REGRESSION ANALYSIS FOR LONGITUDINAL HEMOGLOBIN DATA FOR PREMATURE INFANTS WITH OUTCOMES SUBJECT TO NON-RESPONSE

RAJENDRA NEUPANE

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisisitons et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-84029-8
Our file Notre référence
ISBN: 0-612-84029-8

REGRESSION ANALYSIS FOR LONGITUDINAL HEMOGLOBIN DATA FOR PREMATURE INFANTS WITH OUTCOMES SUBJECT TO NON-RESPONSE

by

Rajendra Neupane

A Practicum Report Submitted to the School of
Graduate Studies in partial fulfillment of
the requirement for the degree of Master
of Applied Statistics

Department of Mathematics and Statistics
Memorial University of Newfoundland

June, 2002

St. John's, Newfoundland, Canada

Abstract

In analyzing longitudinal hemoglobin data for low-birth-weight infants, it is of interest to examine the effects of iron fortification and other covariates such as gender and gestation weeks on the hemoglobin status of the infants over the months. As the hemoglobin data are collected repeatedly over a period of time, the longitudinal correlations of the responses must be taken into account in finding the covariate effects. Further problems get mounted when some of the responses are missing. In this practicum, we conducted a regression analysis after accomodating the longitudinal and missingness nature of the data into account. Also, several non-parametric tests were applied to examine any possible monotonic trend in the longitudinal hemoglobin data.

Acknowledgements

Thanks Goddes *Saraswati* for giving me knowledge.

I wish to thank my supervisor, Professor B.C. Sutradhar for his encouragement, inspiration, suggestions and continuous guidance in completing this practicum.

I sincerely acknowledge the financial support provided by the School of The Graduate Studies and Department of Mathematics and Statistics in the form of Graduate Fellowships and Teaching Assistantships.

I extend my sincere thanks to Professor James Friel from The Department of Biochemistry, Memorial University of Newfoundland for providing the hemoglobin data set that I have analyzed in the practicum.

I am greatful to my parents, brothers and sisters for their love, support and encouragement throughout my university career.

It is my pleasure to thank my friends and well-wishers who encouraged and helped me in the MAS program.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation of the Problem

Very low birth weight infants, defined as weighing less than 1500 gram at birth are known to be at high risk of iron deficiency. These infants are born prematurely and some experience lower than expected birth weight for gestational age. Iron deficiency in the premature infants usually affects their growth and these infants may be prone to various pathological conditions. As a counter measure to this problem, it is standard to feed these infants certain formula fortified with iron. Although the fortification amount is usually determined by comparing with the iron level of full-term infants as a standard measure, the fortification amount may however vary depending on other characterstics of the individual infants such as gender, his or her individual gestation week as well as the initial level of the iron at birth, and so on. This problem of determining the iron intake amount based on other covariates is however not adequately studied in the literature.

To study this important issue, biochemist/nutritionists usually conduct experiments where the necessary data are collected from a group of infants over a period of time. The statistical analysis of such longitudinal data however appears to be quite challenging. This is because in such cases one needs to find the effect of the treatment (iron intake level) and other covariates on responses after taking their longitudinal correlations into account. Further problems may occur when some responses are missing. Based on the above issues, we were motivated to study the problem of determining the effect of iron intake on the health status of the premature infants by applying appropriate statistical methods. This study should be useful for determining the iron needs of individual infants based on their background information, i.e., history of covariates, whereas the current recommendations for iron intake by the Canadian Paediatric Society [2 mg/kg (2 mg of iron intake per 1 kg weight of the infant)] and American Academy of Pediatrics [2 to 3 mg/kg] do not vary based on the covariate information of the individuals.

1.2 Objective of the Practicum

The main objective of the practicum is to use an appropriate statistical model to examine the level of hemoglobin attained by an infant resulting from iron intake over a period of time, conditional on his or her other covariates. Note that as this type of nutritional experiment is usually conducted over a period of time, it is likely that the hemoglobin responses of an individual infant collected over time will be longitudinally correlated. The selected statistical model must accomodate this longitudinal correlation.

When the data are collected longitudinally, it frequently happens that some data for some infants may be partially missing. This makes the longitudinal study further complicated. For example, if the missing values occur completely at random, the available data can be analyzed but the estimation may be difficult as one needs to deal with unbalanced data. If however the responses are missing at random, they require to be imputed which may not be easy in the longitudinal set up. It is also our objective to deal with this type of missing problem while finding the effect of iron intake or the hemoglobin level of the infants.

Next, we will apply certain suitable statistical tests to see whether there is any monotonic longitudinal pattern in hemoglobin levels because of iron intake over the times.

The specific plan of the practicum is as follows:

1. In chapter 2, we provide an exploratory analysis of the covariates and response variable of the Hemoglobin data before we undertake any confirmatory analyses.

2. As the hemoglobin data collected over times are continuous and they are correlated, in chapter 3, we introduce a linear model for autocorrelated hemoglobin responses. This longitudinal analysis is done for complete data, i.e we ignore the missing responses to calculate regression param-

eters and autocorrelations.

3. In chapter 4, we consider certain possible missing mechanisms and carry out the longitudinal analysis after taking the missingness nature of the data into account. More specifically, in this chapter, we analyze the longitudinal data with non-responses under the assumption that they occured completely at random.

4. We continue analyzing longitudinal missing data in chapter 5, mainly under the assumption that the missingness occured at random. Thus the analysis is done based on certain imputations.

5. As it is also important to study the longitudinal pattern in hemoglobin levels, in chapter 6, we conduct certain non-parametric tests in order to test such patterns.

6. We conclude the practicum in chapter 7.

Chapter 2

Background of the Study

2.1 Preamble

Infants of very low birth weight (defined as less than 1500 g) are at high risk for iron deficiency because of low stores of iron at birth (Gortem and Cross, 1964), rapid growth in the erythrocyte mass, which depletes the iron reserves (Worwood, 1977), and uncertainty about their iron requirements (Report by Committe on Nutrition, American Academy of Pediatrics, 1985). The Canadian Pediatric Society (Nutrition Committe, Canadian Pediatric Society, 1981) has recommended that the low-birth-weight infant receive iron in the amount of 2 mg/kg daily, either in formula or as a supplement, from about 2 months of age onward. The American Academy of Pediatrics has recommended 2 to 3 mg/kg daily for very-low-birth-weight infants and has stated that formulas with iron usually contain sufficient supplemental iron.

Note that even though the recommendations by the Canadian Pediatric

Society and American Academy of Pediatrics are available for general fortification purposes, more detailed studies are necessary to improve the iron intake standard. With this in view, Yeung and associates (Yeung et al, 1981) for example, studied the iron status of Canadian infants. Their study however was confined to the term infants only. Later on, Friel et al (1990) examined the iron status of very-low-birth-weight infants given iron-fortified formula during early infancy who were part of a prospective study of the effects of zinc supplements on growth and development. These authors concluded that because of delayed development, very-low-birth-weight infants eat small amounts of cereal and therefore require iron-fortified formula throughout infancy. For some other studies similar to Friel et al (1990), we refer to Ehrenkranz (1993) which shows that there has to be good nutritional management in preterm infants so that there can be sufficient iron supplementation to enhance iron stores and to prevent the development of iron deficiency.

Note however that in studying the effects of fortification, the above studies do not appear to have used the longitudinal correlations that may be present among the hemoglobin responses of the same infant. To be specific, no attempt has been made to understand the longitudinal correlation of the responses of the same infant, which appears to be an important issue to understand the changes in hemoglobin level for an infant. As mentioned in chapter 1, the purpose of our study, unlike the previous work, is to apply a valid statistical approach after taking the longitudinal correlations of the responses into account. Further note that for recommendations for the appropriate amount of fortification, it is also important to take the individual's characterstics into account. This is because, there may be some variable effects between female and male, for example, and among the infants belonging to different gestational age. For this purpose, we plan to conduct a comprehensive longitudinal regression analysis after taking the longitudinal correlations into account. Since the longitudinal data collected over time may also be partially missing, we require to estimate the missing values in such cases before doing further statistical analyses. In our study, we also deal with this type of missing data problem while finding the effect of iron intake on the hemoglobin levels of the infants.

In the next section, we describe in detail a hemoglobin data set collected by Dr. James Friel and his associates during June 1995 to May 1996. The statistical analyses will be given in the subsequent chapters.

2.2 Hemoglobin Data

In order to examine the effect of iron intake on premature children with low hemoglobin level, James Friel and his associates of the Department of Biochemistry, Memorial University of Newfoundland, collected a hemoglobin data set from two different hospitals namely Janeway Child Health Center and Grace General Hospital at St. John's, Newfoundland. More specifically, data were collected from 42 prematurely born infants. To begin with, the hemoglobin data were first collected for these 42 infants within the first week of their birth in June 1995. The data are referred to as baseline hemoglobin data. To study the effect of iron intake, these 42 infants were randomly assigned to two groups namely placebo and treatment groups. The gender and weeks of gestation the infants were also recorded. After collecting the baseline hemoglobin data, the hemoglobin measurements were also recorded longitudinally at 3, 6, 9 and 12 months for all these infants. Note that all of these children were healthy, eligible and had no birth defects at the time of birth, although they were premature. Further note that it was however not possible to collect the hemoglobin data for all 5 time points for each of the 42 infants. This is because some of the responses were subject to non-response or were partially missing, which may have happened because of iron intolerence, blood clotting, refused to eat, left the province, and so on. It was observed that some individuals were missing for one or two times but they continued to take the treatment again. To be specific, 25 observations were completely observed for all 5 time points and the remaining 17 were partially missing, that is, some left after some weeks or months, some joined again and some did not. As mentioned before, the treatment (iron) and placebo were given randomly, irrespective of the gender and the gestational age.

In this study, the main scientific interest is to find the effect of treatment as well as other covariates such as gender and gestational age on the hemoglobin level after taking the longitudinal and missingness nature of the data into account.

2.3 Distribution of the Variables

We now provide the sample distribution for all the fixed covariates (gender, treatment, gestation weeks and baseline hemoglobin level) and the response variable (hemoglobin) collected longitudinally.

The histogram for gender for example, is shown in figure B.1. This shows that among 42 infants, 22 were boys and 20 were girls. The number of individuals in the treatment and placebo groups are shown in figure B.2. In fact, the number in each group was same, that is, 21 in each group. The distribution of the infants according to 3 gestation groups (26-29 weeks, 30-34 weeks and 35-38 weeks) is shown in figure B.3. It was observed that there were 7 infants in the very-low-birth gestation week (26-29), 32 were in the middle group gestation week (30-34) and another 4 were in the last group (35-38). The distribution of baselevel hemoglobin is presented in figure B.4. A small number of infants were found to have either very small or very large hemoglobin levels. A large number of infants (40) were found with 8.0 to 10.0 mg/dl hemoglobin level. We have also plotted the baseline hemoglobin data as opposed to the ordered gestation weeks, which is exhibited in figure B.10. This figure shows that the baseline hemoglobin is higher for the infants with higher gestation week.

The original data set under study is shown in Table A.1 in appendix A. Figure B.5 in Appendix B shows the longitudinal plot of hemoglobin levels for 42 individuals at 5 different time points. Note that although most of the observations were available for all time points, some of them are imputed based on mean imputation technique explained in chapter 5. There is high variations in hemoglobin levels at the first time point, where hemoglobin values range from minimum of 8.0 mg/dl to maximum of 16.0 mg/dl. But from the second time point and on, the variation in hemoglobin value was comparatively smaller than at the first time point.

Figure B.6 shows the plot of baseline hemoglobin, hemoglobin at first time point and hemoglobin at fifth time point to illustrate the effect of treatment and other covariates over time. This figure indicates more variation of hemoglobin level at baseline and first time point, but at the fifth time point, the hemoglobin level appear to be smoother. When the baseline hemoglobin level is compared with the level at time point 5, there appear to be an overall increase at the fifth time point. But the changes were more for those infants with smaller baseline hemoglobin values as compared to other infants with higher hemoglobin values (see Figure B.6).

2.4 Notations

In this section, we develop some notations for the longitudinal hemoglobin data that we explained in section 2.3. These notations will be used for the confirmatory analysis in the subsequent chapters.

Let Y_{it} denote the Hemoglobin level collected at t^{th} time for the $i^{th}(i = 1, 2, \ldots, I)$ individual under study. Also, let $x'_{it} = (x_{it1}, \ldots, x_{itu}, \ldots, x_{itp})$ be the corresponding $p \times 1$ covariate vector. In the present set up, $I = 42$ and $p = 4$. To be specific x_{it1} will denote the Gender (1=male, 0=female) for the i^{th} individual, and x_{it2} will denote the treatment given (1=treatment, 0=Placebo) where iron is taken to be the treatment. Likewise x_{it3} denotes the gestation period of the child expressed in number of weeks. Any baby having gestation period less than 38 weeks is regarded as low gestation period. Finally, x_{it4} denotes the baselevel hemoglobin of the i^{th} individual. Note that although all these covariates are written as time (t) dependent, but in our studies, these covariates are actually time independent, whereas y_{it} really depends on t.

As it is seen from the data presented in Table A.1, all x values were available for all 42 individuals under study. The y values however were not available for all 5 time points. Moreover, the responses $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ for $T = 5$ are correlated as they are reported hemoglobin levels over T consecutive periods. The purpose of this study is to find the effect of covariates x on the responses y, after taking the missingness and the longitudinal nature of the data into account.

Chapter 3

Longitudinal Analysis for Complete Data

3.1 Linear Model for Autocorrelated Responses

As the hemoglobin data collected over time for each of the infants are continuous by nature, one may exploit a linear regression model with autocorrelated error to examine the effect of covariates on the hemoglobin labels recorded for each individuals.

Let y_{it} be the Hemoglobin level recorded at the $t^{th}(t = 1, 2, \ldots, T)$ occasion for $i^{th}(i = 1, \ldots, I)$ infant. Also let $x_{it} = (x_{it1}, \ldots, x_{itu}, \ldots, x_{itp})'$ be the $p \times 1$ vector of covariates corresponding to y_{it}. For the present Hemoglobin data, the covariates are however not time dependent. More specifically, we denote 'intercept' covariate with the code $x_{it1} = 1$, 'gender' by x_{it2} with the code 0 for female and 1 for male. Likewise, we denote the covarite 'formula' by x_{it3} with the code 1 for treatment and 0 for placebo and finally we denote the covariate 'gestation week' by x_{it4} with $26 \leq x_{it4} \leq 38$ and 'baseline Hemoglobin' by x_{it5} with $76 \leq x_{it5} \leq 175$,

Note that $y_{i1}, \ldots, y_{it}, \ldots, y_{iT}$ are the repeated observations collected for the i^{th} infant. Consequently, it is natural to expect that these observations will be correlated. In the longitudinal set up, T is usually small. In the present longitudinal data set up $T = 5$. As y_{it}'s are continuous, one may fit a linear model

$$y_i = X_i\beta + \epsilon_i \tag{3.1}$$

and compute the $p \times 1$ regression vector $\beta = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$ after taking the longitudinal correlations of the observations into account.

In equation (3.1), $y_i = (y_{i1}, \ldots, y_{it}, \ldots, y_{iT})'$ is the $T \times 1$ vector of responses for the i^{th} infant, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{it}, \ldots, \epsilon_{iT})'$ is the corresponding error vector and X_i is the $T \times p$ covariate matrix for the i^{th} individual defined as

$$X_i = \begin{pmatrix} x_{i1}' \\ x_{i2}' \\ \vdots \\ x_{it}' \\ \vdots \\ x_{iT}' \end{pmatrix} \tag{3.2}$$

where x'_{it} is the $1 \times p$ covariate vector as defined earlier.

As far as the error variable is concerned, we may assume that $\epsilon_{it} \sim (0, {\sigma_i}^2)$. Further to accomodate the longitudinal correlations of the responses, we assume that the error vector ϵ_i has the autocorrelation structure given by

$$C(\rho_1, \ldots, \rho_{T-1}) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ & 1 & \rho_1 & \cdots & \rho_{T-2} \\ & & & & \vdots \\ & & & 1 & \rho_1 \\ & & & & 1 \end{pmatrix}, \tag{3.3}$$

(Sutradhar and Das, 1999) where $\rho_l (l = 1, 2, \ldots, T-1)$ denotes the l^{th} lag autocorrelation defined as

$$\rho_l = \frac{Cov(\epsilon_{it}, \epsilon_{i,t+l})}{\sqrt{var(y_{it}) var(y_{it+l})}}, \tag{3.4}$$

Note that the correlation structure (3.3) is suggested by Sutradhar and Das (1999) in a non-linear regression set-up, whereas in our context, the regression model is linear as in equation (3.1). Under the assumption that $y_{it} \sim (x'_{it}\beta, {\sigma_i}^2)$, one can fit the linear model (3.1), i.e.,

$$y_i = X_i \beta + \epsilon_i. \tag{3.5}$$

Note that the autocorrelation structure (3.3) is quite general as it can accomodate the Gaussian type AR(1), MA(1) or exchangeable correlation pattern. For example, if error vector follows Gaussian type AR(1) process, then $\rho_l = \rho^l$, where ρ is a suitable parameter ranging from -1 to $+1$.

3.2 Estimation of Regression Parameters

By using the equation (3.4), we now express the $T \times T$ covariance matrix of the error vector ϵ_i as,

$$\Sigma_i = A_i^{1/2} C(\rho_1, \ldots, \rho_{T-1}) A_i^{1/2}, \tag{3.6}$$

where $A_i = diag[var(\epsilon_{i1}), \ldots, var(\epsilon_{it}), \ldots, var(\epsilon_{iT})] = {\sigma_i}^2 I_T$ with I_T as the $T \times T$ identity matrix. Next, by applying the generalized least square theory, one may minimize the genaralized error sums of square L given by

$$L = \epsilon_i' \Sigma_i^{-1} \epsilon_i \tag{3.7}$$

with respect to β and obtain the complete data based generalized least square estimate for the regression parameter β as

$$\hat{\beta}_{G,c} = (\sum_{i=1}^{I} X_i' \hat{\Sigma_i}^{-1} X_i)^{-1} (\sum_{i=1}^{I} X_i' \hat{\Sigma_i}^{-1} y_i) \tag{3.8}$$

with

$$\hat{\Sigma}_i \;=\; \hat{A}_i^{\;1/2} C(\hat{\rho}_{1,c}, \ldots, \hat{\rho}_{T-1,c}) \hat{A}_i^{1/2} = \hat{\sigma_i}^2 C(\hat{\rho}_{1,c}, \ldots, \hat{\rho}_{T-1,c})$$

3.3 Estimation of Longitudinal Correlations

In equation (3.8), $\hat{\rho}_{l,c}$ denotes the estimator of ρ_l based on the complete data. This estimator has to be consistent for ρ_l which can be obtained using the formula

$$\hat{\rho}_{l,c} = \frac{\sum_{i=1}^{I} \sum_{t=1}^{T-l} z_{it} z_{i,t+1} / I(T-l)}{\sum_{i=1}^{I} \sum_{t=1}^{T} z_{it}^2 / I.T} \tag{3.9}$$

where $z_{it} = (y_{it} - \mu_{it})$ and $\mu_{it} = x_{it}'\hat{\beta}$ and $\sigma_i{}^2$ is estimated by

$$\hat{\sigma}_i^2 = \sum_{t=1}^{T} (y_{it} - x_{it}'\hat{\beta})^2 / T \tag{3.10}$$

and the variance of $\hat{\beta}$ is given by

$$V(\hat{\beta}) = (\sum_{i=1}^{I} X_i' \hat{\Sigma_i}^{-1} X_i)^{-1} \tag{3.11}$$

3.4 Application to Complete Hemoglobin Data.

In this sub-section we now apply the procedure discussed in section 3.1 to analyze the longitudinal hemoglobin data from the children's hospital discussed in chapter 2. To begin with we choose a subset of size 25 out of 42 children with complete information available for all five time points. This means we consider $I = 25$. The objective is to compute the values of $\hat{\beta}_{G,c}$ and $\hat{\rho}_{1,c}, \ldots, \hat{\rho}_{T-1,c}$ for 25 individuals by using the formulas (3.8) and (3.9).

The hemoglobin data for the $i^{th}(i = 1, 2, \ldots, 25)$ child collected at time points $(t = 1, 2, \ldots, 5)$ is denoted by y_{it}. Next as mentioned earlier, we consider all five covariates including the intercept. These covariates are 1. 'intercept' x_{it1}; 2. 'gender' x_{it2}; 3. 'formula' x_{it3}; 4. 'gestation' x_{it4}; 5. 'baselevel hemoglobin' x_{it5}. Here, in general, x_{itu} represents the $u^{th}(u = 1, 2, \ldots, 5)$ covariate for the i^{th} individual. We also examine the autocorrelation structure of the longitudinal responses.

Note that to compute $\hat{\beta}_{G,c}$, one requires to know the values of $\hat{\rho}_{1,c}, \ldots, \hat{\rho}_{T-1,c}$ as well as $\hat{\sigma}_i^2$. To begin with we consider $\hat{\rho}_{1,c} = 0, \ldots, \hat{\rho}_{T-1,c} = 0$ and $\hat{\sigma}_i^2 = 1$ for $(i = 1, 2, \ldots, 25)$ and solve equation (3.8) for $\hat{\beta}_G$. Denote this solution by $\hat{\beta}_{G(1)}$ as the first step estimate of β. This estimate we then use in equation (3.9) to compute the first step estimate of autocorrelations denoted by $\hat{\rho}_{1(1)}, \ldots, \hat{\rho}_{T-1(1)}$. The first step estimate of β is used as well in equation (3.10) to compute $\hat{\sigma}_i^2$. Next the first step values of $\hat{\rho}$ and $\hat{\sigma}_i^2$ are used in equation (3.8) to obtain an improved estimate of β, which in turn is used in equation (3.9) and (3.10) to compute improved estimates of longitudinal correlations and $\hat{\sigma}_i^2$. This constitute a cycle of iterations which we continue until con-

Type of Parameter	Parameter	Estimate	Standard errors
Regression effects $\beta : 5 \times 1$	Intercept	117.187	0.170
	Gender	4.697	0.177
	Treatment	1.850	0.176
	Gestation week	-0.215	0.048
	Baseline Hemo.	0.008	0.004
Auto lag correlation	ρ_1	0.149	
	ρ_2	0.005	
	ρ_3	-0.178	
	ρ_4	-0.401	

Table 3.1: Estimates of Regression and Autocorrelation Parameters for complete Hemoglobin Data from 25 individuals.

vergence in the values of $\hat{\beta}$ and $\hat{\rho}_{1,c}, \ldots, \hat{\rho}_{T-1,c}$. The convergence estimates are summarized in Table 3.1. The standard error of $\hat{\beta}_{G,c}$ are computed by taking the square root of the estimate of diagonal elements of $V(\hat{\beta})$ given by $V(\hat{\beta}) = (\sum_{i=1}^{I} X_i' \hat{\Sigma}_i^{-1} X_i)^{-1}$. They are also shown in Table 3.1.

It is clear from the table 3.1 that the auto lag correlations decrease from moderately positive values to a large negative value. This shows an unusual pattern which does not appear to satisfy the well known lower order such as AR(1) or MA(1) or other correlation models. Rather lag 1 correlation is somewhat positively large, ρ_2 is almost same and other are negatively small.

The intercept effect is found to be significant. The gender effect is also significant. This shows that the change in the hemoglobin level of an individual child depends upon the gender. Here the treatment seems more effective in males than in females. Likewise, the 'treatment' covariate has a positive effect upon the hemoglobin level, i.e., hemoglobin levels increased significantly for the infants those who were given the treatment.

Note that unlike the other covariates, gestation week was found to have negative regression effects (-0.215) on the hemoglobin level. This however does not mean that the infants with larger gestation week had smaller baseline hemoglobin level. That the baseline hemoglobin was more for infants with larger gestation week is shown in figure B.10. But as is seen from figure B.9, the predicted hemoglobin for the infants with smaller gestation week has increased to large extent, whereas the increase in hemoglobin level was moderate or small for the infants with higher gestation week. This resulted in negative effects of gestation week. Once again, it seems from these results that the treatment and other covariates worked favourably for the infants with smaller gestation week as compared to the infants with higher gestation week. Finally, the small positive value (0.008) of baseline hemoglobin indicates that the predicted hemoglobin was higher for the infants with higher baseline hemoglobin.

Chapter 4

Longitudinal Analysis For Incomplete Data Without Any Estimation of Missing Values

4.1 Estimation of Regression Parameters In The Presence of Missing Data

In chapter 3 we estimated regression parameters of a linear model with correlated errors based on complete data. Among 42 children, all together there were 25 children with complete responses for 5 time points. There was at least one observation missing for each of the remaining 17 children. To be specific, 10 individuals had one value missing at the first time point, 3 individuals had one value missing at the third time point, 2 individuals had one value missing at the fifth time point. Likewise, one individual had 2 values missing at the first and second time points. Finally, one individual had 3 values missing at third, fourth and fifth time points. Thus including the group with complete information, there are 6 groups of children with 6 different types of missing cases. Let g denote the g^{th} group and n_g be the number of children in that g^{th} group for $g = 1, 2, \ldots, 6$.

Next $y_{i(1)}$ be the T dimensional vector containing $T = T_1 = 5$ repeated observations for the i^{th} child of the first group. In general $y_{i(g)}$ may be defined as a T_g dimensional vector for the i^{th} child of the g^{th} group. Here $T_g \leq T_1 (= T)$. As there are 2 missing values in the $5^{th}(g = 5)$ group, $y_{i(5)}$ will indicate a vector of dimension $T_5 = 3$. Suppose that $\mu_{i(g)}$ denotes the expectation of $y_{i(g)}$ and $\Sigma_{i(g)}$ denotes the covariance matrix of $y_{i(g)}$. For example, for g=5, the $T_5(= 3)$ dimensional vector $y_{i(5)}$ has the mean

$$\mu_{i(5)} = \begin{pmatrix} \mu_{i3} \\ \mu_{i4} \\ \mu_{i5} \end{pmatrix}$$

as the first two values are missing for the i^{th} child. Consequently we can write $\Sigma_{i(5)}$ as

$$\Sigma_{i(5)} = \begin{pmatrix} \sigma_{i33} & \sigma_{i34} & \sigma_{i35} \\ \sigma_{i43} & \sigma_{i44} & \sigma_{i45} \\ \sigma_{i53} & \sigma_{i54} & \sigma_{i55} \end{pmatrix},$$

where $\Sigma_{i(5)} = A_{i(5)}^{1/2} C_{i(5)} A_{i(5)}^{1/2}$ and $C_{i(5)}$ can be written as,

$$C_{i(5)} = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{pmatrix}.$$

Note however that when responses are missing in a non-monotonic way, then the mean vector, covariance matrix and correlation matrix may be similarly computed by deleting the rows and columns of the $T_g \times T_g$ matrix corresponding to the missing responses. For example, for the third group, the $T_3(= 4)$ dimensional vector $y_{i(3)}$ has mean

$$\mu_{i(3)} = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i4} \\ \mu_{i5} \end{pmatrix}$$

as the third response is missing for the i^{th} child. Consequently we can write the covariance matrix of the third group $\Sigma_{i(3)}$ as

$$\Sigma_{i(3)} = \begin{pmatrix} \sigma_{i11} & \sigma_{i12} & \sigma_{i14} & \sigma_{i15} \\ \sigma_{i21} & \sigma_{i22} & \sigma_{i24} & \sigma_{i25} \\ \sigma_{i41} & \sigma_{i42} & \sigma_{i44} & \sigma_{i45} \\ \sigma_{i51} & \sigma_{i52} & \sigma_{i54} & \sigma_{i55} \end{pmatrix},$$

where $\Sigma_{i(3)} = A_{i(3)}^{1/2} C_{i(3)} A_{i(3)}^{1/2}$ with $C_{i(3)}$ given by

$$C_{i(3)} = \begin{pmatrix} 1 & \rho_1 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_1 & 1 \end{pmatrix}.$$

Let $N(y_{i(g)}|\mu_{i(g)}, \Sigma_{i(g)})$ denote the T_g-dimensional ($T_g \leq T_1 = T = 5$) multinormal density for $y_{i(g)}$ with mean $\mu_{i(g)}$ and covariance $\Sigma_{i(g)}$. One can then write the likelihood function for all individuals as follows:

$$L = \prod_{g=1}^{G} \prod_{i=1}^{n_g} N(y_{i(g)}|\mu_{i(g)}, \Sigma_{i(g)}) \tag{4.1}$$

Note that for the purpose of the estimation of the regression vector β, the maximization of likelihood function (4.1) in fact is equivalent to minimization of the quadratic function given by

$$\sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{i(g)} - \mu_{i(g)})' \Sigma_{i(g)}^{-1} (y_{i(g)} - \mu_{i(g)}) \tag{4.2}$$

where $\mu_{i(g)} = X_{i(g)}\beta$. After some algebra, minimization of the function (4.2) yields the incomplete data based estimator of β as

$$\hat{\beta}_{G,inc} = [\sum_{g=1}^{G} \sum_{i=1}^{n_g} X'_{i(g)} \Sigma_{i(g)}^{-1} X_{i(g)}]^{-1} [\sum_{g=1}^{G} \sum_{i=1}^{n_g} X'_{i(g)} \Sigma_{i(g)}^{-1} y_{i(g)}] \tag{4.3}$$

Note that the construction of the likelihood function (4.1) is quite similar to that of the likelihood function discussed by Krisnamoorthy and Panala (1999) for the estimation of the parameters of multivariate normal distribution with missing cases. The difference between the present case and that of these authors is that while they estimated the parameters in a non-regression set up, we are dealing with the estimation of the parameters in an extended regression set up.

4.2 Estimation of Longitudinal Correlations With Presence of Missing Data

Next, similar to the estimation of β by (3.8), the estimation of β by (4.3) also requires the longitudinal correlations to be known. Note however that unlike in (3.8), the β estimation in (4.3) is done based on unequal number of observations under different groups. The estimation of correlations for this type of unbalanced data is complicated. We however develop a mechanism that we describe below for the purpose of estimation of correlation in such a case.

Let $r_{it} = 1$ if y_{it} is present, otherwise $r_{it} = 0$ for all $i = 1, 2, \ldots, I$ and $t = 1, 2, \ldots, T$. As in the present approach, the missing values are ignored in β estimation, they have to be ignored for the estimation of the lag correlations as well. Consequently, the lag auto correlation can be computed by using a general formula given by

$$\hat{\rho}_l = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T-l} r_{it}r_{i,t+l}z_{it}z_{i,t+l} / \sum_{i=1}^{I}\sum_{t=1}^{T-l} r_{it}r_{i,t+l}}{\sum_{i=1}^{I}\sum_{t=1}^{T} r_{it}z_{it}^2 / \sum_{i=1}^{I}\sum_{t=1}^{T} r_{it}} \tag{4.4}$$

where $z_{it} = (y_{it} - x'_{it}\hat{\beta}_{G,inc})$ if $r_{it} = 1$. For $r_{it} = 0$, it is not necessary to compute z_{it} as this quantity does not contribute towards ρ_l computation. Instead, for simplicity, one can use 0 or some constant value for such z_{it}'s.

The computation of $\Sigma_{i(g)}$ in (4.3) also requires the estimation of σ_i^2, which in the present case is computed by using the formula

$$\hat{\sigma}_{i,inc}^2 = \frac{\sum_{t=1}^{T} r_{it}(y_{it} - x'_{it}\hat{\beta}_{G,inc})^2}{\sum_{t=1}^{T} r_{it}}. \tag{4.5}$$

Further we compute the variance of $\hat{\beta}_{G,inc}$ as

$$V(\hat{\beta}_{G,inc}) = [\sum_{g=1}^{G}\sum_{i=1}^{n_g} X'_{i(g)}\hat{\Sigma_{i(g)}}^{-1}X_{i(g)}]^{-1} \tag{4.6}$$

4.3 Application to Incomplete Hemoglobin Data

In this sub-section we apply the techniques discussed in section 4.1-4.2 to analyze the incomplete longitudinal hemoglobin data without any imputations. Note that all together there were 42 children among whom 25 children had complete responses for 5 time points and the remaining 17 had at least one observation missing. As mentioned in section 4.1, 10 individuals had one value missing at the first time point, 3 had one value missing at the third time point, 2 individuals had one value missing at the fifth time point. Likewise, 1 individual had 2 values missing at the first and second time points and finally one individual had 3 values missing at third, fourth and fifth time points. All these six size groups of responses including the group of individuals with complete information are identified by the values of $g = 1, 2, \ldots, 6$ where g stands for the g^{th} group.

By using the estimating formulas (4.3) for β and (4.4) for autocorrelation structure, we obtain their estimates as reported in Table 4.1

Type of Parameter	Parameter	Estimate	Standard errors
Regression effects $\beta : 5 \times 1$	Intercept	123.3	0.937
	Gender	2.395	0.142
	Treatment	0.646	0.144
	Gestation week	-0.265	0.03
	Baseline Hem.	0.004	0.003
Auto lag correlation	ρ_1	0.232	
	ρ_2	0.008	
	ρ_3	-0.118	
	ρ_4	-0.380	

Table 4.1: The values of estimates of $\hat{\beta}_{u,c}$ for $u = 1, 2, \ldots, 5$ and autocorrelation values $\hat{\rho}_l$ for $l = 1, 2, 3, 4$ for Incomplete Hemoglobin Data from all 42 individuals

From the above table, we can see that the intercept has a significant effect on the model. The gender effect also seems significant which indicates that the increase in hemoglobin level depends upon gender and is higher for males than for females. Also the treatment plays a positive role to increase the hemoglobin level of the individual children. One can see from the above table that the lag correlation values decrease from moderately positive values to large negative values. Clearly, this correlation pattern does not seem to satisfy the well known lower order AR(1) or MA(1) or other correlation models. We are however not concerned about specifying any correlation structure. Rather we have used a robust correlation structure which can be valid for lower as well as higher order correlation process.

When the results of Table 4.1 are compared to that of Table 3.1, the regression estimates are found to be smaller except for the intercept parameter. As we have used information from 42 children from Table 4.1 as opposed to the information from 25 children from Table 3.1, the standard errors for the regression estimates from Table 4.1 are found to be smaller than those of Table 3.1 as expected. Thus the estimates found in Table 4.1 are more efficient than those in Table 3.1.

The interpretation of the results of Table 4.1 are similar to those of Table 3.1. Consequently, we are not repeating the interpretation here.

Chapter 5

Imputation Based Longitudinal Analysis For Incomplete Data

5.1 Estimation of Parameters Under MCAR and MAR Mechanisms

In chapter 4, we have ignored the missing values in estimating the parameters, which one can do provided the missing responses occur completely at random (MCAR). In the present set up, it is however not clear whether the missingness occured following such a simple missing response mechanism. As a remedy, one needs to model the missingness mechanism, which is however very difficult to do. To ease the situation, we assume that the missingness in this data set is dependent on the first response i.e, baseline hemoglobin recorded for the particular child. This mechanism is referred to as missing at random (MAR)

Number of responses present from beginning	t=1	t=2	t=3	t=4	t=5	t=6
1		(2,1) *	(3,1) *	(4,1)	(5,1)	(6,1)
2			(3,2)	(4,2)	(5,2)	(6,2)
3				(4,3) *	(5,3) *	(6,3) *
4					(5,4)	(6,4)
5						(6,5) *
6						

Table 5.1: Missing Pattern

of type I (see Paik, 1997). Under such MAR mechanism, one may consistently impute the missing values by using a sequential mean imputation technique. The sequence of imputation may be clearly identified by following the missing pattern shown in the above table 5.1.

The asteriks on the above table indicate the missing positions for the present data set. Except the missing value at $(6,3)*$, none of the others require sequential imputation. This is because under the columns corresponding to $t = 2, 3, 4$ and 5, there is only one missing position, whereas under the column with $t = 6$, there are two missing positions to fit out. Consequently, to obtain approximately consistent imputed values, for all possible missing values, we simply follow the regular mean imputation technique.

Suppose that for the i^{th} child, we need to impute the value missing at the t^{th} time. Let $\tilde{y}_{it}$ denote this imputed value. Then the simple mean imputation formula for this is given by

$$\tilde{y}_{it} = \frac{\sum_{j\neq i}^{I} y_{jt} r_{jt} I(D_{j,t-1} = D_{i,t-1})}{\sum_{j\neq i}^{I} r_{jt} I(D_{j,t-1} = D_{i,t-1})} \tag{5.1}$$

where y_{jt} is the response at t^{th} time point for the j^{th} child, r_{jt} is the response indicator for the j^{th} child at t^{th} time as in the previous chapters and $I(D_{j,t-1} = D_{i,t-1})$ is the indicator variable which takes value 1 if the covariate history of the j^{th} child upto time $t-1$ is same as the covariate history of the i^{th} child, otherwise $I(D_{j,t-1} = D_{i,t-1}) = 0$. For example, for $i = 26^{th}$ child whose value is missing at second time period ($t = 2$) only, the imputed value was found to be 110.25 which is computed as $\tilde{y}_{26,2} = (y_{6,2} + y_{7,2} + y_{11,2} + y_{19,2})/4 = (139 + 115 + 92 + 95)/4 = 110.25$ i.e., $\tilde{y}_{26,2} = 110.2$. Likewise the imputed value denoted by $\tilde{y}_{42,6}$ for $i = 42^{th}$ child at sixth time point is 134.3. Note that unlike in the previous chapters we have 6 time points here. This is because the baseline hemoglobin is treated as the first response at the first time point, whereas this was treated as a covariate in previous chapters.

Next to compute the regression effects of the covarites for such imputations based data, we first show how to compute the A_i matrix defined in equation (3.6). For the i^{th} child with c missing values, the variance at time point t (t^{th} response is being observed) is calculated by using

$$\hat{\sigma}_i^2 = \sum_{u=1}^{T-c} (y_{iu} - x'_{iu}\hat{\beta})^2/(T-c). \tag{5.2}$$

For $c = 0$, it reduces to $\hat{\sigma}_i^2$ in equation (3.10) exactly.

Next to find the variance of an imputed observation $\tilde{y}_{it}$, say, defined in equation (5.1), we write

$$v(\tilde{y}_{it}) = \frac{\sum_{j\neq i}^{I} var(y_{jt}) r_{jt}^2 [I(D_{j,t-1} = D_{i,t-1})]^2}{[\sum_{j\neq i}^{I} r_{jt} I(D_{j,t-1} = D_{i,t-1})]^2} \tag{5.3}$$

which may be rewritten as

$$v(\tilde{y}_{it}) = \sigma_i^2 / m \tag{5.4}$$

where m denotes the number of individuals whose responses were used to compute $\tilde{y}_{it}$. In equation (5.4), we have used $v(y_{jt}) = \sigma_i^2$, as the $j^{th}(j \neq i)$ individual has the same history as the i^{th} individual. Consequently for the i^{th} individual with one imputed missing value, the diagonal matrix A_i defined in equation (3.9) may be written as

$$A_i = diag[\sigma_i, \ldots, \sigma_i, \tilde{\sigma}_{it}, \ldots, \sigma_i] = diag[\sigma_i, \ldots, \sigma_i, \sigma_i/\sqrt{m}, \ldots, \sigma_i] \tag{5.5}$$

We however estimate this σ_i^2 by pooling the available information from all $m + 1$ individuals including the i^{th} child, and the formula is given by

$$\sigma_i^2 = \frac{\sum_{j\neq i}^{I}\sum_{t=1}^{T} r_{jt}(y_{jt} - x'_{jt}\hat{\beta})^2 + \sum_{t=1}^{T} r_{it}(y_{it} - x'_{it}\hat{\beta})^2}{\sum_{j\neq i}^{I}\sum_{t=1}^{T} r_{jt} + \sum_{t=1}^{T} r_{it}}, \tag{5.6}$$

where the j^{th} child has the same history as the i^{th} child, where $r_{it} = 1$ if the response of the i^{th} child at t^{th} time is observed and 0 otherwise.

Next to compute the auto lag correlation, define $z_{it} = r_{it}y_{it} + (1 - r_{it})\tilde{y}_{it}$ and $w_{it} = [r_{it}\sigma_{it} + (1 - r_{it})\tilde{\sigma}_{it}]^{-1}$. Also define $z_i = (z_{i1}, z_{i2}, \ldots, z_{iT})'$ as a $T \times 1$ vector with individual element z_{it}. As y_{it}, $\tilde{y}_{it}$ and hence z_{it} has the same mean $x'_{it}\beta$, the l^{th} auto lag correlation can be computed as

$$\hat{\rho}_l = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T-l} w_{it}w_{i,t+l}(z_{it} - x'_{it}\beta)(z_{i,t+l} - x'_{it}\beta)/I(T - l)}{\sum_{i=1}^{I}\sum_{t=1}^{T-l} w_{it}^2(z_{it} - x'_{it}\beta)^2/IT} \tag{5.7}$$

Next, the full vector z_i containing original and/or imputed values, along with the values of $\hat{A}_i$ and $\hat{\rho}_l$ are used to estimate the regression coefficient by using the formula

$$\tilde{\beta} = [\sum_{i=1}^{I} X_i'\tilde{\Sigma}_i^{-1}X_i]^{-1}[\sum_{i=1}^{I} X_i'\tilde{\Sigma}_i^{-1}z_i], \tag{5.8}$$

which is similar but different than that of equation (3.8). Here $\tilde{\Sigma}_i$ is the variance of z_i, whereas in equation 3.8, Σ_i is the covariance matrix of y_i with no missing responses.

5.2 Application to Hemoglobin Data based on Imputaion Technique

In this subsection we include the imputed values discussed in subsection 5.1 to analyze the longitudinal hemoglobin data. There were 17 children whose values were missing at least for one time point and the remaining 25 children had the complete information. The missing pattern and the position are presented in Table 5.1. We followed the regular mean imputation technique to obtain approximately consistent imputed values. The imputed values are given in Appendix A Table A.1. We used those imputed values to calculate the regression coefficients and autocorrelation. These estimated values are presented in the following table 5.2.

Type of Parameter	Parameter	Estimate	Standard errors
Regression effects $\tilde{\beta} : 4 \times 1$	Intercept	119.79	0.130
	Gender	2.74	0.137
	Treatment	0.381	0.138
	Gestation week	-0.304	0.028
Auto lag correlation	ρ_1	0.244	
	ρ_2	0.076	
	ρ_3	-0.115	
	ρ_4	-0.178	

Table 5.2: The values of estimates of $\tilde{\beta}$ for $u = 1, 2, \ldots, 4$ and autocorrelation values $\tilde{\rho}_l$ for $l = 1, 2, 3, 4$ for Hemoglobin Data using imputed values.

It will be meaningful to compare the results of Table 5.2 with the results from the previous two tables 3.1 and 4.1. Table 3.1 shows the regression and autocorrelation estimates from complete data only. Similarly, Table 4.1 shows these estimates using the incomplete or observed data available from 42 individuals. Here missing values are ignored under the assumption that the missingness occured completely at random.The regression and autocorrelation estimates in Table 5.2 are more or less similar to table 3.1 and 4.1 but they are found from full data containing both original and imputed values.

Chapter 6

Non-parametric Testing For Longitudinal Monotonic Changes

In the longitudinal studies, it is of interest to estimate the regression effects after taking the longitudinal correlations into account. Also it may be of interest to study the longitudinal changes in hemoglobin levels for an individual over a period of time. To be specific, one may be interested to know whether there is any longitudinal pattern in the responses such as increasing or decreasing trend in hemoglobin levels over the period of time. To examine the presence of any such trend, in this chapter, we apply severel non-parametric tests to the full hemoglobin data obtained after imputations as in the previous chapter.

Recall that in the previous chapters, we have considered time as a stochastic factor so that the hemoglobin observation vector y_i for the i^{th} individual had T- dimensional symmetric distribution with mean $x_i\beta$ and covariance matrix Σ_i(3.5-3.6), where Σ_i was constructed based on longitudinal correlation structure appropriate for repeated data. In this chapter, we introduce a fixed time factor say τ_t and modify the linear model (3.5) as

$$y_{it} = x'_{it}\beta + \tau_t + \epsilon^*_{it}, \tag{6.1}$$

and examine whether there is any trend in the time effects $\tau_1, \tau_2, \ldots, \tau_5$. In modifying the linear model, we now assume that ϵ^*_{it} for $t = 1, 2, \ldots, T$ are independently and identically distributed with median 0 and a scale parameter σ_i^2. Thus, the time factor in equation (6.1) is no longer stochastic, rather the time effect is represented by τ_T only. We now test whether there is any trend in time effects $\tau_1, \tau_2, \ldots, \tau_5$ for the hemoglobin data by applying three distribution free tests.

6.1 Jonckheere-Terpstra Distribution Free Test

Note that this Jonckheere-Terpstra distribution free test procedure (Hollander and Wolfe, (1999), chapter 6, section 6.2, page 202) tests the null hypothesis

$$H_o : \tau_1 = \tau_2 = \ldots = \tau_5 \tag{6.2}$$

against the alternative hypothesis

$$H_1 : \tau_1 \leq \tau_2 \leq \ldots \leq \tau_5, \tag{6.3}$$

with at least one strict inequality. Although this test could be performed under the general regression model (6.1), we however, to begin with, use $\beta = 0$ and examine the time effect only on the responses.

For a selected pair (u, v) so that $1 \leq u < v \leq T$, let U_{uv} be a Mann-Whitney count defined as

$$U_{uv} = \sum_{i=1}^{I} \sum_{j=1}^{I} \phi(Y_{iu}, Y_{jv}), \tag{6.4}$$

where Y_{iu}, for example, is the u^{th} hemoglobin level for the i^{th} child, $\phi(a, b) = 1$ if $a < b$, 0 otherwise. We now add all $T(T-1)/2$ values of U_{uv} and form the J statistic given by

$$J = \sum_{u=1}^{T-1} \sum_{v=u+1}^{T} U_{uv} \tag{6.5}$$

This J statistic is now standardized as

$$J^* = \frac{J - E_o(J)}{\sqrt{var_o(J)}}, \tag{6.6}$$

where $E_o(J)$ and $var_o(J)$ are the expectation and variance of J respectively under the null hypothesis H_o. The formulas for mean and variance are

$$E_o(J) = \frac{N^2 - \sum_{j=1}^{T} n_j^2}{4} \tag{6.7}$$

and

$$var_o(J) = \frac{N^2(2N+3) - \sum_{j=1}^{T} n_j^2(2n_j+3)}{72}, \tag{6.8}$$

where n_j is the number of individuals at j^{th} time point which is I i.e., $n_j = I$ in our case and $N = T \times n_j = TI$.

But if there is any tie among data, the null variance will be slightly different. In such a case, the test statistic is modified as

$$J^{**} = \frac{J - E_o(J)}{\sqrt{var_o^*(J)}} \tag{6.9}$$

where $var_o^*(J)$, the null variance, is different than the $var_o(J)$ in equation (6.8) and it is given by

$$var_o(J) \;=\; \frac{1}{72}[N(N-1)(2N+5) - \sum_{i=1}^{T} n_i(n_i-1)(2n_i+5) - \sum_{j=1}^{g} t_j(t_j-1)(2t_j+5)]$$

$$+\frac{1}{36N(N-1)(N-2)}[\sum_{i=1}^{T} n_i(n_i-1)(n_i-2)][\sum_{j=1}^{g} t_j(t_j-1)(t_j-2)]$$
$$+\frac{1}{8N(N-1)}[\sum_{i=1}^{T} n_i(n_i-1)][\sum_{j=1}^{g} t_j(t_j-1)]. \qquad (6.10)$$

In equation (6.10), g denotes the number of tied groups in the data and t_j is the size of tied group j. Note that an untied observation is considered to be a tied group of size 1. But if there are no ties, then $g = N$ and $t_j = 1$, for $j = 1, 2, \ldots, N$. Furthermore, $\phi(a, b)$ used in the calculation of Mann-Whitney counts U_{uv} is replaced by $\phi^*(a, b) = 1, 1/2, 0$ if $a <, =, or > b$, respectively. If there are no ties, the variance expression in (6.10) reduces to the usual null variance of J as given before in equation (6.8).

6.1.1 Application

For the hemoglobin data set, we have observed a number of situations with ties among the responses. Consequently we compute J^{**} for our purpose and the value of the null test statistics is found to be $J^{**} = 6.87$. At the α level of significance, we reject H_o if $J^{**} \geq j_\alpha$; otherwise do not reject. The constant j_α is chosen to make the type I error probability equal to α. To be specific, at 5 percent level of significance, $|j_\alpha| = 1.96$. Since $J^{**} > j_\alpha$, we conclude that the null hypothesis is rejected and according to the Jonckeere-Terpstra test, there is no trend in hemoglobin data.

6.2 Kendall Distribution Free Test For Independence Based on Signs

6.2.1 Observation Based Test

In this section, we test the longitudinal pattern in a different way than in section 6.1. Following Mann (1945) and Kendall (1962), we test whether 5 longitudinal observations are correlated or not. That is, whether $y_{i1}, \ldots, y_{i5}$ are correlated or not for all $i = 1, 2, \ldots, 42$. Let ρ_i denote this correlation for given i, and we are interested to test the null hypothesis that there is no time effect, i.e.,

$$H_o : \rho_i = 0 \tag{6.11}$$

against the alternative that time as a specific factor positively influencing the responses, i.e.,

$$H_1 : \rho_i > 0 \tag{6.12}$$

To test the above hypothesis in 6.11 versus 6.12, we first write the Kendall sample correlation statistic K_i given by

$$K_i = \sum_{u=1}^{T-1} \sum_{v=u+1}^{T} Q_i[(X_{iu}, Y_{iu}), (X_{iv}, Y_{iv})], \tag{6.13}$$

where for two bivariate observations (X_{iu}, Y_{iu}) and $(X_{iv}, Y_{iv})(1 \leq u < v \leq T)$, Q_i function is defined as

$$Q_i[(a,b),(c,d)] = \begin{cases} 1, & if \quad (d-b)(c-a) > 0 \\ -1, & if \quad (d-b)(c-a) < 0 \\ 0, & if \quad (d-b)(c-a) = 0 \end{cases}$$

Next, by taking $X_{iu} = u$, for $u = 1, 2, \ldots, T$ and $i = 1, 2, \ldots, I$, we reexpress the Kendall sample correlation statistic K_i in (6.13) as

$$K_i = \sum_{u=1}^{T-1} \sum_{v=u+1}^{T} Q_i[(u, Y_{iu}), (v, Y_{iv})], \tag{6.14}$$

which was suggested by Mann (1945) to test for a time trend in the data. In our set up, this is equivalent to a test for a time trend in T longitudinal responses $Y_{i1}, \ldots, Y_{iu}, \ldots, Y_{iv}, \ldots, Y_{iT}$ for the i^{th} individual.

The K_i test statistics in equation (6.13) has Kendall distribution which may be found in any standard non-parametric text book, such as, Hollander and Wolfe, (1999), Table A.30 page 724.

6.2.2 Residual Based Test

In section 6.2.1 we have performed a non-parametric test to examine whether there is any trend in the observations collected longitudinally. Note that as apart from time, the hemoglobin responses may also be affected by treatment and other covariates, to understand any trend because of the time, we now perform a test based on the residuals rather than observations. More specifically, we compute

$$r_{it} = (y_{it} - x'_{it}\hat{\beta})/\hat{\sigma}_i. \tag{6.15}$$

and use them in place of y_{it} in the test developed in the last section.

6.2.3 Application

We calculate the Kendall test statistics $K_i, (i = 1, 2, \ldots, 42)$ from the response of the individuals as well as from the residuals and examine whether ρ_i is rejected or not.

At the α level of significance, we reject H_o if $K_i \geq k_\alpha$; otherwise do not reject, where k_α will be calculated from Table A.30 in Hollander (1999), page 724. Note that the observation based values were found to be
$K_i \equiv$ 3, 8, 3, 10, 5, 0, 8, 2, 2, -2, 6, 2, 3, -2, 5, 3, 8, 2, 5, 0, 6, 8, 10, 4, 4, 3, 8, 8, -5, 8, 4, 4, 1, 2, 1, -4, 6, 6, 8, 8, 6, 6.
Similarly, residuals based K_i's were found to be

$K_i \equiv$ 3, 8, 3, 10, 5, 0, 8, 4, 2, -2, 6, 2, 3, -2, 5, 3, 8, 4, 5, 0, 6, 8, 10, 4, 4, 3, 8, 8, -5, 8, 2, 4, 1, 2, 1, -4, 6, 8, 8, 8, 4, 6.

Further note that at 4 percent level of significance, k_α is found to be $k_\alpha = 8$ for T=5 (see Hollander, 1999, Table A.30 page 724). So by comparing the values of K_i with theoretical k_α values, we find that $K_i > k_\alpha$ holds for 11 individuals out of 42, in the observation group. Likewise, 12 out of 42 individuals had $K_i > k_\alpha$ in residual group. This test leads to the conclusion that there is no highly significant longitudinal correlations among the observations as well as residuals. This conclusion appears to be in agreement with the test performed in the previous section.

6.3 Ranks Based Spearman Distribution-Free Test For Independence

In this section, we test for the longitudinal pattern using the concepts of positive or negative association. Let $(X_{i1}, Y_{i1}), \ldots, (X_{iT}, Y_{iT})$ be a random sample from a continuous bivariate population. To compute the Spearman rank correlation coefficient, we first order the $X_{i1}, \ldots, X_{iT}$ observations from least to greatest and let R_{iu} denote the rank of $X_{iu}, u = 1, 2, \ldots, T$ for the i^{th} individual at the u^{th} time period. Likewise, we order the longitudinal observations $Y_{i1}, \ldots, Y_{iT}$ from least to greatest and let S_{iu} denote the rank of $Y_{iu}, u = 1, 2, \ldots, T$ for the i^{th} individual at u^{th} time period. The Spearman (1904) rank correlation coefficient is defined by

$$\begin{aligned} r_{is} &= \frac{12\sum_{u=1}^{T}\{[R_{iu} - \frac{T+1}{2}][S_{iu} - \frac{T+1}{2}]\}}{T(T^2-1)} \\ &= 1 - \frac{6\sum_{u=1}^{T} D_{iu}^2}{T(T^2-1)}, \end{aligned} \tag{6.16}$$

where $D_{iu} = S_{iu} - R_{iu}, u = 1, 2, \ldots, T$. Note that in our set up, $X_{iu} = u$ for $u = 1, 2, \ldots, T$. Thus $R_{iu} = u$. Now to test whether there is any dependence of the responses on the time $(u = 1, 2, \ldots, T)$, we simply test whether the population correlation between $u(u = 1, 2, \ldots, T)$ and Y_{iu} for a given i is significant or not. Consequently, by putting $R_{iu} = u$ in (6.16) and computing the rank S_{iu} as mentioned above, we find the value of r_{is} in (6.16). The null hypothesis may be written as

$$H_o : \rho_{is} = 0 \tag{6.17}$$

against the alternative,

$$H_o : \rho_{is} > 0, \tag{6.18}$$

where ρ_{is} is the population counterpart of r_{is}.

For the observed data as well as for the residuals, we now compute r_{is} by (6.16). We compare this value with that of its tabulated value (see Hollander, 1999, Table A.31, page 732) and reject the null hypothesis if $|r_{is}| \geq r_{s\alpha}$, $r_{s\alpha}$ being the tabulated value at α level of significance.

Note that for the observed data, the values of $|r_{is}|, i = 1, 2, \ldots, 42$ exceed $r_{s\alpha}$, 11 times. Similarly, for the residuals based test, the values of $|r_{is}|$ exceed $r_{s\alpha}$ in 12 cases out of 42. These results show that the dependence of hemoglobin on the time is not really significant. Thus the conclusions in all three sections remain the same that there is actually no longitudinal monotonic trend in the data. This however does not mean that there were no changes in hemoglobin levels over the time. This is because, as is apparent from Figure B.6, compared to the baseline level, the hemoglobin levels at different times were either higher or lower, indicating clear changes, although there was no specific monotonic trend.

Chapter 7

Concluding Remarks

In the practicum, we have analyzed a hemoglobin data set which is longitudinal by nature. Also the data had missing responses at times. The statistical analysis of such longitudinal data subject to non-response requires careful solution of the methodologies. Following the suggestion of Sutradhar and Das (1999), we have used a general auto-correlation structure in our linear model set up and computed the regression effects efficiently. To compute the covariate effects in the presence of missing values, we have followed Krishnamoorthy and Pannala (1999) as well as the imputation technique used by Paik (1997). We have further studied certain tests for examining possible longitudinal changes in hemoglobin levels. This, we have done using non-parametric tests.

The results of the regression analysis for the complete data were computed based on 25 complete longitudinal observations for 5 time points. For incomplete data, we have used 42 observations under two situations: first, the results were computed from available responses, and second, they were computed based on suitable imputations. In all three cases, it was clear that the predicted hemoglobin level of males was higher as compared to that of females. As the treatment effect was positive, it was clear that the treatment was effective to increase the hemoglobin levels for the infants treated as compared to the placebo group. The baseline hemoglobin levels were higher for the infants with larger gestation week. It however became clear that the hemoglobin level for the infants with lower gestational age eventually increased more compared to the infants with larger gestational age. Finally, the non parametric tests showed that there was no longitudinal pattern (monotonic increasing or decreasing) in the data, although there were changes in hemoglobin levels over the months.

In conclusion, this statistical study should be useful for the scientists to prescribe better recommendation than those are available in the current literature, regarding the iron-intake by the low-birth-weight infants.

Bibliography

1. Committe on Nutrition, American Academy of Pediatrics. (1985). "Nutritional needs of low-birth-weight infants", *Pediatrics*, 75:976-986.

2. Ehrenkranz, R.A. "Iron, Folic Acid, and Vitamin B_{12}", chapter 12, page no. 177 in *Nutritional Needs of the Preterm Infant*, Williams and Wilkins, c1993.

3. Friel, J.K., Andrews, W.L., Matthew, J.D., Long, D.R., Cornel, A.M., Cox, M., and Skinner, C.T. (1990). "Iron status of very-low-birth-weight infants during the first 15 months of infancy", *CAN MED ASSOC J*, 143(8).

4. Gortem, K.M. and Cross, E.R. (1964). "Iron metabolism in premature infants: 2. Prevention of iron deficiency", *Pediatrics*, 64: 509-520.

5. Hollander, M. and Wolfe, D.A. (1999). *Non-parametric Statistical Methods*, John Wiley, New York.

6. Kendall, M.G. (1962). *Rank Correlation Methods*, third edition, London: Griffin.

7. Kirshnamoorthy, K. and Pannala, M.K. (1999). "Confidence Estimation of a normal mean vector with incomplete data", *The Canadian Journal of Statistics*, vol 27, No. 2, 1999, pages 395-407.

8. Mann, H.B. (1945). "Non-parametric tests against trend", *Econometrica.* 13, 245-259.

9. Nutrition Committe, Canadian Pediatric Society. (1981). "Feeding the low birth weight infant", *Can Med Assoc J*, 124: 1301-1310.

10. Paik, M.C. (1997). "The Generalized Estimating Equation Approach When Data Are Not Missing Completely at Random", *Journal of the American Statistical Association*, vol. 92, No. 440, Application and Case Studies.

11. Spearman, C. (1904). "The proof and measurement of association between two things", *Journal of Psychology.* 15:72-101.

12. Sutradhar, B.C. and Das, K. (1999). On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika*, 86, 459-465.

13. Worwood, M. (1997). "The clinical biochemistry of iron", *Semin Hematol*, 14: 3-30.

Appendix A

Hemoglobin Data

```
---------------------------------------------------------------------
 Sn   T1    T2    T3     T4    T5    BHGB  Gender  Formula  Gestation
=====================================================================
  1  122   135   129    135   134     86      0        1        26
  2  100   117   136    124   138     91      1        0        28
  3   87   126   135    128   125     76      0        0        31
  4   80   122   132    139   129     86      1        1        30
  5   86   125   129    127   139     95      0        1        28
  6  139   126   143    137   133    146      1        1        33
  7  115   128   134    134   132    135      1        1        33
  8  127   133   127    124   126    148      1        0        34
  9  100   125   120    124   122    114      0        1        34
 10  138   122   127    128   126    157      0        1        34
 11   92   132   132    135   131    114      1        1        32
 12  106   114   118    122   114    115      0        1        31
 13   96   122   128    118   131    109      0        0        34
 14  110   125   112    136   118    120      1        0        31
 15  124   124   118    121   128    175      0        0        32
 16   99   133   132    134   131    122      1        1        28
 17   87   119   122    116   114     93      0        1        31
 18  119   122   129    128   132    154      1        0        34
 19   95   117   114    131   135    148      1        1        31
 20   86   119   117    115   117     93      0        0        28
 21   99   110   116    125   123    140      0        1        32
```

22	160	107	126	118	115	136	0	0	33
23	103	116	105	123	118	133	0	0	33
24	107	123	142	129	123	97	1	1	27
25	92	113	125	128	140	103	1	0	29
26	110.2*	109	119	117	115	118	1	1	32
27	116.2*	114	128	126	139	101	0	0	31
28	116.2*	112	118	132	139	87	0	0	31
29	102.6*	145	114	130	130	85	0	1	30
30	107.5*	118	129	126	123	99	1	0	33
31	123.0*	111	108	119	127	144	1	0	36
32	112.0*	112	117	112	118	91	1	1	34
33	112.0*	121	128	120	114	91	1	1	34
34	96.0*	108	107	117	110	90	0	0	35
35	123.0*	104	130	136	129	82	1	0	38
36	112	131	120.6*	131	130	101	1	0	31
37	104	119	120.6*	129	123	143	1	0	32
38	107	129	121.6*	130	132	91	0	0	32
39	112	113	131	135	116.0*	103	1	1	35
40	104	110	121	128	123.5*	119	1	0	31
41	97.3*	114.3*	124	123	112	96	0	1	32
42	100	123	124.0*	130.6*	134.3*	88	0	1	30

Table A.1: Hemoglobin Data from Janeway Child Health Center and Grace General Hospital for 42 Children for the Period of 3 Months (June 1995-May 1996) After Birth, With Imputed Values Shown With a '*' Mark.

Appendix B

Graphs

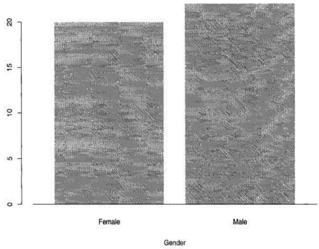

Figure B.1: Histogram of Distribution of Gender as an Explanatory Variable

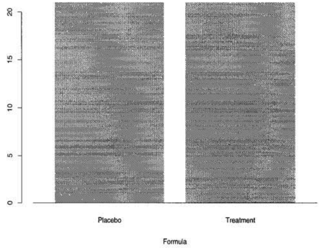

Figure B.2: Histogram of Distribution of Treatment as an Explanatory Variable

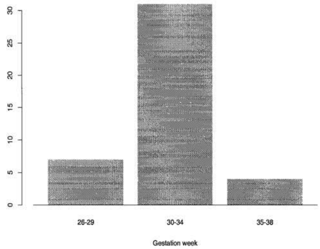

Figure B.3: Histogram of Distribution of Gestation Week as an Explanatory Variable

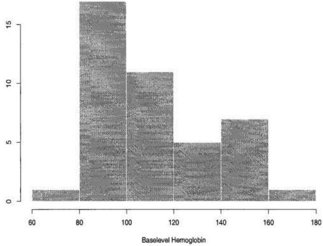

Figure B.4: Histogram of Distribution of Baselevel Hemoglobin as an Explanatory Variable

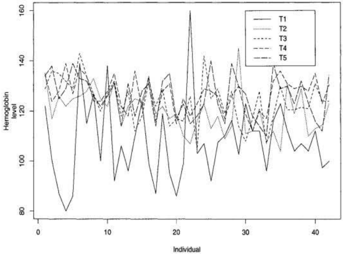

Figure B.5: Longitudinal Plot of Hemoglobin Levels for 42 Individuals at 5 Different Times

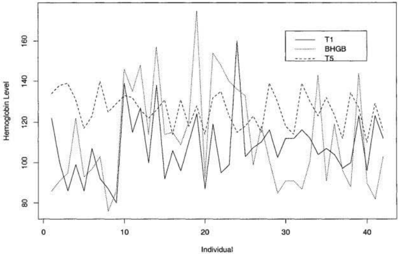

Figure B.6: Plot of Hemoglobin Values For 42 Individuals for Time T1, T5 and Baselevel Hemoglobin

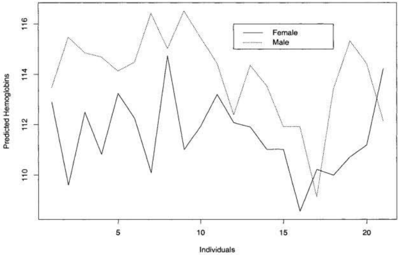

Figure B.7: Plot of Predicted Hemoglobin Values for Males and Females

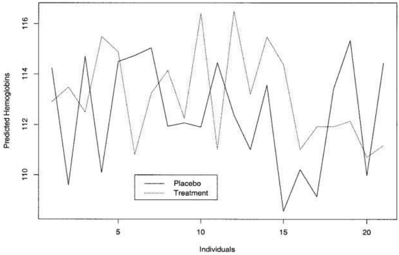

Figure B.8: Plot of Predicted Hemoglobin Values for Treatment Group and Placebo Group

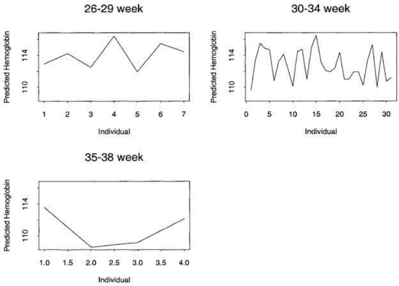

Figure B.9: Plot of Predicted Hemoglobin Values for Different Gestation Week

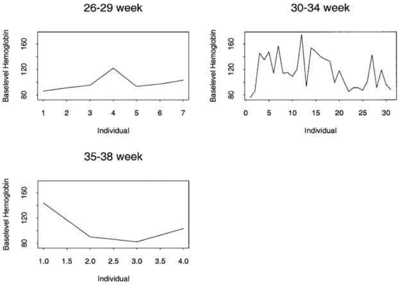

Figure B.10: Plot of Baselevel Hemoglobin Values for Different Gestation Week