FUNCTIONAL DATA ANALYTIC INFERENCE FOR SYSTEMS GOVERNED BY DIFFERENTIAL EQUATIONS WITH APPLICATIONS

Siddhartha Mandal

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2012

Approved by:

Dr. Pranab K. Sen

Dr. Shyamal D. Peddada

Dr. Amy Herring

Dr. Mary Paine

Dr. John S. Preisser

Abstract

SIDDHARTHA MANDAL: FUNCTIONAL DATA ANALYTIC INFERENCE FOR SYSTEMS GOVERNED BY DIFFERENTIAL EQUATIONS WITH APPLICATIONS
(Under the direction of Dr. Pranab K. Sen and Dr. Shyamal D. Peddada)

The objective of this dissertation research is to develop formal statistical methodology for analyzing systems governed by ordinary differential equations (ODE). Ordinary differential equations are commonly used to describe a wide variety of biological and physiological phenomena. They arise in the description of gene regulatory networks, study of HIV dynamics and other infectious diseases and toxicology . This work is motivated by physiologically based pharmacokinetic (PBPK) models in toxicology which are deterministic models used to describe chemical kinetics in human or animal physiology. These models relate the concentration of chemicals in tissues and blood to their rates of change and physiological parameters, such as tissue volume and blood flow, and metabolic parameters among others, through a system of ODEs. Usual strategies of analyzing such models involve non-linear least squares methodology which can potentially be computationally intensive. Often, some of the existing procedures for modeling ODEs do not necessarily account for inter and intra-individual variability that are common in multi-subject experiments. Using functional data analytic methods, in this dissertation research, we provide a formal statistical framework for drawing statistical inferences regarding subject specific and population specific parameters in models governed by a system of ODE. One of the main features of the proposed methodology is to cast the problem in a constrained inferential framework and thus avoid solving the differential equations, which is often challenging and

time consuming. Such a formulation allows for the possibility that all components of the ODE may not adequately describe the underlying biological phenomena. The proposed framework also allows the researcher to estimate both within and between subject variability, while drawing statistical inferences at the individual as well as the population level. We make as few assumptions as possible while taking into account the underlying structure in the data. The proposed framework allows researchers to compare parameters among several populations, such as different dose groups, while adjusting covariates, whether discrete or continuous. Such inferences were not possible until now. We illustrate the proposed methodology using some simulated datasets as well as a real dataset on benzene concentration in exhaled breath.

Acknowledgements

This dissertation work could not have been made possible without the support of a few people, whom I would like to acknowledge. First and foremost I would like to acknowledge my advisors Dr. Pranab K. Sen and Dr. Shyamal D. Peddada for being excellent mentors throughout the period of my dissertation research. Dr. Sen was instrumental in introducing me to the opportunity to work at National Institute of Environmental Health Sciences which has been immensely helpful not only for the research but also for professional development. Also he always encouraged me in research and never let panic creep in even in the toughest of situations, which are ever so common in graduate life. I could approach him at anytime with any sort of problem and he would always listen carefully and advise with a smile on his face.

Dr. Peddada introduced me to the fascinating research problem in toxicology which dealt with a real world problem related to mechanistic modeling. Coming from a statistical background, the problem was particularly novel to me with its mix of statistics and biology. As a mentor, Dr. Peddada was always approachable and showed immense interest in my research. He always found time whenever I had questions or needed help in my research and was extremely supportive and encouraging towards any new ideas. Dr. Peddada also took keen interest in my writing and presentation skills and devoted a lot of time to correct my manuscripts and giving me tips to improve.

I would like to thank my committee members, Dr. Amy Herring, Dr. Mary Paine and Dr. John S. Preisser. They were extremely helpful and appreciative in evaluating my work and provided helpful suggestions to enhance the quality of my work. Both

Dr. Herring and Dr. Preisser took a lot of interest in my work and their questions and criticisms were of great help. Dr. Paine provided valuable inputs regarding the toxicological aspects of the work and it certainly helped me broaden my outlook. The detailed comments she provided have enriched the dissertation greatly.

The time I spent at UNC Chapel Hill and NIEHS has been a wonderful experience and I consider myself lucky to have been part of two such dynamic research environments. I have been lucky to be in company of some wonderful people at both places. I would especially like to thank the library at UNC whose huge journal collections have been my main source of reference. Also the town of Chapel Hill has been a wonderful place to live at.

A special thanks to Prof. Debasis Kundu and my other teachers at IIT Kanpur for helping me build a strong foundation for the research and academic career in statistics. Last but not least I would like to thank my parents and friends for all the support they have shown during the last four years. A special thanks to Sangita, Shyamal, Somenath, Ayon and Debartha for hours spent in wonderful discussions, both funny and serious.

Table of Contents

List of Tables

List of Figures

Chapter 1

MOTIVATION

Many biological and chemical systems or processes can often be described using a system of ordinary differential equations (ODE). Some common areas of application include gene regulatory networks, viral dynamics, modeling of infectious diseases, immunology and toxicology. Researchers developing gene regulatory networks are often interested in studying the changes in mRNA and protein concentrations over time (Polynikis et al. (2009)). The rate of change in the concentration of mRNA of a particular gene (and it's related protein) depends on the instantaneous concentrations, degradation and transcription rates, which can be mathematically expressed through mass balance equations using ODE. The system of ODE involved in a network provides a mathematical model for the gene regulatory network. In immunology, a question of interest is to study the effect of viral infection on the human immune response system. Viral infection induces a response in the system that converts normal lymphocytes to antibody producing cells that resist and destroy viruses. The changes in concentration of these cells over time, which can be modeled using ODE, provide valuable information on immune response (Oprea et al. 2000). Similarly, in the case of HIV infection, one may study the dynamics of viral infection by observing the population of naive cells, infected cells or viral load and their changes over time (Perelson

(2002)). Again ODE can be used to mathematically describe the phenomenon. There are numerous such examples in biological and physical sciences where the underlying processes are described using a system of ODE containing unknown model parameters that are to be statistically estimated. In this dissertation we focus on applications to toxicology where researchers model chemical kinetics, or the flow of chemicals in human or animal body using a system of ODE. Although this work is motivated by an application in toxicology, as described in this dissertation, the methodology developed here can be adapted to other contexts such as those described above.

Humans are exposed to a vast array of compounds, some of which are potentially toxic and even carcinogenic while some are potentially beneficial to human health. In classical toxicology researchers often investigate the toxicity of a chemical by determining the proportion of animals with adverse response (such as tumors or lesions) at a given dose of the chemical. A pharmacologist may conduct similar studies but is often interested in the opposite challenge, namely, to identify the efficacy of a drug. Although such studies are important to determine whether a particular chemical is toxic or beneficial, they do not necessarily explain a chemical's mode of action. For example, when an animal is exposed to a chemical in the classical 2-year bioassay conducted by the National Toxicology Program, a toxicologist can determine whether the animal had a particular lesion or not. From the observed outcome, they cannot determine how the body processed the chemical, although such determinations are extremely important to understand the underlying mechanisms or mode of action of the compound. As noted in Reddy et al. (2005), toxicologists and pharmacologists are often interested in quantitatively investigating factors that determining the processes of absorption, distribution, metabolism and excretion (ADME) in various parts of the body such as, stomach, blood, liver, kidney etc. "Pharmacokinetics may be simply defined as what the body does to the drug" (Benet (1984)).

Suppose a person inhales a volatile chemical compound such as benzene. Then the flow of the chemical through the person's body can be described using the flow diagram in Figure 1.1. The diagram is a schematic representation of human physiology. Note that there exist variations to Figure 1.1 available in the literature to describe kinetics of benzene and other volatile chemicals in general. A flow diagram as shown in Figure 1.1 can be mathematically modeled using a system of ODE (as described in Section 2.1).

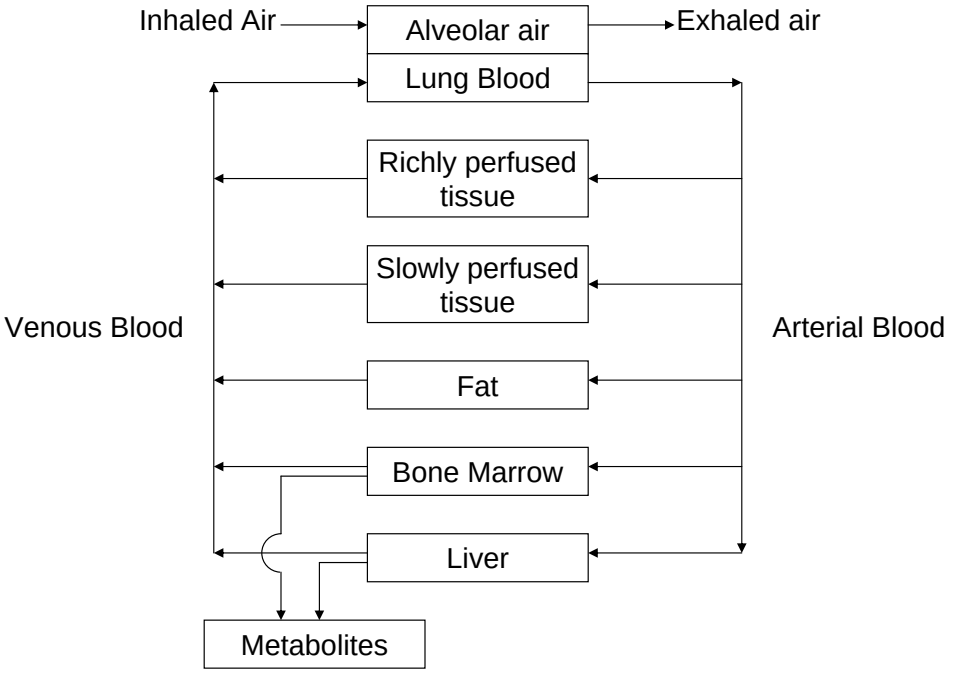

Figure 1.1: A physiologically based pharmacokinetic model for volatile compounds (Travis et al. (1990)).

Using such a model and the available data, a toxicologist is often interested in understanding the chemical kinetics of benzene. Thus the typical problem of interest is to estimate the unknown parameters of the model, along with their uncertainty

estimates, that are specific to subject as well as population. Researchers are often also interested in comparing the model parameters between two or more groups after adjusting for covariates. The goal of this dissertation work is therefore to develop a formal statistical framework and methodology for addressing these issues. Although this dissertation work is motivated by data from toxicology, the methodology is fairly general and can be applied to other contexts, such as those mentioned earlier.

Chapter 2

LITERATURE REVIEW

2.1 Physiologically Based Pharmacokinetic Modeling

Pharmacokinetics of a chemical can be modeled through systems of differential equations that represent mass balance equations for the flow of chemicals. Unlike in usual statistical methodology, the model is specified to describe the derivatives of the response variable and not the response itself. Due to the absence of closed form solutions to the system of differential equations in most complex systems, we do not have an explicit model describing the dependence of the response on the parameters of interest from differential equation models. Hence the need for novel statistical methods for analysis of the class of models governed by differential equations, not necessarily restricted to pharmacokinetic modeling. For the current review, we shall mainly focus on dynamic systems in toxicological and pharmacological areas.

The analysis of chemical kinetics can be compartmental or non compartmental. Non-compartmental methods rely on estimating the total exposure to a chemical. One such non-compartmental technique to estimate the total exposure is to study the area under the curve (AUC) of a concentration-time graph (Denker et al. 2002). For other metrics such as area under the moment curve (AUMC) or mean residence

time (MRT), one may refer to Dunne and King (1989). On the other hand compartmental models try to estimate the concentration-time graph using kinetic models. The main idea behind compartmental pharmacokinetic models is that the animal or human body can be thought of as a collection of interconnected compartments, usually the organs or tissues. Blood serves as the medium of transport for the chemical between these compartments. The chemical flow typically is modeled through first order kinetic mass balance equations that relate the rate of change of the concentration of drug in each compartment to the present concentrations. Usually, these equations are first order linear and/or non-linear differential equations with several parameters. Pharmacokinetic modeling deals with understanding of chemical kinetics through estimation and inference about these parameters. More specifically, population pharmacokinetic modeling deals with studying the sources of variability that affects the pharmacokinetics of a chemical for a group of individuals or a species. The main objectives of population PK analysis is to identify and quantitatively estimate variabilities affecting the pharmacokinetics for a population (Steimer et al. (1994), Davidian and Giltinan (1995)).

In general pharmacokinetic modeling, volume of distribution and clearance are two of the most important parameters that define the kinetics of a chemical while being of biological use. For intravenous administration of a chemical, volume of distribution is the volume in which a chemical is distributed immediately after administration in the physiology. Clearance is the volume of blood/plasma that is cleared of the chemical per unit time. There exist multi-compartment PK models involving multiple volumes of distribution and clearances for each compartment. In spite of being useful to describe the pharmacokinetics of a population, the PK models and parameters are not targeted to reflect the actual physiological and anatomical structures. This gave rise to physiologically based pharmacokinetic (PBPK) modeling. These are models that

incorporate physiological knowledge along with the processes of absorption, distribution, metabolism and excretion (Reddy et al. (2005)). The differential equations defining the model are characterized by parameters that reflect the physiology, for instance tissue volumes or partition coefficients. This makes the model physiologically more realistic or interpretable and eventually facilitates the applicability to varying situations. In cases where the efficacy or toxicity of a chemical compound is highly related to the concentration in the target tissue rather than the plasma concentration, PBPK models are more useful than their PK counterparts (Yang et al. (2004)). Further, one can infer about mechanistic behavior for the chemical of concern through these models (Reddy et al. (2005)). An important use of PBPK models is extension between different dose levels, routes of exposure and even between species (Gargas et al. (2000)). Humans are commonly exposed to chemicals at low dose levels, for instance through occupational exposure for a longer time. Simulating such an experiment in laboratory settings would turn out to be infeasible. In such cases, PBPK models can be used to extend the results from the animal studies using higher doses. In some cases, a mere change in the value of the parameter would be enough while in other cases the model structure might need to be changed (For instance, in case of a pregnant woman, equations for the fetus need to be included).

PBPK models have been developed for several classes of chemical compounds, such as halogenated alkanes, hydrocarbons, aromatic compounds, environmental pollutants and metals to name a few. For example, O'Flaherty et al. (2001) described a PBPK model for Chromium VI, Gentry et al. (2004) described it for arsenic and Kawahara et al. (1999) for the drug digoxin. These models have different characteristics due to the inherent chemical nature of the compounds and consequently the organs involved in the ADME of the compound. The models also vary according to the route of exposure or path of administering the dose (such as inhalation, dermal,

gavage or intravenous). PBPK models are constantly subject to change in model structure to incorporate possible metabolic pathways and metabolite formation to obtain a better understanding of the underlying true phenomena.

The importance of PBPK modeling with respect to public health stems from a number of facts. Chemicals such as lead or arsenic have adverse health effects in humans (Porba et al. (2011)). Pollutants such as PERC or TCE have widespread occupational uses while drugs contain several chemicals which may be beneficial or harmful for the human physiology. PBPK models provide a framework to answer questions about the mechanism of actions of such chemicals (Sweeney et al. (2009)). These models also find use in risk assessment studies since behavior of the chemicals change with dose levels, species or route of exposure (Haddad et al. (2001)). Valid methods describing the chemical actions would be advantageous in categorizing risk associated with these attributes. The widespread application of differential equation models as noted above makes it an important area for development of formal statistical methodology for estimation and testing purposes.

2.1.1 General Model Structure

The examples stated earlier and their modeling approaches have a few things in common. They are all defined by a set of differential equations governing the dynamic processes. The general structure of a model defined through differential equations is illustrated below.

Consider a model with p compartments. Let $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)'$ denote the vector of state variables in the p states. The system of differential equations that describe the rate of change in $\mathbf{Z}$ at time t be given by:

$$\dot{\mathbf{Z}} = \mathbf{F}(\mathbf{Z}, t, \boldsymbol{\theta}), \tag{2.1}$$

where $\dot{\mathbf{Z}} = \frac{d}{dt}\mathbf{Z}$ denotes the rate of change in $\mathbf{Z}$ at time t, $\mathbf{F}(.)$ is a $p \times 1$ vector of known functions, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)'$ is a $m \times 1$ vector of unknown parameters.

Here we provide the differential equations for the PBPK model of benzene that we are trying to analyze using the data from the benzene inhalation experiment. This would help illustrate the general model structure that we have presented. The PBPK model for benzene inhalation is a five compartment model as shown in Figure 1.1. The inhaled benzene gets absorbed into the lung blood and is transported to the various compartments through arterial blood. Benzene is transported back into lungs through venous blood. There is no direct exchange of benzene between compartments. Metabolites form within the metabolizing compartments, bone marrow and liver.

In the model equations, A_j denotes the amount of the chemical in the jth compartment, CV_j is the concentration of the chemical in the venous blood at the jth tissue exit, Q_j denotes blood flow rate, V_j is the volume and p_j is the partition coefficient for the jth compartment. Q_t and Q_{alv} are the total blood flow and the alveolar blood flow respectively. P_b denotes the blood:air partition coefficient. Among the metabolic parameters, V_{max} denotes the maximum metabolism rate and k_m is the concentration at which reaction rate is half of V_{max} for metabolizing sites. C_{art}, C_{ven} and C_{exh} represent the concentration of benzene in arterial blood, venous blood and exhaled breath respectively. The physiological parameters describing the PBPK model for benzene as proposed by Travis et al. (1990) are listed in Table 2.1. Body weight is denoted as bwt in the table.

For non-metabolizing sites, namely richly perfused tissues(rpt), slowly perfused

Table 2.1: Values of physiological parameters for benzene PBPK model for human subjects (Travis et al. (1990)).

Physiological parameter (Units)	Symbol	Value
Total cardiac output (l/min)	Q_t	6.2
Alveolar blood flow (l/min)	Q_{alv}	5
Tissue blood flow rates (l/min)		
Richly perfused tissue	Q_{rpt}	0.44*Q_t
Slowly perfused tissue	Q_{spt}	0.15*Q_t
Fat	Q_{fat}	0.05*Q_t
Bone marrow	Q_{bm}	0.05*Q_t
Liver	Q_{liv}	0.25*Q_t
Tissue volumes (kg)		
Richly perfused tissue	V_{rpt}	0.25*bwt
Slowly perfused tissue	V_{spt}	0.58*bwt
Fat	V_{fat}	0.19*bwt
Bone marrow	V_{bm}	0.05*bwt
Liver	V_{liv}	0.026*bwt
Blood:Tissue partition coefficients (dimensionless)		
Richly perfused tissue	p_{rpt}	1.49
Slowly perfused tissue	p_{spt}	2.03
Fat	p_{fat}	1.49
Bone marrow	p_{bm}	16.22
Liver	p_{liv}	1.49
Blood:Air partition coefficient (dimensionless)	P_b	7.4

tissues(spt) and fat, the differential equations are

$$\frac{dA_j}{dt} = Q_j(C_{art} - CV_j), \quad j \in \{rpt, spt, fat\}. \tag{2.2}$$

(2.2) represents mass balance equations for the three classes of tissues using first order kinetics. Here the rate of change in the amount of benzene in these tissues is directly proportional to the instantaneous concentration in the tissue. Blood flow serves as the rate constant since benzene is being delivered through blood.

Bone marrow (bm) and liver (liv) are possible metabolizing sites for benzene. Along with the usual first order kinetics, these compartments have an extra term that shows non-linear Michaelis-Menten kinetics for the formation of metabolites such as phenol. The equations for these sites and the metabolites (met) are as follows:

$$\frac{dA_j}{dt} = Q_j(C_{art} - CV_j) - \frac{V_{max(j)}CV_j}{(k_{m(j)} + CV_j)}, \quad j \in \{bm, liv\}, \tag{2.3}$$

$$\frac{dA_{met}}{dt} = \frac{V_{max(bm)}CV_{bm}}{(k_{m(bm)} + CV_{bm})} + \frac{V_{max(liv)}CV_{liv}}{(k_{m(liv)} + CV_{liv})}. \tag{2.4}$$

The concentrations in venous and arterial blood are expressed as

$$C_{ven} = \frac{\sum_{j \in \{rpt,spt,fat,bm,liv\}} Q_j CV_j}{Q_t}, \tag{2.5}$$

$$C_{art} = \frac{Q_t C_{ven} + Q_{alv} C_{in}}{Q_t + \frac{Q_{alv}}{p_b}}, \tag{2.6}$$

$$C_{exh} = C_{art}/P_b. \tag{2.7}$$

Concentration in the venous blood at jth tissue exit is expressed as

$$CV_j = \frac{A_j}{V_j p_j}. \tag{2.8}$$

(2.2)-(2.8) represent the PBPK model for benzene inhalation shown in Figure 1.1. We relate this system of ODE with the notation of the general model structure. The amount of benzene in the tissue compartments such as fat, liver etc. is considered as **Z**. The functional dependence (linear and/or non-linear) of the rates of change on the state variables, parameters as shown in (2.2)-(2.8) constitute **F**. In case of PBPK models, the actual number of parameters are large, more than 30 in some cases. Treating all parameters as unknown often renders the problem as inestimable due to sparsity of data. To circumvent this problem, often in case of animal or human PBPK

models, most of the physiological parameters are derived from toxicological literature or previous studies. For example, the blood or tissue volume, blood flow through tissues are similar for humans in general. Hence these physiological parameters are derived from literature or previous studies and held constant while constructing and analyzing the PBPK models. On the other hand some parameters like metabolic constants or Michaelis-Menten constants are not experimentally determined and hence serve as the parameters of interest. In order to analyze the PBPK model, we assume that it is identifiable with respect to these unknown parameters. In the benzene PBPK model, the metabolic parameters V_{max} (in $\mu g/min$) and k_m (in $\mu g/l$) for liver and bone marrow compartments comprise the unknown parameter $\boldsymbol{\theta}$.

We can similarly formulate the gene regulatory network problem in this framework. $\mathbf{Z}$ would represent mRNA and protein concentrations of genes that are measured over time as well as the genes that are not measured but are known to be in the gene regulatory network. The question of interest is to infer about the unknown parameters ($\boldsymbol{\theta}$) which characterize the ODE describing the gene regulatory network. Most often these parameters are rates of the processes of transcription, translation or degradation of mRNA or proteins.

Several other applications such as viral dynamics, infectious disease modeling and immunology use ODE systems to model biological phenomena. We summarize the formulation of these applications in terms of state variables, functions and parameters for a few such areas in Table 2.2.

2.1.2 Issues with Data and Design

Experimental design and data available are important facets of this research. Toxicological studies usually record the incidence of adverse reactions (tumors or lesions)

Table 2.2: Some applications of the proposed methodology.

	State variable	Functions	Parameters
Gene regulatory network	mRNA concentrations or gene expressions	Factors of translation, transcription, degradation etc.	Rates of the factors affecting concentration or expression
Viral dynamics	Density of naive cells, infected cells, dead cells	Factors of infection, death etc.	Rates of viral infection, susceptibility etc.
Infectious diseases	Number of susceptible, infectious and cured people	Factors influencing birth, death, infection	Rates of contact, transition
Immunology	Proportion of centroblasts, centrocytes	Factors of mutation, selection, proliferation	Rates of mutation, selection etc.
Toxicology	Amounts or concentration of chemicals	Factors of chemical processes, eg: absorption, metabolism	Metabolic parameters, Rates of reaction

in exposed subjects, after exposure to a dose of a chemical. On the other hand, studies recording chemical concentrations in the body measure the concentration of the chemical over time in fluids such as blood or urine. Tissue samples are usually not

available, since getting repeated samples from liver, kidney or muscles from human or animal subjects is not feasible. However in certain animal studies, where the animals are sacrificed at the end of the study, tissue samples may be available. Hence in analysis of PBPK models, data are available for at most two or three compartments.

Measurement times are another important aspect of design and data. Chemical concentrations are measured at specified time points after exposure to dose. Compartments that are measured may not have same measurement times. For example, it is easy to get urine measurements more frequently, however getting blood measurements too frequently may be difficult in case of rats and mice. For the same reason, subjects may be measured at only a subset of the measurement times. The frequency of dosing is also important since some studies are single dosage while others are multiple dosage studies. All these factors must be taken into account while planning the analysis of differential equation models.

2.2 Non-linear Least Squares and Bayesian Approaches

In this section we describe some existing methods to analyze PBPK models. One of the most widely used method is the non-linear least squares method as used in Parham et al. (2002). The overall idea of this approach is to minimize the distance between the data and the solution of the system of differential equations. Since the differential equations do not have an explicit solution, the method involves the numerical optimization of the differential equations using algorithms such as Nelder-Mead algorithm or the Runge-Kutta method and obtaining the residual sum of squares between the data and predicted values of the observed compartments. Mathematically, $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \parallel \mathbf{Z} - \mathbf{Z}(\boldsymbol{\theta}) \parallel^2$, where $\mathbf{Z}(\boldsymbol{\theta})$ is the numerical solution of the system given by (2.1) for a particular $\boldsymbol{\theta}$. It is a non-linear optimization problem and there are several computational and statistical challenges associated with this approach.

Notable computational difficulties include lack of convergence or convergence to local solution, computation time, convergence to local solutions due to irregular surface of the residual sum of squares, especially for stiff differential equation systems (Ramsay et al. (2007)). Due to these issues, it is difficult to obtain valid estimates of variability, drawing statistical inferences and testing hypotheses about the parameters of interest.

In the last two decades, significant research has been conducted in the field of non-linear regression in repeated measurements (Davidian and Giltinan 1995). This methodology has been applied to models defined by differential equations (Tornøe et al. 2004). As in the non-linear least squares methodology, these methods also rely on solving the system of differential equations and often using their first order Taylor series expansion as the regression function in a non-linear regression setup. In a mixed effects scenario, a hierarchical model is developed with the data centered around the numerical solution of the differential equations system and information is pooled over all individuals to obtain parameter estimates in a likelihood framework. This method still requires the solution of the differential equations and also is computationally intensive due to the integration over the random effects for each individual.

An alternative to the above approaches is the Bayesian methodology (Bois 2000). Hierarchical models are used to obtain the posterior distributions of the model parameters. There are three levels of modeling in this approach, namely individual model, parameter model and error model. Individual model specifies the distribution of the response variables, usually taken to be normal or lognormal. The solution of the system of differential equations is taken as the location of these distributions while the error model specifies the variability in the individual model. The prior distribution of parameters and the individual model provide the joint distribution of the data and the parameters, which is used to obtain the posterior distributions of

the model parameters. It is clear in this approach that it relies on the accuracy of the system of differential equations to model the phenomenon in the individuals and the population as a whole. Also diagnostics for these models have not been developed yet.

Recently in 2006 the US Environmental Protection Agency (EPA) conducted an international workshop entitled, "The International Workshop on Uncertainty and Variability in PBPK Models" to evaluate the available statistical methodology for analyzing PBPK models. In a publication resulting from the workshop, Barton et al. (2007) concluded that there is a need for a formal statistical methodology for analyzing PBPK models and to derive uncertainty estimates associated with parameter estimates.

2.3 Functional Data Analytic Approaches

Ramsay (1996) introduced a functional data analytic (FDA) method to solve the problem of parameter estimation in differential equation models known as Principal Differential Analysis (PDA). The methodology consists of approximating the state variables in the differential equations through some basis functions. These are linear combinations of some functions of time, such as polynomials or cubic splines. These approximations are made such that they also satisfy some regularization conditions. Usually the approximations are penalized by placing some constraints on the second or higher order derivatives of the approximations. The model parameters are estimated by minimizing the residuals using the data and the approximated values of the response variables. These methods have been mainly used in engineering dynamics problems, such as the constantly stirred tank reactor problems.

On similar lines, a smoothing approach for parameter estimation in differential equation models was proposed by Ramsay et al. (2007). In this paper, the authors

used a basis expansion using B-splines to approximate the state variables like previous instances. However, here the differential equations are treated as a regularization criterion, accompanied by a regularization parameter. So one could decide on how much confidence could be placed on the model itself. This comprised the inner optimization of the method. The outer optimization comprised of minimizing the predicted residuals. Approximate sampling variances of the parameter estimates were provided and behavior of the parameter estimates were studied when the value of the regularization parameter was large.

Poyton et al. (2006), Varziri et al. (2008b) carried forward the work on Principal Differential Analysis by introducing an iterated version. They carried out a simultaneous optimization procedure on the basis parameters as well as the model parameters. To arrive at the properties of the estimates, they used a maximum likelihood approach and denoted the estimator by Approximate Maximum Likelihood Estimator(AMLE). Varziri et al. (2008a) used AMLE in presence of unmeasured states and stationary and non-stationary model disturbances. These methods were shown to work on several engineering problems. However these were primarily for population estimation only. Individual inferences and estimation of variance components were not addressed in these works.

Liang and Wu (2008) approached a similar problem in a slightly different manner. A two stage smoothing approach was employed. They used local polynomial smoothing to approximate the response variables. But there was no regularization of the basis parameters based on the differential equations. Instead, for estimating the model parameters, the estimated values of the response variables and their derivatives were plugged into the differential equation system and the residual error in the differential equation system was minimized. No data points are involved in this stage of the methodology and hence it was named Pseudo Least Squares estimator. So it

was a method which was not exactly the same but close to the measurement error approach. Consistency and asymptotic normality were proved for the estimator under certain regularity conditions. The method was applied to simulated examples and also real data on HIV CD4 cell counts. This methodology requires all the state variables in the system of differential equations to be observable, which is rarely the case in toxicological modeling situations. Also the population estimation methodology was not discussed in this paper. In Table 2.3 we summarize the main features of the above discussed papers and our contribution.

Table 2.3: Main features of some recent papers based on Ramsay et al. (2007) compared to the proposed methodology.

Feature	Varziri et al. (2008b)	Liang and Wu (2008)	Proposed Methodology
1. Allows for unobservable components.	Yes	No	Yes
2. Basis approximation uses the data and the differential equations.	Yes	No	Yes
3. Subject specific inference.	No	Yes	Yes
4. Population based inference.	Yes	No	Yes
5. Separate estimation of variance components.	No	No	Yes

In this paper we extend Ramsay (1996) and Ramsay et al. (2007) methodology to conduct formal statistical inferences by taking into account potential correlations between and within compartments in a subject as well as between and within subject variability. Although the focus of this paper is on analyzing PBPK models, the methodology is sufficiently general and can be applied to other contexts according to the formulations described in Table 2.2.

2.4 Summary

Statistical treatment of differential equation models is a relatively newer area of research with respect to theory and methodology while boasting of a wide range of applications in public health and other fields. Traditional methods of analysis are ridden with problems of several kinds and hence newer methodology based on functional data analysis is an important alternative. Addressing issues such as parameter identifiability and design of experiments are important foundations while developing statistical methodology for these models and accounting for variability and uncertainty provides a complete statistical framework on these basic foundations.

In Chapter 3, the proposed estimation methodology based on functional data analysis is provided with the necessary details and justification for its use. Asymptotic properties are also shown in the same chapter. Chapter 4 describes the statistical methodology to incorporate covariate effects in models defined through systems of differential equations. In Chapter 5, all the proposed methodologies are illustrated using simulated data examples and a real life study based on the same Benzene PBPK model. Chapter 6 describes some of the future research problems that follows from the current work.

Chapter 3

ESTIMATION METHODOLOGY IN DIFFERENTIAL EQUATION MODELS

Modeling of systems of differential equations is different from usual statistical modeling in the sense that the model in consideration does not describe the response variable directly. The mathematical model describes the rate of change in response variables and often the dependence of the response on the parameters are not known explicitly. Some variants may also include algebraic equations that relate the response variables themselves, along with the differential equations. This unique situation demands a novel approach that combines the traditional statistical methods with newer methods of functional data analysis. Motivated by the assessment in Barton et al. (2007), this chapter intends to provide a statistically rigorous framework for analyzing PBPK models. We cast the statistical problem in a general framework by exploiting the functional data analytic methods available in the literature. In this chapter, we present the proposed methodology for estimation of the individual and population parameters and variability in the estimates along with confidence intervals.

3.1 Description

As described earlier, the state variables are related to each other through differential equations. Due to unavailability of explicit solutions, the functional data analytic approach provides an alternative way to solve this problem. The state variables, whether observed or unobserved are approximated using basis functions, which are functions of time. This provides a familiar platform based on regression which is used to estimate the model parameters, first on an individual basis and finally for population parameters. The proposed methodology thus tries to strike a balance between the data available and the system of differential equations that we consider.

Table 3.1 defines the common notations that will be used in this thesis henceforth. Other notations will be defined as they are used.

Table 3.1: Notations used in this work.

Symbol	Meaning
$\mathbf{Z}$	State variables
$\mathbf{Z^o}$	Observable state variables
$\mathbf{Z^u}$	Unobservable state variables
$\mathbf{Y}$	Data on observable state variables
$\boldsymbol{\theta}$	Model parameter
$\boldsymbol{\alpha}$	Basis parameter
$\phi(\Phi)$	Basis function(s)
λ	Regularization parameter
i	Index for individuals
j	Index for states
k	Index for time points

3.2 Methodology

We describe the methodology as two components, individual and population estimation of model parameters.

3.2.1 Individual Parameter Estimation

Let p denote the total number of states (or compartments) in a PBPK model. As often is the case, suppose only d out of p compartments are measurable for the ith subject, $i = 1, 2, .., n$. Suppose $\mathbf{Z}_i(t) = (Z_{i1}(t), Z_{i2}(t), \ldots, Z_{id}(t), Z_{i(d+1)}(t), \ldots, Z_{ip}(t))' = (\mathbf{Z}_i^o(t), \mathbf{Z}_i^u(t))'$ denotes the vector of state variables in the p states at time t, where $\mathbf{Z}^o$ denotes the observable and $\mathbf{Z}^u$ denotes the unobservable part. Let the system of differential equations that describe the rate of change in $\mathbf{Z}_i$ be given by:

$$\dot{\mathbf{Z}}_i = \mathbf{F}(\mathbf{Z}_i, t, \boldsymbol{\theta}_i), \tag{3.1}$$

where $\dot{\mathbf{Z}}_i = \frac{d}{dt}\mathbf{Z}_i$ denotes the rate of change in $\mathbf{Z}_i$ at time t, $\mathbf{F}(.)$ is a $p \times 1$ vector of known functions, $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{im})'$ is the vector of unknown parameters.

Let the observed value of the true state variable $\mathbf{Z}_i^o$ be denoted by $\mathbf{Y}_i$. Thus we have

$$\mathbf{Y}_i = \mathbf{Z}_i^o + \mathbf{e}_i, \tag{3.2}$$

where $\mathbf{e}_i$ is the component that captures the intra-individual variation in the data. In some experimental settings where none of the compartments are observed directly, the observed variable may be a function of the state variables in the model. For instance, if concentration in exhaled breath is the the only observed variable, it can be expressed as a weighted average of the concentrations in the other compartments, which are not observed. Such a case would imply $\mathbf{Y}_i = g(\mathbf{Z}_i^u) + \mathbf{e}_i$, where $g(.)$ is a known function from the specified model.

One may assume that $\mathbf{e}_{ij}$ are independently distributed for all i with mean $\mathbf{0}$ and $a_j \times a_j$ covariance matrix $\Sigma_{ij}(t)$. It is reasonable to assume that the intra-individual correlation between measurements would decrease with amount of separation in time.

Hence as commonly used in time-series models for a parsimonious description of dependence in data (for example, Box, 2008), we let

$$Corr(\mathbf{e}_{ij}(t_1), \mathbf{e}_{ij'}(t_2)) = \begin{cases} \rho_{i(j)}^{|t_1-t_2|} & j = j'; \\ \rho_{i(j,j')}^{|t_1-t_2|} & j \neq j'. \end{cases}$$

Lastly we model inter-individual variability by imposing a hierarchical structure on $\boldsymbol{\theta}_i$. We shall assume that $\boldsymbol{\theta}_i$ are identically and independently distributed with mean $\boldsymbol{\theta}$ and covariance matrix Ψ, which estimates the inter-individual variability.

We begin by estimating parameters for each individual subject. For notational simplicity, we drop the indices for individuals. The functional basis approach (Ramsay et al. 2007) is used to approximate the state variables $\mathbf{Z}$ in (3.1). This amounts to selecting a class of basis functions, such as spline or polynomials, to approximate $\mathbf{Z}_i$. In this article we use B-splines as the basis functions. Let $\tilde{\mathbf{Z}}_{ij}(t), j = 1, \ldots, p$ be the approximate value of compartment j at time t, described through L_j basis functions. Then

$$\tilde{\mathbf{Z}}_{ij}(t) = (\tilde{\mathbf{Z}}_{ij}^o(t), \tilde{\mathbf{Z}}_{ij}^u(t))' = \sum_{l=1}^{L_j} \phi_{ijl}(t)\alpha_{ijl} = \boldsymbol{\phi_{ij}}'(t)\boldsymbol{\alpha_{ij}},$$

where $\boldsymbol{\phi_{ij}}(t)$ is a vector of basis functions and $\boldsymbol{\alpha_{ij}}$ is a vector of unknown basis parameters.

Let $\boldsymbol{\alpha}_i = (\boldsymbol{\alpha}_{i1}, \ldots, \boldsymbol{\alpha}_{ip})'$ and $\boldsymbol{\Phi} = diag(\boldsymbol{\phi_{i1}}'(t), \ldots, \boldsymbol{\phi_{ip}}'(t))$. Hence the approximated state variable for the ith individual is

$$\tilde{\mathbf{Z}}_i(t) = (\tilde{\mathbf{Z}}_{i1}(t), \ldots, \tilde{\mathbf{Z}}_{ip}(t))' = \boldsymbol{\Phi}_i\boldsymbol{\alpha}_i. \tag{3.3}$$

The basis parameter vector $\boldsymbol{\alpha_i}$ are determined such that $\tilde{\mathbf{Z}}_i^o$ closely mimics the data and the state variables $\tilde{\mathbf{Z}}_i$ satisfies the system of differential equations. Hence both $\tilde{\mathbf{Z}}_i$ and $\tilde{\mathbf{Z}}_i^o$ are functions of the nuisance parameters $\boldsymbol{\alpha}_i$.

The objective is to estimate the unknown parameter $\boldsymbol{\theta}_i$ for each individual, however, $\boldsymbol{\alpha}_i$ and ρ_i are unknown nuisance parameters that need to be estimated. We begin by estimating $\boldsymbol{\alpha}_i$, ρ_i and λ_i such that the "distance" between the observed $\mathbf{Y}_i$ and the approximated value $\tilde{\mathbf{Z}}_i^o$ is minimized subject to $\tilde{\mathbf{Z}}_i$ satisfying the underlying differential equations, in similar lines as a ridge regression problem. We can formulate the problem of obtaining the approximation as follows:

$$\min_{\boldsymbol{\alpha}_i,\rho_i} \|\boldsymbol{\Sigma}(\rho_i)^{-1/2}(\mathbf{Y}_i - \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i))\|_2^2 \text{ subject to } \{\boldsymbol{\alpha}_i : \dot{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i = \mathbf{F}(\boldsymbol{\Phi}_i\boldsymbol{\alpha}_i, t, \boldsymbol{\theta}_i)\}$$

Thus, for a given $\boldsymbol{\theta}_i$, we minimize $H(\boldsymbol{\alpha}_i, \rho_i, \lambda_i)$ with respect to $\boldsymbol{\alpha}_i$, ρ_i and λ_i:

$$\begin{aligned} H(\boldsymbol{\alpha}_i, \rho_i, \lambda_i) =& (\mathbf{Y}_i - \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i))'\boldsymbol{\Sigma}(\rho_i)^{-1}(\mathbf{Y}_i - \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i)) \\ &+ \lambda_i \int (\dot{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i - \mathbf{F}(\boldsymbol{\Phi}_i\boldsymbol{\alpha}_i, t, \boldsymbol{\theta}_i))'(\dot{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i - \mathbf{F}(\boldsymbol{\Phi}_i\boldsymbol{\alpha}_i, t, \boldsymbol{\theta}_i))dt. \end{aligned} \tag{3.4}$$

In the above expression, λ_i is a regularization parameter which is estimated in the above objective function. The estimators $\hat{\boldsymbol{\alpha}}_i$, $\hat{\rho}_i$ and $\hat{\lambda}_i$ obtained by minimizing (3.4) are implicit functions of $\boldsymbol{\theta}_i$. Hence the estimated state variables $\widehat{\tilde{\mathbf{Z}}}_i$ and therefore $\widehat{\tilde{\mathbf{Z}}_i^o}$ are implicit functions of $\boldsymbol{\theta}_i$. For notational simplicity, we denote the predicted value of the observable state variable by $\widehat{\tilde{\mathbf{Z}}_i^o}(\boldsymbol{\theta}_i)$. Using the estimators derived by minimizing (3.4), we minimize the following quadratic form to estimate the model parameter $\boldsymbol{\theta}_i$:

$$S(\boldsymbol{\theta}_i) = (\mathbf{Y}_i - \widehat{\tilde{\mathbf{Z}}_i^o}(\boldsymbol{\theta}_i))'\boldsymbol{\Sigma}(\hat{\rho}_i(\boldsymbol{\theta}_i))^{-1}(\mathbf{Y}_i - \widehat{\tilde{\mathbf{Z}}_i^o}(\boldsymbol{\theta}_i)). \tag{3.5}$$

Expressions (3.4) and (3.5) are iteratively optimized until convergence. Using Taylor series expansion of $\hat{\boldsymbol{\theta}}_i$,we obtain an approximate covariance matrix of $\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i$ given by $\left(\frac{d\hat{\boldsymbol{\theta}}_i}{d\mathbf{Z}_i^o}\right)\boldsymbol{\Sigma}_i(\hat{\rho})\left(\frac{d\hat{\boldsymbol{\theta}}_i}{d\mathbf{Z}_i^o}\right)'$, henceforth denoted by $\Gamma(\boldsymbol{\theta}_i)$ in this paper.

3.2.2 Population Parameter Estimation

Using the estimators obtained for each individual subject, we now describe the method to estimate the population parameters. Towards this, as commonly done in non-linear mixed effects models, we assume the following hierarchical model structure for $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_i$.

Assumptions A:

($A.1$) $\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i$ are independently distributed with mean $\boldsymbol{\theta}_i$ and covariance $\Gamma(\boldsymbol{\theta}_i)$.

($A.2$) $\boldsymbol{\theta}_i|\boldsymbol{\theta}$ are i.i.d with mean $\boldsymbol{\theta}$ and covariance Ψ.

Consider the marginal distribution of $\hat{\boldsymbol{\theta}}_i$. Moments of the marginal distribution are

$$\begin{aligned} E(\hat{\boldsymbol{\theta}}_i) &= E_{\boldsymbol{\theta}}(E(\hat{\boldsymbol{\theta}}_i|\boldsymbol{\theta}_i)) = \boldsymbol{\theta}. \\ Cov(\hat{\boldsymbol{\theta}}_i) &= \Psi + \int \Gamma_i(\boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i|\boldsymbol{\theta}) d\boldsymbol{\theta}_i \\ &= \mathbf{V}_{\boldsymbol{\theta}}. \end{aligned}$$

Thus our population level model is,

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta} + \boldsymbol{\delta}_i, \quad Cov(\boldsymbol{\delta}_i) = \mathbf{V}_{\boldsymbol{\theta}}. \tag{3.6}$$

This is a Type III nonlinear marginal model described in Demidenko (2004). Our objective is to estimate the population parameter $\boldsymbol{\theta}$ and the covariance matrix of the estimator of $\boldsymbol{\theta}$. Since $V_{\boldsymbol{\theta}}$ is a function of $\boldsymbol{\theta}$, the classical iterated weighted least squares type methodology is not applicable here. Demidenko (2004) suggests a Total Generalized Estimating equation approach for such a formulation.

Note that the score equation $\frac{\partial l}{\partial \boldsymbol{\theta}} = 0$ reduces to

$$\sum_{i=1}^{n}\Big[\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) + \frac{1}{2}\mathbf{G}'[(\mathbf{V}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \otimes \mathbf{V}_{\boldsymbol{\theta}}^{-1}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})) - vec(\mathbf{V}_{\boldsymbol{\theta}}^{-1})]\Big] = \mathbf{0}, \qquad (3.7)$$

where $\mathbf{G} = \frac{\partial vec(\mathbf{V}_{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}}$. The Fisher information matrix is given by $\mathcal{I}_{\boldsymbol{\theta}} = \mathbf{V}_{\boldsymbol{\theta}}^{-1} + \frac{1}{2}\mathbf{G}'(\mathbf{V}_{\boldsymbol{\theta}}^{-1} \otimes \mathbf{V}_{\boldsymbol{\theta}}^{-1})\mathbf{G}$.

However TGEE is difficult to implement in the present problem, since the explicit form of $\mathbf{V}_{\boldsymbol{\theta}}$ is unknown and a likelihood framework may not be appropriate in this case. Hence we resort to the an empirical Bayes based technique and MINQUE methodology (Rao, 1972) to solve the problem.There exists a well developed literature on MINQUE and has been used in a wide range of contexts. In Zhang et al. (2000) MINQUE based methodology was developed for estimating variance components in non-linear mixed effects models, under heteroscedastic as well as homoscedastic errors. We exploit their methodology for deriving the starting values for $\boldsymbol{\theta}$ and Ψ for solving (3.7) iteratively. Note that MINQUE of Zhang et al. (2000) is itself is an iterative procedure involving iteration between two equations.

$$\begin{aligned}
\hat{\Psi} &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}})' \text{ and} \\
\hat{\boldsymbol{\theta}} &= \Big(\sum_{i=1}^{n}(\hat{\Psi} + \hat{\Gamma}_i)^{-1}\Big)^{-1}\sum_{i=1}^{n}(\hat{\Psi} + \hat{\Gamma}_i)^{-1}\hat{\boldsymbol{\theta}}_i. \qquad (3.8)
\end{aligned}$$

Equivalently it can be implemented as noted below. We start with initial estimates of $\boldsymbol{\theta}$ and Ψ, obtained from

$$\hat{\boldsymbol{\theta}}_{(0)} = \tfrac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_i \text{ and } \hat{\Psi}_{(0)} = \tfrac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{(0)})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{(0)})'.$$

We iterate between the two following steps:

Step 1 : Update estimate of $\boldsymbol{\theta}_i$ as

$$\hat{\boldsymbol{\theta}}_{i,(c+1)} = (\hat{\Gamma}_i^{-1} + \hat{\Psi}_{(c)}^{-1})^{-1}(\hat{\Gamma}_i^{-1}\hat{\boldsymbol{\theta}}_i + \hat{\Psi}_{(c)}^{-1}\hat{\boldsymbol{\theta}}_{(c)}),$$

where $\hat{\Psi}_{(c)}$ and $\boldsymbol{\theta}_{(c)}$ are the estimates from the cth iterate.

Step 2: Update the population parameter estimates as

$$\hat{\boldsymbol{\theta}}_{(c+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_{i,(c+1)} \text{ and } \hat{\Psi}_{(c+1)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{(c+1)})(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_{(c+1)})'.$$

Let $\hat{\Psi}$ and $\hat{\boldsymbol{\theta}}$ denote the estimates upon convergence. Let $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} = \hat{\Psi} + \frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i)$.

3.3 Asymptotic Properties of the Proposed Estimators

To explore the properties of the proposed estimators we need certain regularity assumptions as mentioned in Nagaraj and Fuller (1991).

Assumptions B:

- $(B.1)$ The components of the model function $\mathbf{F}(.)$ are continuous and twice differentiable for $\boldsymbol{\theta}^0 \in$ B, a closed ball, where $\boldsymbol{\theta}^0$ is the true parameter vector.
- $(B.2)$ The matrix of partial derivatives $\mathbf{D}(\boldsymbol{\theta}) = \dfrac{\mathbf{dF}}{\mathbf{d}\boldsymbol{\theta}}$ is of full rank with probability 1 in a neighborhood of $\boldsymbol{\theta}^0$.
- $(B.3)$ The matrix $B_t = H_t^{-1/2}\boldsymbol{\Lambda}(t)'\boldsymbol{\Lambda}(t)H_t^{-1/2}$ and B_t^{-1} converges to positive definite matrices for large t, where $H_t = diag(h_{iit})$ is a sequence of diagonal matrices such that $h_{iit} \to \infty$ as $t \to \infty$.

The iterative individual parameter estimation problem can be viewed as a constrained linear regression problem subject to non-linear constraints dictated by the system of differential equations. Since the observed data is a perturbation of the

approximated observable state variables, we have

$$\mathbf{Y}_i = \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i) + \boldsymbol{\epsilon}_i, \text{ subject to } f(\boldsymbol{\alpha}_i, \boldsymbol{\theta}_i) = 0. \tag{3.9}$$

Owing to the fact that $\mathbf{Z}_i = \boldsymbol{\Phi}_i \boldsymbol{\alpha}_i$, (3.9) can be formulated as

$$\mathbf{Y}_i = \boldsymbol{\Lambda}_i \boldsymbol{\theta}_i^* + \boldsymbol{\epsilon}_i, \text{ subject to } f(\boldsymbol{\theta}_i^*) = 0. \tag{3.10}$$

Then the unconstrained least square estimate $\hat{\boldsymbol{\theta}}_i^* = \boldsymbol{\theta}_i^* + [\boldsymbol{\Lambda}_i' \boldsymbol{\Lambda}_i]^{-1} \boldsymbol{\Lambda}_i' \boldsymbol{\epsilon}_i$.

Lemma 3.1 *(Nagaraja and Fuller, 1991): Under Assumptions B,* $H_t^{1/2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i) = O_p(1)$*, where* $\boldsymbol{\theta}_i$ *is the true parameter value for the ith individual.*

Hence as the number of time points increase, the individual parameter estimates $\hat{\boldsymbol{\theta}}_i$ is consistent for the true individual parameter value $\boldsymbol{\theta}_i$.

The individual parameter estimates obtained are $\hat{\boldsymbol{\theta}}_i$, $i = 1, \ldots, n$. Using delta method, we obtain an approximate covariance matrix of the estimator as $\Gamma(\boldsymbol{\theta}_i)$. Corresponding estimator of the covariance is $\Gamma(\hat{\boldsymbol{\theta}}_i)$. Since $\hat{\boldsymbol{\theta}}_i$ is consistent, by delta method, $\Gamma(\hat{\boldsymbol{\theta}}_i)$ is consistent for $\Gamma(\boldsymbol{\theta}_i)$, since $\Gamma(.)$ is a continuous function. This can be easily shown since $f(.)$ is continuous and $\hat{\boldsymbol{\theta}}_i$ is a continuous function of $\mathbf{Y}_i$.

We make certain model assumptions for the population parameter estimation in Assumption A in Section 3.2.2. Therefore the marginal covariance of $\hat{\boldsymbol{\theta}}_i$ is

$$\begin{aligned} E[\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}]' &= E[\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i + \boldsymbol{\theta}_i - \boldsymbol{\theta}][\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i + \boldsymbol{\theta}_i - \boldsymbol{\theta}]' \\ &= E[\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i][\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i]' + E[\boldsymbol{\theta}_i - \boldsymbol{\theta}][\boldsymbol{\theta}_i - \boldsymbol{\theta}]' \\ &= E[\Gamma(\boldsymbol{\theta}_i)] + \Psi. \end{aligned} \tag{3.11}$$

From convolution of density functions and Assumption (A.2), $E[\Gamma(\boldsymbol{\theta}_i)]$ is a function of $\boldsymbol{\theta}$ only. Hence $\hat{\boldsymbol{\theta}}_i$ are independent and identically distributed random variables.

The iterative population parameter estimation involves calculating $\hat{\boldsymbol{\theta}}$ and $\hat{\Psi}$. Let $\mathbf{W}_i = \Big(\sum_{i=1}^{n}(\Psi+\Gamma_i)^{-1}\Big)^{-1}(\Psi+\Gamma_i)^{-1}$ and $\hat{\mathbf{W}}_i = \Big(\sum_{i=1}^{n}(\hat{\Psi}+\hat{\Gamma}_i)^{-1}\Big)^{-1}(\hat{\Psi}+\hat{\Gamma}_i)^{-1}$. Then $\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n}\hat{\mathbf{W}}_i\hat{\boldsymbol{\theta}}_i$ and $\tilde{\boldsymbol{\theta}} = \sum_{i=1}^{n}\mathbf{W}_i\hat{\boldsymbol{\theta}}_i$. To prove the consistency of $\hat{\boldsymbol{\theta}}$, we first prove the consistency of $\tilde{\boldsymbol{\theta}}$ and then appeal to Slutsky's theorem.

From the Noether's condition for weighted least square estimators, if

(i) $\max_{1\le j\le n}[w'_{nj}(\mathbf{W}'_n\mathbf{W}_n)^{-1}w_{nj}] \to 0$ as $n\to\infty$ and

(ii) $\lim_{n\to\infty} n^{-1}(\mathbf{W}'_n\mathbf{W}_n) = \mathbf{V}^*$, finite and positive definite,

then $\tilde{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ for large n.

The weights $\mathbf{W}_i$ need to be estimated in our case. Hence we study the behavior of the variance components that serve as weights.

Theorem 3.1 *Under Assumptions A and B,*

(a) $\hat{\Psi} \xrightarrow{pr} \Psi$ *as* $t, n\to\infty$

(b) $\frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i) \xrightarrow{pr} E[\Gamma(\boldsymbol{\theta}_i)]$ *as* $t,n\to\infty$

Proof : *(a)* $\hat{\Psi}$ is the sum of square of residuals in the marginal individual parameter estimates.

$$\begin{aligned}
\hat{\Psi} &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i-\hat{\boldsymbol{\theta}})' \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}+\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}+\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})' + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' \\
&\qquad + (\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})' + (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})'\}. \qquad (3.12)
\end{aligned}$$

Consider the first term in (3.12).

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta})' \\
=&\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i+\boldsymbol{\theta}_i-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i+\boldsymbol{\theta}_i-\boldsymbol{\theta})' \\
=&\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)(\hat{\boldsymbol{\theta}}_i-\boldsymbol{\theta}_i)'+\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{\theta}_i-\boldsymbol{\theta})(\boldsymbol{\theta}_i-\boldsymbol{\theta})'+\frac{1}{n}\text{cross products.} \qquad (3.13)
\end{aligned}
$$

The second term in (3.13) converges to Ψ due to Assumption (A.2) as $n \to \infty$. The rest of the terms converge to 0 as $t \to \infty$ due to the individual level parameter consistency and as $n \to \infty$. Hence $\hat{\Psi} \xrightarrow{pr} \Psi$ as $n, t \to \infty$.

(b) To prove *(b)* it is enough to show that for any two $m \times 1$ vector $\mathbf{a}_1$ and $\mathbf{a}_2$, $\left|\mathbf{a}_1'\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i)-E[\Gamma(\boldsymbol{\theta}_i)]\right]\mathbf{a}_2\right| \xrightarrow{pr} 0$ as $t, n \to \infty$.

$$
\begin{aligned}
&\left|\mathbf{a}_1'\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i)-E[\Gamma(\boldsymbol{\theta}_i)]\right]\mathbf{a}_2\right| \qquad (3.14)\\
=&\left|\mathbf{a}_1'\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i)-\frac{1}{n}\sum_{i=1}^{n}\Gamma(\boldsymbol{\theta}_i)\right]\mathbf{a}_2+\mathbf{a}_1'\left[\frac{1}{n}\sum_{i=1}^{n}\Gamma(\boldsymbol{\theta}_i)-E[\Gamma(\boldsymbol{\theta}_i)]\right]\mathbf{a}_2\right| \\
=&|Z_1+Z_2| \\
\leq&|Z_1|+|Z_2|.
\end{aligned}
$$

Consider $Z_1 = \mathbf{a}_1'\left[\frac{1}{n}\sum_{i=1}^{n}\{\Gamma(\hat{\boldsymbol{\theta}}_i)-\Gamma(\boldsymbol{\theta}_i)\}\right]\mathbf{a}_2$. Recall for each i, $\Gamma(\hat{\boldsymbol{\theta}}_i)-\Gamma(\boldsymbol{\theta}_i)$ are marginally independent and identically distributed random variables and $\Gamma(\hat{\boldsymbol{\theta}}_i)-\Gamma(\boldsymbol{\theta}_i)=o_p(1)$ for large t. Consider a random variable

$$
U_i=\begin{cases}\mathbf{a}_1'\{\Gamma(\hat{\boldsymbol{\theta}}_i)-\Gamma(\boldsymbol{\theta}_i)\}\mathbf{a}_2 & \text{if} \quad |\mathbf{a}_1'\{\Gamma(\hat{\boldsymbol{\theta}}_i)-\Gamma(\boldsymbol{\theta}_i)\}\mathbf{a}_2|<\epsilon \\ 0 & \text{otherwise}\end{cases}
$$

Note that $\sum_{i=1}^{n} P(U_i \neq \mathbf{a}_1'\{\Gamma(\hat{\boldsymbol{\theta}}_i) - \Gamma(\boldsymbol{\theta}_i)\}\mathbf{a}_2) < \infty$. By Khintchine equivalence lemma, Strong Law of Large Numbers(SLLN) holds for both sequences of random variables or none. For large values of t, $E(U_i) = 0$. Hence we can apply Khintchine Strong Law of Large Numbers on U_i for large n. Since SLLN holds for $U_i, i = 1, \ldots, n$, it holds for $\mathbf{a}_1'\{\Gamma(\hat{\boldsymbol{\theta}}_i) - \Gamma(\boldsymbol{\theta}_i)\}\mathbf{a}_2$. Hence $|Z_1| \xrightarrow{pr} 0$.

From Assumption (A.2), $\boldsymbol{\theta}_i$ are i.i.d random variables. By Strong Law of Large numbers $\frac{1}{n}\sum_{i=1}^{n}\Gamma(\boldsymbol{\theta}_i) \xrightarrow{pr} E[\Gamma(\boldsymbol{\theta}_i)]$, which implies $|Z_2| \xrightarrow{pr} 0$. Hence (3.14) converges in probability to 0 as $n, t \to \infty$.

Hence using Slutsky's theorem and Theorem A.1, the proposed population variance estimate $\widehat{\mathbf{V}}_{\boldsymbol{\theta}} = \hat{\Psi} + \frac{1}{n}\sum_{i=1}^{n}\Gamma(\hat{\boldsymbol{\theta}}_i)$ is consistent for $\mathbf{V}_{\boldsymbol{\theta}}$ as $t, n \to \infty$.

3.4 Variations in Design of Experiments

Toxicological studies often differ with respect to design of the study. Such varying situations need to be accounted for in the proposed methodology for the differential equation models fitted to the data. Consider an experiment where the response variable for each individual is measured at different times. This is a common occurrence in both animal and human studies. For instance, blood measurements from mice may be drawn at different time points for groups of mice. Consequently, for the ith individual in the study, the measurement time points are $\{t_1, \ldots, t_{n_i}\}$. The individual parameter estimation procedure would thus be based on the individual measurement sets only. The intra-individual correlation structures ($\boldsymbol{\Sigma}_i$) would be appropriately modified to retain the same structure only with different orders ($n_i \times n_i$). Hierarchical model assumptions about $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_i$ remain unchanged. The proposed methodology would apply in the same way for this situation due to the formulation of the problem based only on the model parameters.

A different (but not exclusive) situation that might arise in such studies is the measurement of multiple compartments with different observation times for each compartment. For example, a study might be observing chemical concentrations in both exhaled breath and blood. Exhaled breath measurements are usually more easily available than blood observations. These situations can be readily incorporated in the provided formulation. In this case the data for an individual would be $(\mathbf{Y}_{blood}, \mathbf{Y}_{exh})'$, where $\mathbf{Y}_{blood}$ represents the concentrations in blood and $\mathbf{Y}_{exh}$ are the concentrations in exhaled breath at respective measurement times. Further, this situation may warrant the choice of different basis functions for the two compartments. Hence we can visualize and tackle different design situations arising in toxicological studies in context of differential equation models with the proposed methodology.

3.5 Discussion

Modeling systems of differential equation plays an important role especially in context of pharmacokinetic and toxicokinetic modeling. Although such models are widely used in a variety of contexts, as noted in the recent EPA workshop and the resulting publication (Barton et al. 2007) there does not exist a well developed statistical methodology for drawing inference on the model parameters. This research, which exploits the functional data analytic approach of Ramsay et al. (2007), takes the first step towards a formal statistical theory and methodology. Specifically, an important contribution of the proposed methodology is that it accounts for; (i) inter and intra-individual variability, (ii) the dependence within subjects and between compartments/states. Secondly, our methodology overcomes the computational burden of usual strategies by avoiding the problem of numerically solving a system of differential equations. This in turn also alleviates situations where the differential equations explain the behavior of the chemical better in some compartments than others. In

Chapter 5, the proposed methodology has been illustrated using simulated data examples and real data example based on a benzene inhalation experiment.

Chapter 4

ACCOMMODATING COVARIATES IN DIFFERENTIAL EQUATION MODELS

Most toxicological studies investigating the response to chemicals in human or animals also record covariate information on the individuals in the study. These may include variables such as age, dose or weight. A question of interest to researchers would be whether and how these variables affect the chemical phenomena. Often chemical kinetics differ with changes in dosage or age of the subject. Some chemical processes might be expressed more when the dose is low than higher doses. In this chapter we try to provide approaches for covariate inclusion and testing in context to the differential equation models described in the earlier chapters.

4.1 Existing Literature

Models governed by systems of differential equations can be visualized as a form of non-linear regression problem with an unknown functional form of the regression at

the response. Hence we first explore the literature for methods of covariate inclusion and testing in the non-linear regression problem and then focus on the methods for the same in differential equation modeling.

4.1.1 Non-linear Regression Problem

In usual strategies for estimation in differential equation models, the problem of interest is treated as a non-linear parameter estimation problem, where the dependence of response on the parameters is the solution to the system of differential equations. Consider a non-linear regression problem,

$$\mathbf{Y} = \mathbf{f}(\mathbf{U}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{f}$ represents the non-linear regression function. Here $\mathbf{U}$ represents the covariates. Usually in such a problem, the functional form $\mathbf{f}$ is known. The estimation procedures involve minimizing the residual sum of squares,

$$SS = (\mathbf{Y} - \mathbf{f}(\mathbf{U}, \boldsymbol{\theta}))'(\mathbf{Y} - \mathbf{f}(\mathbf{U}, \boldsymbol{\theta})), \tag{4.2}$$

with respect to $\boldsymbol{\theta}$. The estimate of $\boldsymbol{\theta}$ thus depends on the covariates $\mathbf{U}$. However the dependence is an implicit one and thus testing for the covariate effects is difficult. One has to compare models by inclusion and exclusion of covariates, using likelihood ratio statistics. Note that the usual model selection criterion like AIC would not work in such situations due to the absence of a nested model structure.

A more recent methodology, namely varying coefficient models, developed for incorporation of covariate effects was proposed by Hastie and Tibshirani (1993) and later used by Sentürk and Müller (2006). They developed this methodology in the context of linear regression models. More recently, Cui et al. (2009) developed the

varying coefficient model methodology to solve the problem of covariate adjusted non-linear regression.

Originally, Hastie and Tibshirani (1993) proposed the varying coefficient model to incorporate effects of latent covariates in the parameters of the linear regression problem. The framework suggested that the coefficients in a linear regression model be allowed to vary as smooth functions of some covariates. So the modified linear regression model in a simple Gaussian univariate case can be represented as

$$Y = X_1\beta_1(R_1) + \ldots + X_p\beta_p(R_p) + \epsilon, \tag{4.3}$$

where X and R are two kinds of covariates in the model, explicit and implicit. The response is modeled as a linear function of the explicit covariates but the coefficients are functions of the implicit ones. So the parameters ($\beta(.)$) represent an interaction between these two classes of covariates. The estimation procedure in this case used a penalized spline approach to estimate the unknown functions ($\beta(.)$) which served as estimates for effects of the explicit covariates (X) adjusted for the implicit covariates (R). The paper also highlighted some general models that can be described through the same formulation.

Sentürk and Müller (2006) and Cui et al. (2009) approached the problem from the response modification angle. According to them, the response variable and the predictors are modified by the latent covariates through multiplicative factors. Mathematically, the general non-linear model in this varying coefficient framework can be

expressed as

$$
\begin{aligned}
Y &= f(X, \beta) + \epsilon, \\
\tilde{Y} &= \psi(R)Y, \\
\tilde{X}_r &= \phi_r(R)X_r,\ r = 1, \ldots, q,
\end{aligned}
\tag{4.4}
$$

where Y is an unobservable response, $X = (X_1, \ldots, X_q)'$ is an unobservable predictor, β is the unknown model parameter, f is a known continuous function and $\tilde{Y}$ and $\tilde{X}_r$ are the actual observable response and covariates. Further, $\psi(.)$ and $\phi_r(.)$ are the unknown distorting functions of the observed variable R. The methodology involves estimation of the distorting function non-parametrically by regressing the predictors and response on the distorting covariate, under some restricting conditions on the expected values of $\psi(R)$ and $\phi_r(R)$. The predicted response ($\hat{\tilde{Y}}$) and predictors ($\hat{\tilde{X}}_r$) are subsequently used in a non-linear parameter estimation framework to minimize a L_2 norm under (4.4).

All these state of the art methodologies in the area of covariate adjustment model the response in terms of covariates. Also in case of the non-linear problem mentioned in Cui et al. (2009), the functional form of the regression function is given. Further, the distorting functions are multiplicative in nature which might not be a right choice in all situations. More importantly in all the cases, the dependence of the response on the covariates is explicitly known. Also the model parameters here directly measure the effect of the covariates on the response. Although these methods are effective ways to deal with covariate effects in non-linear regression models, alternative methods must be explored to study covariate effects in models defined by system of differential equations due to the peculiarities in the model structure and formulation that distinguishes them from the usual non-linear regression model.

4.1.2 Covariate Effects in PBPK Modeling

The focal point of a physiologically based pharmacokinetic model is a system of differential equations that define the kinetics of a chemical in the physiology over time. Covariates affecting the kinetics often take a secondary position in the analysis. In a non-linear mixed effects framework or Bayesian approach to estimate parameters in a system of differential equations, the key element is the numerical solution to the ODE system. To include covariates in the analysis, the dependence of the system on covariates must be explicitly known. This is often not the case in PBPK models.

Models defined by differential equations and in the special case of PBPK models, knowledge about the parameters determines the system completely. However, in the analysis of covariate effects in PBPK models, the effect of covariates on the parameters of interest are not often analyzed, even though this seems to be the more intuitive way of differentiating between PBPK models for different groups based on their covariate value. Most methods investigate the effect of covariates on the response variable through a mixed effects or Bayesian framework. We look at a few references to illustrate the state of the art in incorporation of covariate effects in PBPK models, using both the non-linear least squares and the Bayesian approaches.

Joerger (2012) in a recent paper on pharmacokinetic modeling reviews the latest approach to include covariates in a PBPK model for analysis using non-linear mixed effects modeling techniques. The paper centers around the pharmacokinetic modeling of anti-cancer drugs. Covariates are extremely important in cancer studies to provide a more accurate modeling while taking into account the variability induced by the variation in the covariate values. Some major covariates in pharmacokinetic studies for anti-cancer drugs include weight, gender, glomerular filtration rate and body surface area. The paper lists explicitly the relationships between the pharmacokinetic parameters and some of the covariates (both categorical and continuous) of interest

for specific anti-cancer drugs. For example, in the pharmacokinetics of busulfan in children, the clearance parameter (CL) is related to weight (WT) in the following way :

$$CL = 4.04L/h/20kg \cdot (WT/20)^{0.74}.$$

In case of a drug pemetrexed, the clearance parameter (CL) is mathematically related to the glomerular filtration rate (GFR) according to the following :

$$CL = 43 + 47.2 \cdot (GFR/92.6).$$

These structural dependencies of the parameters on the covariate are fed into the non-linear mixed effects framework through the solution of the system of the differential equations. As is evident from the above examples, specific covariates are explicitly included in the PBPK model itself. Thus solving the system numerically would provide an implicit functional dependence on the covariates. This approach to covariate modeling is feasible only if the mathematical relationships are well known from literature. For a more general covariate analysis of any system of differential equations, this approach would not be applicable due to lack of knowledge about the dependencies.

The Bayesian methodology for analyzing PBPK models also take covariates into account in a similar way. In a recent publication, Mörk et al. (2009) describe a Bayesian analysis of a washin-washout PBPK model for acetone. The authors use a Bayesian hierarchical model to study the chemical kinetics for acetone in human physiology. The PBPK model for acetone used in this paper describes the kinetics of acetone through multiple tissue compartments and observes the concentration of acetone in arterial blood and exhaled breath. Some of the covariates considered were

body weight, height and endogenous acetone levels. These covariates featured explicitly in the system of differential equations describing the PBPK model. The Bayesian model assumptions used the solution of the differential equations as the location for the distribution of the data points. Hence as in the non-linear mixed effects approach, the covariate effects are only present if one knows the actual mathematical formulation of the covariates in the PBPK model. Problems would arise if the covariates are not present in the original system of differential equations.

In both the approaches described above, testing for particular covariate effects is impossible due to the peculiarities in the formulation of the problem. In contrast to a regression problem ($\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$), there are no parameters that describe the effect of a covariate in a differential equation model. This makes it difficult to test whether the covariates have an effect on the response. Further, if one is interested in testing the effects of covariates on the parameters of interest in the PBPK model, these approaches are not appropriate. Hence tests of hypothesis on the effects of covariates have not been formulated in case of models defined by system of differential equations and is a relevant research problem.

To summarize, testing of covariate effects in non-linear models has been an especially challenging problem often due to lack of explicit structural form of dependence. This deficiency is more stark in the case of models dictated by differential equations. Especially in case of physiologically based pharmacokinetic models where the unknown parameters are usually metabolic parameters or rate constants, the dependence on covariates of interest are more difficult to infer. This calls for a more structured methodology to test for covariate effects in PBPK models, and models governed by differential equations in general. In this chapter, we propose a method for covariate inclusion, estimation of the model parameters while accounting for the covariates and testing of hypothesis about covariate effects in differential equation

models, in light of the functional data analytic approach used in Chapter 3. The methodology has been illustrated using simulated data examples and a real dataset from a benzene inhalation study.

4.2 Problem Description

Models governed by differential equations are inherently different from usual statistical models such as regression models in terms of covariate inclusion. Consider a simple linear regression model given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{4.5}$$

Here covariates $\mathbf{X}$ are included explicitly in the model and the effect of the covariates are expressed through the parameters $\boldsymbol{\beta}$. This model thus makes it convenient to test effects of the covariates on the response $\mathbf{Y}$. However in case of the models we have been considering, neither the parameters nor the covariates are as in (4.5). Consider the model

$$\dot{\mathbf{Z}} = \mathbf{F}(\mathbf{Z}, t, \boldsymbol{\theta}). \tag{4.6}$$

The model parameter $\boldsymbol{\theta}$ in (4.6) are physiological or metabolic parameters and not indicative of the covariate effects. Another important aspect to be noted here is that there are no explicit covariates in the differential equation model (4.6). These are the two main questions that motivate the work done in this chapter.

The main objectives of this section is to explore methodology to incorporate covariates in differential equation models, specifically physiologically based pharmacokinetic (PBPK) models and develop methodology to investigate effect of covariates on

the parameters and/or response variables, based on the functional data analytic estimation methodology proposed in Chapter 3. This would facilitate the formulation for measuring the effects of covariates in a sound statistical framework without having to solve or simulate the system of differential equations.

4.2.1 Incorporation of Covariates

The primary question is the formulation of the problem as to how the covariates feature in the model. The covariates may affect the parameters and/or response. In this chapter, we develop a methodology where the model parameters are dependent on the covariates. This would implicitly mean that the response or state variables are dependent on the covariates only through the model parameters. Suppose there are q covariates to be considered. Suppose $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{im})'$ denote the $m \times 1$ model parameter, $\boldsymbol{\eta}_i$ be a vector of order $m(q+1) \times 1$ and U_i be the $(q+1) \times 1$ covariate vector for the ith individual, taking into account the intercept term also. Let $\{g : \mathbb{R}^{m(q+1)} \to \mathbb{R}^m\}$ be a one-one link function between $\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_i$. The proposed formulation implies

$$\boldsymbol{\theta}_i = g(U_i, \boldsymbol{\eta}_i). \tag{4.7}$$

Hence the entire problem setup now focuses on $\boldsymbol{\eta}_i$. The simplest model that we can adopt is when $g(.)$ is linear, that is $\boldsymbol{\theta}_i = diag(U_i, \ldots, U_i)\boldsymbol{\eta}_i = \mathbf{U}_i\boldsymbol{\eta}_i$. Often in PBPK models, the parameters of interest are metabolic parameters or rate constants which are always strictly positive. Hence the dependence of model parameters on the covariates need to be modeled differently. One such formulation can be $log(\theta_{ik}) = U_i\eta_{ik}$, where θ_{ik} and η_{ik} are the kth component of $\boldsymbol{\theta}_i$ and $\boldsymbol{\eta}_i$ respectively. This approach to the problem provides a simple yet intuitive approach to address questions

regarding the effect of covariates on model parameters.

The model in (4.6) is now expressed as

$$\dot{\mathbf{Z}}_i = \mathbf{F}(\mathbf{Z}_i, t, g(U_i, \boldsymbol{\eta}_i)). \tag{4.8}$$

The complete state variable $\mathbf{Z}$ is comprised of observable ($\mathbf{Z}^o$) and unobservable state variables ($\mathbf{Z}^u$). Data on the observable state variables for the ith individual are denoted by $\mathbf{Y}_i$. Hence, for each individual $\mathbf{Y}_i = \mathbf{Z}_i^o + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i$ denotes the intra-individual error for the ith individual with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}(\rho_i)$. As in the previous chapter we impose a structure on $\boldsymbol{\Sigma}(\rho_i)$ to obtain a parsimonious representation of intra-individual variability.

Using basis function expansion for the state variables $\mathbf{Z}_i$, we obtain a similar individual parameter estimation procedure as described in Section 3.2.1. Let the basis functional expansion of $\mathbf{Z}_i$ be denoted by $\tilde{\mathbf{Z}}_i = \boldsymbol{\Phi}_i\boldsymbol{\alpha}_i$, where $\boldsymbol{\Phi}_i$ is a function of time. Hence both $\tilde{\mathbf{Z}}_i^o$ and $\tilde{\mathbf{Z}}_i^u$ re functions of $\boldsymbol{\alpha}_i$. We have to minimize the distance between the data and the approximation while ensuring that the approximations minimize the error in the differential equations. Therefore, the objective functions for estimation of the transformed parameters are

$$\begin{aligned} S_1(\boldsymbol{\alpha}_i, \rho_i, \lambda_i) =& (\mathbf{Y}_i - \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i))'\boldsymbol{\Sigma}(\rho_i)^{-1}(\mathbf{Y}_i - \tilde{\mathbf{Z}}_i^o(\boldsymbol{\alpha}_i)) \\ &+ \lambda_i \int (\dot{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i - \mathbf{F}(\boldsymbol{\Phi}_i\boldsymbol{\alpha}_i, t, g(U_i, \boldsymbol{\eta}_i)))'(\dot{\boldsymbol{\Phi}}_i\boldsymbol{\alpha}_i - \mathbf{F}(\boldsymbol{\Phi}_i\boldsymbol{\alpha}_i, t, g(U_i, \boldsymbol{\eta}_i)))dt \end{aligned} \tag{4.9}$$

and

$$S_2(\boldsymbol{\eta}_i) = (\mathbf{Y}_i - \widehat{\tilde{\mathbf{Z}}_i^o}(\boldsymbol{\eta}_i))'\boldsymbol{\Sigma}(\hat{\rho}_i(\boldsymbol{\eta}_i))^{-1}(\mathbf{Y}_i - \widehat{\tilde{\mathbf{Z}}_i^o}(\boldsymbol{\eta}_i)). \tag{4.10}$$

In the individual estimation procedure, the intermediate parameter estimates, $\hat{\boldsymbol{\alpha}}_i$, $\hat{\rho}_i$ and $\hat{\lambda}_i$ are all functions of $\boldsymbol{\eta}_i$ and U_i. Hence the predicted value of $\tilde{\mathbf{Z}}_i^o$ is a function of $\boldsymbol{\eta}_i$. We minimize Equations (4.10) to obtain the estimated parameters, denoted as $\hat{\boldsymbol{\eta}}_i$, which are functions of the individual data $\mathbf{Y}_i$ and the individual covariates U_i. Hence using delta method, we have

$$\hat{\boldsymbol{\eta}}_i(\mathbf{Y}_i, U_i) \simeq \hat{\boldsymbol{\eta}}_i(\mathbf{Z}_i^o, U_i) + (\mathbf{Y}_i - \mathbf{Z}_i^o)\frac{d\hat{\boldsymbol{\eta}}_i(\mathbf{Y}_i, U_i)}{d\mathbf{Y}_i}\Big|_{\mathbf{Y}_i=\mathbf{Z}_i^o}.$$
$$Cov(\hat{\boldsymbol{\eta}}_i(\mathbf{Y}_i, U_i)) \simeq \Big[\frac{d\hat{\boldsymbol{\eta}}_i(\mathbf{Y}_i, U_i)}{d\mathbf{Y}_i}\Big|_{\mathbf{Y}_i=\mathbf{Z}_i^o}\Big] Cov(\mathbf{Y}_i) \Big[\frac{d\hat{\boldsymbol{\eta}}_i(\mathbf{Y}_i, U_i)}{d\mathbf{Y}_i}\Big|_{\mathbf{Y}_i=\mathbf{Z}_i^o}\Big]'.$$

Corresponding estimate of the actual model parameters is $\hat{\boldsymbol{\theta}}_i = g(U_i, \hat{\boldsymbol{\eta}}_i)$ and by delta method, it's asymptotic covariance is given by $[g'(U_i, \hat{\boldsymbol{\eta}}_i)]Cov(\hat{\boldsymbol{\eta}}_i, \mathbf{U}_i)[g'(U_i, \hat{\boldsymbol{\eta}}_i)]'$, where $g'(.)$ is the derivative of $g(.)$ with respect to $\boldsymbol{\eta}_i$.

We assume that the conditional distribution of $\hat{\boldsymbol{\eta}}_i|\boldsymbol{\eta}_i$ have mean $\boldsymbol{\eta}_i$ and covariance $\big(\frac{d\hat{\boldsymbol{\eta}}_i}{d\mathbf{Z}_i^o}\big)\hat{\boldsymbol{\Sigma}}_i\big(\frac{d\hat{\boldsymbol{\eta}}_i}{d\mathbf{Z}_i^o}\big)'$ (denoted by $\Gamma(\boldsymbol{\eta}_i, U_i)$), which depends on the covariate U_i for the ith individual and the true individual parameter $\boldsymbol{\eta}_i$. Also to indicate that the individuals are sampled from a common population, we assume that the true individual parameter $\boldsymbol{\eta}_i$ has a distribution with mean $\boldsymbol{\eta}$ and covariance matrix $\mathbf{W}$. Hence the marginal distribution of $\hat{\boldsymbol{\eta}}_i$ is centered at $\boldsymbol{\eta}$ and has a covariance matrix $\mathbf{W} + E_{\boldsymbol{\eta}}[\Gamma(\boldsymbol{\eta}_i, U_i)]$, which is a function of the population level parameter $\boldsymbol{\eta}$ and the individual covariate value U_i. Hence marginally the estimated individual parameter values are independent but not identically distributed. As for the actual model parameters $\boldsymbol{\theta}_i$, they are not identically distributed as in the previous chapter.

We shall use the proposed methodology to perform population estimation on the modified parameters $\boldsymbol{\eta}_i$. Consider the following model for $\hat{\boldsymbol{\eta}}_i$.

$$\hat{\boldsymbol{\eta}}_i = \boldsymbol{\eta} + \boldsymbol{\zeta}_i, \tag{4.11}$$

where $\boldsymbol{\zeta}_i$ are independent with mean $\mathbf{0}$ and covariance $\mathbf{W} + E_{\boldsymbol{\eta}}[\Gamma(\boldsymbol{\eta}_i, U_i)]$. For notational simplicity, we denote $E_{\boldsymbol{\eta}}[\Gamma(\boldsymbol{\eta}_i, U_i)]$ by $\Omega_i(\boldsymbol{\eta})$.

Stacking the n linear models for $\hat{\boldsymbol{\eta}}_i$ we get

$$\begin{bmatrix} \hat{\boldsymbol{\eta}}_1 \\ \vdots \\ \hat{\boldsymbol{\eta}}_n \end{bmatrix} = \mathbf{1}_n \otimes \boldsymbol{\eta} + \boldsymbol{\zeta}, \tag{4.12}$$

where $\boldsymbol{\zeta}$ is the vector containing $\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n$. Further,

$$E[\boldsymbol{\zeta}] = \mathbf{1}_n \otimes \mathbf{0}.$$

$$Cov[\boldsymbol{\zeta}] = \begin{bmatrix} \mathbf{W} + \Omega_1(\boldsymbol{\eta}) & 0 & \cdots & 0 \\ 0 & \mathbf{W} + \Omega_2(\boldsymbol{\eta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{W} + \Omega_n(\boldsymbol{\eta}) \end{bmatrix}$$

$$= \mathbf{I} \otimes \mathbf{W} + \begin{bmatrix} \Omega_1(\boldsymbol{\eta}) & 0 & \cdots & 0 \\ 0 & \Omega_2(\boldsymbol{\eta}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_n(\boldsymbol{\eta}) \end{bmatrix}$$

$$= \mathbf{V}_1 + \mathbf{V}_2.$$

The MINQUE theory for estimation of variance components provides an estimator for $\mathbf{W}$ and $\boldsymbol{\eta}$, which are the population level parameters. We use the iterative estimation

methodology described in Chapter 3 to obtain the population estimators, $\hat{\mathbf{W}}$ and $\hat{\boldsymbol{\eta}}$. Estimator of the covariance of $\hat{\boldsymbol{\eta}}$ is $\hat{\mathbf{W}} + \frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}_i$, denoted by $\widehat{\mathbf{V}}$. The methodology is illustrated in Chapter 5 using simulated data examples and a real data example from the benzene inhalation experiment.

4.2.2 Asymptotic Theory for the Proposed Estimators

First, we consider the asymptotic theory for individual parameter estimation. It can be formulated in a similar vein as in Chapter 3 as a constrained non-linear parameter estimation problem. The consistency of the estimated individual parameter values $(\hat{\boldsymbol{\eta}}_i)$ and the associated covariance matrix is $\Gamma(\boldsymbol{\eta}_i, U_i)$ is shown under the Assumptions B.

Assumptions B:

- $(B.1)$ The components of the model function $\mathbf{F}(.)$ are continuous and twice differentiable for $\boldsymbol{\eta}^0 \in$ B, a closed ball, where $\boldsymbol{\eta}^0$ is the true parameter vector.
- $(B.2)$ The matrix of partial derivatives $\mathbf{D}(\boldsymbol{\eta}) = \dfrac{\mathbf{dF}}{\mathbf{d}\boldsymbol{\eta}}$ is of full rank with probability 1 in a neighborhood of $\boldsymbol{\eta}^0$.
- $(B.3)$ The matrix $B_t = H_t^{-1/2}\Lambda(t)'\Lambda(t)H_t^{-1/2}$ and B_t^{-1} converges to positive definite matrices for large t, where $H_t = diag(h_{iit})$ is a sequence of diagonal matrices such that $h_{iit} \to \infty$ as $t \to \infty$.

Lemma 4.1 *Under Assumptions B,* $H_t^{1/2}(\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i) = O_p(1)$ *as* $t \to \infty$.

Using Lemma 2.1 and delta method, the individual covariance matrix $\Gamma(\hat{\boldsymbol{\eta}}_i, U_i)$ converges to $\Gamma(\boldsymbol{\eta}_i, U_i)$ in probability.

Lemma 4.2 $\hat{\Omega}_i \xrightarrow{pr} \Omega_i$ *as* $t \to \infty$ *for all* i.

Proof To prove this lemma we need the following compactness condition on the individual covariance matrices.

$$E\Big[sup_{|\epsilon|<\delta}\|\Gamma'(\boldsymbol{\eta}_i+\epsilon,U_i)-\Gamma'(\boldsymbol{\eta}_i,U_i)\|\Big] \to 0, \text{ as } \delta \to 0,$$

where $\Gamma'(.)$ represents the derivative of $\Gamma(.)$ with respect to $\boldsymbol{\eta}_i$. The above condition requires that $\left\|\frac{\partial^2\Gamma(\boldsymbol{\eta}_i,U_i)}{\partial\boldsymbol{\eta}_i^2}\right\|$ is bounded. Under these conditions,

$$\begin{aligned}
\hat{\Omega}_i &= E(\Gamma(\hat{\boldsymbol{\eta}}_i,U_i)) \\
&= E[\Gamma(\hat{\boldsymbol{\eta}}_i-\boldsymbol{\eta}_i+\boldsymbol{\eta}_i,U_i)] \\
&= E[\Gamma(\boldsymbol{\eta}_i,U_i)] + E[E((\hat{\boldsymbol{\eta}}_i-\boldsymbol{\eta}_i)\Gamma'(\boldsymbol{\eta}_i,U_i))|\boldsymbol{\eta}_i] \\
&\to E[\Gamma(\boldsymbol{\eta}_i,U_i)] \text{ as } t\to\infty \text{ (Using Lemma 2.1)} \\
&= \Omega_i
\end{aligned}$$

We now explore the large sample theory for the estimator of the population parameter $\boldsymbol{\eta}$ under the model specified by (4.12). We rewrite the model as

$$\mathbf{y}^* = \mathbf{M}^*\boldsymbol{\eta} + \boldsymbol{\zeta}^*, \tag{4.13}$$

where

$$\mathbf{y}^* = (\mathbf{V}_1+\mathbf{V}_2)^{-1/2}(\hat{\boldsymbol{\eta}}_1,\ldots,\hat{\boldsymbol{\eta}}_n)',$$

$$\mathbf{M}^* = (\mathbf{V}_1+\mathbf{V}_2)^{-1/2}[\mathbf{I}:\cdots:\mathbf{I}]' = \sum_{i=1}^{n}(\mathbf{W}+\Omega_i)^{-1/2}$$

and

$$\boldsymbol{\zeta}^* = (\mathbf{V}_1+\mathbf{V}_2)^{-1/2}\boldsymbol{\zeta}.$$

We assume here that $\boldsymbol{\eta}$ is independent of $\boldsymbol{\zeta}^*$ for developing the theory. Under Model 4.13, the generalized least square estimate of $\boldsymbol{\eta}$ is given by

$$\begin{aligned}\tilde{\boldsymbol{\eta}} &= (\mathbf{M}^{*\prime}\mathbf{M}^{*})^{-\mathbf{1}}\mathbf{M}^{*\prime}\mathbf{y}^{*} \\ &= \boldsymbol{\eta} + (\mathbf{M}^{*\prime}\mathbf{M}^{*})^{-\mathbf{1}}\mathbf{M}^{*\prime}\boldsymbol{\zeta}^{*} \\ &= \boldsymbol{\eta} + (\sum_{i=1}^{n}(\mathbf{W}+\Omega_i)^{-1})^{-1}\sum_{i=1}^{n}(\mathbf{W}+\Omega_i)^{-1}\boldsymbol{\zeta}_i\end{aligned}$$

Let $\mathbf{m}_i^*$ be the ith column of $\mathbf{M}^*$. If the variance components are known in this framework, under the following conditions,

$$\max_{1\le i\le n}\left[\mathbf{m}_i^{*\prime}(\sum_{i=1}^{n}(\mathbf{W}+\Omega_i)^{-1})\mathbf{m}_i^{*}\right] \to 0 \text{ as } n\to\infty, \text{ and}$$
$$\lim_{n\to\infty} n^{-1}(\sum_{i=1}^{n}(\mathbf{W}+\Omega_i)^{-1}) = \mathbf{G}(\text{finite and positive definite}),$$

the estimate of $\boldsymbol{\eta}$, denoted by $\tilde{\boldsymbol{\eta}}$ is asymptotically normal and

$$\sqrt{nt}(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta}) \xrightarrow{D} \mathcal{N}(\mathbf{0},\mathbf{G}^{-1})$$

However in the covariate setup explained earlier, the variance components $\mathbf{V}_1$ and $\mathbf{V}_2$ are unknown functions of $\boldsymbol{\eta}$ and the covariates U_i, and is estimated by $\widehat{\mathbf{V}}$. Consider the estimator $\hat{\boldsymbol{\eta}}$ with the plug-in estimator of the covariance.

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta} + \Big(\sum_{i=1}^{n}(\widehat{\mathbf{W}}+\hat{\Omega}_i)^{-1}\Big)^{-1}\sum_{i=1}^{n}(\widehat{\mathbf{W}}+\hat{\Omega}_i)^{-1}\boldsymbol{\zeta}_i$$

Theorem 4.1 *: Under the model in (4.7), let $\hat{\boldsymbol{\eta}}$ be the proposed estimator. Under the above Assumption B, $\sqrt{nt}(\hat{\boldsymbol{\eta}}-\tilde{\boldsymbol{\eta}}) \xrightarrow{pr} 0$ as $nt\to\infty$.*

Proof :

$$\hat{\boldsymbol{\eta}} - \tilde{\boldsymbol{\eta}} = \Big[\sum_{i=1}^{n} (\widehat{\mathbf{W}} + \widehat{\Omega}_i)^{-1})^{-1} - \sum_{i=1}^{n} (\mathbf{W} + \Omega_i)^{-1})^{-1} \Big] [\mathbf{I}' : \cdots : \mathbf{I}'] \widehat{\mathbf{V}}^{-1} \mathbf{V}^{1/2} \boldsymbol{\zeta}^* + \Big(\sum_{i=1}^{n} (\mathbf{W} + \Omega_i)^{-1} \Big)^{-1} [\mathbf{I}' : \cdots : \mathbf{I}'] [\widehat{\mathbf{V}}^{-1} - \mathbf{V}^{-1}] \mathbf{V}^{1/2} \boldsymbol{\zeta}^*.$$

We need two conditions for the result to hold :

(i) Largest eigenvalue of $\mathbf{V}\widehat{\mathbf{V}}^{-1}$ are $O_p(1)$.

(ii) Largest eigenvalue of $\Big(\sum_{i=1}^{n} (\mathbf{W} + \Omega_i)^{-1} \Big) \Big(\sum_{i=1}^{n} (\widehat{\mathbf{W}} + \hat{\Omega}_i)^{-1} \Big)^{-1}$ are $o_p(1)$.

To evaluate these conditions we need to prove the following:

$$\max_{1 \le i \le n} \|\Omega_i^{-1} \hat{\Omega}_i - \mathbf{I}\| \to 0 \text{ as } t, n \to \infty.$$

Consider any two non-null vectors $\mathbf{x}$ and $\mathbf{y}$.Then,

$$\begin{aligned}
&\mathbf{x}' \Big[\Omega_i^{-1} \hat{\Omega}_i - \mathbf{I} \Big] \mathbf{y} \\
\simeq &\mathbf{x}' \Big[\Omega_i^{-1} \{ \Omega_i + E[(\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i) \Gamma'(\boldsymbol{\eta}_i)] \} - \mathbf{I} \Big] \mathbf{y} \\
= &\mathbf{x}' E \Big[E \{ (\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i) \Gamma'(\boldsymbol{\eta}_i) | \boldsymbol{\eta}_i \} \Big] \mathbf{y} \\
= &\mathbf{x}' E \Big[E \{ (\hat{\boldsymbol{\eta}}_i - \boldsymbol{\eta}_i) | \boldsymbol{\eta}_i \} \Gamma'(\boldsymbol{\eta}_i) \Big] \mathbf{y} \\
\to &0. \text{ [Using Lemma 2.1]}
\end{aligned}$$

This implies $(\mathbf{W} + \Omega_i)^{-1} (\widehat{\mathbf{W}} + \hat{\Omega}_i) - \mathbf{I}_p \xrightarrow{pr} \mathbf{0}$. Hence both conditions (i) and (ii) hold and the theorem holds.

4.2.3 Testing Covariate Effects

Consider a study on multiple individuals that measures concentration of a chemical in blood. We are interested in fitting a physiologically based pharmacokinetic (PBPK) model to these data. The individuals in the study may have different values of the covariates such as age, weight, gender or dose. Our objective is to test whether the model parameter (and consequently the PBPK model) varies with the values of the covariates. For example, if $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_f$ represent the model parameters for male and female subjects, a question of interest would be to test whether they are equal. In case of a continuous covariate such as dose, one might be interested to test whether the parameters change with dose.

Suppose we are interested in testing whether the parameters are dependent on the $1 \times q$ covariate vector U_i. Let $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{im})$ be a $m \times 1$ model parameter vector for the ith individual. The hypothesized model with a linear link function is

$$\theta_{ik} = \eta_{ik}^0 + U_i \boldsymbol{\eta}_{ik}^*, \quad k = 1, \ldots, m \text{ and } i = 1, \ldots, n.$$

Hence,

$$\begin{aligned} \boldsymbol{\theta}_i &= \boldsymbol{\eta}_i^0 + diag(U_1, \ldots, U_i)\boldsymbol{\eta}_i^* \\ &= \boldsymbol{\eta}_i^0 + \mathbf{U}_i \boldsymbol{\eta}_i^*, \end{aligned}$$

where $\boldsymbol{\eta}_i^*$ is of order $mq \times 1$.

We assume that the true parameters $\boldsymbol{\eta}_i$ are independent and identically distributed with mean $\boldsymbol{\eta} = (\eta^0, \boldsymbol{\eta}^*)'$ and covariance matrix $\mathbf{W}$. Accordingly, a general test of hypothesis that no covariate effect is present can be written as $H_0 : L\boldsymbol{\eta} = 0$ against $H_a : L\boldsymbol{\eta} \neq 0$, where L is an appropriately chosen matrix. Consider the marginal distribution of $\hat{\boldsymbol{\eta}}_i$, which has mean $\boldsymbol{\eta}$ and covariance $\mathbf{W} + \Omega_i$. The test statistic based

on the $\hat{\boldsymbol{\eta}}_i$s would test whether these estimated parameter values are coming from a distribution with a mean that has $\boldsymbol{\eta}^*$ as 0.

The test statistic for such a test is given by

$$F = \frac{(L\hat{\boldsymbol{\eta}} - L\boldsymbol{\eta})'(\widehat{Cov}(\hat{\boldsymbol{\eta}}))^{-1}(L\hat{\boldsymbol{\eta}} - L\boldsymbol{\eta})}{mq},$$

where $\widehat{Cov}(\hat{\boldsymbol{\eta}}) = \hat{\mathbf{W}} + \frac{1}{n}\sum_{i=1}^{n}\hat{\Omega}_i$. Under H_0,

$$F = \frac{\hat{\boldsymbol{\eta}}' L'(\widehat{Cov}(\hat{\boldsymbol{\eta}}))^{-1} L\hat{\boldsymbol{\eta}}}{mq}$$
$$\overset{H_0}{\sim} F_{mq,n-mq}.$$

We shall reject H_0 at $100\alpha\%$ level of significance if the observed value of the test statistic is greater than $F_{mq,n-mq}(\alpha)$.

4.3 Discussion

We have presented a methodology to evaluate effects of covariates on models governed by differential equations. We extend the functional data analytic estimation methodology presented in Chapter 3 to a covariate testing framework. Both estimation and testing of hypotheses for the parameters are developed in the presence of covariates. We overcome the usual difficulties in testing for covariate effects in such complicated models where the explicit form of dependence of response on the covariates is not available. Since the solution of the system of differential equations is not required in the proposed methodology, the covariates do not need to be included in the model. A linear dependence of the model parameters on the covariates is used in this work. Due to this approach, tests of hypotheses about covariate effects can be performed more directly than in the approaches for non-linear regression. However, a non-parametric

methodology for testing covariate effects may be explored if no specific form of the functional dependence is assumed.

Chapter 5

DATA EXAMPLES : SIMULATED AND REAL DATA

In this section we present all the data examples, both simulated and real, to illustrate the methodologies described in the previous chapters. The estimation methodology described in Chapter 3 is illustrated using simulated data from the benzene PBPK model and the benzene inhalation data. The methodology presented in Chapter 4 is illustrated using simulated data from a two compartment pharmacokinetic model and is also tested on the benzene inhalation data example.

5.1 Simulated Example : Based on Benzene PBPK Model

In this section, we present simulated data based on the PBPK model of benzene described earlier. The simulated example is designed according to the real data example used in this paper. We simulated a random sample of four subjects such that each subject is exposed to $161\mu g/m^3$ benzene through inhalation for two hours, following which the subjects were provided with clean breathing air. Concentration of benzene in exhaled breath at 5, 15, 30, 60, 90, 120 and 150 minutes post-exposure is the response variable for each individual. The PBPK model is described in Section 2.1 and Equations (2.2)-(2.8) represent the differential equation model of interest.

We obtain population parameter estimates of the metabolic parameters along with estimates of their variability.

We treat the maximum metabolic rates in liver and bone marrow as unknown parameters. So $\boldsymbol{\theta} = (V_{max(liv)}, V_{max(bm)}, k_{m(liv)}, k_{m(bm)})'$ and it's true population value is taken as $\boldsymbol{\theta}^0 = (387, 80, 1.2, 17)'$. We assume that the individual parameter values $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} \mathbf{N_4}\left(\boldsymbol{\theta}^0, diag(50, 3, .01, .1) + J\right)$, where J is a matrix of 1's.

Concentration in venous blood ($\mathbf{X}_i$) is obtained by solving the benzene PBPK model equations given by Equations (2.2)-(2.8) using $\boldsymbol{\theta}_i$ as the parameter. Finally the data $\mathbf{Y}_i$ is generated from $N_7(\mathbf{X}_i, \mathbf{R})$. Here $\mathbf{R}$ is an intra-individual covariance structure. We choose the (t, t')th element of $\mathbf{R}$ as $r_{t,t'} = 5 * (0.2)^{|t-t'|}$. We look at 200 datasets consisting of four individuals each.

Linear combinations of nine B-splines of order four are used to approximate the concentration of benzene in each compartment and the metabolites concentration. Simulated annealing is applied to obtain the basis and model parameter estimates for each individual. We use the iterative methodology described in Section 3.2.2 to obtain the population parameter estimates and the corresponding variability estimates.

5.1.1 Results of Simulation Study

Table 5.1: Population parameter estimation results

	$V_{max(liv)}$	$V_{max(bm)}$	$k_{m(liv)}$	$k_{m(bm)}$
Units	$\mu g/min$	$\mu g/min$	$\mu g/l$	$\mu g/l$
True value	387.5	80	1.2	17
Estimate	386.4	80.05	1.33	12.99
Rel. MSE	1.13	0.67	0.87	0.76
Rel. Bias	-0.58	0.05	0.11	-0.27

Results of the simulation study, summarized in Table 5.1, suggest that the proposed methodology has a small relative bias and relative MSE (relative to true value

of the parameter) for a sample size as small as 5.

The estimated coverage probability for a 90% joint confidence region of $\boldsymbol{\theta}$ centered at $\hat{\boldsymbol{\theta}}$ is 0.915 (s.e.= 0.0197). This suggests that even with a sample size as small as five the proposed methodology yields reasonably accurate confidence regions for the population parameter.

5.2 Real Data : Benzene Inhalation Experiment

Benzene is an ubiquitous chemical reported to be carcinogenic to humans and animals. It is an important study chemical due to its extensive industrial usage and production leading to widespread occupational exposure. Certain sources of non-occupational exposure have also been identified, such as automobile exhaust and cigarette smoke. Epidemiological evidence suggests an increased incidence of leukemia due to benzene and its metabolites. It is of interest to us to investigate the mode of action of benzene in human physiology. Several pharmacokinetic models have been proposed to model the flow of benzene (Travis et al. (1990), Woodruff and Bois (1993)). These are compartmental models with main tissues and metabolizing sites serving as the compartments and blood acting as the mode of delivery within these tissues. The five tissue groups are (1) Richly perfused tissues, (2) Slowly perfused tissues, (3) Fat, (4) Liver and (5) Bone marrow. In case of a benzene pharmacokinetic model, it is common practice to include bone marrow as a separate tissue compartment since it is a potential metabolizing site for a carcinogen like benzene. A schematic representation of one such model is shown in Figure 1.1.

The differential equations describing the benzene PBPK model are given by (2.2)-(2.8). We are using data from a benzene inhalation experiment on four individuals where each individual was exposed to certain concentration of benzene through inhaled air for two hours. At the end of two hours, benzene exposure was stopped and

subjects were provided clean breathing air. Benzene concentrations (in $\mu g/m^3$) were measured in exhaled breath at 5, 15, 30, 60, 90, 120 and 150 minutes post exposure. We use the benzene PBPK model mentioned earlier. The data are shown in Figure 5.1.

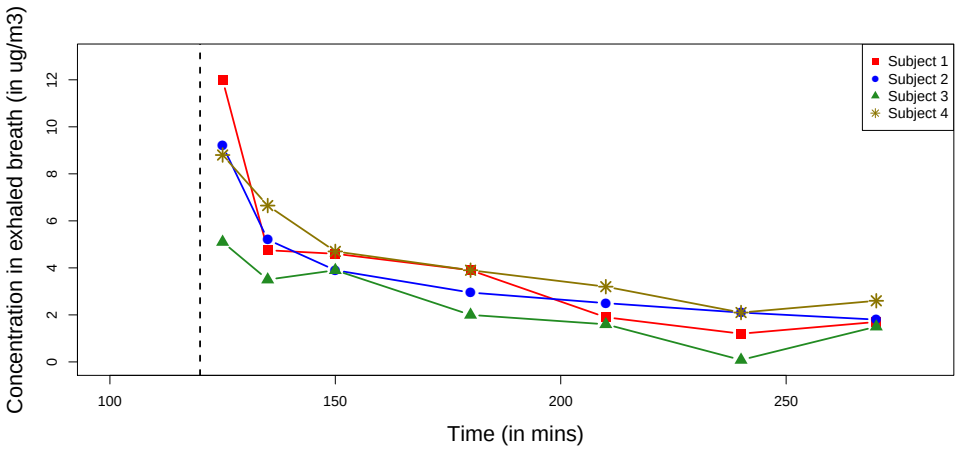

Figure 5.1: Exhaled breath data for benzene inhalation experiment. Concentration of benzene in exhaled breath (in $\mu g/m^3$) was measured post-exposure. The black dotted line represents the exposure stoppage time of 120 minutes.

Our objective of interest is to estimate and infer about the parameters describing the physiologically based pharmacokinetic model using the exhaled breath concentration data.

5.2.1 Method and Results

For individual parameter estimation, nine B-splines of order four are used for approximating each of the six compartments in the model. The observed time points are used as the knots for fitting splines. The regularization parameter, λ, is taken to be unknown and estimated for each individual within the individual estimation methodology. Using the estimated $\hat{\boldsymbol{\theta}}_i$, we obtain the following fits from the solution of the system of differential equations. In Table 5.2, we present the estimated parameter

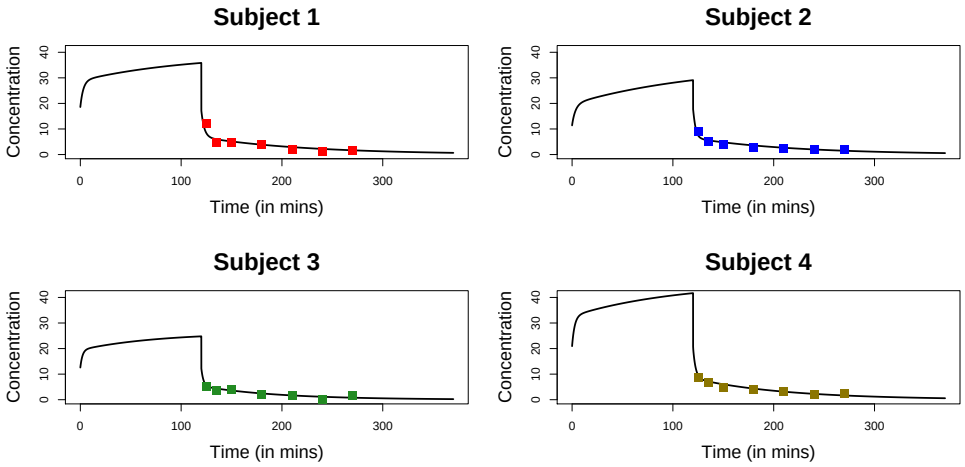

Figure 5.2: Individual parameter fits showing the estimated exhaled breath concentration (in $\mu g/m^3$) of benzene (solid black lines) for the four individuals obtained by solving the differential equations with the estimated individual parameter values.

values for each individual in the study.

Table 5.2: Estimated values of the metabolic parameters for each individual in the study.

	$V_{max(liv)}$	$V_{max(bm)}$	$k_{m(liv)}$	$k_{m(bm)}$
Individual 1	503.45	99.19	1.28	3.59
Individual 2	146.82	20.3	0.88	62.67
Individual 3	399.95	99.13	0.26	0.49
Individual 4	499.45	98.37	2.45	1.45

The estimated values of the intra-individual correlation coefficients ρ_i for the four subjects are 0.146, 0.2434, 0.623 and 0.025. Estimated values of λ for the four individuals are 2.38×10^{-5}, 1.98×10^{-4}, 1.06×10^{-4} and 1.37×10^{-6}. The population parameter estimates obtained are as follows:

$\hat{V}_{max(liv)} = 387.41\ \mu g/min$, $\hat{V}_{max(bm)} = 79.25\ \mu g/min$, $\hat{k}_{m(liv)} = 1.22\ \mu g/l$, $\hat{k}_{m(bm)} = 17.04\ \mu g/l$.

$$\widehat{\mathbf{V}_{\boldsymbol{\theta}}} = \begin{pmatrix} 20556.87 & 4957.96 & -307.30 & 471.16 \\ 4957.96 & 1292.01 & -78.07 & 77.02 \\ -307.30 & -78.07 & 4.77 & -7.08 \\ 471.16 & 77.02 & -7.08 & 190.22 \end{pmatrix}.$$

Consider an individual of weight 130 *lbs* given an exposure concentration of 161 $\mu g/m^3$ for two hours. The post exposure population curve for such an individual with the prediction intervals at observed time points are shown in Figure 5.3.

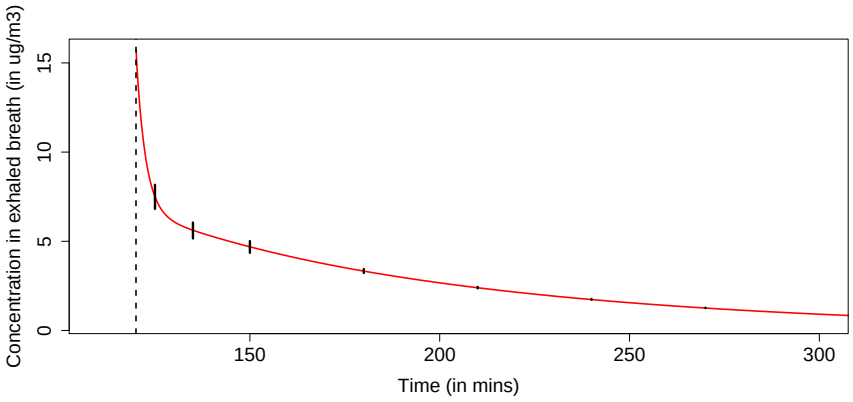

Figure 5.3: Population fitted exhaled breath concentration of benzene with 95% prediction intervals. The solid curve is obtained from the solution of the system of differential equations in (2.2)-(2.8) using the value of the population parameter estimate. The vertical lines represent the prediction intervals.

We analyze the predicted behavior of benzene as explained by the predicted system of differential equations, obtained through the described methodology, in Figure 5.4. We consider the post exposure concentration of benzene in the different compartments of the PBPK model and exhaled breath across time with the solution of the PBPK model using the estimated population parameters.

We investigate a few important features of the predicted model. For all the compartments and exhaled breath, the concentration reaches a peak at 120 minutes and

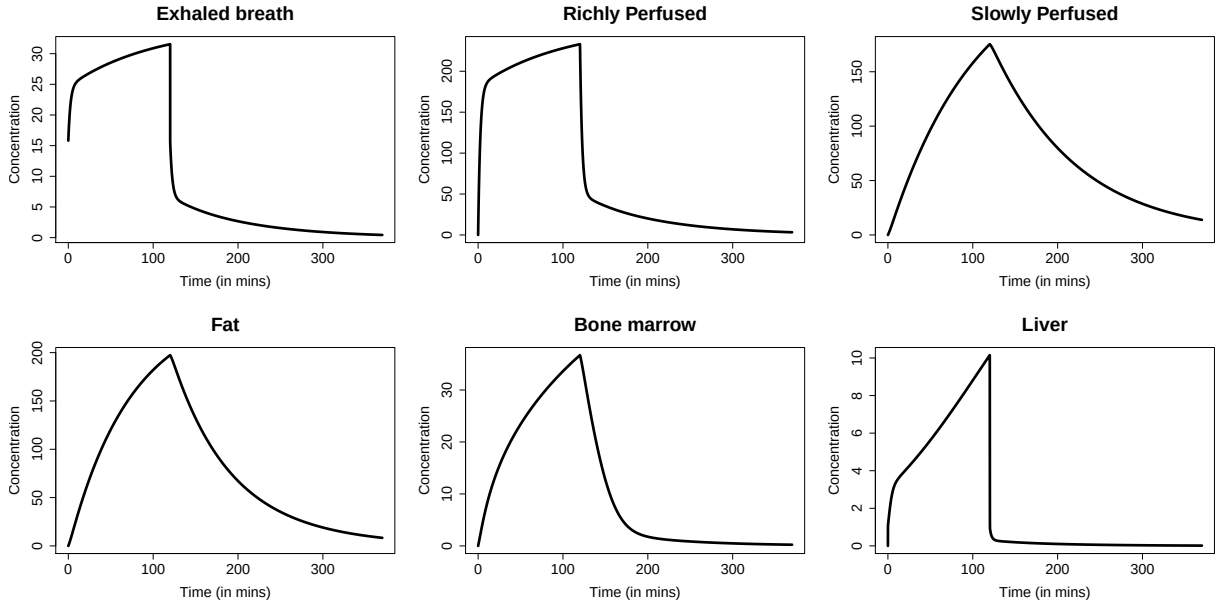

Figure 5.4: Predicted compartmental concentrations (in $\mu g/m^3$) over time. These plots are obtained by solving the differential equation model given by (2.2)-(2.8) with the estimated population model parameter estimates.

decreases post exposure. These plots give us an idea as to how and where the benzene is being processed. Both fat and slowly perfused compartments show a slow decrease in concentration than others, indicating an affinity of benzene towards these class of tissues. As for the metabolizing sites, liver and bone marrow, the concentrations decrease rapidly. This could indicate formation of metabolites of benzene in these two sites. Also the rate of metabolization appears to be faster in liver than in bone marrow. The ratio of the estimates of V_{max} to those of k_m for the two metabolizing sites is 317.5 for liver and 4.66 for bone marrow. This might indicate different enzymatic processes and activity in the two sites. The information here provides an insight into the kinetic behavior of benzene which was one of our main objectives. Further data on enzymatic reactions or metabolite concentrations could enhance the quality of the inference in the given setup.

5.3 Simulated Data : Covariate Effects in a Compartmental Model

We present here a simulated data example based on a compartmental pharmacokinetic model to study the effect of covariates and illustrate the estimation methodology presented in Chapter 4.

5.3.1 Pharmacokinetic Model Description

The model used for illustration is a two compartment pharmacokinetic model represented by Figure 5.5. The chemical is being absorbed through Compartment 1. Only

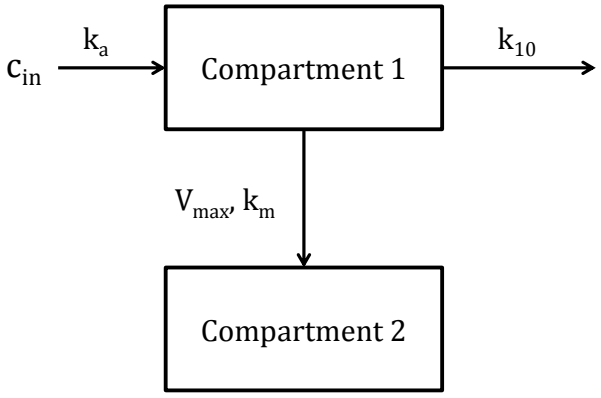

Figure 5.5: A two compartment pharmacokinetic model with linear and non-linear kinetics.

a fraction of the given chemical concentration gets absorbed in Compartment 1 at a rate of k_a. The chemical gets removed from Compartment 1 at a rate k_{10}. Further, there is a formation of metabolites according to a non-linear Michaelis-Menten kinetics in Compartment 2.

The system of differential equations describing the compartmental model shown in Figure 5.5 is given by (5.1). The rates of change in concentration of chemical in the two compartments (c_1, c_2) and the excreted concentration (c_3) are given by

$$\left.\begin{aligned} \text{Compartment 1} &: \frac{dc_1}{dt} = k_a c_{in} - k_{10} c_1 - \frac{V_{max} c_1}{k_m + c_1}, \\ \text{Compartment 2} &: \frac{dc_2}{dt} = \frac{V_{max} c_1}{k_m + c_1}, \\ \text{Excreted chemical} &: \frac{dc_3}{dt} = k_{10} c_1, \end{aligned}\right\} \quad \text{ODE Model} \qquad (5.1)$$

where c_{in} represents the exposure concentration which is non-zero up to four hours of exposure and zero after that. The parameters k_{10} and k_a are assumed to be known and fixed at 0.2 and 0.9 respectively. The metabolic parameters V_{max} and k_m comprise the unknown parameters ($\boldsymbol{\theta}$).

5.3.2 Study Design for Simulations

For simulating a dataset based on covariates, we design a study for $n(= 5)$ individuals. Each of these n individuals are subjected to one of the four exposure concentrations (3, 5, 7, 10 mg/L) of interest. A continuous exposure is provided for four hours and after that the exposure is stopped. The concentration in Compartment 1 is observed both during and after exposure for each subject at 1, 1.5, 2, 3, 4, 5 and 7 hours from the beginning of the study. The concentration in Compartment 1 is the only observable quantity. To build a covariate effect of exposure concentration, we assume that the unknown parameters are functions of the exposure concentration ($excon$) for

each individual,

$$\left.\begin{aligned} log(V_{max(i)}) &= \eta_{0v} + \eta_{1v} * excon_i \\ log(k_{m(i)}) &= \eta_{0k} + \eta_{1k} * excon_i. \end{aligned}\right\} \quad \text{Covariate effects.} \tag{5.2}$$

The true population parameter values for $\boldsymbol{\eta} = (\eta_{0v}, \eta_{1v}, \eta_{0k}, \eta_{1k})'$ is $(0.50, 0.20, 1.00, 0.20)'$. Figure 5.6 shows the behavior of the observable state variable under the above covariate model for different exposure concentrations. The new individual parameters $(\boldsymbol{\eta}_i)$

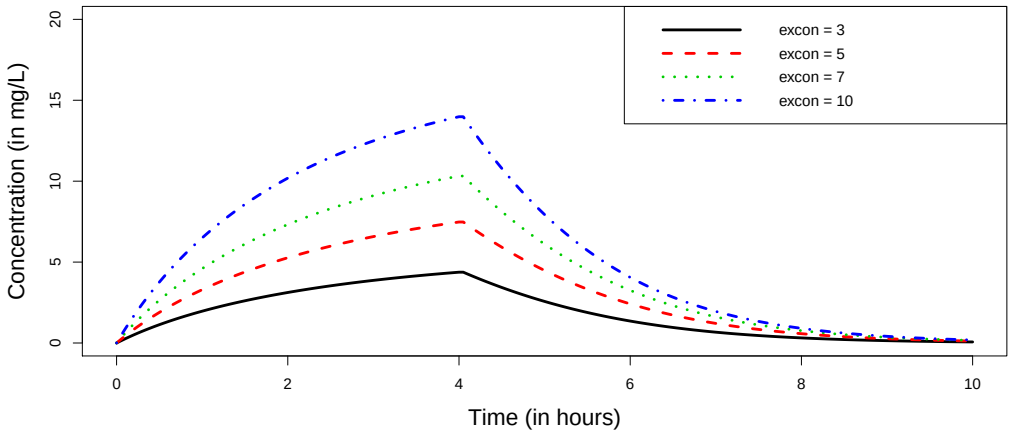

Figure 5.6: The simulated behavior of the observable state variable (Compartment 1) for different exposure concentrations under a log-linear covariate model for the two compartment pharmacokinetic model. Exposure concentrations (excon) are in units of mg/L.

are generated from a four variate normal distribution with mean $(0.50, 0.20, 1, 0.20)'$ and covariance matrix $\mathbf{W}$ where

$$\mathbf{W} = \begin{pmatrix} 0.01 & 0.001 & 0.005 & 0 \\ 0.001 & 0.01 & 0 & 0.005 \\ 0.005 & 0 & 0.05 & 0.001 \\ 0 & 0.005 & 0.001 & 0.01 \end{pmatrix}.$$

The generated parameter values ($\boldsymbol{\eta}_i$) are used in (5.2) to solve the system of ODEs (5.1) to obtain a location parameter for the response distribution. To introduce intra-individual variability, a random multivariate normal error term with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$ of order seven is added to the mean response for each individual. The $(k, k')th$ element of $\boldsymbol{\Sigma}$ is

$$\Sigma_{kk'} = \begin{cases} \sqrt{s_k} & k = k', \\ \sqrt{(s_k s_{k'})} 0.1^{|t_k - t_{k'}|} & k \neq k', \end{cases}$$

where $k, k' = 1, \ldots, 7$ and $\mathbf{s} = (s_1, \ldots, s_7)' = (.05, .3, .3, .3, .4, .001, .00001)'$. The simulations were repeated for 100 datasets, with five individuals in each, based on the same study design.

5.3.3 Results of the Simulated Example

Individual and population parameter estimation is carried out according to the methodology described in Chapter 4. Linear combinations of nine B-spline functions of order four are used to approximate the state variables. The results of the population estimation are summarized in Table 5.3. The estimated coverage probability for a 90%

Table 5.3: Population parameter estimation results in simulation of covariate effects.

	η_{0v}	η_{1v}	η_{0k}	η_{1k}
True value	0.50	0.20	1.00	0.20
Estimate	0.507	0.213	0.984	0.21
Rel. MSE (%)	1.5	1.72	3.16	1.40
Rel. Bias (%)	1.46	6.43	-1.56	4.82

joint confidence region of $\boldsymbol{\eta}$ centered at $\hat{\boldsymbol{\eta}}$ is 0.925 (s.e.= 0.0263). This suggests that even with a sample size as small as five the proposed methodology yields reasonably accurate confidence regions for the population parameter.

5.4 Benzene Inhalation Experiment : Covariate Analysis

We consider the experiment conducted by Yu (1995) on Benzene inhalation to study the effect of covariates on kinetics of benzene. We consider the PBPK model given by (2.2)-(2.8) as the system for modeling the kinetic process. Recall that the PBPK model does not specifically mention any covariates that may affect the parameters involved. For the purpose of illustration, we take exposure concentration as a continuous covariate in this experiment.

In this experiment, four individuals were given four different exposure concentrations of benzene through inhaled air for two hours, following which the exposure was stopped and concentration of benzene was measured in exhaled breath at specific time points. Benzene concentrations (in $\mu g/m^3$) were measured in exhaled breath at 5, 15, 30, 60, 90, 120 and 150 minutes post exposure. We are interested in inferring about the parameters in the PBPK model while adjusting for covariates. We consider the metabolic parameters $V_{max(liv)}$ and $V_{max(bm)}$ as the unknown parameters in the PBPK model. All other parameters are assumed to be known.

We assume the following model to incorporate covariates in the PBPK model.

$$\left.\begin{aligned} V_{max(liv)} &= \eta_{01} + \eta_{11} * excon, \\ V_{max(bm)} &= \eta_{02} + \eta_{12} * excon, \end{aligned}\right\} \tag{5.3}$$

where $excon$ represents exposure concentration. Hence the new parameter to be estimated is $\boldsymbol{\eta} = (\eta_{01}, \eta_{11}, \eta_{02}, \eta_{12})'$ for each of the four individuals. The individual estimation results are shown in Table 5.4.

We use the individual parameter estimates to perform a population estimation for $\boldsymbol{\eta}$. Using the iterative algorithm mentioned in Chapter 4 and Chapter 5, we obtain the population parameter estimates as $\hat{\eta}_{0v} = 387.57$, $\hat{\eta}_{1v} = 0.0093$, $\hat{\eta}_{0k} = 80.02$ and

Table 5.4: Individual parameter estimation results for benzene data with dose as a continuous covariate.

	η_{0v}	η_{1v}	η_{0k}	η_{1k}
Individual 1	499.85	0.0059	99.85	0.0049
Individual 2	150.45	0.0048	20.22	0.0070
Individual 3	399.99	0.0048	100.00	0.0051
Individual 4	499.98	0.0216	100.02	0.0043

$\hat{\eta}_{1k} = 0.0053$.

The estimated population variability is given by the matrix

$$\begin{pmatrix} 20405.80 & 0.504 & 4725.55 & -0.14 \\ 0.504 & 7.565275 \times 10^{-5} & 0.090 & -1.35 \times 10^{-5} \\ 4725.55 & 0.090 & 1192.03 & -0.033 \\ -0.14 & -1.35 \times 10^{-5} & -0.033 & 2.87 \times 10^{-5} \end{pmatrix}.$$

The population predicted curve for a typical subject with weight of 130 *lbs* given an exposure concentration of 161.5 $\mu g/m^3$ under the same study scheme is shown in Figure 5.7 with 95% prediction intervals at the observed time points.

In order to test whether the covariate effects estimated in this real example are significant, we use the Wald type test developed in Chapter 4. The test of hypothesis can be written as

$$H_0 : \begin{pmatrix} \eta_{1v} \\ \eta_{1k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

We can rewrite the null hypothesis as $H_0 : L\boldsymbol{\eta} = 0$, where $L = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. The test statistic is $F = \frac{(L\hat{\boldsymbol{\eta}} - L\boldsymbol{\eta})'(\widehat{Cov}(\hat{\boldsymbol{\eta}}))^{-1}(L\hat{\boldsymbol{\eta}} - L\boldsymbol{\eta})}{2}$. The observed value of F under H_0 is 1.495. Comparing with the null distribution which is $F_{2,4-2}$, observed F is

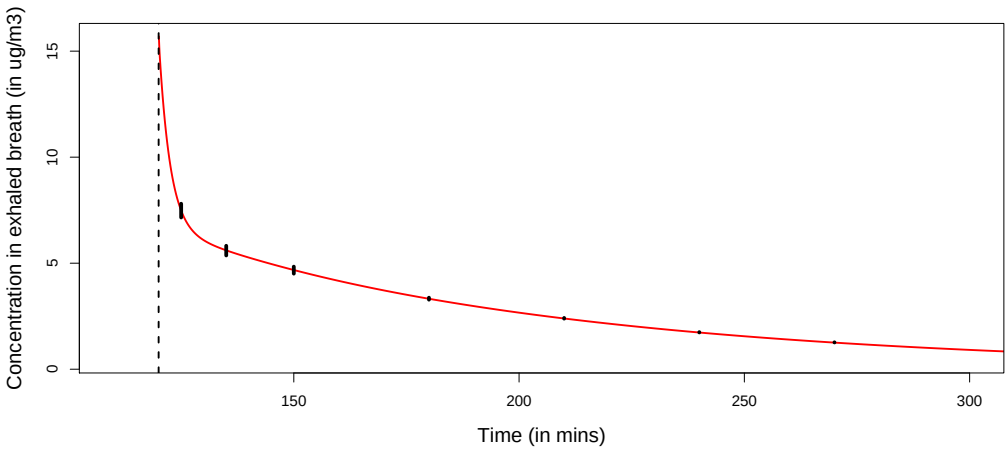

Figure 5.7: Population predicted exhaled breath concentration of benzene along with 95% prediction intervals for a typical person given an exposure concentration of $161.5\mu g/m^3$ of benzene. The vertical black solid lines represent the 95% prediction intervals.

statistically insignificant at 95% level of significance. Hence in light of the given sample and linear covariate model, we may conclude that exposure concentration does not affect the metabolic parameters significantly. However, in cases where more data are available, we may consider more detailed models with more parameters to map the relation between the model parameters and covariates.

5.5 Discussion

In this chapter, numerical results from the simulated examples and the real data applications have been presented. First, we summarize the findings from the numerical examples to illustrate the methodology developed in Chapter 3. From the results of both the simulated data example based on the benzene PBPK model and the real benzene inhalation data, we observe that the individual parameter estimation provides close basis approximations to the observed data points. We obtained estimates of both intra and inter-individual variability along with the individual and population

parameter estimates. We also construct valid prediction intervals for the population curve at the observed time points thus providing a complete statistical framework for this problem. The methodology does away with the need to solve the system of differential equation in order to obtain parameter estimates. This reduces computation time considerably and also takes into account the fact that the differential equations may not capture underlying biological phenomenon accurately. Using the population parameter estimates, we have shown the prediction for the chemical kinetics of benzene in observable as well as unobservable compartments using the population parameter estimates. The methodology has been shown to work well for sample sizes as small as four and this is an advantage in many toxicological studies, where number of subjects are small.

The proposed methodology also provides a foundation for inclusion and estimation of covariate effects and thus provides a feasible alternative to existing methodologies for models governed by system of differential equations. This has been illustrated through the simulated data examples and the real benzene inhalation data. We presented simulated examples based on a compartmental model using covariates. In case of a single covariate, the results provide small relative mean squared errors and biases for the true population values. Hence we can capture the covariate effects well in case they are truly affecting the parameter values and hence the differential equation model. The methodology when applied to the benzene inhalation data with a linear model of dependence, does not reveal any significant effect of exposure concentration on the maximum metabolic rates for bone marrow and liver.

Chapter 6

FUTURE RESEARCH DIRECTIONS

Keeping with the present and planned work on developing methodology for modeling systems of differential equations, we present in this chapter some directions for future research some of which can serve as ideas for postdoctoral work. The objective of this dissertation was to develop a formal methodology for analyzing systems governed by differential equations which can be applied to a variety of areas. The current work tries to estimate and infer about the population model parameters taking into account the intra and inter-individual variability that may exist in the data. The underlying system/phenomenon was represented using a multi-compartment system of ordinary differential equations (ODE). Individuals were observed for measurements on one or more of the compartments over time. The methodology consisted of individual parameter estimation and population parameter estimation using the estimated individual parameter values under a hierarchical model structure. The inclusion and testing of covariates in models defined by differential equations is also accomplished in this work. A functional data analytic methodology has been developed in this work motivated by physiologically based pharmacokinetic modeling which enhances

the literature on analysis of such models. The statistical framework allows for inference avoiding the solution for differential equations which is novel in case of PK and PBPK modeling situations. Toxicologists can use the methodology to infer about individuals separately as well as the population while making minimal assumptions about the distributions of the data even for small sample sizes, while obtaining valid variance component estimates. Apart from toxicology and pharmacology, the proposed methodology has potential for use in several other fields that involve modeling using differential equations to infer about individual and population patterns.

The methods developed here are for a general class of models defined by a system of ordinary differential equations. Often in modeling of biological, chemical or environmental phenomena, other classes of differential equations such as partial differential equations (PDE), stochastic differential equations (SDE) or time delayed differential equations are used for mathematical modeling. Each of these classes of equation have different structures and hence the statistical methodology for analyzing these systems needs attention. The functional data analytic methodology is yet to be extended to apply to such situations and hence serves as an important methodological area of research.

Often complex networks comprise of several interconnected modules where each module can be modeled using systems of differential equations. For instance, in studying the pharmacokinetics of a pregnant mother, one has to take into account the chemical kinetics in the fetus. Also in metabolic pathways, different chemical processes take place simultaneously or in a synchronized manner. For example, in different metabolizing sites of the human body, several enzymes metabolize various chemicals. In order to achieve a better understanding of chemical kinetics and mechanism of action, these information need to be incorporated along with pharmacokinetics. Developing a statistical methodology for such problems require the synthesis

of results from the differential equations systems defining different modules in the system.

Physiologically based pharmacokinetic modeling in itself presents ample opportunities for methodological research. Design of studies for PBPK model analysis is one of the valid questions. There are several unobserved compartments with one or two observable response variables in a typical pharmacokinetic study. This often poses a problem in analyzing high dimensional ODEs especially in genetic models. More research is needed to develop better designs for analysis using such models.

Exposure to harmful chemicals is often occupational and long-term in nature for human subjects. In some other cases, there may also be exposure to mixture of chemicals over a certain period of time. For instance, exposure to pollutants in air involves exposure to different kinds of chemical. Also there may exist multiple routes of exposure like dermal, ingestion and inhalation. These situations call for more realistic and complex models that account for such varied conditions. The proposed methodology need to be modified for analyzing such models and infer about the relative harms being caused by these exposures.

Analysis of gene regulatory networks (GRN) is a flourishing area of research. Ordinary differential equations serve as a major technique in modeling regulatory networks. A recent paper (Polynikis et al. 2009) compares different modeling approaches for modeling GRN. Simplified ODE models based on quasi-steady-state assumption of mRNA concentrations and non-linear Hill functions are used to describe the processes of translation, transcription and degradation. A sample ODE system from (Polynikis et al. 2009) is shown below. For each gene i, two ODEs are used to describe the rate of change in transcribed mRNA concentration (r_i) and the rate of change in the

translated protein (p_i).

$$\text{Transcription}: \frac{dr_i}{dt} = F(f_i^R(p_1), \ldots, f_i^R(p_n)) - \gamma_i r_i,$$
$$\text{Translation}: \frac{dp_i}{dt} = f_i^P(r_i) - \delta_i p_i,$$

where $i = 1, \ldots, n$. The functions $f_i^R(p_i)$ describe the dependence of mRNA concentration on protein concentration and are usually non-linear. Translation is described by the function $f_i^P(r_i)$. The other terms represent the degradation of mRNA and protein. The structure of the problem lends itself perfectly to the functional data analytic methodology for estimation of the parameters in this model. However, the high dimensionality of the problem needs to be balanced with the data available since gene expression profiles may not be available for all genes. The methodology developed in this dissertation may be used to identify or reconstruct the regulatory networks involving the candidate genes, taking into account the covariates that may occur in a genetic study.

Similarly, study of viral dynamics poses a problem that can be modeled using systems of differential equations and hence is a potential area of application for the proposed methodology. Consider a group of subjects being treated for influenza. We might be interested in studying the dynamics of the infection by flu virus and its effects on the human physiology from the start of treatment till remission or death. A system of differential equations can be formulated to capture the rates of change in the densities of naïve cells, infected cells and viral load. The infection process involves naïve cells getting infected by the virus, and these infected cells may die or recover over time. These processes may be affected by the rates of infection, rate of death and/or the factors determining the proliferation of the virus. Individual immune response and treatment received in form of medications also have an effect

on infections. Developing and analyzing such ODE based models in such fields would help in enhancing and validating the methodology developed in this dissertation.

References

Barton, H., Chiu, W., Woodrow, S., Andersen, M., Bailer, A., Bois, F., Dewoskin, R., Hays, S., Johanson, G., Jones, G., N.and Loizou, Macphail, R., Portier, C., Spendiff, M. and Tan, Y. (2007). Characterizing uncertainty and variability in physiologically based pharmacokinetic models: state of the science and needs for research and implementation. *Toxicological Sciences*, **99** 395–402.

Benet, L. (1984). Pharmacokinetics: Basic principles and its use as a tool in drug metabolism. *Drug Metabolism and Drug Toxicity*, **199**.

Bois, F. (2000). Statistical analysis of clewell et al. pbpk model of trichloroethylene kinetics. *Environmental Health Perspectives*, **108** 307–316.

Cui, X., Guo, W., Lin, L. and Zhu, L. (2009). Covariate-adjusted nonlinear regression. *Annals of Statistics*, **37** 1839–1870.

Davidian, M. and Giltinan, D. (1995). *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall.

Demidenko, E. (2004). *Mixed Models Theory and Applications*. Wiley-Interscience.

Denker, A. E., Morelli, G., Vessey, L., Li, S., Yuan, J., Dunbar, S., Lewis, N., Taggart, W. and Wagner, J. (2002). Pharmacokinetics of digoxin in healthy subjects receiving taranabant, a novel cannabinoid-1 receptor inverse agonist. *Advances in Therapy*, **26** 230–240.

Dunne, A. and King, P. (1989). Estimation of noncompartmental parameters: A technical note. *Journal of Pharmacokinetics and Pharmacodynamics*, **17** 131–137.

Gargas, M., Tyler, T., Sweeney, L., Corley, R., Weitz, K., Mast, T., Paustenbach, D. and Hays, S. (2000). A toxicokinetic study of inhaled ethylene glycol ethyl ether acetate and validation of a physiologically based pharmacokinetic model for rat and human. *Toxicology and Applied Pharmacology*, **165** 63–73.

Gentry, P., Covington, T., Mann, S., Shipp, A., Yager, J. and Clewell,

H. I. (2004). Physiologically based pharmacokinetic modeling of arsenic in the mouse. *Journal of Toxicology and Environmental Health, Part A*, **67** 43–71.

Haddad, S., Beliveau, M., Tardif, R. and Krishnan, K. (2001). A pbpk modelling-based approach to account for interactions in the health risk assessment of chemical mixtures. *Toxicological Sciences*, **63** 125–131.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55** pp. 757–796.

Joerger, M. (2012). Covariate pharmacokinetic model building in oncology and its potential clinical relevance. *The AAPS Journal*, **14** 119–132.

Kawahara, M., Sakata, A., Miyashita, T., Tamai, I. and Tsuji, A. (1999). Physiologically based pharmacokinetics of digoxin in mdr1a knockout mice. *Journal of Pharmaceutical Sciences*, **88** 1281–1287.

Liang, H. and Wu, H. (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, **103** 1570–1583.

Mörk, A.-K., Jonsson, F. and Johanson, G. (2009). Bayesian population analysis of a washinwashout physiologically based pharmacokinetic model for acetone. *Toxicology and Applied Pharmacology*, **240** 423 – 432.

Nagaraj, N. and Fuller, W. (1991). Estimation of the parameters of linear time series models subject to nonlinear restrictions. *Annals of Statistics*, **19** 1143–1154.

O'Flaherty, E. J., Kerger, B., Hays, S. and Paustenbach, D. (2001). A physiologically based model for the ingestion of chromium(iii) and chromium(vi) by humans. *Toxicological Sciences*, **60** 196–213.

Oprea, M., Van Nimwegen, E. and Perelson, A. (2000). Dynamics of one-pass germinal center models: Implications for affinity maturation. *Bulletin of Mathematical Biology*, **62** 121–153.

Parham, F., Matthews, H. and Portier, C. (2002). A physiologically based pharmacokinetics model of p,p'-dichlorodiphenylsulfone. *Toxicology and Applied Pharmacology*, **181** 153–163.

Perelson, A. (2002). Modeling viral and immune system dynamics. *Nature Reviews Immunology*, **2** 28–36.

Polynikis, A., Hogan, S. and Bernardo, M. (2009). Comparing different ode modeling approaches for gene regulatory networks. *Journal of Theoretical Biology*, **261** 511 – 530.

Porba, R., Ga, P., Porba, M., Juchniewicz, J. and Andrzejak, R. (2011). Relation between occupational exposure to lead, cadmium, arsenic and concentration of cystatin c. *Toxicology*, **283** 88 – 95.

Poyton, A., Varziri, M., McAuley, K., McLellan, P. and Ramsay, J. O. (2006). Parameter estimation in continuous-time dynamic models using principal differential analysis. *Comput. Chem. Eng.*, **30** 698–708.

Ramsay, J., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B*, **69(5)** 741–796.

Ramsay, J. O. (1996). Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society, Series B*, **58** 495–508.

Reddy, M., Yang, R., Andersen, M. and Clewell, H. (2005). *Physiologically Based Pharmacokinetic Modeling: Science and Applications*. Wiley and Sons : New York, USA.

Sentürk, D. and Müller, H. G. (2006). Inference for covariate-adjusted regression via varying coefficient models. *Annals of Statistics*, **34** 654679.

Steimer, J. L., Vozeh, S. and Racine-Poon, A. (1994). *The Population Approach: Rationale, Methods, and Applications in Clinical Pharmacology and Drug Development*, vol. 110, chap. 15. Berlin-Heidelberg: Springer-Verlag, 404–451.

Sweeney, L., Kirman, C., Gargas, M. and Dugard, P. (2009). Contribution of trichloroacetic acid to liver tumors in perchloroethylene (perc)-exposed mice. *Toxicology*, **260** 77–83.

Tornøe, C., Agersø, H., Jonsson, E., Madsen, H. and Nielsen, H. (2004).

Non-linear mixed-effects pharmacokinetic/pharmacodynamic modelling in nlme using differential equations. *Computer methods and program in biomedicine*, **76** 31–40.

Travis, C., Quillen, J. and Arms, A. (1990). Pharmacokinetics of benzene. *Toxicology and Applied Pharmacology*, **102** 400–420.

Varziri, M., McAuley, K. and McLellan, P. (2008a). Parameter estimation in continuous-time dynamic models in the presence of unmeasured states and nonstationary disturbance. *Ind. Eng. Chem. Res.*, **47** 380–393.

Varziri, M., McAuley, K. and McLellan, P. (2008b). Selecting optimal weighting factors in ipda for parameter estimation in continuous-time dynamic models. *Comput. Chem. Eng.*, **32** 3011–3022.

Woodruff, T. and Bois, F. (1993). Optimization issues in physiological toxicokinetic modeling: a case study with benzene. *Toxicology Letters*, **69** 181–196.

Yang, R., Dennison, J., Andersen, M., Ou, Y., Liao, K. and Reisfeld, B. (2004). *Physiologically based pharmacokinetic and pharmacodynamic modeling*. John Wiley & Sons, Inc.

Yu, R. (1995). *Measurement of Partitioning of Benzene in Breath and Urinary Metabolites from Benzene Exposure in Humans*. Ph.D. thesis, Rutgers University.

Zhang, J., Peddada, S. and Rogol, A. (2000). Estimation of parameters in nonlinear regression models. *Statistics for the 21st Century, Eds. C. R. Rao and G. Szekely* 459–483.