2.1)$$

and we define $t_0 = a$ and $t_{n+1} = b$. Because of the continuity of g(t) and its first two derivatives at internal knots, there are restrictions among the coefficients $\{a_i, b_i, c_i, d_i\}$, and such representation is redundant in practice.

The value-second derivative representation provide a more convenient way to specify a NCS with the value $g(t_i)$ and the second derivative $g^{''}(t_i)$ at each knots.[17] Let $g_i = g(t_i)$ be the NCS on interval $[a, b]$ with fixed knots $t_i, i = 1, ..., n$ such that $a < t_1 < t_2 < ... < t_n < b$. Define $g_i = g(t_i)$ and $\gamma_i = g^{''}(t_i)$. From the property of NCS we know that $\gamma_1 = \gamma_i = 0$. Define two matrix g and γ as $g = [g_1, g_2, ..., g_n]^T$ and $\gamma = [\gamma_2, \gamma_2, ..., \gamma_{n-1}]^T$. Then we can define the $n \times (n-2)$ matrix Q and the $(n-2) \times (n-2)$ matrix R. Let $h_i = t_{i+1} - t_i$ for i = 1,...,n-1, then matrix Q is constructed as

$$q_{j-1,j} = h_{j-1}^{-1},\ q_{j,j} = h_{j-1}^{-1} - h_j^{-1}\ and\ q_{j+1,j} = h_j^{-1} j = 2, ..., n-1$$

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

R is a symmetric matrix constructed with elements $r_{i,j}$:

$$r_{i,i} = \frac{1}{3}(h_{i-1} + h_i)\ for\ i = 2, ..., n-1,$$
$$r_{i,i+1} = r_{i=1,i} = \frac{1}{6}h_i\ for\ i = 2, ..., n-1,$$

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

By theorem, a NCS can be specified if and only if the relationship $Q^T g = R\gamma$ is satisfied.[17] Let g be the cubic curve on interval $[t_L, t_R]$, and define $g(t_L) = g_L$, $g(t_R) = g_R$, $g^{''}(t_L^+) = \gamma_L$, $g^{''}(t_R^-) = \gamma_R$, $h = t_R - t_L$. Then we have

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

because $g^{''}(t)$ is linear and $g^{'''}(t)$ is constant on the interval.

By taking integration with respect to t and plugging in the expressions $g(t_L) = g_L$, $g(t_R) = g_R$, we have the expression of g as:

$$g(t) = \frac{(t - t_L)g_R + (t_R - t)g_L}{h} - \frac{1}{6}(t - t_L)(t_R - t)\{(1 + \frac{t - t_L}{h})\gamma_R + (1 + \frac{t_R - t}{h})\gamma_L\} \quad (2.2)$$

This representation has the same form as in (2.1). Given the condition that $Q^T g = R\gamma$, γ can be obtained as

$$\gamma = R^{-1} Q^T g$$

Thus, all the parameters in (2.2) can be obtained from the observed data. Consequently, the value of the NCS at ant point t can be expressed with matrix g and γ.

In order to illustrate the application of NCS in practice, we provide an example R code. Suppose given a dataset called "data.office", SBP measure is named "sbp_mmhg" and the corresponding measurement time is named "bp_time". We could fit a NCS between SBP and time using the code below.

```
>z.s <- ns(data.office$bp_time,
          knots=times(c('08:36','10:22','12:05','13:52')),
          Boundary.knots = times(c('06:54','15:32')))
>NCS1 <- lm(sbp_mmhg ~ z.s,data = data.office)
```

The function *"ns()"* is used to generate the basis matrix for the NCS with respect to the measurement time in the dataset. The interval is set to be 06:54am and 15:32pm and the internal knots are selected to be 08:51am, 10:20am, 12:02pm and 13:43pm. The basis matrix is stored in the object "z.s", and one can directly apply the basis matrix to a linear model to fit the NCS between SBP and the measurement time. The fitted NCS model was stored in the object "NCS1" and can be further applied into other statistical models such as mixed-effects model.

2.3 Smoothing Splines

Compared to fitting polynomial curves with NCS, smoothing splines is a more general method to estimate the relationship. In this section, we describe a method developed by Yuedong Wang (2011) to model smoothing splines based on reproducing kernel Hilbert space.[18]

2.3.1 Reproducing Kernel Hilbert Space (RKHS)

On a linear space E, a nonnegative function $||\cdot||$ is called a norm if it satisfies
(i)$||f|| = 0$ only when f = 0;
(ii)$||\alpha f|| = |\alpha| \cdot ||f||$;
(iii)$||f + g|| \leq ||f|| + ||g||$;

And an inner product is a mapping $(\cdot, \cdot)$: $E \times E \to \mathbb{R}$ which satisfies
(i)$(f, g) = (g, f)$;
(ii)$(\alpha f + \beta g, h) = \alpha(f, h) + \beta(g, h)$;
(iii)$(f, f) \geq 0$ and $(f, f) = 0$ only when $f = 0$;

Thus, a norm could be defined by an inner product as $||f|| \triangleq \sqrt{(f, f)}$, and an linear space together with an inner product is called an inner product space.

If every Cauchy sequence in the space E converges to an element in E, then the space E is called complete. A Hilbert space $(\mathcal{H})$ is a complete inner product space defined by the norm.

The most common example of a Hilbert space is a Euclidean space of three dimensions $(\mathbb{R}^3)$ with an inner product defined as

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1y_1 + x_2y_2 + x_3y_3$$

The inner product satisfies the three properties above, and a norm can be defined accordingly as $||x|| = \sqrt{x \cdot x}$. Given the completeness of Euclidean space, the Euclidean space with an inner product is a complete inner product space, thus is a Hilbert space.

Let $\mathcal{H}$ denote a Hilbert space and the elements of $\mathcal{H}$ are real-valued functions f: $\mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is an random set. For any fixed $x \in \mathcal{X}$, define

$$\mathcal{L}_x(f) \triangleq f(x), f \in \mathcal{H},$$

then $\mathcal{L}_x$: $\mathcal{H} \to \mathbb{R}$ is called the *evaluational functional*. A Hilbert space is called a reproducing kernel Hilbert space(RKHS) if every evaluational functional $\mathcal{L}_x$ is continuous. $\mathcal{L}_x$ is

continuous when $\mathcal{L}_x f_n = f_n(x) \to f(x) = \mathcal{L}_x f$

Apply the Riesz representation theorem, there exists an unique $R_x \in \mathcal{H}$ such that

$$\mathcal{L}_x(f) = (R_x, f)$$

Thus, R_x itself is function: $\mathcal{X} \to \mathbb{R}$. Define $R_x(Z)$ as a bivariate function of x and z such that $R(x, z) \triangleq R_x(z)$, $x, z \in \mathcal{X}$. Here, the bivariate function $R(x, z)$ is called the reproducing kernel of the RKHS $\mathcal{H}$. It can be proven that the reproducing kernel is nonnegative definite:

$$R(x, z) = R_x(z) = \mathcal{L}_x(R_z) = (R_x, R_z) \geq 0$$

For any $\alpha_1, ..., \alpha_n \in \mathbb{R}$ and $x_1, ..., x_n \in \mathcal{X}$,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j R(x_i, x_j) = R(\sum_{i=1}^{n} \alpha_i x_i, \sum_{j=1}^{n} \alpha_j x_j) \geq 0$$

Thus, for any $x_1, ..., x_n$, the matrix $|R(x_i, x_j)|_{n \times n}$ is nonnegative definite.

Denote $\mathcal{S}$ as a subspace of a Hilbert space $\mathcal{H}$, then the subspace $\mathcal{S}$ is a Hilbert space if $\mathcal{S}$ is closed. The orthogonal complement of can be defined:

$$\mathcal{S}^{\perp} \triangleq \{f \in \mathcal{H} : (f, g) = 0 \ for \ all \ g \in \mathcal{S}\}$$

Thus for all elements $f \in \mathcal{H}$, f can be decomposed as $f = g + h$, where $g \in \mathcal{S}$ and $h \in \mathcal{S}^{\perp}$.

Meanwhile, $\mathcal{H}$ is decomposed as $\mathcal{H} = \mathcal{S} \bigoplus \mathcal{S}^{\perp}$

Suppose $\mathcal{H}$ is RKHS and $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$, then the RK has the property that

$$R(x, z) = R((x_0, x_1), (z_0, z_1)) = R_0(x_0, z_0) + R_1(x_1, z_1)$$

2.3.2 RKHS for Smoothing Spline

Consider the general smoothing spline regression(SSR) model

$$y_i = f(x_i) + \epsilon_i,\ i = 1, ..., n$$

where y_i is the observation of the function f given the covariates x_i, and ϵ_i is the independent random error with mean zero and variance σ^2. f is the functional form of the smoothing spline. The Sobolev space is defined as:

$$W_2^m[a,b] = \{f : f, f^{'}, ..., f^{(m-1)} \textit{ are absolutely continuous}, \int_a^b (f^{(m)})^2 dx \leq \infty\}$$

If the smoothing spline f has a domain of $\mathcal{X} = [a,b]$, and $f \in W_2^m[a,b]$, then the the smoothing spline estimate $\hat{f} \in W_2^m[a,b]$ is the solution to the penalized least squares(PLS):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx \tag{2.3}$$

The Sobolev space $W_2^m[a,b]$ is an RKHS $\mathcal{H}$ and one can define an inner product as

$$(f,g) = \sum_{\nu=0}^{m-1} f^{(\nu)}(a)g^{(\nu)}(a) + \int_a^b f^{(m)}(x)g^{(m)}(x)dx$$

Define a decomposition of $W_2^m[a,b]$ as $W_2^m[a,b] = \mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$, where $\mathcal{H}_0$ is the subspace of all the (m-1) order polynomials and $\mathcal{H}_1$ contains the orthonormal complement of $\mathcal{H}_0$.

$$\mathcal{H}_0 = span\{1, (x-a), ..., (x-a)^{(m-1)}/(m-1)!\}$$
$$\mathcal{H}_1 = span\{f : f^{(\nu)}(a) = 0, \nu = 0, ..., m-1, \int_a^b (f^{(m)})^2 dx < \infty\}$$

Thus, the corresponding RKs of the RKHS $W_2^m[a,b] = \mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ are

$$R_0(x,z) = \sum_{\nu=1}^{m} \frac{(x-a)^{(\nu-1)}}{(\nu-1)!}\frac{(z-a)^{(\nu-1)}}{(\nu-1)!}$$
$$R_1(x,z) = \int_a^b \frac{(x-u)_+^{(m-1)}}{(m-1)!}\frac{(z-u)_+^{(m-1)}}{(m-1)!}du$$

Here, $(x)_+ = max\{x, 0\}$. Denote P_1 as the orthogonal projection from a function onto $\mathcal{H}$. For any $f \in \mathcal{H}$, f can be decomposed as $f = f_0 + f_1$, where $f_0 \in \mathcal{H}_0$, $f_1 \in \mathcal{H}_1$. Thus, the roughness penalty term ca be expressed as

$$\int_a^b (f^{(m)})^2 dx = ||f_1||^2 = ||P_1 f||^2$$

The PLS in (2.3) can be written as:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||P_1 f||^2 \tag{2.4}$$

There is no penalty applied to functions is f_0. The example of polynomial splines suggests that RHKS could be used to construct the more general smoothing spline with the following conditions:

(i)The model space of f is an RKHS $\mathcal{H}$;

(ii)The model space can be decomposed as $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$, where functions $f_0 \in \mathcal{H}_0$ are not penalized;

(iii)A penalty term as $\lambda||P_1 f||^2$.

2.3.3 Application in General Smoothing Splines

A more general SSR model is that

$$y_i = \mathcal{L}_i f + \epsilon_i,\ i = 1, ..., n$$

where $\mathcal{L}_i$ are bounded linear functions that $\mathcal{L}_i \in \mathcal{H}$. This is useful when the observation of $f(x)$ is made through linear functions of $f(x)$, such as $f^{'}(x_i)$. (2.) is a special case of (2.) where $\mathcal{L}_i$ is the evaluational fuctionals at design points(covariates) such that $\mathcal{L}_i f = f(x_i)$. $\mathcal{L}_i$ is are bounded according to the definition of RKHS.

In the more general smoothing spline model, the estimate of f, denoted as $\hat{f}$, minimizes the PLS in a form as

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathcal{L}f)^2 + \lambda||P_1 f||^2$$

The domain of f is an arbitrary set $\mathcal{X}$ and the model space is an RKHS $\mathcal{H}$ on $\mathcal{X}$ with RK $R(x, z)$, which can be decomposed as $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$. Equivalently, the function can be

decomposed as $f = f_0 + f_1$. The subspace $\mathcal{H}_0$ is a finite dimensional space, and its basis functions are consisted of $\{\phi_\nu(x), \nu = 1, ..., p\}$. $\mathcal{H}_0$ is usually called the null space because the functions $f_0 \in \mathcal{H}_0$ are not penalized. Separately, f_0 and f_1 are the projection of f onto $\mathcal{H}_0$ and $\mathcal{H}_1$. Thus, the magnitude of f_1 can be measured with $||P_1 f||$, which represents the departure from the null space and can be used to assess the appropriateness of the model. And λ is used to trade off between the goodness of fit and the appropriateness of the model.

Applying the Riesz representation theorem, there exists a unique representer $\eta_i \in \mathcal{H} = W_2^m[a, b]$ such that

$$\mathcal{L}_i f = f(x_i) = (\eta_i, f)$$
$$\eta_i(x) = (\eta_i, R_x) = \mathcal{L}_i R_x = \mathcal{L}_{i(z)} R_1(x, z)$$

$\mathcal{L}_{i(z)}$ is the evaluation functional at x_i applied to a function of z.
Let $\xi_i = P_1 \eta_i$, then ξ_i is the projection of η_i onto $\mathcal{H}_1$. It can be proved that P_1 is self-adjoint that

$$(P_1 g, h) = (P_1 g, P_1 h + \{h - P_1 h\}) = (P_1 g, P_1 h) = (P_1 g + \{g - P_1 g\}) = (g, h)$$

Given $R(x, z) = R_0(x, z) + R_1(x, z)$, it can be shown that

$$\xi_i(x) = (\xi_i, R_x) = (P_1 \eta_i, R_x) = (\eta_i, P_1 R_x) = \mathcal{L}_{i(z)} R_1(x, z)$$

Thus, the projection of the representer η_i onto $\mathcal{H}_1$, ξ_i, can be calculated by applying the operator $\mathcal{L}_{i(z)}$ to R_1. The inner product of ξ is defined as

$$(\xi_i, \xi_i) = \mathcal{L}_{i(x)} \xi_j(x) = \mathcal{L}_{i(x)} \mathcal{L}_{j(z)} R_1(x, z)$$

Recall that $\{\phi_\nu(x),\ \nu = 1, ..., p\}$ is the basis function for $\mathcal{H}_0$ where there is no penalty, define T as a $n \times p$ matrix and Σ as a $n \times n$ matrix that

$$T = \{\mathcal{L}_i \phi_\nu\}_{i=1\ \nu=1}^{n\ p}$$
$$\Sigma = \{\mathcal{L}_{i(x)} \mathcal{L}_{j(z)} R_1(x, z)\}_{i,j=1}^{n}$$

Then, assume that

$$\hat{f}(x) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i \xi_i(x) + \rho$$

where $\rho \in \mathcal{H}_1$, and $(\rho, \xi_i) = 0$ for i = 1,...,n. $\sum_{\nu=1}^{p} d_\nu \phi_\nu(x)$ is the projection of $\hat{f}$ onto $\mathcal{H}_0$, denoted as $P_0 f$. $\sum_{i=1}^{n} c_i \xi_i(x)$ is the projection of $P_1 \hat{f}$ onto the subspace spanned by $\{\xi_i(x),\ i = 1, ..., n\}$. Recall that $\xi_i = P_1 \eta_i \in \mathcal{H}_1$, then $\zeta_i = \eta_i - \xi_i \in \mathcal{H}_0$. ζ_i and ξ_i are the projection of the unique representer $\eta_i \in \mathcal{H}$ onto $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Therefore,

$$\mathcal{L}_i \rho = (\eta_i, \rho) = (\xi_i, \rho) + (\zeta_i, \rho) = 0 + 0 = 0$$

Denote $\boldsymbol{y} = (y_1, ..., y_n)^T$ and $\hat{\boldsymbol{f}} = (\mathcal{L}_1 \hat{f}, ..., \mathcal{L}_n \hat{f})^T$ respectively as the observation vector and fitted value vector. Let $\boldsymbol{d} = (d_1, ..., d_p)^T$ and $\boldsymbol{c} = (c_1, ..., c_p)^T$, then we have

$$\hat{\boldsymbol{f}} = T\boldsymbol{d} + \Sigma \boldsymbol{c}$$

In addition, $||P_1 f||^2 = ||\sum_{i=1}^{n} c_i \xi_i + \rho||^2 = \boldsymbol{c}^T \Sigma \boldsymbol{c} + ||\rho||^2$. Hence, the PLS can be reformed as

$$\frac{1}{n}||\boldsymbol{y} - T\boldsymbol{d} - \Sigma \boldsymbol{c}||^2 + \lambda(\boldsymbol{c}^T \Sigma \boldsymbol{c} + ||\rho||^2)$$

Thus, the PLS is minimized only when $\rho = 0$.

The Kimeldorf-Wahba representer theorem is: *Given T is of full column rank. Then the PLS in* (2.4) *has a unique minimized from*[19]

$$\hat{f}(x) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i \xi_i(x)$$

At $\rho = 0$, the PLS term is

$$\frac{1}{n}||\boldsymbol{y} - T\boldsymbol{d} - \Sigma \boldsymbol{c}||^2 + \lambda \boldsymbol{c}^T \Sigma \boldsymbol{c}$$

where $\boldsymbol{c}$ and $\boldsymbol{d}$ are coefficients that needed to be estimated from the observed data. Apply QR decomposition to T, we have $T = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix}$, where (Q_1, Q_2) is an orthogonal matrix and Q_1, Q_2 are respectively $n \times p, n \times (n-p)$ matrices, and R is an upper triangular invertible $p \times p$ matrix.

Define $M = \Sigma + n\lambda I$, it can be shown that the PLS has a solution that the fitted values $\hat{f}$ has a form

$$\hat{f} = T\boldsymbol{d} + \sigma\boldsymbol{c} = \boldsymbol{y} - n\lambda\boldsymbol{c} = H(\lambda)\boldsymbol{y}$$

where

$$H(\lambda) \triangleq I - n\lambda Q_2(Q_2^T M Q_2)^{-1}Q_2^T$$

Computing details are given in the book. It is proved that given the observation $(x_i, y_i)^T, i = 1, ..., n$, the fitted values $\hat{\boldsymbol{y}}$ could be calculated with a specified λ.

2.3.4 Linear Mixed-Effects Model

Linear mixed-effects model is an extension of linear regression model, with both fixed effects and random effects included.[20] When the outcome is repeatedly measured within each research subject, it is not proper to assume the observations are independent with other, and the data has a longitudinal structure and a mixed-effects model should be fitted. Fixed effect is the variation that is explained by independent predictors, and random effect is the variation not explained by the predictors. For example, when studying the effect of placental malperfusion on repeatedly measured BPs, the variation due to placental malperfusion is fixed effects and the variation within each research subject is random effect. The linear mixed-effects model has a form

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, i = 1, 2, ..., n$$

where $\boldsymbol{y}$ is the vector of continuous outcome, $\boldsymbol{X}$ are the design matrix, $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{Z}$ is the matrix of covariates, $\boldsymbol{b}$ is the vector of random effects and $\boldsymbol{\epsilon}$ is the vector of residuals. For example, $\boldsymbol{X}$ contains the value of placental malperfusion and $\boldsymbol{Z}$ is the matrix of subject indicators. It is assumed that $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ follow normal distribution with mean zero and variance of matrix $\mathcal{D}$ and $\mathcal{R}_i$, respectively. The random effects $\boldsymbol{b}_i$ are independent with the residual error $\boldsymbol{\epsilon}_i$ for the same subject.

2.3.5 Semiparametric Mixed-Effects Models

In the motivating dataset of our study, the outcome variable is the three repeatedly measured BP for each women. For each measurement, both the systolic blood pressure (SBP) and diastolic blood pressure (DBP) were recorded as well as the measurement time. For each women, the covariate of primary interest is "placental malperfusion" which is a binary variable coded as 0 or 1, with 1 representing women with placental malperfusion and 0 representing women without placental malperfusion. Other binary covariates includes "preeclampsia" (1 for cases and 0 for non-cases), "race" (1 for black and 0 for others), "family history of hypertension" (1 for cases and 0 for non-cases) and "family history of heart disease" (1 for cases and 0 for non-cases). There are two continuous variables, "age" and "body mass index (BMI) at pregnancy".

For our repeatedly measured data, the semiparametric linear mixed-effects models has the form as

$$y_{ij} = \boldsymbol{S}_{ij}^T\boldsymbol{\beta} + \mathcal{L}_{ij}f + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i + \epsilon_{ij},\ i = 1, 2, ..., m;\ j = 1, ..., n_i,$$

where y_{ij} ($i = 1, ..., n$ and $j = 1, 2, 3$) represents the jth BP measurements for ith woman. Here, n is number of subjects in the dataset. $\boldsymbol{\beta}$ is the coefficient vector of the fixed effects and $\boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{D})$ are the random effects and $\boldsymbol{D}$ is the covariance matrix. In our data, the fixed effects are the covariates, and the random effects are the variation of baseline between each women. $\mathcal{L}_{ij}f = f(t_{ij})$ is the smoothing spline function in regards to the time points t_{ij}. ϵ_{ij} is the random error and $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{in_i})^T \sim \mathcal{N}(0, \sigma^2\Lambda_i)$, where $\Lambda_i = I_{n_i}$. The PLS can be further written as

$$\frac{1}{n}\sum_{i=1}^{m}(y_i - \boldsymbol{S}_i^T\boldsymbol{\beta} - \mathcal{L}_i f)^T W_i(y_i - \boldsymbol{S}_i^T\boldsymbol{\beta} - \mathcal{L}_i f) + \lambda||P_1 f||^2,\ n = \sum_{i=1}^{m} n_i, \tag{2.5}$$

where $W^{-1} = ZDZ^T + \Lambda$, $Z = diag(Z_1, ..., Z_m)$ and $Z_i = (z_i 1, ..., z_{in_i})$. The minimizer $\hat{f}(t)$ for the PLS(2.5) is

$$\hat{f}(t) = \sum_{\nu=1}^{p} d_\nu\phi_\nu(t) + \sum_{i=1}^{n} C_i\xi_i(t) = \mathcal{L}_i f$$

where $\{\phi_\nu(t), \nu = 1, ..., p\}$ are the basis functions of $\mathcal{H}_0$ with no penalty imposed. And the PLS term can be reformed as

$$\sum_{i=1}^{m}(y_i - S_i^T\beta - T_i^T d - \Sigma_i^T c)^T W_i(y_i - S_i^T\beta - T_i^T d - \Sigma_i^T c) + n\lambda c^T \Sigma c$$

The random effect can be estimated[21] as

$$\hat{\boldsymbol{b}} = DZ^T W(\boldsymbol{y} - S\beta - T\boldsymbol{d} - \Sigma c)$$

On the other hand, the linear mixed-effects model can be written as

$$y_i = S_i^T\beta + T_i^T\boldsymbol{d} + \mu_i + Z_i^T\boldsymbol{b}_i + \epsilon_i = X_i\alpha + \mu_i + Z_i^T\boldsymbol{b}_i + \epsilon_i \tag{2.6}$$

where $\mu_i \sim \mathcal{N}(0, \frac{\Sigma}{n\lambda})$, $\boldsymbol{b}_i \sim \mathcal{N}(0, \boldsymbol{D})$, $\epsilon_i \sim \mathcal{N}(0, \sigma^2 I_{n_i})$ and $\Sigma = (R_1(t_i, t_j))$. $S_i^T\beta$ is the fixed effect, $Z_i^T\boldsymbol{b}_i$ is the random effect and $T_i^T\boldsymbol{d} + \mu_i$ are related to the smoothing splines. This method could turn the fitted smoothing splines into a form of linear mixed-effect model, and fitting the linear mixed-effect model is equivalent to fitting the smoothing splines.

2.4 Adopted Models

The purpose of this thesis is to describe the effect of placental malperfusion on maternal BP 8 to 10 years after delivery, adjusting for other covariates.

At first, the missingness of BP measurements and measurement time was checked, and only complete cases were included in the final dataset. The distribution of BP measurements were checked with QQ-plots and the outliers were excluded from the data. Descriptive statistics were created for the covariates to check if the study design is balanced.

Then, we studied the effects of the covariates on BP without adjusting the measurement time. Linear mixed-effects model was used to fit the regression model for our longitudinal data. Two models were fitted step by step to investigate the random effects and fixed effects.

Model I:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

Here, Y_{ij} denotes the jth measurement of blood pressure for the ith subject. μ_i denotes the mean blood pressure for the ith subject, which has a normal distribution with mean μ and variance τ^2, denoted as $\mu_i \sim \mathcal{N}(\mu, \tau^2)$. μ is the global mean of blood pressure among the study population and τ^2 is the variation of blood pressure measurements among subjects. ϵ_{ij} is the error term for the jth measurement of blood pressure for the ith subject, which follows a normal distribution with mean zero and variance σ^2, denoted as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. σ^2 denotes the variation of blood pressure measurements within each subject.

Model II:

$$Y_{ij} = \alpha_i + \beta X_i + \epsilon_{ij}$$

Here, X_i denotes the covariates of the ith subject fitted in the model and β denotes the coefficients of the covariates. α_i denotes the variation of blood pressure measurements between each subjects, which has a normal distribution as $\alpha_i \sim \mathcal{N}(0, \tau^2)$. ϵ_{ij} denotes variation of blood pressure measurements within each subject, which has a normal distribution as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

After that, we used natural cubic spline to adjust for the relationship between BP measurements and measurement time. Then combined the NCS into linear mixed-effects model to fit the relationship between BP and the covariates of interest, with adjusting for the time.

Model III:

$$Y_{ij} = \alpha_i + \beta X_i + g(t_{ij}) + \epsilon_{ij}$$

with $g(t_{ij})$ representing the NCS.

Lastly, we used smoothing spline to provide a more flexible adjustment for the association between BP and measurement time. A semiparametric linear mixed-effects model was fitted as demonstrated in the following.

Model IV:

$$Y_{ij} = \alpha_i + \beta X_i + \mathcal{L}f + \epsilon_{ij}$$

where $\mathcal{L}f = f(t_{ij})$ represents the smoothing spline term. As illustrated in chapter 2, Model IV could be fitted by equation (2.6).

In model III and model IV, most of the terms have the same interpretation as in model II except for the spline term.

3.0 Data Analysis

3.1 Description of the Data

There were 498 subjects in the original dataset. After excluding subjects with missingness in their BP measurement times, a total of 367 subjects were included in our study from the office visit dataset where each subject had 3 repeated measurements of both SBP and DBP as well as the time for each BP measurement. The normality of the distribution of BP measures were examined by QQ-plot in R and shown in Figure 1.

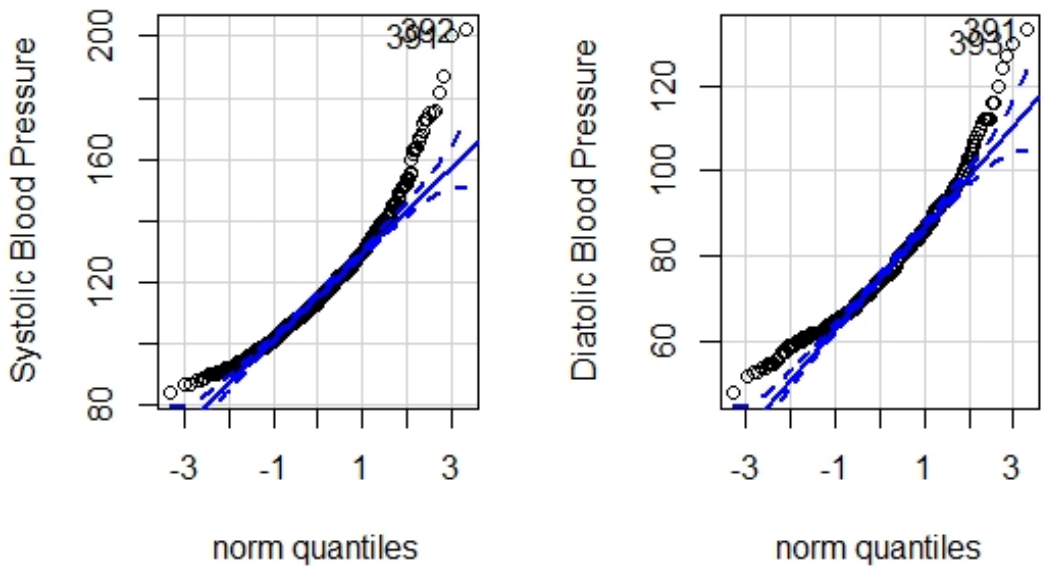


Figure 1: QQ Plot Blood Pressure Measures before Excluding Outliers

It is obvious that there were heavy tails in the distribution of both SBP and DBP. After excluding the outliers (data points outside 1.5 times the interquartile range above the upper quartile and bellow the lower quartile), there were 351 subjects in our final dataset and the distribution of the BP measures were very close to normal distribution (Figure 2).

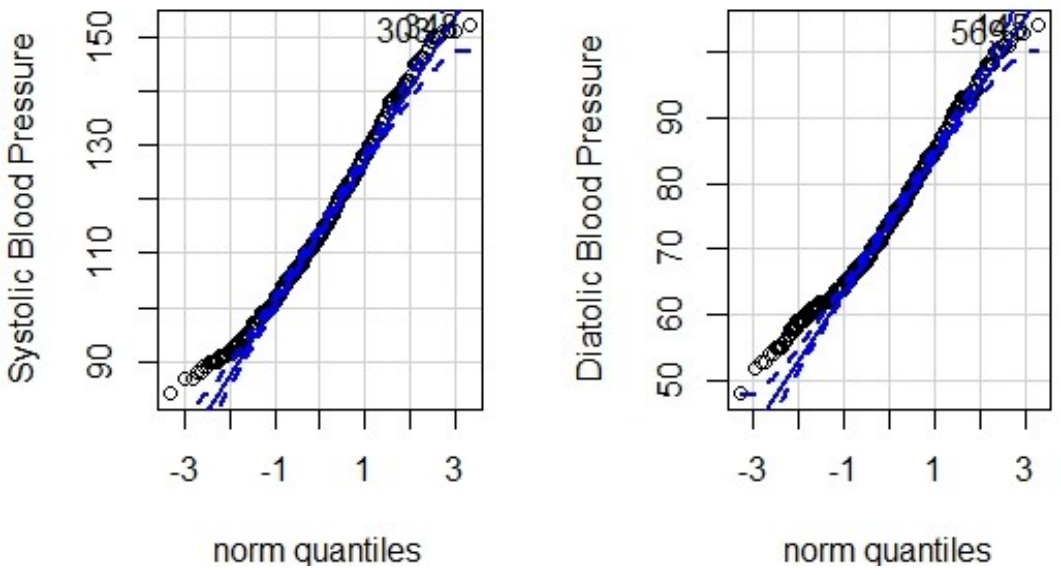
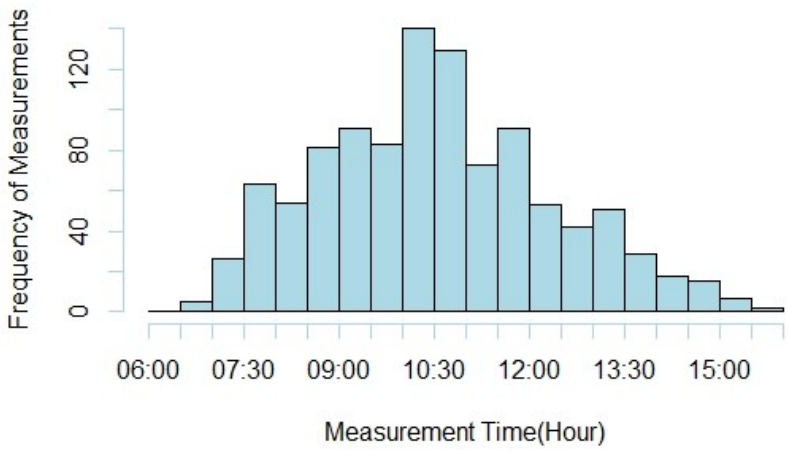
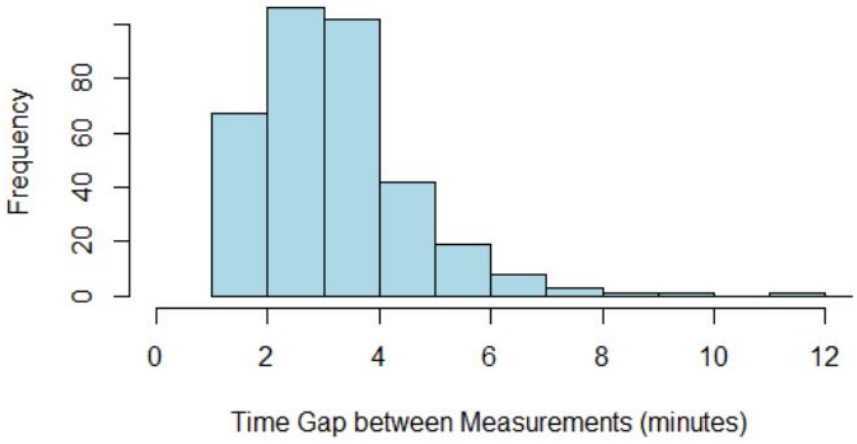


Figure 2: QQ Plot Blood Pressure Measures after Excluding Outliers

For each subject, the three measurements of BP were performed sequentially in a short time period, with approximately one minute apart. The measurement times ranged from 06:53am to 15:34pm and the distribution was shown in Figure 3.1 while figure 3.2 shows the distribution of the measurement time gap for each subject. The time gap was calculated as the gap between the first measurement and the last measurement for each subject. Most of the time gaps were smaller than six minutes, thus, for the convenience of our study, the measurement times of BP were estimated with the mean of the three time records for each subject.



(1) Histogram of Blood Pressure Measurement Time



(2) Histogram of Measurement Time Range for Each Subject

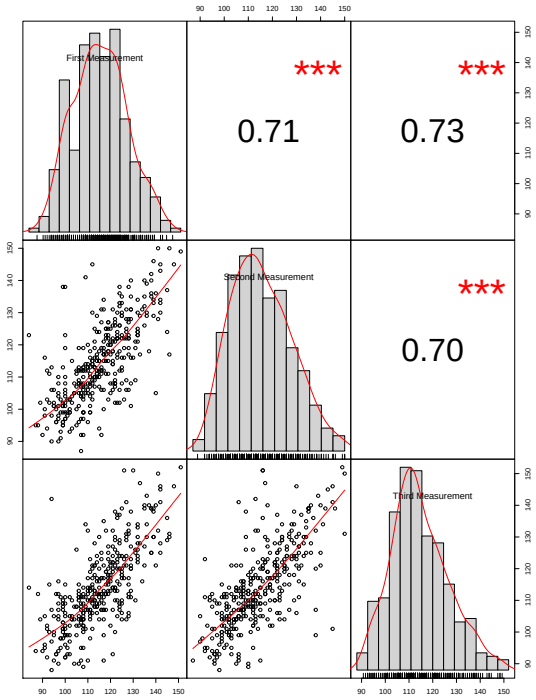
Figure 3: Distribution of Blood Pressure Measurement Time

Besides fitting the time trend of BP, we were also interested in the effect of malperfusion, preeclampsia, race (demorace), BMI at pregnancy (prepregbmi), age, family history of hypertension outside of pregnancy (famhbp) and family history of heart disease (famheart) on BP and malperfusion was of primary interest. Table 1 summarizes the descriptive statistics for these covariates of interest. The number of records shows how many subjects have observed data for this covariate.

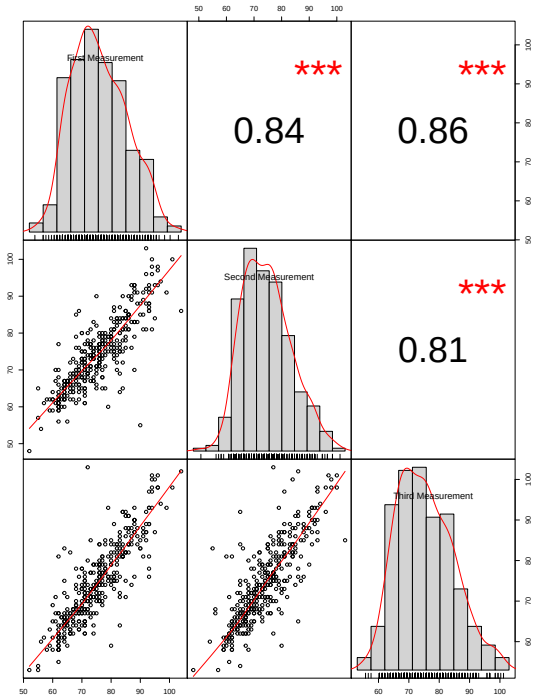
Table 1: Descriptive Statistics on Covariates of Interest

Covariates	Number of Records (Proportion)	Mean (SD) / n (%)
Malperfusion	351 (100%)	117 (33.3%)
Preeclampsia	351 (100%)	45 (12.8%)
Race(black)	351 (100%)	98 (27.9%)
Family History of Hypertension	331(94.3%)	189 (53.8%)
Family History of Heart Disease	305 (86.9%)	72 (20.5%)
BMI at Pregnancy (kg/m^2)	247 (70.4%)	26.1 (6.2)
Age (years)	347 (98.9%)	38.0 (6.0)

The Pearson correlation between the three measurements were calculated to examine the relationship between the three repeated measures(Figure 4). The results showed that the correlation between the three repeated measurements were exchangeable for both sbp and dbp, indicating that the order of the measurements does not influence the measurement result and it was reasonable to treat the three measurement time as an identical one.



(1) correlation of SBP



(2) correlation of DBP

Figure 4: Correlation among Blood Pressure Measurements

3.2 Application of the Linear Mixed-Effects Model

In model I, the variation of blood pressure measurements within each subject was estimated to be 49 for sbp and 16 for dbp, and the variation between each subjects had an estimated value of 121 for sbp and 80 for dbp. From the results, we can tell that the majority of the variation came from the variation of different subjects, and the repeated measurements for each subject were consistent.

Table 2 and Table 3 summarized the coefficients, p-values, between subject variations(τ^2) and within subject variations (σ^2) of each univariable models in model II for sbp and dbp in the office visit data.

Table 2: Univariable Linear Mixed-Effects Model for SBP

Covariates for Model II	Coefficient	P-value	τ^2	σ^2
Malperfusion	0.73	0.58	121.36	49.01
Preeclampsia	7.09	0.0001	115.83	49.01
Race(black)	-0.37	0.61	121.36	49.07
Age	0.16	0.13	119.79	49.32
BMI at Pregnancy	0.73	< 0.0001	93.56	51.66
Family History of Hypertension	2.81	0.03	120.55	49.26
Family History of Heart Disease	1.90	0.23	120.94	48.76

Table 3: Univariable Linear Mixed Model for DBP

Covariates for Model II	Coefficient	P-value	τ^2	σ^2
Malperfusion	1.77	0.09	79.99	16.22
Preeclampsia	4.39	0.003	78.53	16.22
Race(black)	-0.57	0.31	80.61	16.24
Age	0.13	0.11	80.04	15.99
BMI at Pregnancy	0.44	< 0.0001	73.35	15.74
Family History of Hypertension	3.01	0.004	80.62	14.94
Family History of Heart Disease	2.28	0.07	79.28	17.13

It can be noticed that the variation within subjects are smaller than variation between subjects, and the variation of DBP is much small than SBP. Malperfusion alone was not significant in predicting either SBP or DBP. Among the covariates of interest, preeclampsia, BMI at Pregnancy and family history of hypertension were identified to be significant in the univariable models.

Table 4 and Table 5 summarized the coefficients, p-values, between subject variations (τ^2) and within subject variations (σ^2) of the multivariate models in model II for SBP and DBP in the office visit data.

Table 4: Multivariate Mixed-Effects Model for SBP

Covariates for Model II	Coefficient	P-value	τ^2	σ^2
Malperfusion	2.08	0.18	96.45	52.32
Preeclampsia	5.61	0.01		
Race(black)	2.95	0.12		
Age	0.14	0.32		
BMI at Pregnancy	0.66	< 0.0001		
Family History of Hypertension	0.41	0.80		
Family History of Heart Disease	0.25	0.90		

Table 5: Multivariate Mixed-Effects Model for DBP

Covariates for Model II	Coefficient	P-value	τ^2	σ^2
Malperfusion	2.94	0.02	69.94	15.92
Preeclampsia	2.60	0.13		
Race(black)	2.85	0.06		
Age	0.16	0.15		
BMI at Pregnancy	0.48	< 0.0001		
Family History of Hypertension	1.07	0.42		
Family History of Heart Disease	2.17	0.17		

In the multivariate model, BMI at pregnancy was significant for predicting both SBP and DBP. Malperfusion was significant in the multivariate model for DBP and preeclampsia was significant for in the multivariate model for SBP. Women with malperfusion were estimated to have 2.08 (95% CI [-0.98, 5.14]) mmHg higher SBP and 2.94 (95% CI [0.45, 5.43]) mmHg higher DBP than woman without malperfusion lesion. Even though family history of hypertension and heart disease were pretty significant in the univariable models, they are not significant in the multivariate model for both SBP and DBP. The between subject variations were both much greater that the within subject variation of blood pressure measurements.

3.3 Application of the Mixed-Effects Model with Natural Cubic Splines

In this section, Model III was fitted with five knots for both SBP and DBP, and plotted in Figure 5.

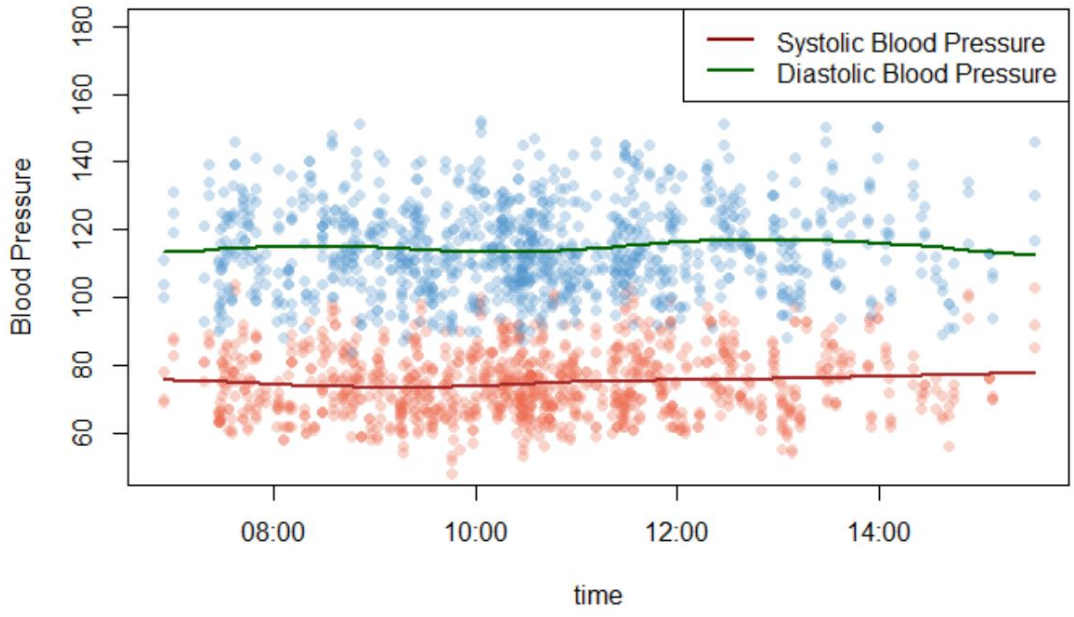
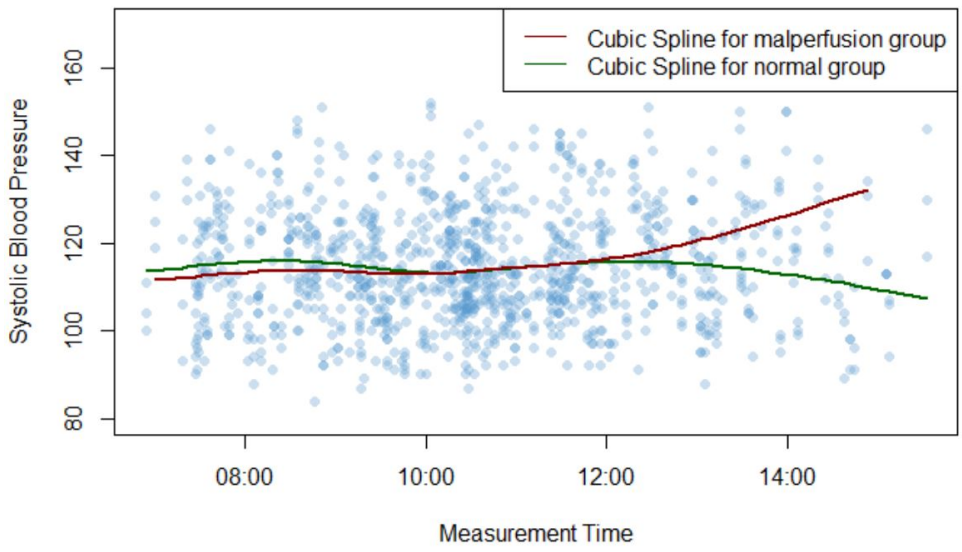
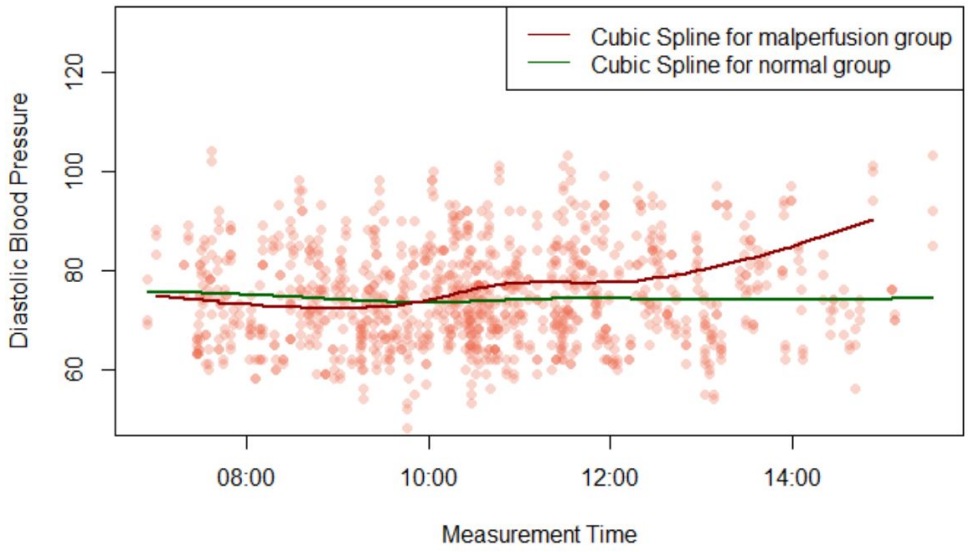


Figure 5: Natural Cubic Splines Plot for Blood Pressure over the Measurement Time

As shown in the plot, both SBP and DBP appeared to be slightly increased in the afternoon of a day. However, the time trend of BP was not obvious. To explore the effect of malperfusion on the time trend of BP, the NCS plots were plotted by different malperfusion groups.(Figure 6)



(1) Systolic Blood Pressure



(2) Diastolic Blood Pressure

Figure 6: Natural Cubic Splines Plot of Blood Pressure with Grouping of Malperfusion

For both SBP and DBP, the lines for malperfusion group and non-malperfusion group were very close to each other in the morning. However, the two lines began to diverge in the afternoon, and those who had malperfusion during pregnancy had higher BP measurements than those did not have malperfusion.

The coefficients, p-values, between subject variations (τ^2) and within subject variations (σ^2) of the full models in Model III for SBP and DBP are summarized in Table 6 and Table 7, respectively.

Table 6: Multivariate Mixed-Effects Model for SBP with NCS

Covariates for Model III	Coefficient	P-value	τ^2	σ^2
Malperfusion	1.91	0.23	97.41	52.32
Preeclampsia	5.37	0.01		
Race(black)	2.90	0.13		
Age	0.16	0.25		
BMI at Pregnancy	0.67	< 0.0001		
Family History of Hypertension	0.43	0.79		
Family History of Heart Disease	0.45	0.82		

Table 7: Multivariate Mixed-Effects Model for DBP with NCS

Covariates for Model III	Coefficient	P-value	τ^2	σ^2
Malperfusion	3.06	0.02	71.10	15.92
Preeclampsia	2.89	0.10		
Race(black)	2.52	0.11		
Age	0.17	0.14		
BMI at Pregnancy	0.48	< 0.0001		
Family History of Hypertension	0.95	0.47		
Family History of Heart Disease	2.27	0.16		

After adding the NCS to adjust measurement time, there is no remarkable change in the estimated coefficients and the p-values for each covariates. The variation within and between each women are also relatively similar with the model without NCS. Women with placental malperfusion are estimated to have 1.91 (95% CI [-1.20, 5.02]) and 3.06 (95% CI [0.52, 5.59]) mmhg higher SBP and DBP, respectively.

3.4 Application of Mixed-Effects Model with Smoothing Splines

In this section, we apply the method of smoothing splines into the linear mixed-effect model to fit the semiparametric linear mixed-effects model. ModelIV

Table 8: Semiparametric Mixed-Effects Model for SBP

Covariates for Model IV	Coefficient	P-value	τ^2	σ^2
Malperfusion	2.14	0.17		
Preeclampsia	5.69	0.01		
Race(black)	2.82	0.13		
Age	0.15	0.30	96.92	52.32
BMI at Pregnancy	0.66	< 0.0001		
Family History of Hypertension	0.36	0.82		
Family History of Heart Disease	0.26	0.89		

Table 9: Semiparametric Mixed-Effects Model for DBP

Covariates for Model IV	Coefficient	P-value	τ^2	σ^2
Malperfusion	3.08	0.02		
Preeclampsia	2.87	0.10		
Race(black)	2.54	0.10		
Age	0.18	0.12	69.66	15.92
BMI at Pregnancy	0.48	< 0.0001		
Family History of Hypertension	0.96	0.47		
Family History of Heart Disease	2.20	0.16		

The results from the smoothing splines is also quite similar with the estimate from Model II and III. After adding the SS to adjust measurement time, women with placental malperfusion are estimated to have 2.14 (95% CI [-0.94, 5.21]) and 3.08 (95% CI [0.59, 5.57]) mmhg higher SBP and DBP, respectively.

4.0 Conclusions

From the results of the fitted model, we conclude that BMI at pregnancy was the most significant predictor for BP 8-10 years after delivery in the window study. Considering both statistical results and pathophysiology mechanism, even though placental malperfusion was only statistically significant for DBP at the commonly used level of 0.05, the influence of placental malperfusion should not be ignored to evaluate the risk of post-delivery CVD for women.

Both NCS and SS provide powerful tools to adjust for the time trend of BP measurements in our dataset. NCS is easier to fit and combine in the linear mixed-effects model. However, NCS need to manually select knots or specify the degree of freedom, and one could get different results with different settings in the NCS model.

Smoothing spline is a more general and flexible method. SS will fit in regards as all the time points, thus no knots needs to be selected. It could fit more than the cubic polynomial splines in NCS, and it could combine the smoothness penalty into a semiparametric linear mixed-effect model to fit the data. However, this method is much more complicated than NCS, and it is harder to implement in software and takes longer time to compute.

Based on the analytical results, we conclude that placental malperfusion is a prognostic biomarker of DBP at about 8-10 years after delivery. "Preeclampsia" and "BMI at pregnancy" are very significant in our study, indicating that health provider should use these clinical profiles to consider interventions to reduce the risk of hypertension and consequential health impacts.

Appendix

Example R Code

```
#Read in the cleaned dataset
data.office<-read.csv('data_office_order_final1053.csv')
library(lme4)
library(lmerTest)


#Model I

#Fit model I for SBP
glm1.s <- lmer(sbp_mmhg ~ (1 | Window_ID), data=data.office)

#Fit model I for DBP
glm1.d <- lmer(dbp_mmhg ~ (1 | Window_ID), data=data.office)


#Model II

#Fit multivariate model II for SBP
glm2.s<-lmer(sbp_mmhg ~ malperfusion  + demorace +prepregbmi + age
    + preeclampsia + famhbp
          +famheart + (1 | Window_ID), data=data.office)
```

```
#Fit multivariate model II for DBP
glm2.d<-lmer(dbp_mmhg ~ malperfusion  + demorace +prepregbmi + age
    + preeclampsia + famhbp
            +famheart + (1 | Window_ID), data=data.office)

#The code for multivariate model II can be easily transformed for
   fitting univariable model
#II by changing the covariates, for example, univariate model for
   SBP with malperfusion
glm3.s<-lmer(sbp_mmhg ~ malperfusion + (1 | Window_ID), data=data.
   office)



#Model III

library(splines)
library(chron)#package to deal with time variable

#Fit the natural cubic splines for SBP
z.s<-ns(data.office$mean_bp_time,
        knots=chron::times(c('08:51:40','10:20:00','12:02:00','
           13:43:00')),
        Boundary.knots = chron::times(c('06:54:00','15:32:00')))
NCS1<-lm(sbp_mmhg ~ z.s,data = data.office)

#Fit the natural cubic splines for DBP
z.d<-ns(data.office$mean_bp_time,
        knots=chron::times(c('08:51:40','10:20:00','12:02:00','
           13:43:00')),
        Boundary.knots = chron::times(c('06:54:00','15:32:00')))
```

```
NCS1.d<-lm(dbp_mmhg ~ z.s,data = data.office)

#Fit model III for SBP
ncsglm1.s<-lmer(sbp_mmhg ~ malperfusion  + z.s + prepregbmi +
   demorace +
            preeclampsia + age + famhbp + famheart + (1 | Window_
               ID), data=data.office)

#Fit model III for DBP
ncsglm1.d<-lmer(dbp_mmhg ~ malperfusion  + z.d + prepregbmi +
   demorace +
            preeclampsia + age + famhbp + famheart + (1 | Window_
               ID), data=data.office)



#Model IV

#The semiparametric mixed-effects model (slm) function is built in
    the package "assist"
library(assist)

#Fit model IV for SBP
#data1 is data.office after excluding all missing values because
   slm() function does not allow
#missing value
slm1.s<- slm(sbp_mmhg~mean_bp_time + malperfusion + demorace +
   prepregbmi +
             preeclampsia + age + famhbp + famheart,
             rk=cubic(chron::times(mean_bp_time)),
             random=list(Window_ID=~1), data=data1)
```

```
#Fit model IV for DBP
slm1.d<- slm(dbp_mmhg~mean_bp_time + malperfusion + demorace +
    prepregbmi +
                preeclampsia + age + famhbp + famheart,
                rk=cubic(chron::times(mean_bp_time)),
                random=list(Window_ID=~1), data=data1)
```

Bibliography

[1] Dena Ettehad, Connor A Emdin, Amit Kiran, Simon G Anderson, Thomas Callender, Jonathan Emberson, John Chalmers, Anthony Rodgers, and Kazem Rahimi. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. *The Lancet*, 387(10022):957–967, 2016.

[2] Kenneth D Kochanek, Sherry L Murphy, Jiaquan Xu, and Elizabeth Arias. Deaths: final data for 2017. *National Vital Statistics Reports*, 68(9), 2019.

[3] Catherine M Loria, Kiang Liu, Cora E Lewis, Stephen B Hulley, Stephen Sidney, Pamela J Schreiner, O Dale Williams, Diane E Bild, and Robert Detrano. Early adult risk factor levels and subsequent coronary artery calcification: the cardia study. *Journal of the American College of Cardiology*, 49(20):2013–2020, 2007.

[4] Janet W Rich-Edwards, Abigail Fraser, Deborah A Lawlor, and Janet M Catov. Pregnancy characteristics and women's future cardiovascular health: an underused opportunity to improve women's health? *Epidemiologic reviews*, 36(1):57–70, 2014.

[5] Paul K Whelton, Robert M Carey, Wilbert S Aronow, Donald E Casey, Karen J Collins, Cheryl Dennison Himmelfarb, Sondra M DePalma, Samuel Gidding, Kenneth A Jamerson, Daniel W Jones, et al. 2017 acc/aha/aapa/abc/acpm/ags/apha/ash/aspc/nma/pcna guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the american college of cardiology/american heart association task force on clinical practice guidelines. *Journal of the American College of Cardiology*, 71(19):e127–e248, 2018.

[6] Carlene MM Lawes, Stephen Vander Hoorn, Anthony Rodgers, et al. Global burden of blood-pressure-related disease, 2001. *The Lancet*, 371(9623):1513–1518, 2008.

[7] Heather M Johnson, Carolyn T Thorpe, Christie M Bartels, Jessica R Schumacher, Mari Palta, Nancy Pandhi, Ann M Sheehy, and Maureen A Smith. Undiagnosed hypertension among young adults with regular primary care use. *Journal of hypertension*, 32(1):65, 2014.

[8] Rebecca N Baergen. *Manual of Benirschke and Kaufmann's pathology of the human placenta*. Springer Science & Business Media, 2005.

[9] R Kelly, C Holzman, P Senagore, J Wang, Y Tian, MH Rahbar, and H Chung. Placental vascular pathology findings and pathways to preterm delivery. *American journal of epidemiology*, 170(2):148–158, 2009.

[10] Alfredo M Germain, Jorge Carvajal, Marta Sanchez, Guillermo J Valenzuela, Harumi Tsunekawa, and Benedicto Chuaqui. Preterm labor: placental pathology and clinical correlation. *Obstetrics & Gynecology*, 94(2):284–289, 1999.

[11] Yeon Mee Kim, Emmanuel Bujold, Tinnakorn Chaiworapongsa, Ricardo Gomez, Bo Hyun Yoon, Howard T Thaler, Siegfried Rotmensch, and Roberto Romero. Failure of physiologic transformation of the spiral arteries in patients with preterm labor and intact membranes. *American journal of obstetrics and gynecology*, 189(4):1063–1069, 2003.

[12] Roberto Romero, Sudhansu K Dey, and Susan J Fisher. Preterm labor: one syndrome, many causes. *Science*, 345(6198):760–765, 2014.

[13] Puja K Mehta, Margo Minissian, and C Noel Bairey Merz. Adverse pregnancy outcomes and cardiovascular risk factor management. In *Seminars in perinatology*, volume 39, pages 268–275. Elsevier, 2015.

[14] T Yee Khong, Eoghan E Mooney, Ilana Ariel, Nathalie CM Balmus, Theonia K Boyd, Marie-Anne Brundler, Hayley Derricott, Margaret J Evans, Ona M Faye-Petersen, John E Gillan, et al. Sampling and definitions of placental lesions: Amsterdam placental workshop group consensus statement. *Archives of pathology & laboratory medicine*, 140(7):698–713, 2016.

[15] Michael A Weber, Jan IM Drayer, Dina K Nakamura, and Frederic A Wyle. The circadian blood pressure pattern in ambulatory normal subjects. *The American journal of cardiology*, 54(1):115–119, 1984.

[16] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

[17] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press, 1993.

[18] Yuedong Wang. *Smoothing splines: methods and applications*. CRC Press, 2011.

[19] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[20] Andrzej Gałecki and Tomasz Burzykowski. Linear mixed-effects model. In *Linear Mixed-Effects Models Using R*, pages 245–273. Springer, 2013.

[21] Yuedong Wang. Mixed effects smoothing spline analysis of variance. *Journal of the royal statistical society: Series b (statistical methodology)*, 60(1):159–174, 1998.