UNIVARIATE POLYTOMOUS ORDINAL REGRESSION ANALYSIS WITH APPLICATION TO DIABETIC RETINOPATHY DATA

DENNIS WILLIAM BATTEN

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UNIVARIATE POLYTOMOUS ORDINAL REGRESSION ANALYSIS WITH APPLICATION TO DIABETIC RETINOPATHY DATA

by

Dennis William Batten

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

January, 2000

St. John's, Newfoundland

Abstract

There are many situations in practice where one is interested to find the regression effects of the covariates on polytomous responses. Furthermore, there are situations where polytomous responses are ordinal by nature. These types of data are commonly analyzed by exploiting the well-known probit and cumulative logit models. These methods, however, require the introduction of certain cut-points to distinguish ordered categories of the polytomous responses, and these cut-points are required to be estimated consistently, which may not be easily obtained. In the practicum, we use a recently developed non-cut-point based cumulative logit model to resolve this estimation problem. The regression analysis chosen in the practicum was motivated by a need for a refined analysis of a diabetes data set used in the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR). The practicum discusses the advantages and disadvantages of the existing as well as the new techniques. The non-cut-point based approach was found to give the best fit to the diabetes data, with easy interpretation of the regression estimates.

Acknowledgments

I wish to thank my supervisor, Dr. B. C. Sutradhar for his encouragement, helpful comments and suggestions, and most of all for the endless amount of time he made available for me.

I acknowledge the Department of Mathematics and Statistics for the financial support in the form of Teaching Assistantships and Sessional Lecturer.

I am also grateful to my parents for their encouragement and support throughout my university career.

Finally, I am grateful of the support from my twin brother Douglas who has been with me every step of the way.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation of the Problem

Analyzing multinomial ordinal data is important in practice. One of the main scientific interests in such problems is to find the effect of the covariates on ordered responses. Consider, for example, a medical problem with regard to diabetes, where the responses such as severity of diabetic retinopathy may be explained as a function of associated covariates. There exist some studies where this type of data are analyzed to understand the effects of the treatments and other covariates on the severity of diabetic retinopathy. We refer to the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) as one such study. This study contains complete records of 720 younger onset Type 1 diabetics. The data contains information pertaining to numerous covariates such as duration of diabetes and glycosylated hemoglobin level. The objective is to investigate the effect of such covariates on the ordered responses labeled as 'none', 'mild', 'moderate' and 'proliferative' categories

which are indicators for the severity of diabetic retinopathy. As these ordered responses are in order from best to worst, methods to find the effect of covariates after distinguishing these categories are required.

There exists several statistical approaches to analyze the multinomial ordinal types of data described above. For example, we refer to McCullagh (1980), Stram et al. (1988), Walker and Duncan (1967), Williams and Grizzle (1972), and Williamson et al. (1995). These authors discuss cumulative logit models to analyze the multinomial ordinal data. In such models, cumulative probabilities are defined as a function of the covariates and certain cut-points, where the cut-points distinguish the adjacent categorical responses. More specifically, the cumulative logits are defined so that the cut-points follow an (increasing) order restriction. For a comprehensive discussion on the basic development of this type of model, we refer to Agresti (1990, section 9.4) among others.

A second approach based on the probit model, available to analyze multinomial ordinal data, has been presented by Aitchison and Silvey (1957), Ashford and Sowden (1970), Gurland, Lee and Dahm (1960), Harville and Mee (1984) and Kim (1995). As the probabilities are based on the covariates and cut-points, this approach is similar to the logit model. The main difference between the probit and logit model is that the probit model uses the standard normal cumulative distribution function as the cumulative link function, whereas the logit model uses the binary logistic function as its cumulative link function.

The above mentioned commonly used methods (cumulative logit and probit approaches) to analyze multinomial ordinal data, may however run into

difficulties in estimating the cut-point parameters associated with such procedures. More specifically, as the cut-points must follow an order restriction, the estimation procedures such as the Newton Raphson iteration technique available to estimate the values of these cut-points does not guarantee that this order restriction will be maintained. Recently, in a multivariate set up, Sutradhar and Kovacevic (2000) [see also Das and Sutradhar (1999)] have proposed an alternative approach which, unlike the probit and logit models, avoids the use of the cut-points in modeling the multinomial ordinal regression data. More specifically, they use a non-cut-point based cumulative probability model which distinguishes the ordered categories in a natural way. Note, however, that in the present practicum, we deal with the univariate ordinal polytomous model which is a special case of the multivariate model introduced by Sutradhar and Kovacevic (2000).

1.2 Objective of the Practicum

As mentioned above, in the present practicum, we deal with a univariate case and simplify the multivariate procedure of Sutradhar and Kovacevic (2000) to analyze univariate ordinal regression data. One of the strong motivations to deal with such univariate case came from the fact that an exploratory analysis of WESDR data did not appear to show any difference between the left and right eye's retinopathy level. Moreover, we wish to examine the performance of all three methods, including the new non-cut-point based procedure, to analyze a univariate problem.

The specific plan of the practicum is as follows:

1. In chapter 2, we provide an exploratory analysis of the covariates and response variables of the WESDR data, which will be helpful in developing the appropriate ordinal regression model for further statistical analysis.

2. In chapter 3, the probit model will be exploited to analyze the relationship between the ordinal response and associated covariates. Throughout the chapter a detailed description of the model will be presented along with the likelihood estimation procedure used to obtain the estimates for the parameters involved in the model. The goodness of fit of the model to the data will also be investigated.

3. In chapter 4, we will review the cumulative logit model as an alternative model to the probit model, to analyze the ordered categorical data. To be specific, we first show how this model can be developed and then we discuss the likelihood estimation for the parameters of the model. Similar to chapter 3, the goodness of fit of the model to the data will also be provided.

4. In chapter 5, we argue for a newly suggested non-cut-point based multinomial logistic approach to explore the relationship between the ordinal response variable and associated covariates. In this chapter, unlike chapters 3 and 4, we exploit the generalized estimating equation approach for the estimation of the parameters of the model. Further we investigate the goodness of fit of the model to the data, and provide comparison of all three approaches with regard to fitting the models to the data.

5. The conclusion of the practicum is provided in chapter 6. We also provide some remarks on future research in this area.

Chapter 2

Background of the Problem

2.1 Severity of Diabetic Retinopathy

Diabetes is one of the most serious diseases which causes extreme suffering and is a leading cause of death by disease. Currently 1.5 million Canadians have been diagnosed with diabetes and an additional .75 million are suspected to have the disease, but are unaware of it. Diabetes is a disorder which does not allow the body to utilize sugar obtained from the foods we eat. Diabetics are categorized into one of two types, Type 1 and Type 2 diabetes. A person is identified as having Type 1 or insulin-dependent diabetes when the Pancreas either stops producing or produces very little insulin. A person typically develops Type 1 diabetes before the age of 30. Type 2 diabetes is marked as being an older onset diabetes, usually after the age of 30, and is non-insulin-dependent. Type 2 diabetes is identified when either the Pancreas does not produce enough insulin or when the insulin produced by the Pancreas is not being used by the body.

Although the symptoms for Type 1 and Type 2 diabetes are very similar, they develop much faster and are more disastrous for the Type 1 diabetic. These symptoms include frequent urination, unusual thirst, extreme hunger, unusual weight loss, extreme fatigue, irritability, nausea, vomiting, blurred vision, and others. Costing an estimated 5-6 billion dollars to the Canadian health care system, people still fail to recognize diabetes as a serious disease. In fact, diabetes is known to significantly increase the risk of heart disease, kidney disease, non-traumatic amputation, impotence, and is also the leading cause of adult blindness.

The increased blood sugar levels caused by diabetes is known to damage both small and large blood vessels in the body. Damaged blood vessels within the eye will cause impaired or loss of vision referred to as Retinopathy. When diabetes is the cause of Retinopathy it is referred to as Diabetic Retinopathy and is mainly present with Type 1 Diabetes. It is known that among 86 percent of people diagnosed with early onset diabetes who have went blind, the only contributing factor of their blindness was retinopathy. **(source: Canadian Diabetes Association, http://www.diabetes.ca)**

Although treatments such as insulin injections, and proper diet and exercise plans have been developed to create more comfortable living for diabetics, no cure exists. Early detection of diabetes is important to halt or prevent some of the complications that arise from diabetes such as Retinopathy. In order to find a better clinical remedy more understanding about how other factors such as age, sex, and duration of diabetes contribute to the disease is necessary. For this purpose, many clinical organizations, in particular in USA and Canada, are continually engaged in biomedical research concerning

this disease.

One such study has been recently done by Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR). This data set was analyzed by Williamson et al (1995), among others, to understand the effects of associated covariates on the severity of Retinopathy on both the left and right eyes. Note that as the socioeconomic conditions are similar for both USA and Canada, the results obtained from a USA study should be useful for Canadian diabetes researchers as well. Turning back to the WESDR, Williamson et al (1995) have used a cut-point based polytomous approach and exploited suitable estimating equations to find the covariate effects as mentioned above. We however, will take a simpler statistical approach to analyze such a data set. But, before we go for details, we now explain the variables involved in the study and examine the nature of these variables through an exploratory analysis.

This data set contains records of 996 younger onset Type 1 diabetics, of which complete records are present for 720 of these persons. A 10-point ordinal scale increasing from none to worst was used to grade the severity of Diabetic Retinopathy for both left and right eyes. Altogether four ordered categories were considered [cf Williamson et al (1995)] and they are: none, mild, moderate, and proliferative. In total, information was collected on 17 covariates: 1. right eye macular edema; 2. left eye macular edema; 3. right eye refractive error; 4. left eye refractive error; 5. right eye intra-ocular pressure; 6. left eye intra-ocular pressure; 7. age at diagnosis of diabetes; 8. duration of diabetes; 9. glycosylated hemoglobin level; 10. systolic blood pressure; 11. diastolic blood pressure; 12. body mass index; 13. pulse rate; 14. sex; 15. proteinuria; 16. doses of insulin per day; 17. type of county of

residence.

2.2 Exploratory Analysis

In this section we provide an exploratory analysis for the diabetes data set analyzed by Williamson et al (1995), which will depict how the response variable as well as the covariates are behaving. Although, information on 17 covariates was collected in the original data set, we will only consider the so called marginal covariates and they are: 1. duration of diabetes; 2. glycosylated hemoglobin level; 3. diastolic blood pressure; 4. proteinuria; 5. sex; 6. right or left eye macular edema. Note that these six marginal covariates were also chosen by Williamson et al (1995). These authors, however, had one more association covariate (doses of insulin per day) in their analysis, which we do not include in our study as we are examining properties of the marginal variables in this section.

In the following subsections, we exhibit various exploratory graphs for the response as well as covariables and discuss the patterns to understand the effect of the covariates on the responses.

2.2.1 Response Variable

The histograms for the right and left eye retinopathy level for the 720 persons with complete records is shown in Figure A.1. The four ordered categories: none, mild, moderate, and proliferative represented by 0-1, 1-2, 2-3, and 3-4 respectively are given along the horizontal axis. The number of subjects in each category is indicated on the vertical axis. It appears that there are same

number of individuals in the none and mild categories under both left and right eyes. Under the moderate and proliferative categories there are fewer individuals as compared to the other two categories. Between the moderate and proliferative groups, the moderate group appears to contain almost twice the observations as compared to the proliferative group. Note that it is not only that the none and mild groups have the same number of individuals under both eyes, the overall distribution of the individuals appear to be the same under both left and right eyes. For the left eye there are 268 individuals in the none category, 277 individuals in the mild category, 127 individuals in the moderate category, and 48 in the proliferative category. The right eye contains 275 individuals in the none category, 270 individuals in the mild category, 128 individuals in the moderate category, and 47 in the proliferative category.

2.2.2 Explanatory Variables

Figures A.2, A.3, and A.4 show the distributions of the duration of diabetes. In all three Figures the duration of diabetes (in years) is given on the horizontal axis and is divided into six equal groups representing 10 years for each group. The number of observations in each of the six intervals is given on the vertical axis. Figure A.2 exhibits the distribution of the duration of diabetes under the assumption that other covariates are held fixed at suitable levels. It appears that a large number of subjects suffer from diabetes even after a period of ten years. In Figure A.3 and A.4 we look at the distributions of the covariates separately for the left and right eye respectively, and record the number of individuals under all four ordered categories: none,

mild, moderate, and proliferative. We see from Figure A.3 that in each of the four ordered categories the histograms strongly reflect the overall picture displayed in Figure A.2. A comparison between the left eye (Fig A.3) and the right eye (Fig A.4) clearly shows a remarkable resemblance between the histograms for the duration of diabetes for each category under both left and right eyes.

The second covariate, Glycosylated Hemoglobin Level, is graphically depicted in Figures A.5, A.6, and A.7. Glycosylated Hemoglobin Level is given on the horizontal axis and the number of observations is given on the vertical axis. From Figure A.5 it appears that the distribution is symmetric with nearly 500 of the 720 subjects having a Glycosylated Hemoglobin Level in the middle range. Figures A.6 and A.7 presents a closer look at the behavior of Glycosylated Hemoglobin Level under each of the four ordered categories for both left and right eyes. These pictures again show strong similarities to the general picture displayed for Glycosylated Hemoglobin Level given in Figure A.5. We again note the close resemblance of the left eye (Fig A.3) and the right eye (Fig A.4) with regard to the distribution of Glycosylated Hemoglobin Level under each of the four ordered categories.

The next covariate we explore is Diastolic Blood Pressure. This covariate is displayed in Figures A.8, A.9, A.10, with Diastolic Blood Pressure given along the horizontal axis the number of observations on the vertical axis. The histogram, displayed in Figure A.8, appears to have to normal curve shape centering around 75. The distribution of Diastolic Blood Pressure for each of the four ordered categories for both left eye and right eyes exhibited in Figures A.9 and A.10 also follow the normal shape. Note that the left eye

(Fig A.9) and the right eye (Fig A.10) produce almost identical histograms with regard to the distribution of Diastolic Blood Pressure, under each of the four ordered categories.

The distribution of the fourth covariate, Proteinuria, is displayed in Figure A.11, A.12, and A.13. Proteinuria is a dichotomous variable, that is, it is either absent or present in the individuals, which is shown by two bars for the respective groups given along the horizontal axis. The number of observations is indicated on the vertical axis. It appears from the histogram displayed in Figure A.11, the histogram for the overall distribution of Proteinuria, that Proteinuria is absent in approximately 85 % of the individuals. Figures A.12 and A.13 provide information about the presence of Proteinuria within each of the four ordered categories for both left and right eyes. The histograms for the distribution of Proteinuria under each of the four ordered categories strongly reflect the histogram for the overall distribution of Proteinuria presented in Figure A.11, for both left and right eyes. Again, we note the strong resemblance between the left eye (Fig A.12) and the right eye (Fig A.13), with regard to the distribution of Proteinuria, under each of the four ordered categories.

Figures A.14, A.15, and A.16 are graphical representations concerning the fifth covariate, gender. Gender is indicated on the horizontal axis, and the corresponding number of observations is indicated on the vertical axis. Figure A.14 which contains all 720 individuals gives clear indication that there is nearly the same number of female subjects as male subjects. Further, this representation of equal number of males and females is reflected within each of the four ordered categories: none, mild, moderate, and proliferative.

The above observation suggests that retinopathy level does not appear to be gender sensitive. Note that the histograms appear virtually unchanged from the left eye (Fig A.15) to the right eye (Fig A.16), with regards to gender within each of the four ordered categories.

Finally, we consider the covariate, Macular Edema displayed in Figures A.17, A.18, A.19, and A.20. This covariate represents a measurement taken directly from the eyes and therefore two separate measurements are taken from each individual one for the left eye (Left Eye Macular Edema) and the other for the right eye (Right Eye Macular Edema). The histograms for the left eye is depicted in Figures A.17 and A.18, while the graphs for the right eye are shown in Figures A.19 and A.20. For all four Figures, presence of Macular Edema is indicated on the horizontal axis while the number of observations under each of the two groups is shown on the vertical axis. The overall histograms for Left Eye Macular Edema (Fig A.17) and Right Eye Macular Edema (Fig A.19) suggest that Macular Edema is present in 10 % of individuals. The break down of the number of individuals under each of the four ordered categories is shown in Figure A.18 that corresponds to Figure A.17 for the Left Eye Macular Edema. Similarly, the break down of the number of individuals under each of the four ordered categories is shown in Figure A.20 which corresponds to Figure A.19 for the Right Eye Macular Edema. A comparison of the histograms between the left eye (Fig A.18) and right eye (Fig A.20) again reveals the similarities between the left eye and right eye, within each of the four ordered categories.

From the above discussion it is clear that the distribution of individuals with regard to both response and covariables under each of the four ordered

categories appear to be almost identical for the left and right eyes. Consequently we have decided to study, in details, about the effects of the selected covariates on one response variable only, namely the Right Eye Retinopathy Level.

Note however that the regression analysis for a univariate ordered categorical variable (such as Right Eye Retinopathy Level) is not adequately discussed in the literature. Some of the existing methods are:

1. Latent Gaussian process based categorical approach.

2. Linear cumulative Logits approach.

Further note that the above two approaches distinguish the adjacent categories with the introduction of an appropriate cut-point, which may not be easy to estimate consistently. In the next two chapters we review these two approaches in detail and apply these methods to the diabetes data set discussed above. In chapter 5, we provide a new approach to this problem and apply a newly suggested multinomial logistic approach, which, unlike the approaches discussed in chapter 3 and 4, does not require the introduction of any cut points directly. But, unlike the existing approaches, the cumulative logits are nonlinear which is relatively slightly more difficult to interpret as compared to the linear cumulative logit. When this interpretation problem is weighted against the non-cut-point based advantage, the new approach appears to be superior to the existing approaches.

Chapter 3

Latent Process Based Probit Analysis

In this approach, we assign the ith ($i = 1, \ldots, N$) individual into one of several categories, where its category is determined based on its own interval on the real axis of an unobservable latent variable. To be more specific, let Y represent an underlying continuous latent response variable and Y_i is the value of Y for the ith individual. Also suppose that although Y_i is unobservable, an interval that contains Y_i is known. Assign the numbers $1, \ldots, M$, respectively, to the M ordered categories. As Y_i itself is not observable, the ith individual is observed to belong to a category number h, say $Z_i = h$, through the interval relationship of Y_i given by

$$\alpha_{h-1} < Y_i < \alpha_h \Leftrightarrow Z_i = h \tag{3.1}$$

where $h \in \{1, \ldots, M\}$. In (3.1) $\alpha_0 = -\infty, \alpha_M = +\infty$ and $\alpha_1, \ldots, \alpha_{M-1}$ are unknown boundary points that define a partitioning of the real line into M intervals. Thus, when the realized value of Y_i belongs to the hth interval, we say that $Z_i = h$.

Under our assumptions, the probability-mass function of $Z_1, \ldots, Z_N$ is

$$\begin{aligned} P(z_1, \ldots, z_N) &= pr\{Z_i = z_i (i = 1, \ldots, N)\} \\ &= pr\{\alpha_{z_i - 1} < Y_i < \alpha_{z_i} (i = 1, \ldots, N)\} \end{aligned} \tag{3.2}$$

where $Y_i \sim N(x_i^T \beta, \sigma^2)$. Here, $x_i = (x_{i1}, \ldots, x_{iu}, \ldots, x_{ip})^T$ is the $p \times 1$ covariate vector for the ith individual, β is the $p \times 1$ regression vector and σ^2 is the variance of Y_i discussed. Note that for the diabetes data discussed in chapter 2, $p = 6$, and i varies from 1 to $N = 720$. Finally these N individuals are independent. It then follows from (3.2) that

$$L(\beta, \sigma^2) = \prod_{h=1}^{M} \prod_{i_h=1}^{N_h} \int_{\alpha_{h-1}}^{\alpha_h} \phi\left(\frac{y_{i_h} - x_{i_h}^T \beta}{\sigma}\right) dy_{i_h} \tag{3.3}$$

where N_h $(h = 1, \ldots, M)$ identifies the number of individuals in the hth category. Here i_h indicates the ith individual belongs to the hth category. Then y_i and x_i for the ith individual are re-expressed as y_{i_h} and x_{i_h} provided the ith individual belongs to the hth category. In (3.3), $\phi(\cdot)$ denotes the probability density function (pdf) of the standard normal variable $\frac{y_{i_h} - x_{i_h}^T}{\sigma}$.

3.1 Estimation of the Parameters

As the likelihood function is available, the parameters may be estimated by maximizing the likelihood function itself. Note however that the likelihood estimation for all the parameters β, σ and $\alpha_1, \ldots, \alpha_{M-1}$ is quite involved as we have to solve $M+p$ likelihood estimating equations which is usually done by applying the Newton Raphson iteration technique. Further there is no guarantee that such likelihood solutions will ensure the restriction

$$\hat{\alpha}_1 < \hat{\alpha}_2 < \ldots < \hat{\alpha}_{M-1} \tag{3.4}$$

where $\hat{\alpha}_h$ denotes the likelihood estimate of α_h. Nevertheless, in this section, we attempt to obtain the maximum likelihood estimates of all the parameters and examine whether the restriction (3.4) is satisfied or not for the α parameter. Further, this approach will be compared later on with the other two methods to be discussed in the next two chapters. All these will be done in connection with the analysis of the Wisconsin diabetic retinopathy data set. Turning back to the likelihood estimation method, we first write the log-likelihood function of (3.3) as

$$l = \log L = \sum_{h=1}^{M} \sum_{i_h=1}^{N_h} \log \left[\Phi\left(\frac{\alpha_h - x_{i_h}^T \beta}{\sigma} \right) - \Phi\left(\frac{\alpha_{h-1} - x_{i_h}^T \beta}{\sigma} \right) \right] \tag{3.5}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (cdf) of the standard normal variable $\frac{y_{i_h} - x_{i_h}^T}{\sigma}$. This log-likelihood function (l) will be maximized to obtain the estimates of α_h ($h = 1, \ldots, M-1$), β as well as σ^2. The

goodness of fit of the method to the data will also be explored.

Note that for convenience of writing the first and second order partial derivatives of the log-likelihood function, we define two functions as follows.

Let

$$W_{i_h}(h, h-1) = \frac{\phi\left(\alpha_h - x_{i_h}^T\beta\right)}{\Phi\left(\alpha_h - x_{i_h}^T\beta\right) - \Phi\left(\alpha_{h-1} - x_{i_h}^T\beta\right)} \tag{3.6}$$

and

$$W_{i_{h+1}}(h, h+1) = \frac{\phi\left(\alpha_h - x_{i_{h+1}}^T\beta\right)}{\Phi\left(\alpha_{h+1} - x_{i_{h+1}}^T\beta\right) - \Phi\left(\alpha_h - x_{i_{h+1}}^T\beta\right)}. \tag{3.7}$$

be these two functions, where $\Phi(\alpha_0 - x_{i_{h+1}}^T) = 0$ and $\Phi(\alpha_M - x_{i_{h+1}}^T) = 1$, as $\alpha_0 = -\infty$ and $\alpha_M = +\infty$.

3.1.1 Step 1: Estimation of the Regression Parameters

For suitable initial values of α_h and σ^2, the likelihood estimate of β is obtained by solving the iterative equation

$$\hat{\beta}(r+1) = \hat{\beta}(r) - \left[\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right)^{-1}\right]_r \left[\frac{\partial l}{\partial\beta}\right]_r \tag{3.8}$$

where,

$$\frac{\partial l}{\partial\beta} = -\frac{1}{\sigma}\sum_{h=1}^{M}\sum_{i_h=1}^{N_h}\left[Wi_h(h, h-1) - W_{i_{h+1}}(h, h+1)\right] X_{i_h}$$

and

$$\frac{\partial^2 l}{\partial\beta\partial\beta^T} = \frac{1}{\sigma^2}\sum_{h=1}^{M}\sum_{i_h=1}^{N_h}\Big[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) - W_{i_{h+1}}(h,h+1)$$
$$(\alpha_{h-1} - x_{i_h}^T\beta) - \{Wi_h(h,h-1) - W_{i_{h+1}}(h,h+1)\}^2\Big]X_{i_h}X_{i_h}^T$$

with $\hat{\beta}(r)$ as the estimated value of β on the rth trial and the expression $[\cdot]_r$ is evaluated at $\hat{\beta}(r)$. In (3.8), $\frac{\partial l}{\partial\beta}$ is the $p \times 1$ vector of first order partial derivatives and $\frac{\partial^2 l}{\partial\beta\partial\beta^T}$ is the $p \times p$ second order partial derivative matrix.

Note that the estimation of β could be done jointly along with the estimation of the other parameters, namely $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{M-1})^T$ and σ^2, by solving $p + M$ estimating equations. But, this is quite involved algebraically as well as numerically. As a remedy, we have chosen to estimate β, $\alpha_h (h = 1, \ldots, M-1)$ and σ^2 following a three step procedure. Once, for given values of α_h and σ^2, the convergent values of β's are obtained by solving (3.8), they will be used in section 3.1.2 to obtain improved estimates for $\alpha_h (h = 1, \ldots, M-1)$. This will be referred to as the second step, first step being the estimation of β. Next the estimates of β and $\alpha_h (h = 1, \ldots, M-1)$ from steps 1 and 2 are used to obtain an improved value of σ^2 in section 3.1.3. Once improved estimates of $\alpha_h (h = 1, \ldots, M-1)$ and σ^2 are obtained in steps 2 and 3, they are used in step 1 to improve the β estimate further. This cycle of iterations continues until convergence is obtained for all parameters β, $\alpha_h (h = 1, \ldots, M-1)$ and σ^2.

3.1.2 Step 2: Estimation of the Cut-Points

Note that for an initial value of σ^2, and the β estimate from step 1, the likelihood estimates of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{M-1})^T$ is obtained by solving the iterative equation

$$\hat{\alpha}(r+1) = \hat{\alpha}(r) - \left[\left(\frac{\partial^2 l}{\partial\alpha\partial\alpha^T}\right)^{-1}\right]_r \left[\frac{\partial l}{\partial\alpha}\right]_r \tag{3.9}$$

where $\hat{\alpha}(r)$ is the estimated value of α on the rth iteration and the expression $[\cdot]_r$ is evaluated at $\hat{\alpha}(r)$, $\dfrac{\partial l}{\partial\alpha}$ is the $M-1\times 1$ vector of first derivatives of the likelihood function with respect to α and $\dfrac{\partial^2 l}{\partial\alpha\partial\alpha^T}$ is the $M-1\times M-1$ second derivative matrix. Further, in equation (3.9),

$$\frac{\partial l}{\partial\alpha_h} = \sum_{i_h=1}^{N_h} Wi_h(h, h-1) - \sum_{i_{h+1}=1}^{N_{h+1}} W_{i_{h+1}}(h, h+1),$$

and

$$\begin{aligned}\frac{\partial^2 l}{\partial\alpha_h\partial\alpha_{h'}} &= \sum_{i_h=1}^{N_h}\left[\frac{1}{\sigma^2}Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) - \frac{1}{\sigma}W_{i_h}^2(h,h-1)\right] \\ &- \sum_{i_{h+1}=1}^{N_{h+1}}\left[\frac{1}{\sigma^2}W_{i_{h+1}}(h,h+1)(\alpha_h - x_{i_{h+1}}\beta) - \frac{1}{\sigma}W_{i_{h+1}}^2(h,h+1)\right],\end{aligned}$$

for all $h = 1, \ldots, M-1$, where $W(\cdot)$ are defined in (3.6) and (3.7). Note that it is clear from the above second derivatives that for $|h - h'| = 1$, it reduces

to

$$\frac{\partial^2 l}{\partial \alpha_h \partial \alpha_{h'}} = \sum_{i_{h+1}=1}^{N_{h+1}} \frac{1}{\sigma} W i_h(h, h-1) W_{i_{h+1}}(h, h+1),$$

otherwise,

$$\frac{\partial^2 l}{\partial \alpha_h \partial \alpha_{h'}} = 0.$$

As the $\alpha_h (h = 1, \ldots, M-1)$ values are the cut-points distinguishing the adjacent categories, and because no order restrictions are imposed on the α_h's in the traditional likelihood approaches, such as in the estimating equation (3.9), there is no guarantee that estimated values of $\alpha_h (h = 1, \ldots, M-1)$ will maintain the restriction $\alpha_1 < \alpha_2 < \ldots < \alpha_{M-1}$. Further, there may be multiple solutions or roots for these α_h parameters because of the possibility of local maxima for such a high dimensional likelihood surface that we are considering here. Nevertheless, for simplicity we use the likelihood estimating equation (3.9) to obtain the cut-point estimates for the diabetes data discussed in chapter 2. These estimates will be compared with the estimates obtained by a similar procedure in chapter 4, and with estimates from a new non-cut-points based procedure to be discussed in chapter 5.

3.1.3 Step 3: Variance Component Estimation

Similar to the likelihood estimating equations for β and α, we write the likelihood estimating equation for σ^2 as

$$\hat{\sigma}^2(r+1) = \hat{\sigma}^2(r) - \left[\left(\frac{\partial^2 l}{\partial \sigma^4}\right)^{-1}\right]_r \left[\frac{\partial l}{\partial \sigma^2}\right]_r \tag{3.10}$$

where,

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= \frac{1}{2}(\sigma^2)^{-\frac{3}{2}}\Big[-\sum_{h=1}^{M}\sum_{i_h=1}^{N_h}[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) \\ &- W_{i_{h+1}}(h,h+1)(\alpha_{h-1} - x_{i_h}^T\beta)]\Big]\end{aligned}$$

and

$$\begin{aligned}\frac{\partial l}{\partial \sigma^4} = \frac{3}{4}(\sigma^2)^{-\frac{5}{2}}\Big[&- \sum_{h=1}^{M}\sum_{i_h=1}^{N_h}[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) \\ &- W_{i_{h+1}}(h,h+1)(\alpha_{h-1} - x_{i_h}^T\beta)]\Big] \\ &+ \frac{1}{2}(\sigma^2)^{-\frac{3}{2}}\Big[-\sum_{h=1}^{M}\sum_{i_h=1}^{N_h}\Big[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta)\Big[\frac{(\alpha_h - x_{i_h}^T\beta)^2}{2\sigma^4} \\ &- \frac{1}{2\sigma^2} + \frac{1}{2\sigma^3}\Big[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) \\ &- W_{i_{h+1}}(h,h+1)(\alpha_{h-1} - x_{i_h}^T\beta)\Big]\Big] - W_{i_{h+1}}(h,h+1)(\alpha_{h-1} - x_{i_h}^T\beta) \\ &\Big[\frac{(\alpha_{h-1} - x_{i_h}^T\beta)^2}{2\sigma^4} - \frac{1}{2\sigma^2} + \frac{1}{2\sigma^3}\Big[Wi_h(h,h-1)(\alpha_h - x_{i_h}^T\beta) \\ &- W_{i_{h+1}}(h,h+1)(\alpha_{h-1} - X_{i_h}\beta)\Big]\Big]\Big]\Big]\end{aligned}$$

with $\hat{\sigma^2}(r)$ is the estimated value of σ^2 on the rth iteration. The expression $[\cdot]_r$ is evaluated at $\hat{\sigma^2}(r)$.

With regard to the estimation of σ^2, it should be pointed out that if the actual value of σ^2 is close to zero, then the above iteration equation (3.10)

may yield a negative estimate. One may however, use restricted maximum likelihood estimation method or EM algorithm to obtain a non-negative estimate of σ^2. For the diabetes data set however, the maximum likelihood estimate $\hat{\sigma^2}$ itself was found to be positive.

3.2 Probit Analysis of Diabetes Data

In this section we demonstrate the application of the procedure discussed earlier in this chapter by analyzing the data from the Wisconsin Epidemiologic Study of Diabetic Retinopathy (WESDR) discussed in chapter 2. For the purpose we recall all six covariates (i.e. $p = 6$), chosen in chapter 2, to examine their effects on the categorical responses for the right eye retinopathy. These covariates are 1. duration of diabetes $x_{\cdot 1}$; 2. glycosylated hemoglobin level $x_{\cdot 2}$; 3. diastolic blood pressure $x_{\cdot 3}$; 4. proteinuria $x_{\cdot 4}$; 5. sex $x_{\cdot 5}$; 6. right eye macular edema $x_{\cdot 6}$. Here , in general $x_{\cdot u}$ represents the uth ($u = 1, \ldots, 6$) covariate for an individual. Note that as the ith ($i = 1, \ldots, N = 720$) individual belongs to one of the $M = 4$ ordered categories: none, mild, moderate and proliferative, there are $M - 1 = 3$ cut-points explained by $\alpha_1 < \alpha_2 < \alpha_3$ separating the adjacent categories.

Applying the three step iterative procedure given in section 3.1 we obtain the estimates for the unknown scale parameter σ^2, the six regression coefficients $\beta_1, \beta_2, \ldots, \beta_6$ and the three cut-point parameters α_1, α_2, and α_3. These estimates are shown in Table 3.1. In the same table, we also provide the standard errors of these estimates that were obtained from the observed information matrix $-[\psi(\sigma)]^{-1}$, $-[\psi(\alpha)]^{-1}$, and $-[\psi(\beta)]^{-1}$ respectively, where

$$\psi(\sigma) = \frac{\partial^2 l}{\partial \sigma^4}, \quad \psi(\alpha) = \frac{\partial^2 l}{\partial \alpha \partial \alpha^T}, \quad \text{and} \quad \psi(\beta) = \frac{\partial^2 l}{\partial \beta \partial \beta^T}.$$

Table 3.1: Latent Process Based Probit Model Estimates for the Diabetes data

Type of parameter	Parameter	Estimate	Standard errors
Cut-point	α_1	8.66	0.2608
	α_2	14.41	0.1506
	α_3	18.95	0.1733
Regression	β_1	0.281	0.0100
	β_2	0.249	0.0393
	β_3	0.028	0.0044
	β_4	2.964	0.3338
	β_5	-0.340	0.2191
	β_6	6.440	0.4300
Scale	σ^2	5.48	0.0543

It is clear from Table 3.1 that all six covariates except x_5 (sex) have influential effects on the severity of diabetic retinopathy. As far as the sex covariate is concerned, it does not appear to have any influence on the severity of the diabetic retinopathy as $\hat{\beta}_5 = -.340$ with the large standard error

0.2191. This result however contradicts the findings of Williamson et al (1995) as they found that the sex covariate also was influential. As for the estimation of the cut-points is concerned, $\hat{\alpha}_1$, $\hat{\alpha}_2$, and $\hat{\alpha}_3$, clearly met the order restrictions. One of the reasons for this perhaps is that the latent variable had high variation with $\hat{\sigma}^2 = 5.48$, which may have contributed in the first instance to separate out the cut-points clearly, although the cut-points can be close to each other even though $\hat{\sigma}^2$ is high.

As for the remaining five covariates, the covariate $x_{.6}$ (right eye macular edema) seems to be the most influential followed by $x_{.4}$ (proteinuria). The covariates $x_{.1}$ (duration of diabetes) and $x_{.2}$ (glycosylated hemoglobin level) appear to behave similar effects and their effects seem to be less significant than $x_{.4}$, while $x_{.3}$ has the effect does not seem to be very significant.

We should however note here that the convergence values of the parameters shown in Table 3.1 were quite dependent on proper selection of initial values for all the estimates of the parameters α and β. For an assumed value of $\sigma^2 = 4.0$, we searched for the maximum of the likelihood function by trial and error method until the global maximum zone was found. We have then selected initial estimates for the α and β parameters from that zone. These initial estimates, $\hat{\alpha} = (14.79, 17.75, 19.95)$ and $\hat{\beta} = (-0.37, 0.16, 0.14, -4.33, 1.58, 10.88)$ were used in steps 1 and 2 to obtain a temporary convergent set of estimates for α and β. These convergent values of α and β were then used in step 3 to obtain an improved estimate of σ^2. Next, this improved estimate of σ^2 was used in steps 1 and 2 to further improve the estimate of β and α. This cycle of iterations continues until convergence for all three parameters β, α and σ^2. It was found that the

convergence was achieved in 8 such iterations.

Remark that at times, when the three step iterations were performed based on poor selection of initial estimates (such as estimates yielding local maximum) of the parameter, we observed many difficulties, such as; $\hat{\alpha}$'s were not maintaining the order restrictions, and convergence was never achieved.

3.2.1 χ^2 Goodness of Fit

In this chapter we have argued to fit a probit model to the diabetic data. One may however, try to fit other similar but different models to analyze the same data set. For example, we consider a linear cumulative logit model in chapter 4 and a new non-cut-point based multinomial logistic approach in chapter 5 to fit the same diabetic data. It then raises a natural concern, to choose the best model among all possible models that one may wish to fit. A remedy to this concern is to examine the goodness of fit of each model to the data. With this in view, we will evaluate an appropriate goodness of fit statistic under three different models considered in this chapter as well as in chapters 4 and 5.

An appropriate goodness of fit statistic for fitting the probit model to the data is

$$S_1 = \sum_{i=1}^{N} \left[\sum_{h=1}^{M} (\hat{p}_{ih} - p_{ho})^2 \right] \tag{3.11}$$

where p_{ho} is the observed proportion for the ith individual falling into the hth ($h = 1, \ldots, M$) category, and $\hat{p}_{ih}$ is the estimated proportion for the

ith ($i = 1, \ldots, N$) individual to fall into the hth category under the probit model. Here $\hat{p}_{ih}$ is computed by

$$\hat{p}_{ih} = \Phi\left(\frac{\hat{\alpha}_h - x_i^T\hat{\beta}}{\hat{\sigma}}\right) - \Phi\left(\frac{\hat{\alpha}_{h-1} - x_i^T\hat{\beta}}{\hat{\sigma}}\right) \tag{3.12}$$

for $i = 1, \ldots, N$ and $h = 1, \ldots, M$. In (3.11), $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution, where $\Phi(-\infty) = 0$ and $\Phi(+\infty) = 1$. Further in (3.11), the observed proportions are calculated from the data just by counting the number of individuals falling into each of the four categories. For example, if the ith individual is observed to fall into the first category then for all $i = 1, \ldots, N_1$ $p_{iho} = N_1/N$. The observed proportions for the four ordered categories are; $p_{i1o} = 0.38$, $p_{i2o} = 0.38$, $p_{i3o} = 0.18$, $p_{i4o} = 0.06$. Remark that in place of observed proportions one could test any other possible population proportions that may be justified by scientific reasoning.

Note that under the probit model the test statistic in (3.11) has asymptotically χ^2 distribution with degrees of freedom $N - (\overline{M-1} + p + 1)$, where $p + 1$ is the number of regression parameters plus the scale parameter, and $\overline{M-1}$ is the number of cut-points. For the probit model the test statistic S_1 was evaluated as $S_1 = 161.78$ with 710 degrees of freedom (df). Our calculated value of S_1 is less than $\chi^2_{459} = 409.3804 < \chi^2_{710}$ indicating that the data agrees with the null hypothesis.

3.2.2 Display of Squared Error Distances

Let $\hat{d}_i = \sum_{h=1}^{M} (\hat{p}_{ih} - p_{iho})^2$ denote the sum of squares of the differences between the estimated and observed proportions. Remark that, a similar but different statistic could be constructed by using standardized distances $(\hat{p}_{ih} - p_{iho})/\sqrt{var(\hat{p}_{ih})}$ instead of ordinary distances $\hat{p}_{ih} - p_{iho}$. Note however that it is common in practice to compute $var(\hat{p}_{ih}) = p_{ih}q_{ih}/N$ under the null hypothesis. Since these variances, $p_{iho}q_{iho}/N$, are the same for all individuals, $\hat{d}_i$ remains to be an appropriate statistic to make any necessary comparisons and conclusions concerning the fit of the model given in this chapter, as will as the models to be presented in the next two chapters. It is clear that if $\hat{d}_i$ is small then the model based estimated proportions for the ith individual is close to the hypothesized proportions indicating a good fit for the individual. Consequently, to have an overall idea about the fitting of the model, we display all these $\hat{d}_i$ $(i = 1\ldots., 720)$ in Fig 3.1.

It is clear from Fig 3.1 that most of the squared error distances are fairly close to zero. As the magnitude of the squared error distances increase, the number of individuals decline sharply. This pattern is what one would expect to see if the model fits the data well. In the contrary if the model does not fit the data well, one would observe fewer individuals with squared error distances close to zero.

Two other figures similar to Fig 3.1 will be constructed for the other two procedures in chapters 4 and 5. These figures along with the values of the test statistics S_1, S_2 (computed in chapter 4) and S_3 (computed in chapter 5) respectively will be used to compare the three models in deciding the best fitted model among the three.

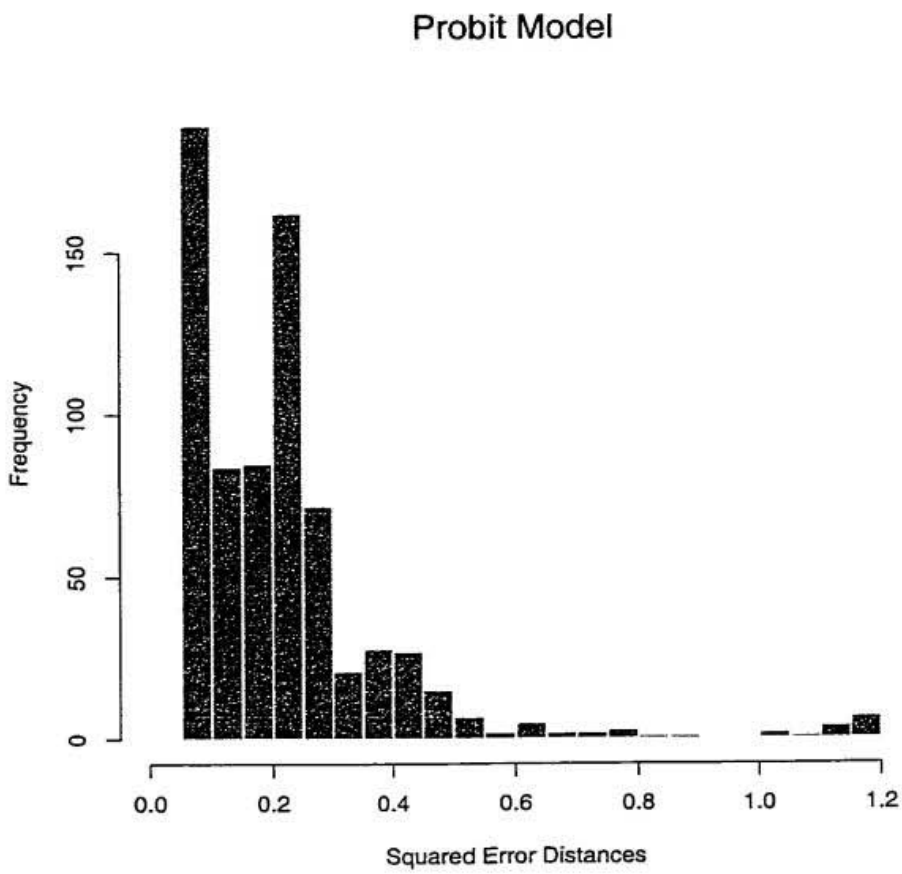

Figure 3.1: Display of Squared Error Distances for the Probit Model

Chapter 4

Linear Cumulative Logit Analysis

In this chapter as well as in chapter 5 we maintain the same notation used for the latent process based probit approach in chapter 3. Let Z_i be the ordered categorical response for the ith ($i = 1, \ldots, N$) individual. Also let Y_{ih} denote the random response variable for the ith person that belongs to the hth ($h = 1, \ldots, M$) category. It then follows that

$$y_{ih} = \begin{cases} 1 & \text{, if } z_i = h \\ \\ 0 & \text{, otherwise} \end{cases}$$

for $h = 1, \ldots, M-1$. Next denote the cumulative probability of Z_i up to the hth category by

$$\gamma_i(h) = F_h(x_i) = Pr(Z_i \leq h | X_i = x_i) \tag{4.1}$$

where $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is the $p \times 1$ covariate vector for the ith person irrespective of his or her category. Now, consider a link function $g(\cdot)$ for this cumulative probability given by

$$\text{logit}(F_h(x_i)) = g(\gamma_i(h)) = \alpha_h - x_i^T \beta \tag{4.2}$$

where β is the $p \times 1$ vector of regression coefficients. In equation (4.2) α_h $(h = 1, \ldots M-1)$ is known as the cut-point between the hth and $(h-1)$th categories. More specifically, as $\gamma_i(h) \geq \gamma_i(h-1)$ always, the logit relationship in (4.2) indicates that α_h has to satisfy the restriction $\alpha_1 < \alpha_2 < \ldots < \alpha_{M-1}$. Further it follows that

$$\gamma_i(h) = \frac{e^{\alpha_h - x_i^T \beta}}{1 + e^{\alpha_h - x_i^T \beta}}. \tag{4.3}$$

Recall that for the diabetes data described in chapter 2, $p = 6$, $M = 4$, and $N = 720$, where these N individuals are independent. The likelihood function is written as

$$L(\beta) = \prod_{i_1=1}^{N_1} p_{i_1}(1,0) \cdots \prod_{i_h=1}^{N_h} p_{i_h}(h, h-1) \cdots \prod_{i_M=1}^{N_M} p_{i_M}(M, M-1) \tag{4.4}$$

where N_h $(h = 1, \ldots, M)$ is the number of individuals that we observed in the hth $(h = 1, \ldots, M)$ category with regard to the diabetic data set. In (4.4), $p_{i_h}(h, h-1)$ is the probability of the ith individual falling into the hth category based on the covariate information of the individual. This probability is expressed as

$$p_{i_h}(h, h-1) = \gamma_{i_h}(h) - \gamma_{i_h}(h-1) \tag{4.5}$$

where i_h indicates the ith individual belongs to the hth category, and $\gamma_{i_h}(h)$ and $\gamma_{i_h}(h-1)$ are given by

$$\gamma_{i_h}(h) = \frac{e^{\alpha_h - x_{i_h}^T \beta}}{1 + e^{\alpha_h - x_{i_h}^T \beta}} \quad , \qquad \gamma_{i_h}(h-1) = \frac{e^{\alpha_{h-1} - x_{i_h}^T \beta}}{1 + e^{\alpha_{h-1} - x_{i_h}^T \beta}}$$

with x_{i_h} written for x_i under the condition that the ith individual belongs to the hth category. Note that as the γ_{i_h}'s are the cumulative probabilities, $\gamma_{i_h}(0) = 0$ and $\gamma_{i_h}(M) = 1$.

4.1 Estimation of the Parameters

Recall that the maximum likelihood estimation method was used in chapter 3 in order to find estimates for the unknown parameters for the probit model. As the likelihood function is also available under the logit model, we exploit the similar likelihood estimating equation approach (as in chapter 3) and obtain the estimates of the parameters.

The proposed logit model given in (4.2) will require us to solve $(\overline{M-1}+p)$ likelihood estimating equations as we have to find estimates for β and $\alpha_1, \ldots, \alpha_{M-1}$. These likelihood estimating equations will be solved by applying the Newton Rhapson iteration technique as before. Remark that the restriction placed on $\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{M-1}$ given in (3.4) namely $\tilde{\alpha}_1 < \tilde{\alpha}_2 < \ldots < \tilde{\alpha}_{M-1}$ may not be achieved when traditionally non-restricted likelihood estimating method is used as an estimation technique. Here $\tilde{\alpha}$ is the estimated value of α based on the logit model. Now, maximizing the likelihood function given in (4.4) is equivalent to maximizing the log-likelihood function given by

$$l = \log L = \sum_{h=1}^{M} \sum_{i_h=1}^{N_h} \log p_{i_h}(h, h-1) \tag{4.6}$$

where $p_{i_h}(h, h-1) = \gamma_{i_h}(h) - \gamma_{i_h}(h-1)$.

Note that for analyzing bivariate ordinal polytomous data, some authors, for example Williamson et al (1995) have used the generalized estimating equation approach to obtain the estimates of the cut-points as well as regression parameters. This is because the likelihood method is quite cumbersome in the bivariate set up.

4.1.1 Step 1: Estimation of the Regression Parameters

To find the likelihood estimate for β namely $\tilde{\beta}$ we first choose some initial value for α_h $(h = 1, \ldots, M-1)$ and then solve the following iterative equation.

$$\bar{\beta}(r+1) = \bar{\beta}(r) - \left[\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right)^{-1}\right]_r \left[\frac{\partial l}{\partial\beta}\right]_r \tag{4.7}$$

where,

$$\frac{\partial l}{\partial\beta} = -\sum_{h=1}^{M}\sum_{i_h=1}^{N_h} \frac{1}{p_{i_h}(h,h-1)} \left[\gamma_{i_h}(h)(1-\gamma_{i_h}(h)) - \gamma_{i_h}(h-1)(1-\gamma_{i_h}(h-1))\right] X_{i_h}$$

and

$$\frac{\partial^2 l}{\partial\beta\partial\beta^T} = \sum_{h=1}^{M}\sum_{i_h=1}^{N_h} \frac{1}{p_{i_h}^2(h,h-1)} \left[a_{h,h-1} - b_{h,h-1} - c_{h,h-1}\right]$$

with,

$$\begin{aligned}
a_{h,h-1} &= \gamma_{i_h}(h)\left[\gamma_{i_h}(h-1)\gamma_{i_h}(h) - \gamma_{i_h}(h-1)\right] \\
b_{h,h-1} &= \gamma_{i_h}^2(h)\left[\gamma_{i_h}(h)\{1-\gamma_{i_h}(h)\} - \gamma_{i_h}(h-1)\{1+\gamma_{i_h}(h-1)\} + 2\gamma_{i_h}(h-1)\gamma_{i_h}(h)\right] \\
c_{h,h-1} &= \gamma_{i_h}(h-1)\left[-\gamma_{i_h}(h)\gamma_{i_h}(h-1) + \gamma_{i_h}^2(h)\right]\}
\end{aligned}$$

In (4.7), $\bar{\beta}(r)$ is the estimated value of β on the rth trial and the expression $[\cdot]_r$ is evaluated at $\bar{\beta}(r)$.

As mentioned in chapter 3 one can estimate β jointly along with the estimation of the α parameters. Instead, for reasons outlined in chapter 3, we have chosen to estimate β and α_h $(h = 1, \ldots, M-1)$ using a two step approach, step 2 being discussed below.

4.1.2 Step 2: Estimation of the Cut-Points

Using the estimates for β obtained in step 1, as the initial values for $\tilde{\beta}$, the likelihood estimates of $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{M-1})$ are obtained by solving the iterative equation.

$$\tilde{\alpha}(r+1) = \tilde{\alpha}(r) - \left[\left(\frac{\partial^2 l}{\partial\alpha\partial\alpha^T}\right)^{-1}\right]_r \left[\frac{\partial l}{\partial\alpha}\right]_r \tag{4.8}$$

where,

$$\begin{aligned}\frac{\partial l}{\partial\alpha_h} &= \sum_{i_h=1}^{N_h} \frac{1}{p_{i_h}(h, h-1)}\gamma_{i_h}(h)(1-\gamma_{i_h}(h)) \\ &- \sum_{i_{h+1}=1}^{N_{h+1}} \frac{1}{p_{i_{h+1}}((h+1), h)}\gamma_{i_{h+1}}(h)(1-\gamma_{i_{h+1}}(h))\end{aligned}$$

and

$$\frac{\partial^2 l}{\partial\alpha_h\partial\alpha_{h'}} = \sum_{i_h=1}^{N_h} \frac{1}{p_{i_h}(h, h-1)} d_{h,h-1} - \sum_{i_{h+1}=1}^{N_{h+1}} \frac{1}{p_{i_{h+1}}(h+1, h)} e_{h,h-1}$$

with,

$$\begin{aligned}d_{h,h-1} &= \gamma_{i_h}(h)(1-\gamma_{i_h}(h))\left[-\gamma_{i_h}(h-1) - \gamma_{i_h}^2(h-1) + 2\gamma_{i_h}(h)\gamma_{i_h}(h-1)\right] \\ e_{h,h-1} &= \gamma_{i_{h+1}}(h)(1-\gamma_{i_{h+1}}(h))\left[\gamma_{i_{h+1}}(h+1) - 2\gamma_{i_{h+1}}(h+1)\gamma_{i_{h+1}}(h) + \gamma_{i_{h+1}}^2(h)\right]\end{aligned}$$

for all $h = 1, \ldots, M-1$. Note that for $|h - h'| = 1$, the second derivatives may be expressed in a simpler form given by,

$$\frac{\partial^2 l}{\partial \alpha_h \partial \alpha_{h'}} = -\sum_{i_{h+1}=1}^{N_{h+1}} \frac{1}{p^2_{i_{h+1}}(h+1,h)} \gamma_{i_{h+1}}(h)(1-\gamma_{i_{h+1}}(h))\gamma_{i_{h+1}}(h+1)(1-\gamma_{i_{h+1}}(h+1))$$

whereas for $|h-h'| \neq 1$,

$$\frac{\partial^2 l}{\partial \alpha_h \partial \alpha_{h'}} = 0.$$

In equation (4.8) $\tilde{\alpha}_{(r)}$ is the estimated value of α on the rth iteration and the expression $[\cdot]_r$ is evaluated at $\tilde{\alpha}(r)$. Also, in equation (4.8) $\frac{\partial l}{\partial \alpha}$ is the $M-1 \times 1$ vector of first derivatives and $\frac{\partial^2 l}{\partial \alpha \partial \alpha^T}$ is the $M-1 \times M-1$ second derivative matrix.

We remark that in the manner similar to that of chapter 3, the likelihood equation for β (4.7) and α (4.8) will be solved in two steps as mentioned before, where these two steps constitute a cycle. This cycle of iterations continue until convergence.

Note that, in estimating the cut-point parameters $\alpha_1, \ldots, \alpha_{M-1}$ one expects that these estimates maintain the order $\tilde{\alpha}_1 < \tilde{\alpha}_2 < \ldots < \tilde{\alpha}_{M-1}$, but there is no guarantee that these likelihood estimates of α_h's will maintain this order restriction.

4.1.3 Some Remarks on Choosing the Initial Estimates

To choose initial estimates for the unknown α and β parameters, it is natural to make an attempt to use the estimates given in Williamson et al (1995), although their analysis is done for the two variable case while the present study

deals with the one variable case. More specifically, the bivariate correlation

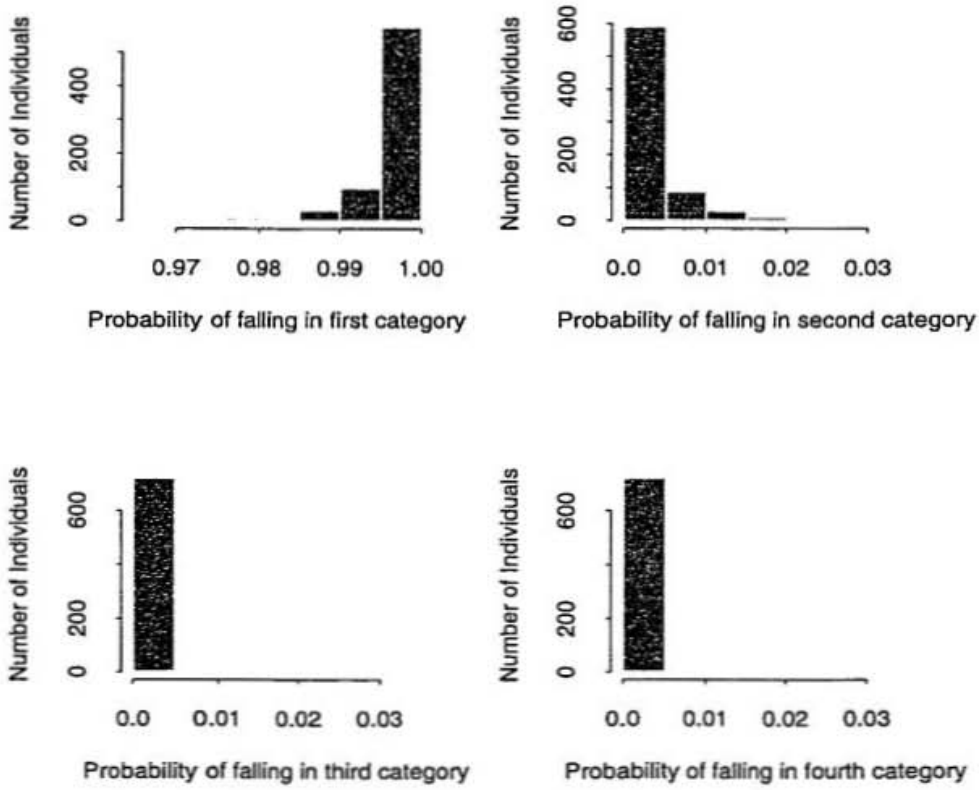

Figure 4.1: Display of Estimated proportions based on the Cumulative Logit Model

may not have large influence on these common parameters α and β between the two variables. Furthermore, these are only initial values to start the iterative procedure explained in the last section.

As we explain below, we however, find that Williamson et al's (1995)

estimates when used in our univariate case fall short in producing estimated proportions those should be in agreement with the observed proportions. To be precise, we first used their estimates for α ($\alpha_1 = -.823, \alpha_2 = 1.58, \alpha_3 = 3.67$) and β ($\beta_1 = -0.129, \beta_2 = -0.0913, \beta_3 = -0.0421, \beta_4 = -0.901, \beta_5 = 0.338, \beta_6 = -0.138$) and calculated the probabilities for the ith ($i = 1, \ldots, 720$) individual to fall into each of the four ordered categories. These probabilities under each category for all 720 individuals are shown in Figure 4.1. It is clear from the figure that an individual appears to have expected probability close to one for the first category but his/her probability to be in any other category appears to be almost zero. As the observed proportions for an individual belonging to these four categories are .38, .38, .18, .06, there does not appear any agreement between these observed proportions and estimated proportions given in Figure 4.1.

Further, when these estimates, provided by Williamson et al (1995) were used as initial estimates, we were not able to obtain convergent estimates. This is not unlikely as the probabilities computed based on such estimates for an individual to belong to the last three categories were found to be extremely low. As in the probit analysis discussed in the last chapter, there does not appear any easy alternative way to choose appropriate initial values for the parameters $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, $\tilde{\alpha}_3$ that might lead to the global maximization of the likelihood function (4.4) with respect to these parameters. As a remedy, we adapted a trial and error method to choose initial values of $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, $\tilde{\alpha}_3$ such that the expected proportions reflect the observed proportions to a good extent. Note that in calculating such expected proportions, we used β estimates as in Williamson et al (1995) as, unlike the estimation of α values,

it is straight forward to obtain consistent estimates of β based on any suitable initial values.

To be more precise, for a selected individual, we use his/her covariate information along with Williamson et al's β estimates to solve equation 4.3 for α_h $(h = 1, \ldots, M-1)$ taking $\gamma_i(h)$ $(h = 1, \ldots, M-1)$ to be the observed cumulative proportions for the individual to belong to the hth or lower category. For example the observed cumulative proportion for the first category is .38, therefore, we set $\gamma_i(1) = .38$ in equation 4.3. Now, using Williamson et al's β vector $(\beta' = [-0.129, -0.0913, -0.0421, -0.901, 0.338, -0.138])$ and the selected individuals covariate information along with $\gamma_i(1)$, we arrived at a starting value for α_1. Similarly, as the $\gamma_i(h)$'s are cumulative probabilities, we took $\gamma_i(2) = .76$ and $\gamma_i(3) = .94$, and solved for α_2 and α_3 respectively. These initial values of $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, $\tilde{\alpha}_3$ were used to compute the probabilities for each of the four categories for all 720 individuals. Next we computed the average probability for each of the four categories, which we denoted by $\bar{p}_1$, $\bar{p}_2$, $\bar{p}_3$ and $\bar{p}_4$. A comparison was made between $\bar{p}_h$ $(h = 1, \ldots, M)$ and the observed proportions. As expected, to begin with, there were some differences between $\bar{p}_h$ and the observed proportions, which lead us to select improved values of $\tilde{\alpha}_h$'s to minimize such differences. This procedure was repeated until the values of α_1, α_2 and α_3 produced $\bar{p}_h$'s that reflect the observed proportions to a good extent. These values then became the initial estimates of the cut-point parameters namely $\tilde{\alpha}_1$, $\tilde{\alpha}_2$ and $\tilde{\alpha}_3$.

4.2 A Limited Simulation Study to Verify Initial Cut-Points

To verify the validity of the cut-point estimates obtained in the previous section, we now conduct a limited simulation study. Note however that, as indicated before, since β is a non-restricted regression vector, it was not necessary to examine the performance of β estimates through the simulation study.

To generate the multinomial response data using ordered initial α values obtained as in section 4.1.3, Williamson et al's (1995) β values, and the covariates available for all 720 individuals, we calculate

$$p_{i1} = \frac{e^{\tilde{\alpha}_1 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_1 - x_i^T\tilde{\beta}}}, \quad p_{i2} = \frac{e^{\tilde{\alpha}_2 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_2 - x_i^T\tilde{\beta}}} - \frac{e^{\tilde{\alpha}_1 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_1 - x_i^T\tilde{\beta}}},$$
$$p_{i3} = \frac{e^{\tilde{\alpha}_3 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_3 - x_i^T\tilde{\beta}}} - \frac{e^{\tilde{\alpha}_2 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_2 - x_i^T\tilde{\beta}}}, \quad p_{i4} = 1 - \frac{e^{\tilde{\alpha}_3 - x_i^T\tilde{\beta}}}{1+e^{\tilde{\alpha}_3 - x_i^T\tilde{\beta}}}$$

for the ith individual. The initial values of α's were

$$\tilde{\alpha}_1 = -8.5, \quad \tilde{\alpha}_2 = -6.5, \quad \tilde{\alpha}_3 = -5.0 \tag{4.9}$$

We supply these probabilities and use the IMSL subroutine RNMTN to generate the multinomial response such as $[1,0,0,0]$ or $[0,1,0,0]$ or $[0,0,1,0]$ or $[0,0,0,1]$. Here $[1,0,0,0]$, for example, indicates that the ith person belongs to the first category. We have done it for 720 individuals.

Next these responses are used along with the covariates, to estimate α by (4.8). The simulation was carried out 2000 times and the average value of $\tilde{\alpha}_1$, $\tilde{\alpha}_2$ and $\tilde{\alpha}_3$ were found to be

$$\tilde{\alpha}_1 = -8.05, \quad \tilde{\alpha}_2 = -5.86, \quad \tilde{\alpha}_3 = -4.67 \tag{4.10}$$

which appear to agree to a good extent with the initial chosen value of $\tilde{\alpha}_1$, $\tilde{\alpha}_2$ and $\tilde{\alpha}_3$ shown in (4.9).

4.3 Fitting Cumulative Logit Model to the Diabetes Data

In this section we analyze the data from the WESDR to illustrate the application of the proposed method given in this chapter. Again, recall from chapter 2 that the six covariates (i.e. $p = 6$) used to explain severity of right eye diabetic retinopathy are 1. duration of diabetes $x_{\cdot 1}$; 2. glycosylated hemoglobin level $x_{\cdot 2}$; 3. diastolic blood pressure $x_{\cdot 3}$; 4. proteinuria $x_{\cdot 4}$; 5. sex $x_{\cdot 5}$; 6. right eye macular edema $x_{\cdot 6}$, where $x_{\cdot v}$ denotes the vth ($v = 1, \ldots, 6$) covariate for an individual. In this study an individual is assumed to belong to one of the $M = 4$ possible categories: none, mild, moderate and proliferative with a suitable probability . To distinguish the adjacent categories there are $M - 1 = 3$ cut-point parameters α_1, α_2, α_3 which need to be estimated. Further we require these cut-point parameters to hold the order restriction $\alpha_1 < \alpha_2 < \alpha_3$.

To obtain estimates for all nine unknown parameters: six regression and three cut-point parameters, we exploit the 2 steps iterative procedure discussed in section 4.1. More specifically, based on some initial values of α and β we solve the iterative equation (4.7) until convergence is achieved for $\tilde{\beta}$. Now, suppling these new convergent values of β and the initial values of $\tilde{\alpha}$ to step 2 of the procedure in section 4.1, equation (4.8) is solved until convergence is achieved for $\tilde{\alpha}$. These new improved values of $\tilde{\alpha}$ from step 2 and the improved values of $\tilde{\beta}$ from step 1, are then used in step 1 to obtain a new set of improved estimates for β. This cycle of iterations continues until convergence is obtained between cycles. The final estimates for the cut-point parameters α_1, α_2, α_3 and the regression coefficients $\beta_1, \beta_2, \ldots, \beta_6$ are shown in Table 4.1. We also report the standard errors of these estimates that were obtained by using the observed information matrix $-[\psi(\alpha)]^{-1}$ and $-[\psi(\beta)]^{-1}$ respectively for the estimates of α and β.

$$\psi(\alpha) = \frac{\partial^2 l}{\partial \alpha \partial \alpha^T}, \quad \text{and} \quad \psi(\beta) = \frac{\partial^2 l}{\partial \beta \partial \beta^T}.$$

All covariates ($x_{\cdot v}$) appear to have a significant contribution for explaining severity of diabetic retinopathy. Remark that, in this approach the covariate for sex ($x_{\cdot 5}$) also has influential effect on the severity of diabetic retinopathy as opposed to the method of chapter 3 which concluded that the sex covariate was not influential. Further, the most influential covariate was found to be $x_{\cdot 6}$ (right eye macular edema) followed by $x_{\cdot 4}$ (proteinuria), $x_{\cdot 5}$ (sex), $x_{\cdot 1}$ (duration of diabetes), $x_{\cdot 2}$ (glycosylated hemoglobin level) and $x_{\cdot 3}$ (diastolic blood pressure). This pattern in the behavior of the covariate estimates

Table 4.1: Linear Cumulative Logit Model Estimates

Type of parameter	Parameter	Estimate	Standard errors
Cut-point	α_1	-8.14	0.09460
	α_2	-6.62	0.09349
	α_3	-5.20	0.14890
Regression	β_1	-0.1290	0.0000325
	β_2	-0.0902	0.0005216
	β_3	-0.0400	0.0000172
	β_4	-0.9006	0.0012251
	β_5	0.3395	0.0017232
	β_6	-1.3790	0.0013490

appears to be the same as that produced by the Probit Analysis of chapter 3 except for the sex covariate $x_{.5}$. Note that the estimates for the regression parameters displayed in table 4.1 are almost identical to those of Williamson et al (1995).

As far as the cut-points are concerned, we can clearly see from table 4.1 that these estimates $(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)$ meet the order restriction of $\tilde{\alpha}_1 < \tilde{\alpha}_2 < \tilde{\alpha}_3$ as in Williamson et al (1995), but they are quite different than those estimates provided by these authors. Further there is no standard errors available in Williamson et al (1995) for these estimates.

4.3.1 χ^2 Goodness of Fit

In this section we investigate the goodness of fit of the logit model to the diabetes data. This will examine whether the logit model is an appropriate model to describe the relationship between the covariates and severity of diabetic retinopathy. Further, the goodness of fit statistic for the logit model will be compared to that of the probit model to aid in deciding which of the two models appear to have the best fit to the data. These two goodness of fit statistics will be recalled again in chapter 5 in order to compare them with the new non-cut-point based approach presented.

For the current cumulative logit model, we may use the goodness of fit statistic

$$S_2 = \sum_{i=1}^{N} \left[\sum_{h=1}^{M} (\tilde{p}_{ih} - p_{ho})^2 \right] \tag{4.11}$$

to test the fitting of this model to the data. In (4.11) p_{ho} is the observed proportion for the ith individual to fall into the hth ($h = 1, \ldots, M$) category, and these p_{ho} are: ($p_{1o} = .38$, $p_{2o} = .38$, $p_{3o} = .18$, $p_{4o} = .06$,) as in chapter 3. Now, in (4.9) $\tilde{p}_{ih}$ is the estimated proportion for the ith ($i = 1, \ldots, N$) individual to fall into the hth category under the cumulative logit model. More specifically, for the current logit model, these $\tilde{p}_{ih}$ are given by

$$\tilde{p}_{ih} = \omega_i(h) - \omega_i(h-1) \tag{4.12}$$

where,

$$\omega_i(h) = \frac{e^{\bar{\alpha}_h - x_i^T\bar{\beta}}}{1 + e^{\bar{\alpha}_h - x_i^T\bar{\beta}}}, \quad \omega_i(h-1) = \frac{e^{\bar{\alpha}_{h-1} - x_i^T\bar{\beta}}}{1 + e^{\bar{\alpha}_{h-1} - x_i^T\bar{\beta}}}$$

for $i = 1, \ldots, N$ and $h = 1, \ldots, M$. Further, in equation (4.12) $\omega_i(0) = 0$ and $\omega_i(M) = 1$.

Remark that the goodness of fit statistic S_2 in (4.11) is quite similar to that of S_1 in equation (3.11). The only difference is, $\bar{p}_{ih}$ is the estimated proportion for the ith ($i = 1, \ldots, N$) individual to fall into the hth category under the cumulative logit model and $\hat{p}_{ih}$ is the estimated proportion for the ith ($i = 1, \ldots, N$) individual to fall into the hth category under the probit model.

The test statistic S_2 (4.11) for testing the closeness of the estimated proportions under the logit model to the observed proportions, has asymptotically χ^2 distribution with degrees of freedom $N - (\overline{M-1} + p)$. It is clear that as compared to the probit model, we now have one less parameter, yielding the degrees of freedom $N - (\overline{M-1} + p)$. For the logit model the test statistic S_2 was evaluated as $S_2 = 126.55$ with 711 degrees of freedom (df). Our calculated value of S_2 is less than $\chi^2_{459} = 409.3804 < \chi^2_{711}$ indicating that the data agrees with the null hypothesis.

In chapter 3, we investigated the fit of probit model as a reasonable model for explaining severity of diabetic retinopathy in the diabetes data set. For the probit model we calculated the test statistic S_1 given in equation (3.11) to be $S_1 = 161.78$. As $S_1 = 161.78$ is greater than $S_2 = 126.55$ it appears that the logit model provides an improved fit to the data over the probit

model.

4.3.2 Display of Squared Error Distances

For a deeper insight with regard to the goodness of fit of the cumulative logit model to the diabetes data, we calculated the squared distances between the model based proportions and the observed proportions. These distances for the ith ($i = 1, \ldots, 720$) individual are calculated by $\tilde{d}_i = \sum_{h=1}^{M} (\tilde{p}_{ih} - p_{iho})^2$. If the logit model provides a good fit to the data we expect to observe a small value for $\tilde{d}_i$ for the ith ($i = 1, \ldots, 720$) individual. In contrary, a large value of $\tilde{d}_i$ will indicate that the logit model is providing a poor fit to the data. These squared distances $\tilde{d}_i$ for all individuals ($i = 1, \ldots, 720$) are displayed in figure 4.2.

This figure clearly shows that the majority of the individuals have a squared error distance that is fairly close to zero, which in turn shows that the probability for an individual to fall in the four categories reflect the observed proportions indicating that the cumulative logit model is providing a good fit to the diabetes data.

A comparison of the squared error distances for the probit model with the squared error distances for the logit model provides additional information to support the notion that the logit model appears to provide a better fit to the data than the probit model. A comparison of figure 3.1 and figure 4.1 demonstrates this fact, as the logit model exhibits more squared error distances closer to zero. As well, the logit model appears to have lower squared distances overall. The largest squared error distance for the probit model is nearly 1.2 while it is less then .8 for the logit model.

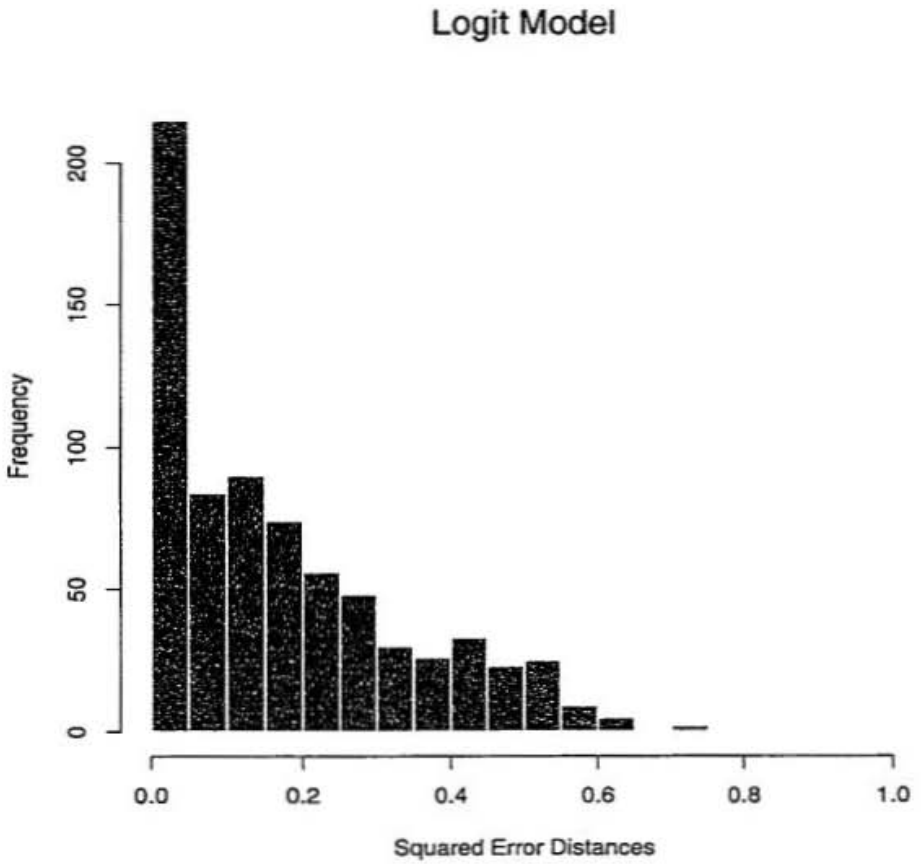

Figure 4.2: Display of Squared Error Distances for the Logit Model

Chapter 5

Non-Cut-Point Based Multinomial Logistic Approach

In this chapter we apply a new multinomial logistic approach to model ordinal categorical responses, which was recently suggested by Das and Sutradhar (1999) in connection with bivariate ordinal polytomous data analysis and by Sutradhar and Kovacevic (2000) in a multivariate set up. Unlike the probit and the cumulative logit approaches discussed in chapter 3 and 4 respectively, this approach does not require any cut-points at all. This is a big improvement over the existing procedures, as all the parameters become non-restricted regression parameters which may be consistently estimated by a suitable method such as the generalized estimating equation approach. To be more specific, although after a lengthly trial and error search, we were able to find cut-point estimators under both of the probit and cumulative logit models, there is however no guarantee that these types of cut-point estimates will maintain the order restriction which is inherent in the model. In other

words, the cut-point estimates will be consistent or will maintain the order restriction only when suitable ordered restricted estimation is exploited. This is however extremely complicated for the logistic multinomial data. Note that as it will be described in this chapter, the current procedure maintains the order nature of the data without any introduction of the cut-points.

We now describe the present model for the univariate case following Sutradhar and Kovacevic (2000). Maintaining the same notation as in chapters 3 and 4, let Z_i be the ordered categorical response for the ith ($i = 1, \ldots, N$) individual. Thus Z_i can take on values of $1, \ldots, M$ following the cumulative probability

$$\gamma_i(h) = F_h(x_i) = Pr(Z_i \leq h | X_i = x_i) \tag{5.1}$$

where h indicates the hth ($h = 1, \ldots, M$), and $X_i^T = (X_{i1}, X_{i2}, \ldots, X_{ip})$ is the $p \times 1$ covariate vector for the ith individual. Recall that the cumulative probabilities shown in (5.1) is exactly the same as the cumulative probabilities defined in (4.1). Further this cumulative probability, by using the polytomous logistic regression, may be written as

$$\begin{aligned} \gamma_i(h) &= Pr(Z_i \leq h) \\ &= \frac{\sum_{l=1}^{h} e^{x_{il}^T \beta_l}}{\sum_{u=1}^{M} e^{x_{iu}^T \beta_u}}, \end{aligned} \tag{5.2}$$

for $h = 1, \ldots, M$. In (5.2) β_u $(u = 1, \ldots, M)$ is the $p \times 1$ vector of regression parameters corresponding to the uth category, and $x_{iu}^T = (x_{iu1}, x_{iu2}, \ldots, x_{iup})$ is the $p \times 1$ covariate vector for the ith individual that belongs to the uth category. Note, however, that in the present data set, the covariates for the ith individual remains the same irrespective of the category. That is $x_{iu}^T = x_i^T$. Consequently, in what follows we use x_i^T for x_{iu}^T for all $u = 1, \ldots, M$. Further, in (5.2), without any loss of generality, we assume that $\beta_M = 0$.

It then follows that the multinomial logistic marginal probability that $Z_i = h$ is given by

$$\begin{aligned} Pr(Z_i = h) &= Pr(Y_{ih} = 1) \\ &= \pi_{ih} \\ &= \frac{e^{x_i^T \beta_h}}{\sum_{u=1}^{M} e^{x_i^T \beta_u}}. \end{aligned} \tag{5.3}$$

In (5.3), Y_{ih} $(h = 1, \ldots, M-1)$ is the dichotomous random response variable for the ith $(i = 1, \ldots, N)$ individual that belongs to the hth $(h = 1, \ldots, M)$ category. More specifically, it follows that Z_i and Y_{ih} are connected through the following relation

$$y_{ih} = \begin{cases} 1 & , \text{if } z_i = h \\ 0 & , \text{otherwise} \end{cases}$$

Note that the logits of the cumulative marginal probabilities are given by

$$\text{logit}(\gamma_i(h)) = \log\left[\frac{\sum_{l\leq h} e^{x_i^T\beta_l}}{\sum_{l>h}^{M} e^{x_i^T\beta_l}}\right] \tag{5.4}$$

which is rather in log odds ratio form. Further, the logits of the marginal probabilities are given by

$$\text{logit}(\pi_{ih}) = \log\left[\frac{e^{x_i^T\beta_h}}{\sum_{u=1}^{M} e^{x_i^T\beta_u}}\right] \tag{5.5}$$

for $h = 1, \ldots, M-1$. Note that, the logits for the cumulative probabilities given in (5.4) are quite similar to their corresponding logits for the marginal probabilities shown in (5.5). In the marginal case, the logits are the log of the odds of an exponential function for an ordinal category versus a sum of the similar exponential functions for the remaining categories, whereas in the cumulative case, the logits are the log of the odds for a sum of the exponential functions up to an ordinal category versus the sum of the similar exponential functions for the remaining categories. Thus, as it happens in the multinomial logistic case, we do not have any linear logits (in covariates) either for the cumulative margins or for the margins themselves. The non-linear logits in the present approach are, however, easy to interpret.

5.1 The Estimating Equations For Regression Parameters

For the cut-point based probit and cumulative logit models, we have exploited the likelihood estimation for the regression and cut-point parameters. Since the present approach does not suffer from the complexity of involving cut-points, we choose to use the generalized estimating equation approach. This estimating approach, however, is technically much easier than the likelihood estimating approach.

To obtain estimates for the regression parameter vector $\beta = (\beta_1^T, \ldots, \beta_h^T, \ldots, \beta_{M-1}^T)$, where $\beta_h = (\beta_{h1}, \ldots, \beta_{hj}, \ldots, \beta_{hp})^T$ we observe that β is present in all $\pi_i = (\pi_{i1}, \ldots, \pi_{ih}, \ldots, \pi_{iM-1})^T$ for $i = 1, \ldots, N$, where $E(Y_{ih}) = \pi_{ih}$ (5.3). Consequently, to construct the estimating equations for β, we minimize a suitable weighted distance vector, where the distance vector is given by $Y_i - \pi_i$, for the ith $(i = 1, \ldots, N)$ individual, $Y_i = (Y_{i1}, \ldots, Y_{ih}, \ldots, Y_{iM-1})^T$ being the observation vector and $\pi_i = E(Y_i)$. More specifically we estimate β by solving

$$N^{-\frac{1}{2}} \sum_{i=1}^{N} D_i^T V^{-1} (Y_i - \pi_i) = 0 \tag{5.6}$$

where

$$V_i = \text{var}(Y_i)$$

$$= \begin{pmatrix} \pi_{i1} & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & \pi_{ih} & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \pi_{iM-1} \end{pmatrix} - \pi_i \pi_i^T \tag{5.7}$$

Further in equation (5.6)

$$\begin{aligned} D_i^T &= \frac{\partial \pi_i^T}{\partial \beta} \\ &= V_i \otimes X_i \end{aligned} \tag{5.8}$$

where D_i is an $(M-1) \times (M-1)$ matrix, and X_i is the $p \times 1$ covariate vector and $\otimes$ denotes the Kronecker product.

Note that the estimating equation (5.6) is usually referred to as the quasi-likelihood estimating equation [cf Miller et al (1993), McCullagh (1983)].

5.1.1 Newton Rhapson Iteration Technique

Estimates for the $M-1$ regression parameter vectors is obtained by solving equation (5.6). This solution denoted by β^*, may be obtained by the Newton Raphson iterative technique. For some initial value of β^* we solve the following iterative equation

$$\beta^*(r+1) = \beta^*(r) + \left[\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right]_r^{-1} \left[\sum_{i=1}^{N} D_i^T V_i^{-1} (Y_i - \pi_i)\right]_r \tag{5.9}$$

where $[\cdot]_r$ denotes that the expression within the bracket is evaluated at $\beta^*(r)$. Further, it follows that $N^{\frac{1}{2}}(\beta^* - \beta)$ is asymptotically multivariate normal with zero mean and covariance matrix V_β given by

$$V_\beta = \lim_{N\to\infty} N\left(\sum_{i=1}^{N} D_i^T V_i^{-1} D_i\right)^{-1}. \tag{5.10}$$

This covariance matrix V_β of β^* may be consistently estimated by using β^* for β in obtained from (5.9).

5.2 Multinomial Logistic Analysis of Diabetes Data

In this section we illustrate the application of the new non-cut-point based procedure introduced by Sutradhar and Kovacevic (2000) by reanalyzing the data from the WESDR. The present data set, which was introduced in chapter 2, contains a large number of covariates, six of which were considered in the probit analysis of chapter 3 and the cumulative logit analysis in chapter 4. We continue to consider these same six covariates in the present analysis. For the purpose of the analysis of this section we recall all six covariates to study their effectiveness in explaining severity of right eye diabetic retinopathy. These covariates are 1. duration of diabetes $x_{\cdot 1}$; 2. glycosylated hemoglobin level $x_{\cdot 2}$; 3. diastolic blood pressure $x_{\cdot 3}$; 4. proteinuria $x_{\cdot 4}$; 5. sex $x_{\cdot 5}$; 6. right eye macular edema $x_{\cdot 6}$. Here, in general $x_{\cdot w}$ represents the wth ($w = 1, \ldots, 6$) covariate for an individual.

The application of the multinomial logistic model (discussed earlier in this chapter) to the WESDR data, will require the estimation of $p \times (M-1) = 6 \times 3 = 18$ regression coefficients denoted by $\beta^* = (\beta_1^{*T}, \ldots, \beta_h^{*T}, \ldots, \beta_{M-1}^{*T})$ with $M = 4$. To compute this β^* we apply the estimating equation approach of Section 5.1. Supplying suitable initial values of $\beta^* = \beta^*(0) = (\beta_{1_0}^{*T}, \ldots, \beta_{h_0}^{*T}, \ldots, \beta_{(M-1)_0}^{*T})$ to the iterative equation (5.9), convergence was achieved for all β parameters except β_{16}, the regression coefficient for right eye macular edema, for the 'none' category. Note that, even though the estimates provided by Das and Sutradhar (1999) were obtained from bivariate analysis, they may still be used as suitable initial values of β^* for the present univariate approach. These values are: $\beta_{1_0}^{*T} = (-0.1261,\ 0.0182,\ 0.0122,\ -0.8868,\ 0.5066,\ -0.5302)$, $\beta_{2_0}^{*T} = (0.0530,\ 0.0175,\ -0.0012,\ -0.2371,\ 0.1321,\ -2.02341)$, $\beta_{3_0}^{*T} = (0.0511,\ 0.0075,\ -0.0017,\ 0.3112,\ -0.4283,\ 0.5704)$. As for the sixth covariate, convergence was obtained for β_{26} and β_{36}, but not for β_{16}, this prompted us to investigate the reason why problems were occurring when trying to obtain a solution for this covariate under the 'none' category as opposed to the other two categories, namely the 'mild' and 'moderate' categories. For the purpose, we decided to take a detailed look at the WESDR data set. The data for sixth covariate $x_{\cdot 6}$ (right eye macular edema) is a series of 0's and 1's only, where 1 represents that macular edema is present in the right eye and 0 indicates that macular edema is not present in the right eye. Of the 720 individuals it was observed that 33 of them showed macular edema present in the right eye, which means that there are 33 1's and 687 0's representing this covariate. A more through inspection revealed that none of these 33 1's were present with the 275 individuals who belong

to the 'none' category, while two 1's showed up in the 'mild' category, 14 in the 'moderate' category and 17 were observed in the proliferative category. As this covariate consists of only 0's for the 'none' category, it becomes clear that there is no unique solution for β_{16}.

To shed further light into this non-convergence problem, we also conducted a search procedure to select a possible estimate of β_{16} by computing the distance function

$$d^* = \{d_1^2 + \ldots + d_l^2 + \ldots + d_{18}^2\}^{\frac{1}{2}} \tag{5.11}$$

for many possible values of β_{16}, namely $-50 \leq \beta_{16} \leq 3$, while the values for all other regression estimates were kept fixed at their convergent values. In (5.11), d_l $(l = 1, \ldots, 18)$ is the value of the lth element of the 18×1 vector in (5.6). It was observed that d^* was decreasing to zero as the value of β_{16}^* was decreasing to $-\infty$. Also, it was observed that the corresponding variance of the estimate of β_{16} was getting larger as β_{16}^* was getting smaller. Consequently, the covariate $x_{.6}$ under the 'none' category, appears to contribute nothing to the change in the response variable. Because of this, in calculating the goodness of fit of the model to the data, we will use $\beta_{16} = 0$. Also, we perform a separate analysis, excluding altogether the sixth covariate. The final estimates of β except for β_{16} are shown in Table 5.1, along with their standard errors obtained from (5.10)

The present model provides a separate set of regression coefficients for each of the categories of 'none', 'mild' and 'moderate' for explaining severity of diabetic retinopathy, as shown in table 5.1. But, the probability for an

individual to fall in the hth category depends on all 18 regression coefficients,

Table 5.1: Non-Cut-Point Based Multinomial Logistic Model Estimates

Category	Parameter	Estimate	Standard errors
NONE	β_1	-0.2767	0.0658
	β_2	0.0745	0.5828
	β_3	0.0268	0.0067
	β_4	-2.3475	0.4237
	β_5	0.9784	0.3210
	β_6	-	-
MILD	β_1	-0.0377	0.0204
	β_2	0.1314	0.0584
	β_3	0.0047	0.0067
	β_4	-1.6306	0.4153
	β_5	0.6525	0.3324
	β_6	-4.1458	1.3067
MODERATE	β_1	0.0099	0.0176
	β_2	0.1543	0.0526
	β_3	-0.0009	0.0059
	β_4	-1.1392	0.3523
	β_5	-0.2226	0.2968
	β_6	-1.5781	0.4254

as the probability for an individual to belong to the hth category is determined by

$$\pi_{ih} = \frac{e^{x_i^T \beta_h}}{\sum_{u=1}^{M} e^{x_i^T \beta_u}}.$$

Thus, the interpretation regarding the significance of β's is different than it usually is for linear or non-linear simple multiple regression problems. This is, however, generally true that a small value of the regression coefficient for any covariate under a given category, as compared with any larger value of the same covariate under another category, will indicate its poor influence in determining the probability for the individual to belong to that particular category. For example, consider the value of β_1^* under all three categories of 'none', 'mild', 'moderate'. As under the 'none' category β^* has large negative value along with its (relatively) small standard error as compared to its values in the other two categories, the contribution of this covariate is naturally significant in yielding a large probability for an individual falling into the 'none' category.

It is clear from Table 5.1 that all covariates ($x_{\cdot w}$) appear to be necessary to explain the severity of diabetic retinopathy. Note however, the contribution pattern of the covariates does not appear to be the same from category to category. More specifically in the 'none' category, the influential covariates are duration of diabetes $x_{\cdot 1}$, diastolic blood pressure $x_{\cdot 3}$, proteinuria $x_{\cdot 4}$, and sex $x_{\cdot 5}$, where as glycosylated hemoglobin level $x_{\cdot 2}$ does not appear to be an influential covariate under this category. This means that $x_{\cdot 2}$ does not

contribute to make the probability larger for an individual to belong to the 'none' group. Further, for the same reasons discussed earlier in this chapter, the regression effect of the sixth covariate, right eye macular edema $x_{.6}$, is not reported. In the 'mild' category all covariates appear to be important to the model except for diastolic blood pressure $x_{.3}$. As far as the 'moderate' category is concerned, only half of the covariates appear to be influential. These covariates are: glycosylated hemoglobin level $x_{.2}$, proteinuria $x_{.4}$ and right eye macular edema $x_{.6}$.

5.2.1 χ^2 Goodness of Fit

In this chapter we have proposed a non-cut-point based multinomial logistic approach to model severity of diabetic retinopathy. The purpose of this Subsection is to investigate the goodness of fit of this model to the diabetes data, in order to make inferences as to whether this model is an adequate model to describe the diabetes data set. Once an appropriate statistic is calculated we will compare it with the goodness of fit statistics for the probit model of chapter 3 and the cumulative logit model of chapter 4. This comparison will assist in choosing the 'best' fit of all three models proposed.

An appropriate goodness of fit statistic for testing the fit of the current non-cut-point based multinomial logistic model to the data is given by

$$S_3 = \sum_{i=1}^{N} \left[\sum_{h=1}^{M} (p^*_{ih} - p_{ho})^2 \right] \tag{5.12}$$

which is a comparable statistic to the goodness of fit statistic S_1 (3.11) for

the probit model and S_2 (4.11) for the cumulative logit model. In (5.12), p^*_{ih} denotes the estimated proportion for the ith ($i = 1, \ldots, N$) individual to belong to the hth category under the non-cut-point based multinomial logistic model, which is computed by

$$p^*_{ih} = \frac{e^{x_i^T \beta^*_h}}{\sum_{u=1}^{M} e^{x_i^T \beta^*_u}} \tag{5.13}$$

for $i = 1, \ldots, N$ and $h = 1, \ldots, M$. Note that, in (5.12) p_{ho} which is the same as in the equations for S_1 in (3.11) and S_2 in (4.11), is the observed proportions. The values of p_{ho} are: $p_{1o} = .38$, $p_{2o} = .38$, $p_{3o} = .18$, and $p_{4o} = .06$.

Recall from chapters 3 and 4, that the goodness of fit statistics S_1 and S_2 differ from each other through the estimating formulas for the proportions $\hat{p}_{ih}$ and $\tilde{p}_{ih}$ respectively. Here, $\hat{p}_{ih}$ is the estimated proportion for the ith ($i = 1, \ldots, N$) individual to fall in the hth category under the probit model and $\tilde{p}_{ih}$ is the estimated proportion for the ith ($i = 1, \ldots, N$) individual to fall in the hth category under the cumulative logit model. In the same manner, the goodness of fit statistic S_3 differs from S_1 and S_2, as it is defined based on different estimated proportions than used for S_1 and S_2. More specifically, in (5.12), p^*_{ih} is computed by (5.13), which is quite different than the formulas for $\hat{p}_{ih}$ and $\tilde{p}_{ih}$ discussed in chapters 3 and 4 respectively.

The test statistic S_3 for testing the fit of the non-cut-point based multinomial logistic model to the data given in (5.12), has asymptotically χ^2 distribution with $N - (\overline{M-1} \times p)$ degrees of freedom. Under the current non-

cut-point based model the statistic S_3 has the value $S_3 = 100.74$ with 702 degrees of freedom. As $\chi^2_{702} > \chi^2_{459} = 409.3804$, the value of S_3 indicates that the model is an appropriate fit to the data.

In chapters 3 and 4 we concluded that the probit model and cumulative logit model respectively, are both appropriate models explaining severity of diabetic retinopathy for the diabetes data. Further, we decided in chapter 4 that the cumulative logit model provided an improved fit over the probit model since $S_1 = 161.78$ is greater than $S_2 = 126.55$. Now, we observe that the value of the test statistic $S_3 = 100.74$ for the non-cut-point based multinomial logistic model is less than the values of both of the statistics for probit model ($S_1 = 161.78$) and the cumulative logit model ($S_2 = 126.55$). Consequently, the non-cut-point based multinomial logistic model provides the best fit to the data among the three competitors.

5.2.2 Display of Squared Error Distances

To gain further insight regarding the fit of the non-cut-point based multinomial logistic model to the data, a graphical display of the squared error distances between the model based proportions and the observed proportions is shown in Figure 5.1. In the manner similar to that of chapters 3 and 4, these distances for the ith (i=1, ..., N) individual are calculated by $d_i^* = \sum_{h=1}^{M} (p_{ih}^* - p_{iho})^2$. For every $i = 1, \ldots, 720$ we expect to observe values of d_i^* close to zero if the proposed model is providing a good fit to the data.

It is clear from Figure 5.1, that for a large number of individuals, the value of their squared error distances are very close to zero. Thus the model based estimated proportions, for an individual to fall in one of the four ordered

categories are in agreement with the observed proportions for an individual to fall into one of the four categories. This verifies the adequacy of the non-cut-point based multinomial logistic model in fitting the diabetes data.

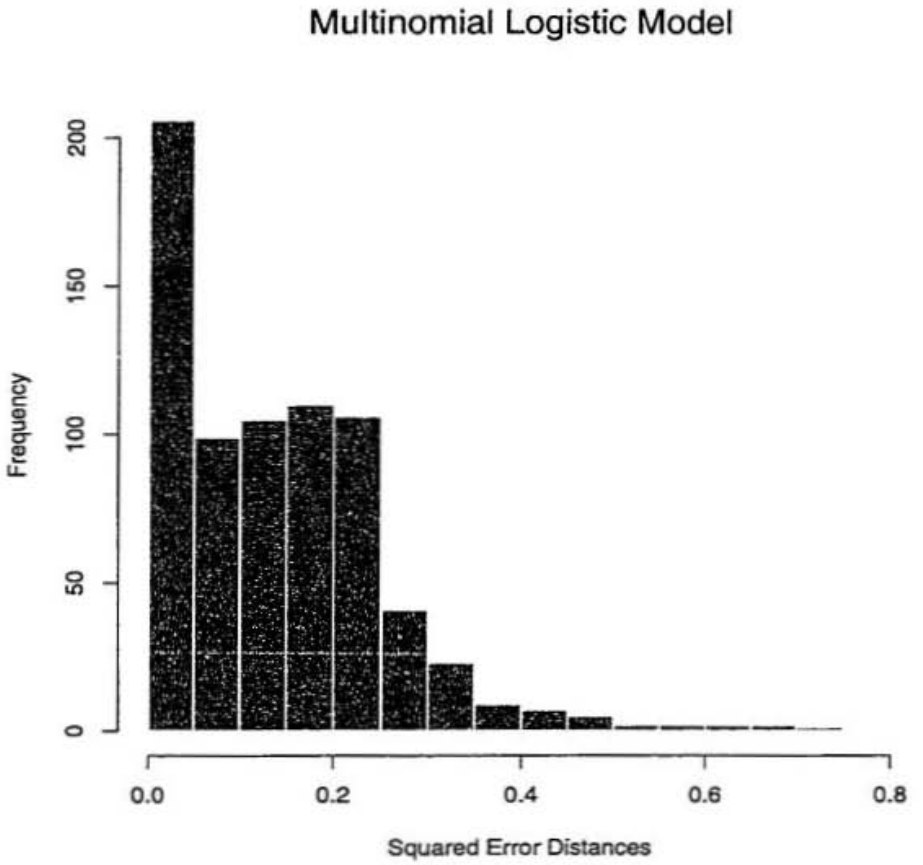

Figure 5.1: Display of Squared Error Distances for the Multinomial Logistic Model

Furthermore, when this histogram is compared with the histograms in Figure 3.1 and Figure 4.2 it is clear that Figure 5.1 exhibits the largest

number of squared error distances close to zero, along with the smallest number of large values of squared error distances. Consequently, the non-cut-point based multinomial logistic model fits the data best as compared to it's other two competitors.

5.3 Fitting a Reduced Model

As it was discussed in the previous section, there was no convergent solution for β_{16} which is the effect of the sixth covariate under the 'none' category. A justification for this non-convergence problem was also provided in the same section. As the values of the sixth covariate never varied under the 'none' category, it was natural to explore the convergence problem. As a remedy, in this section, we consider modeling severity of diabetic retinopathy based on one less covariate, that is, the covariate x_6 (right eye macular edema) will be omitted as we fit the non-cut-point based multinomial logistic model to the data. Further, we will give the final estimates for this reduced model and investigate the goodness of fit of this model to the data. A comparison will then be made only between the non-cut-point based multinomial logistic model based on six covariates and the non-cut-point based multinomial logistic model based on five covariates.

As we now have $p = 5$ covariates, we require the estimation of $p \times (M - 1) = 5 \times 3 = 15$ regression coefficients denoted by $\beta^{**} = (\beta_1^{**T}, \ldots, \beta_h^{**T}, \ldots, \beta_{M-1}^{**T})$ with $M = 4$ categories. We exploit the same methods provided in section 5.1 to obtain the estimates for these 15 unknown β parameters. For the purpose recall equation (5.6), and compute the 15 values using the iterative

equation (5.9). The only difference is that in (5.9), we computed 18 values of β denoted by β^* and here we denote the 15 estimates of the components of β by β^{**}. Note that, the dimension of D_i matrix was adjusted to reflect this change in the number of regression parameters. Also, the formulas appropriate for the V_i matrices were adjusted.

Convergence was obtained for all 15 regression parameters in six iterations. These estimates along with their standard errors are given in Table 5.2. It is clear from Table 5.2 that these estimates and their standard errors are very close to the estimates and standard errors (displayed in Table 5.1) for the non-cut-point based multinomial logistic model based on six covariates. More specifically, within each of the ordered categories, the covariates that were deemed non-influential for the analysis based on six covariates, appear to maintain their patterns for the analysis based on five covariates.

Similar to the analysis provided for the non-cut-point based model for all six covariates, a goodness of fit statistic is provided and a graphical display of the squared error distances is displayed in Figure 5.2 for the model based on five covariates. The goodness of fit statistic for the reduced model denoted by S_3^*, was evaluated as $S_3^* = 94.97$ which is is found to be less than $S_3 = 100.74$. Note that since a model with more covariates is expected to produce a smaller value for the goodness of fit statistic, the value of $S_3 = 100.74 > S_3^* = 94.97$ appears to indicate a problem with the larger model which we already explained in the last section as a possible effect of the sixth covariate in general.

The histogram of the squared error distances in Figure 5.2 appears to be quite similar to that of Figure 5.1. The only difference is that in Figure 5.1,

there appear to be a few squared error distances which are large in magnitude, which may be a result of the non-convergence problem of the sixth covariate.

Table 5.2: Non-Cut-Point Based Multinomial Logistic Model Estimates (Reduced Model)

Category	Parameter	Estimate	Standard errors
NONE	β_1	-0.2881	0.0767
	β_2	0.0750	0.0606
	β_3	0.0241	0.0069
	β_4	-2.4745	0.4495
	β_5	1.0712	0.3343
MILD	β_1	-0.0468	0.0185
	β_2	0.1258	0.0564
	β_3	0.0022	0.0062
	β_4	-1.7387	0.3644
	β_5	0.7548	0.3084
MODERATE	β_1	0.0069	0.0187
	β_2	0.1592	0.0590
	β_3	-0.0040	0.0065
	β_4	-1.1901	0.3827
	β_5	-0.1638	0.3290

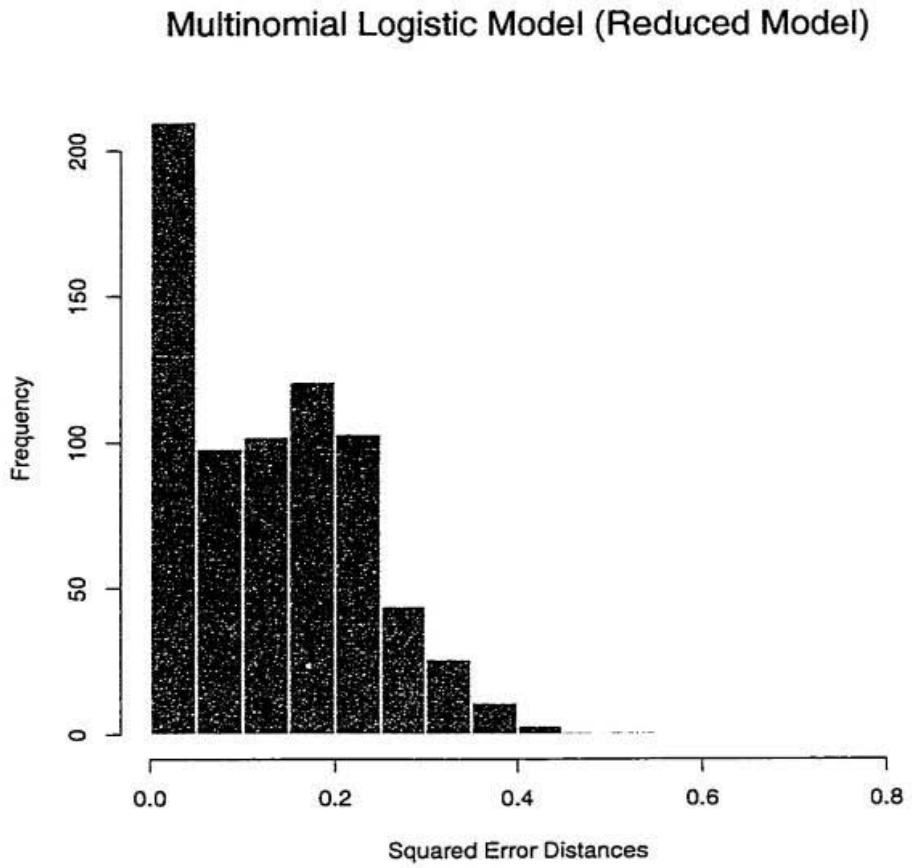

Figure 5.2: Display of Squared Error Distances for the Multinomial Logistic Model (Reduced Model)

Chapter 6

Concluding Remarks

Many studies in the scientific field involve analyzing multinomial ordinal data. Such studies involve investigating the effect of covariates on ordinal responses. When conducting an analysis on multinomial ordinal data a common problem arises in selecting a model that can adequately distinguish the ordered responses. A standard practice is to utilize models that implement parameters known as cut-points to distinguish the adjacent ordered categories (responses).

In chapters 3 and 4, we discussed in details two such models that require the inclusion of cut-point parameters namely the probit model and the cumulative logit model respectively. A serious problem one faces when implementing cut-point based procedures is that the estimates of the cut-points of such models must follow an order restriction. More specifically, if the cut-points are thought to be in increasing order, then it is required that the cut-point estimate that distinguishes category 1 form category 2 must be less than the cut-point estimate that distinguishes category 2 from

category 3 and so on. But the existing non-restricted estimation procedures used to estimate the restricted parameters of the probit and logit models do not guarantee that the order restriction will be maintained. Although, in this practicum we were able to successfully estimate these cut-point parameters, it was however not without any difficulty. The successful estimation of these parameters relied heavily on initial estimates which were already close to maximizing the likelihood surface. Further, with regard to the cumulative logit model, additional measures including a small simulation were conducted to find suitable initial estimates, which is expensive as it required extra efforts. Moreover, in general, there is no guarantee that such searches will always be successful.

In chapter 5, we looked for a suitable resolution to this cut-point problem. An obvious remedy was to find a model which does not rely on any cut-points to distinguish the adjacent ordered categories. One such model is the non-cut-point based multinomial logistic model proposed by Sutradhar and Kovacevic (2000) [see also Das and Sutradhar (1999)]. This model uses a logistic cumulative probability model, where probability depends only on the regression parameters. More specifically, each category is described by its own set of regression coefficients, as opposed to the other approaches which had only one set of regression coefficients.

It is not enough to explore models based on the ease of their computations, but to find a model which provides an adequate fit to the data as well. Therefore, in the chapters we investigated the goodness of fit of the different models to the diabetes data. It was concluded that among the three models described in the practicum, the non-cut-point based multino-

mial logistic model provides the best fit to the diabetes data. The goodness of fit was measured by a suitable statistic based on observed and expected proportions.

Note that the ordinal analysis presented in the practicum requires complete information on the covariates of all respondents. It may, however, be the case that information on some or all covariates may be missing for some respondents. For these types of missing information cases, one needs to develop suitable methodology in addition to taking care of the cut-points. This problem requires further investigation which is beyond the scope of the present practicum.

Bibliography

[1] Agresti, A. (1990), *Categorical Data Analysis,* John Wiley & Sons, New York.

[2] Aitchison, J., and Silvey, S. D. (1957), "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika,* 44, 131-140.

[3] Ashford, J. R., and Sowden, R. R. (1970), "Multivariate Probit Analysis" *Biometrics,* 26, 535-546.

[4] Das, K., and Sutradhar, B. C., (1999), " Analyzing Bivariate Polytomous Data: A Marginal Multinomial Logistic Approach," *Submitted for publication.*

[5] Gurland, J., Lee, I., and Dahm, P. A. (1960), "Polychotomous Quantal Response in Biological Assay," *Biometrics,* 16, 382-398.

[6] Harville, D. A., and Mee, R. W. (1984), "A Mixed-Model Procedure for Analyzing Ordered Categorical Data," *Biometrics,* 40, 393-408.

[7] Kim, K. (1995), "A Bivariate Cumulative Probit Regression Model for Ordered Categorical Data," *Statistics in Medicine,* 14, 1341-1352.

[8] McCullagh, P. (1980), "Regression Models for Ordinal Data (with discussion)," *J. Roy. Statist. Soc.*, B42, 109-142.

[9] McCullagh, P. (1983), "Quasi-likelihood Functions," *Ann. Statist.*, 11, 59-67

[10] Miller, M. E., Davis, C. S., and Landis, J. R. (1993), "The Analysis of Longitudinal Polytomous Data: Generalized Estimating Equations and Connections With Weighted Least Squares," *Biometrics,* 49, 1033-1044.

[11] Stram, D. O., Wei, L. J., and Ware, J. H. (1988), "Analysis of Repeated Ordered Categorical Outcomes with Possibly Missing Observations and Time-Dependent Covariates," *J. Amer. Statist. Assoc.*, 83, 631-637.

[12] Sutradhar, B. C., and Kovacevic, M., (2000), "Analyzing Ordinal Lognitudial Survey Data: Generalized Estimating Equations Approach," *To appear in Biometrics.*

[13] Walker, S. H., and Duncan, D. B. (1967), "Estimation of the Probability of an Event as a Function of Several Independent Variables," *Biometrika,* 54, 167-179.

[14] Williams, O. D., and Grizzle (1972), "Analysis of Contingency Tables having Ordered Response Categories," *J. Amer. Statist. Assoc.*, 67, 55-63.

[15] Williamson, J. M., Kim, K., and Lipsitz (1995) "Analysing Bivariate Ordinal Data Using a Global Odds Ratio," *J. Amer. Statist. Assoc.*, 90, 1432-1437.

Appendix A

Graphs

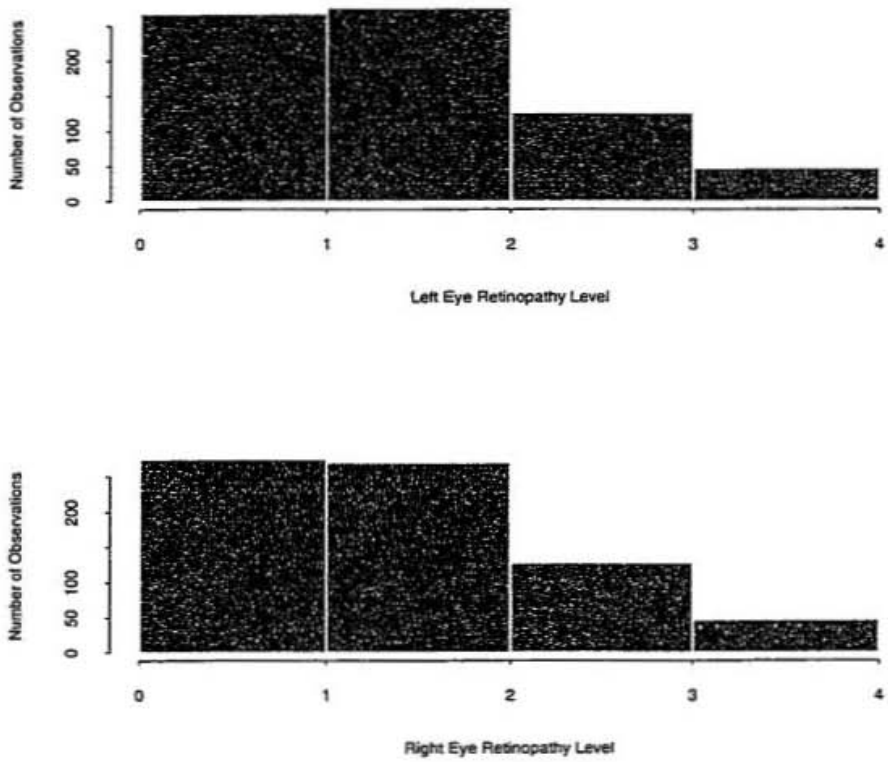

Figure A.1: Histogram of the Distribution of the Response Variable Left Eye and Right Eye Retinopathy Levels

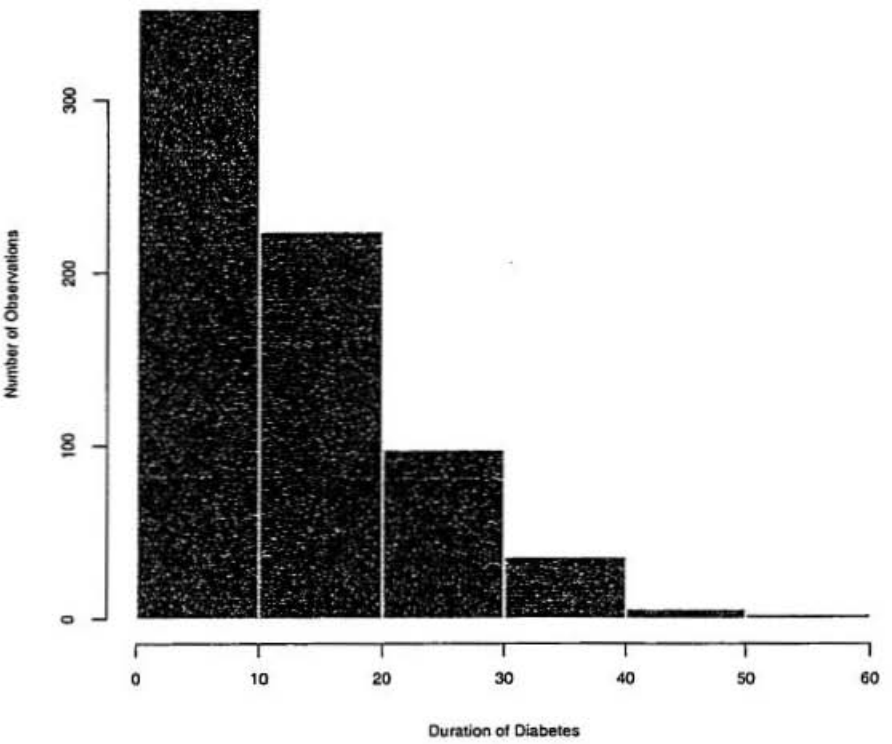

Figure A.2: Histogram of the Distribution of the Covariate Duration of Diabetes

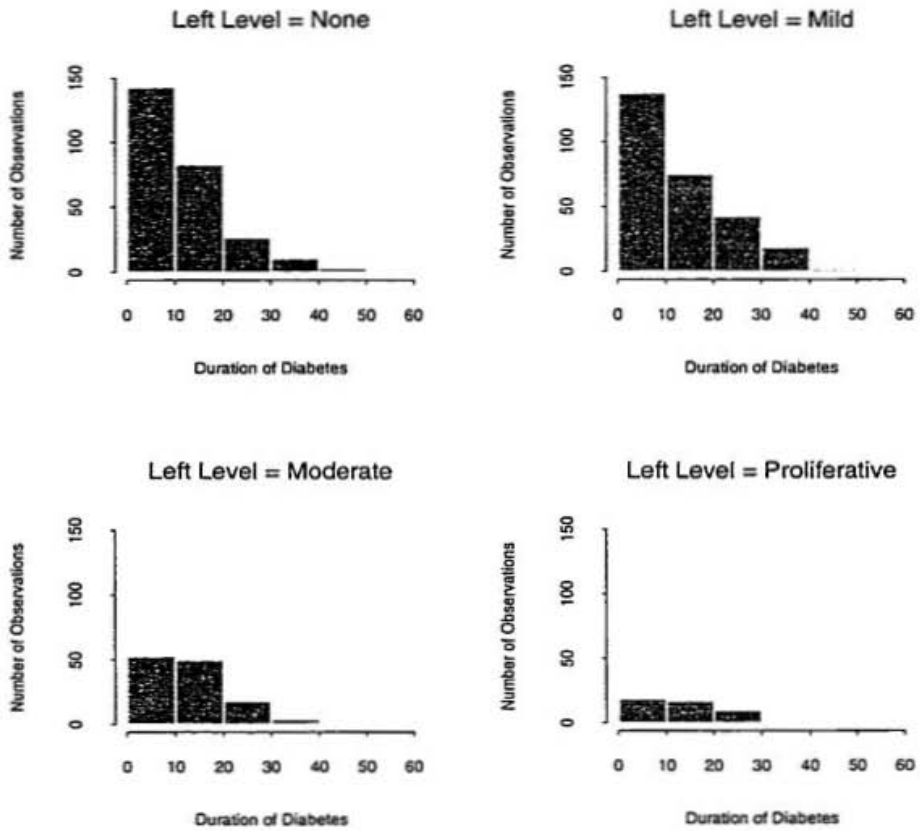

Figure A.3: Histogram of the Distribution of the Covariate Duration of Diabetes within each of the ordered categories for the Left Eye

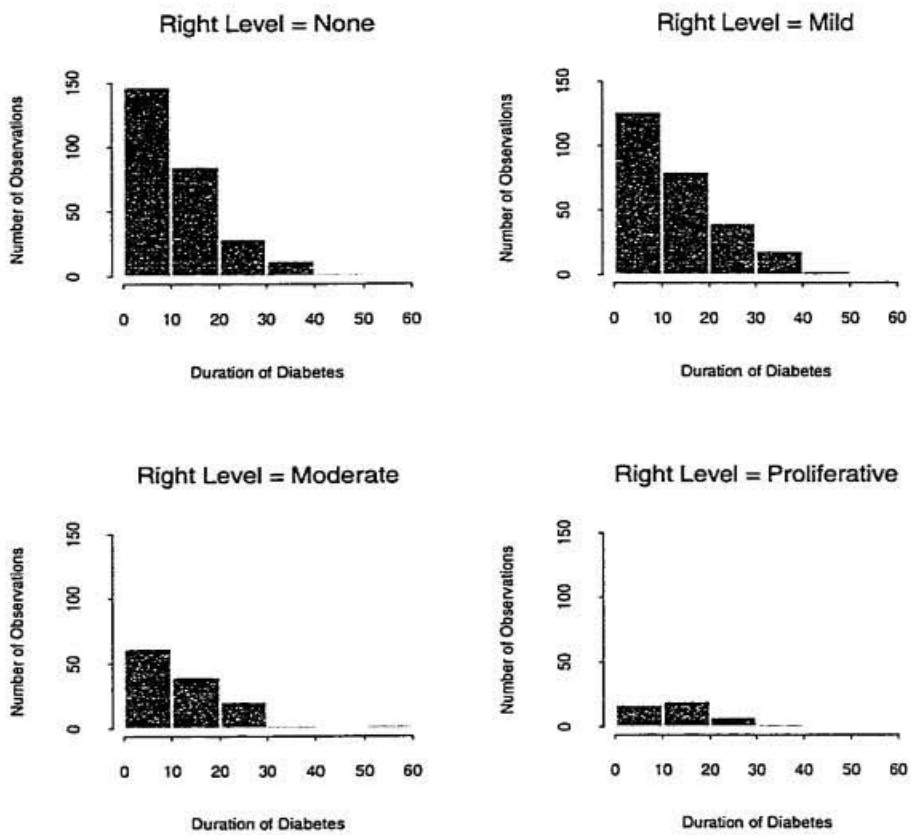

Figure A.4: Histogram of the Distribution of the Covariate Duration of Diabetes within each of the ordered categories for the Right Eye

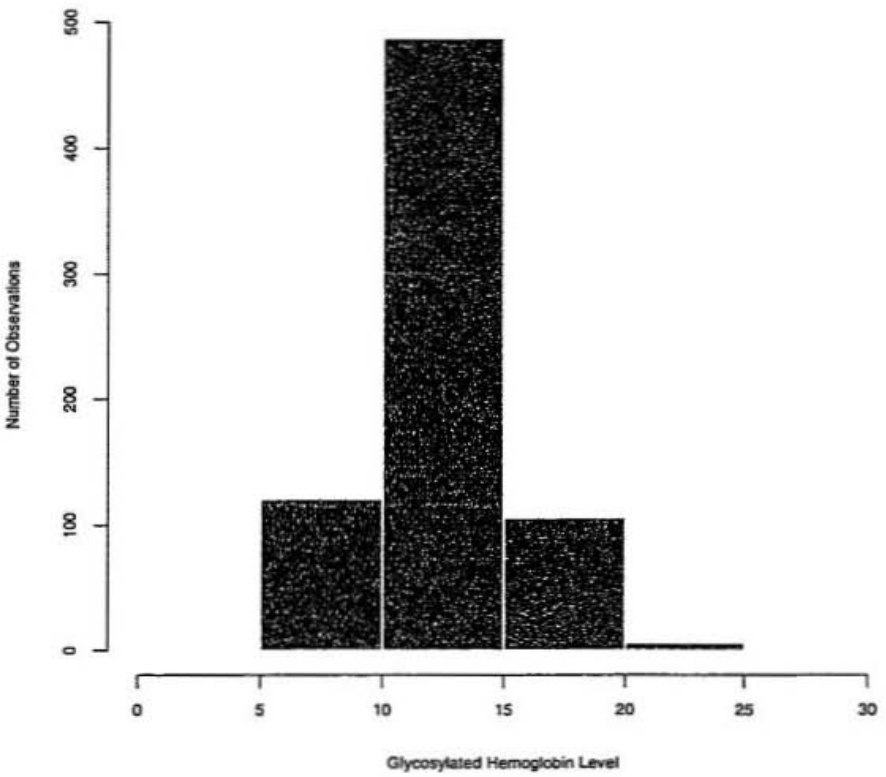

Figure A.5: Histogram of the Distribution of the Covariate Glycosylated Hemoglobin Level

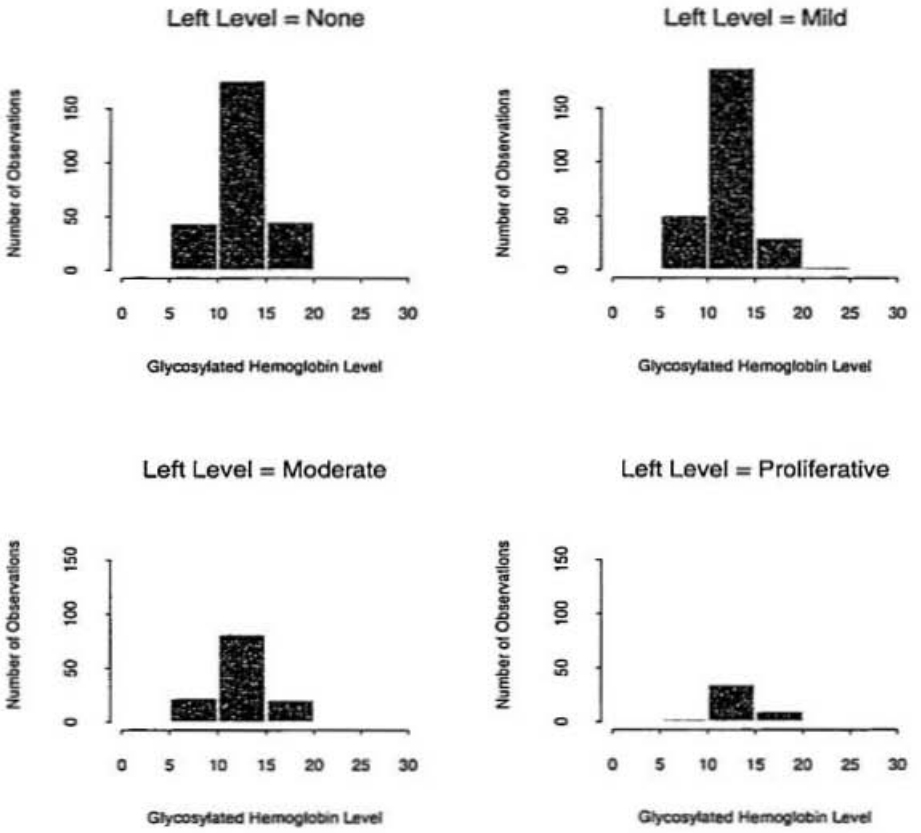

Figure A.6: Histogram of the Distribution of the Covariate Glycosylated Hemoglobin Level within each of the ordered categories for the Left Eye

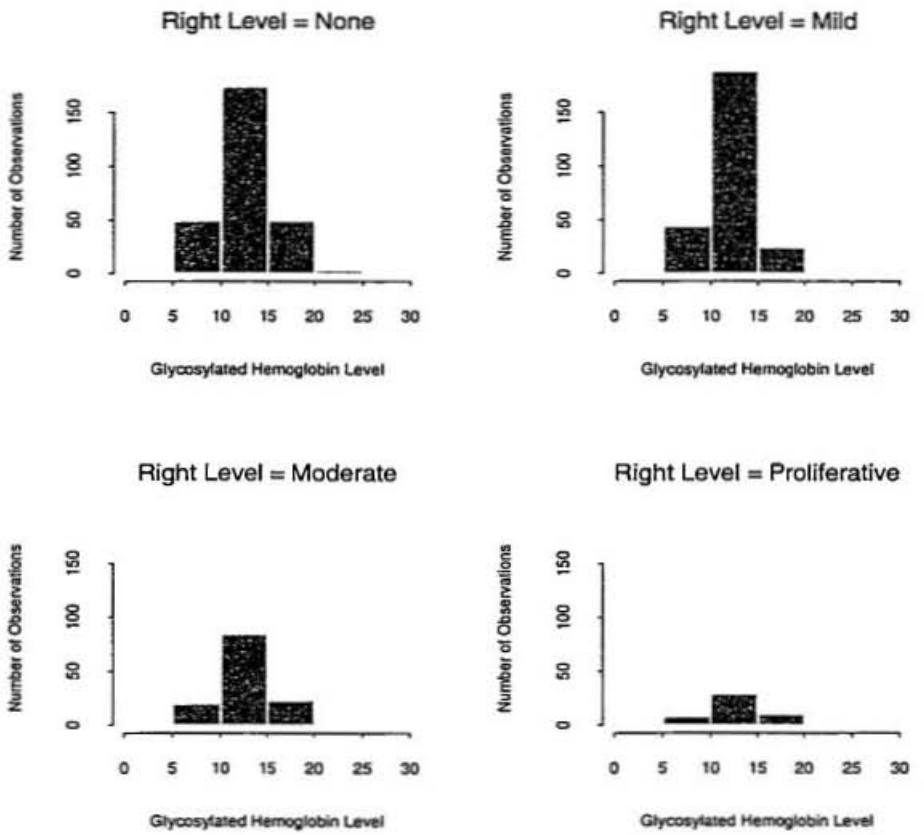

Figure A.7: Histogram of the Distribution of the Covariate Glycosylated Hemoglobin Level within each of the ordered categories for the Right Eye

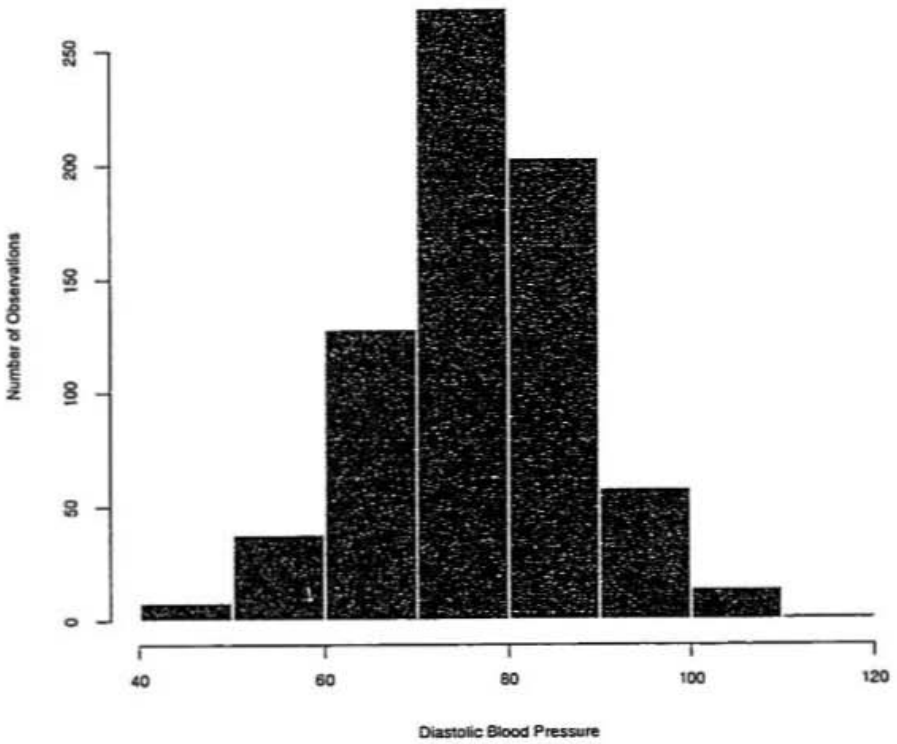

Figure A.8: Histogram of the Distribution of the Covariate Diastolic Blood Pressure

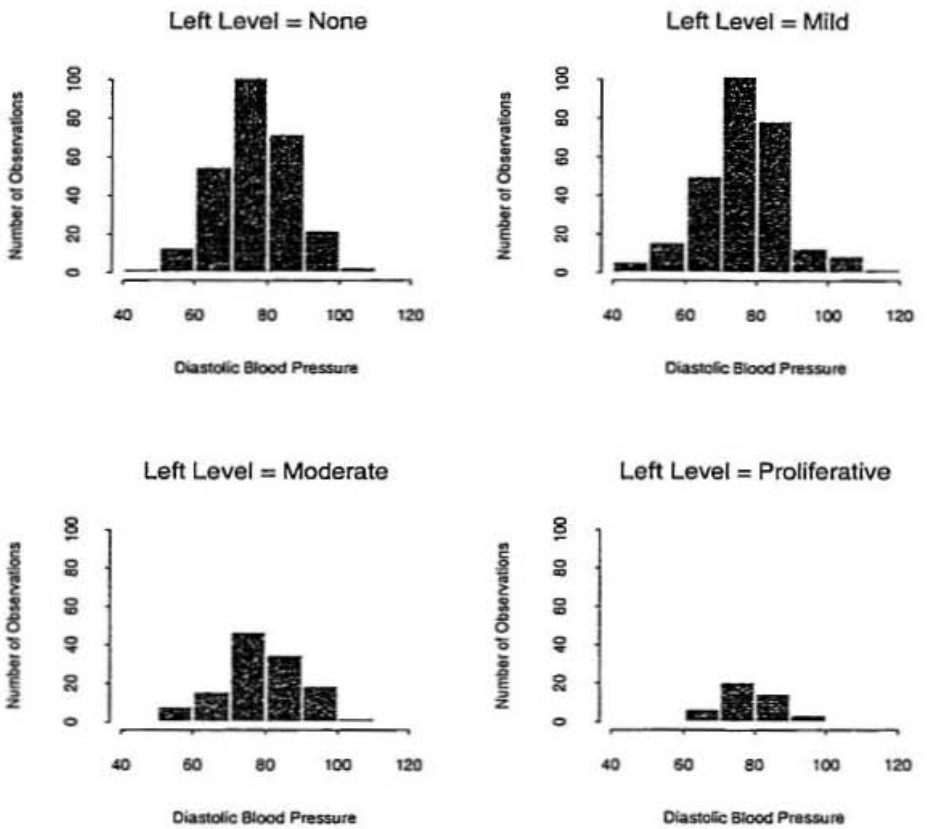

Figure A.9: Histogram of the Distribution of the Covariate Diastolic Blood Pressure within each of the ordered categories for the Left Eye

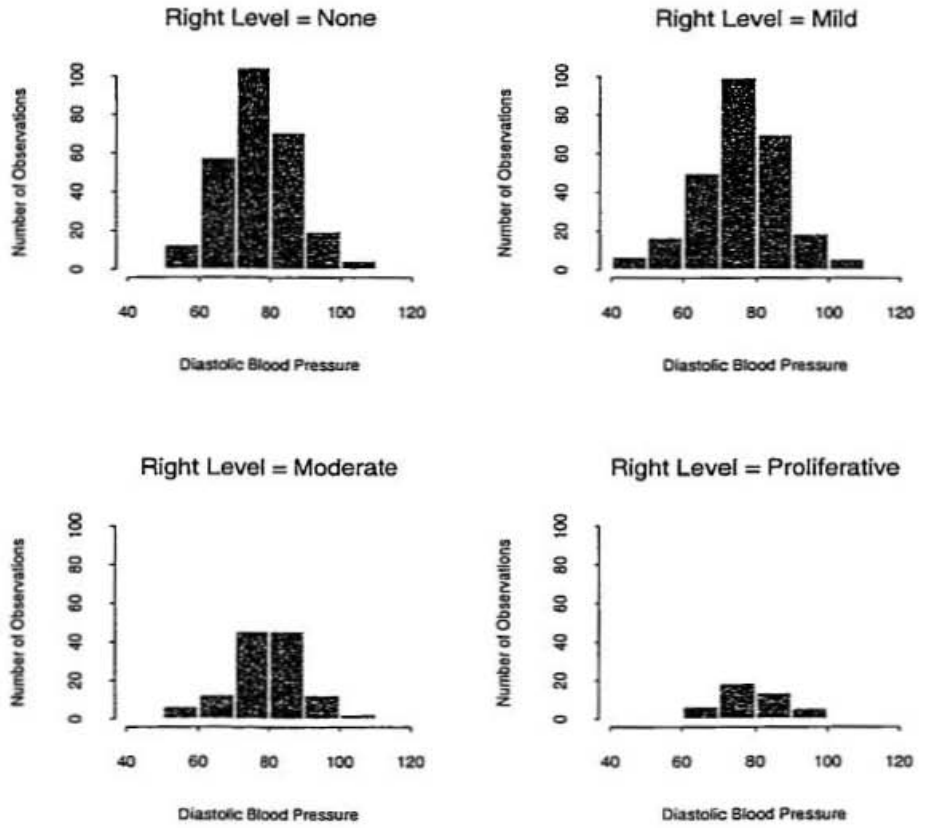

Figure A.10: Histogram of the Distribution of the Covariate Diastolic Blood Pressure within each of the ordered categories for the Right Eye

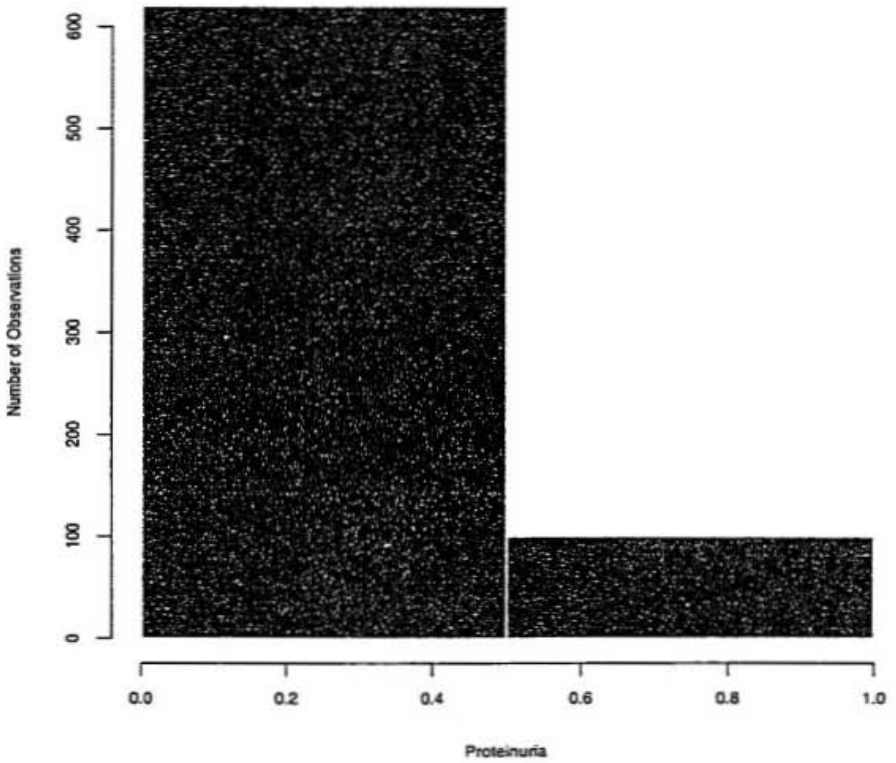

Figure A.11: Histogram of the Distribution of the Covariate Proteinuria

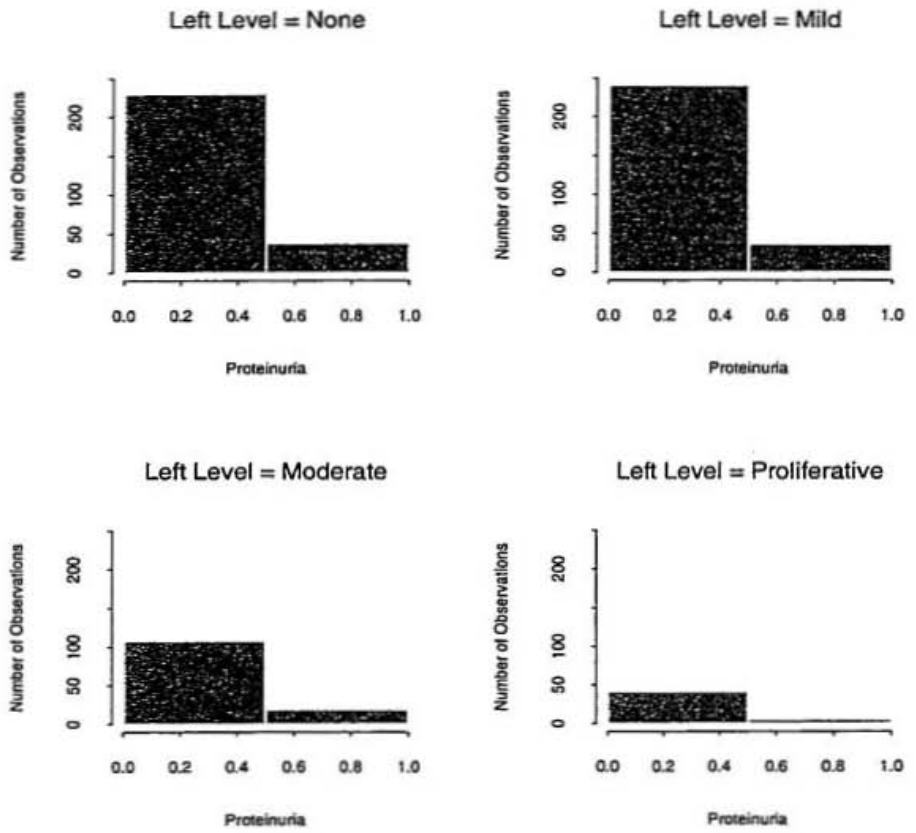

Figure A.12: Histogram of the Distribution of the Covariate Proteinuria within each of the ordered categories for the Left Eye

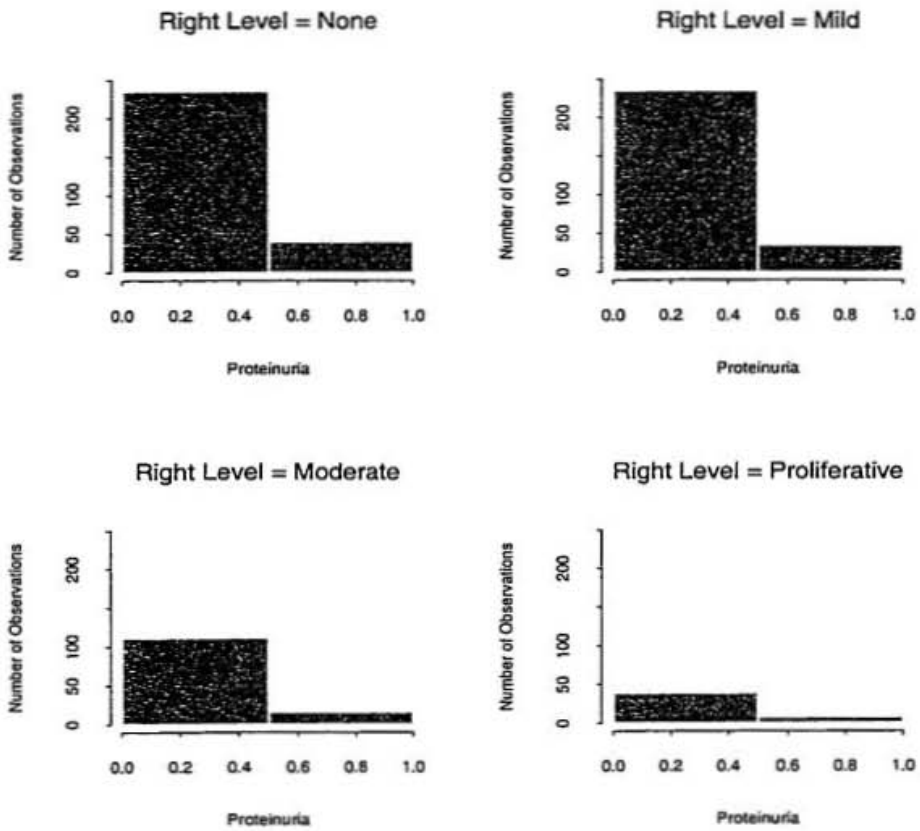

Figure A.13: Histogram of the Distribution of the Covariate Proteinuria within each of the ordered categories for the Right Eye

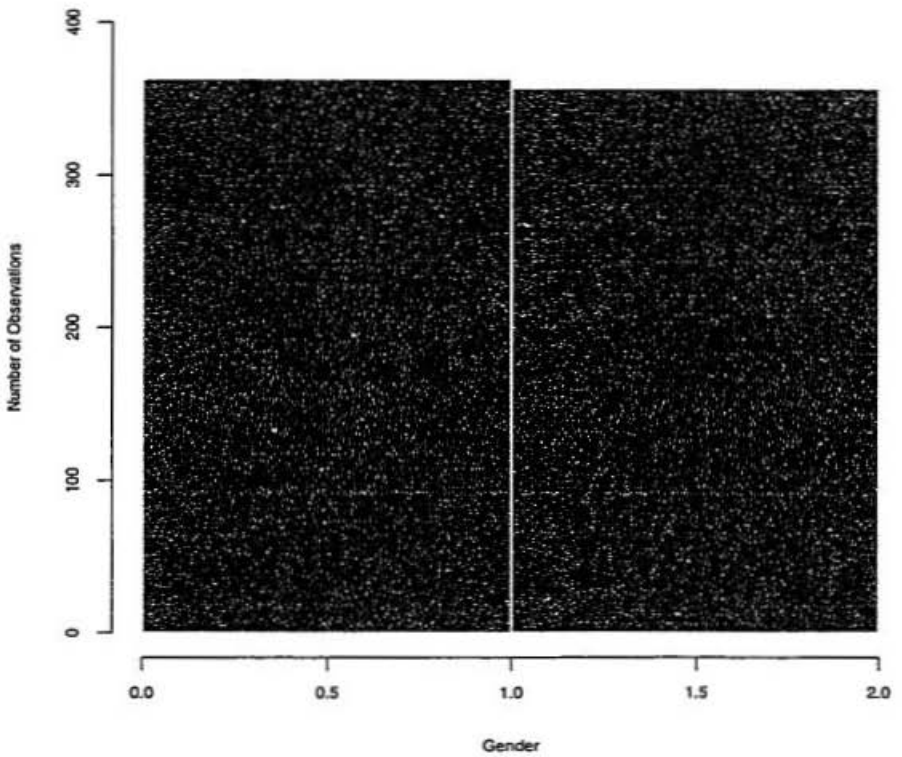

Figure A.14: Histogram of the Distribution of the Covariate Gender

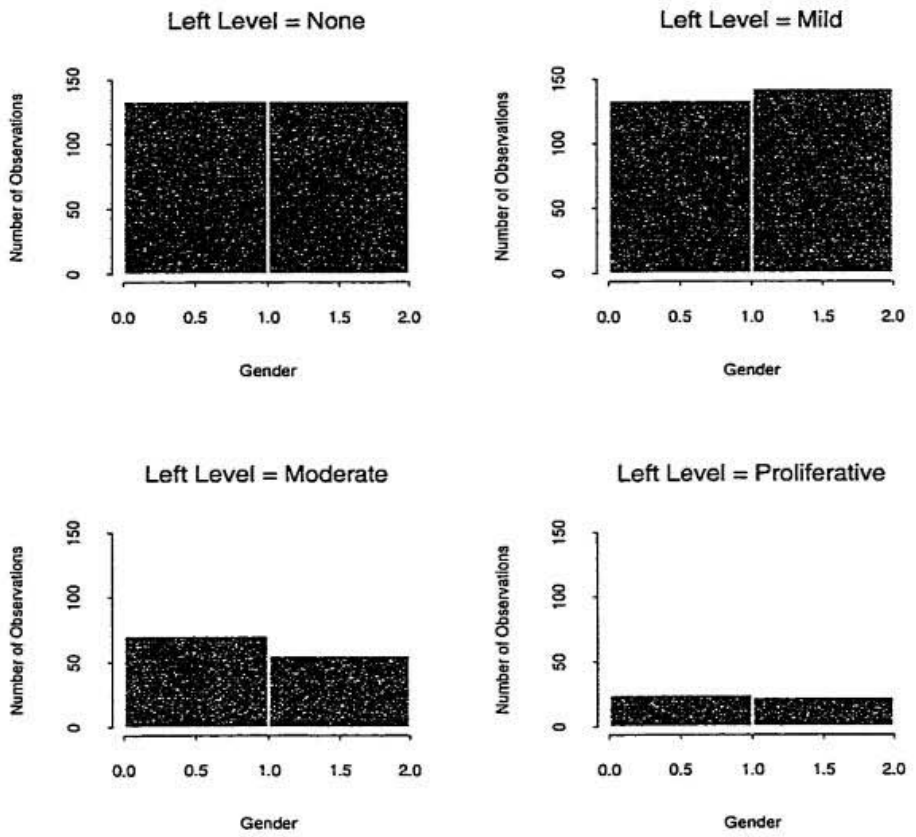

Figure A.15: Histogram of the Distribution of the Covariate Gender within each of the ordered categories for the Left Eye

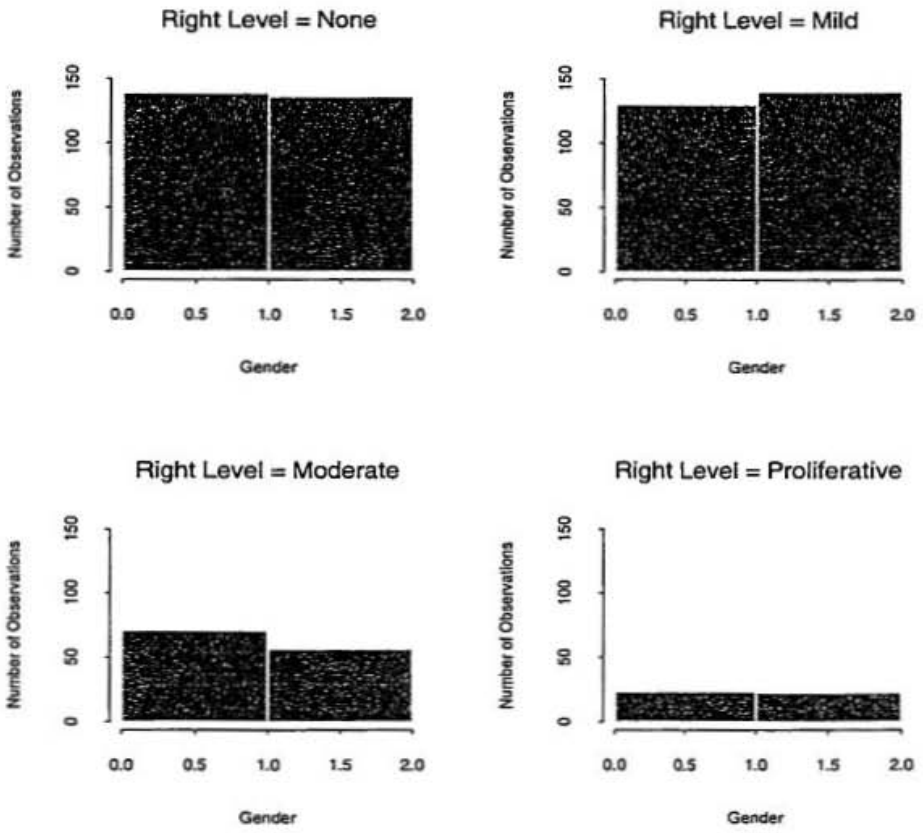

Figure A.16: Histogram of the Distribution of the Covariate Gender within each of the ordered categories for the Right Eye

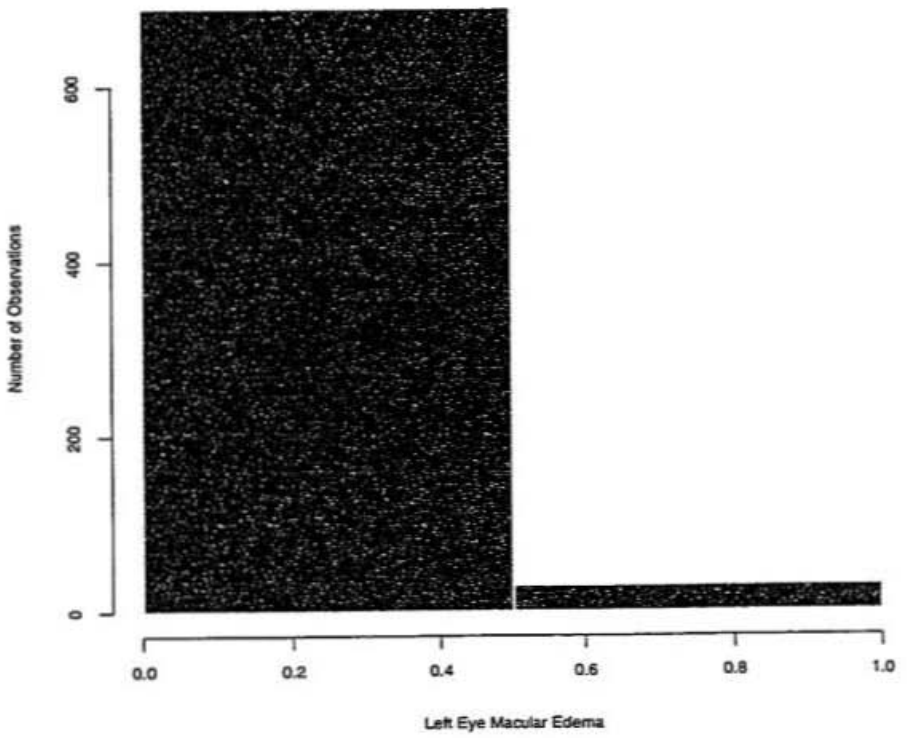

Figure A.17: Histogram of the Distribution of the Covariate Left Eye Macular Edema

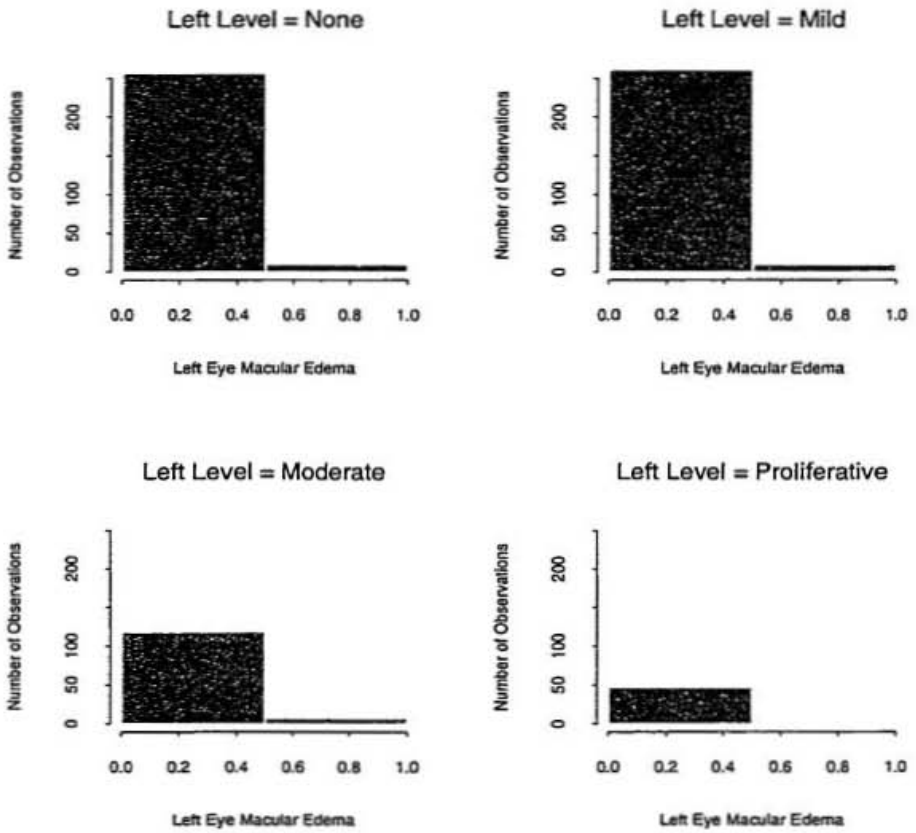

Figure A.18: Histogram of the Distribution of the Covariate Left Eye Macular Edema within each of the ordered categories

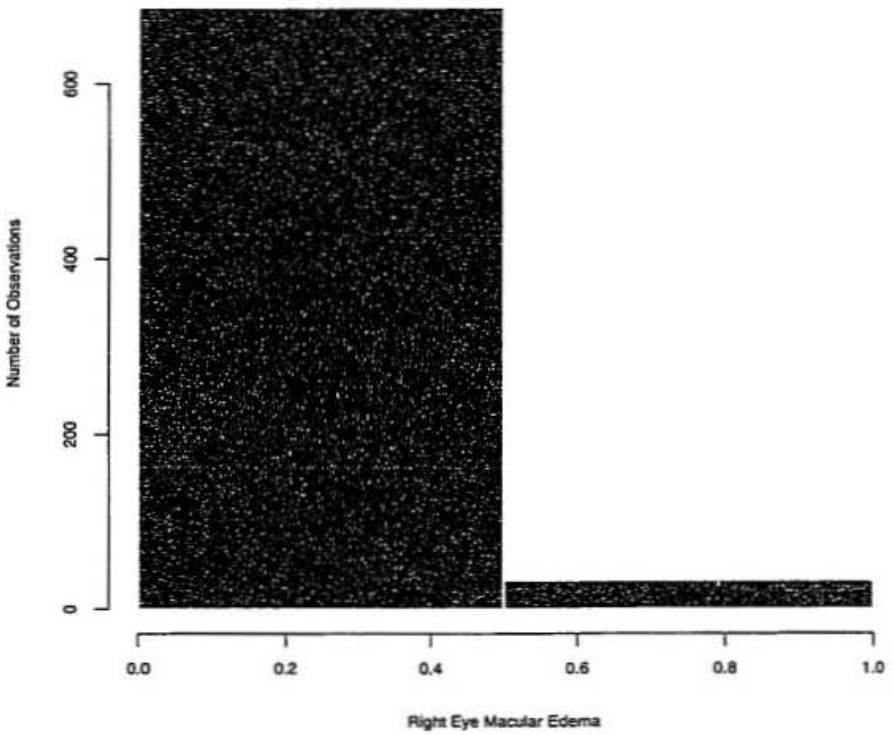

Figure A.19: Histogram of the Distribution of the Covariate Right Eye Macular Edema

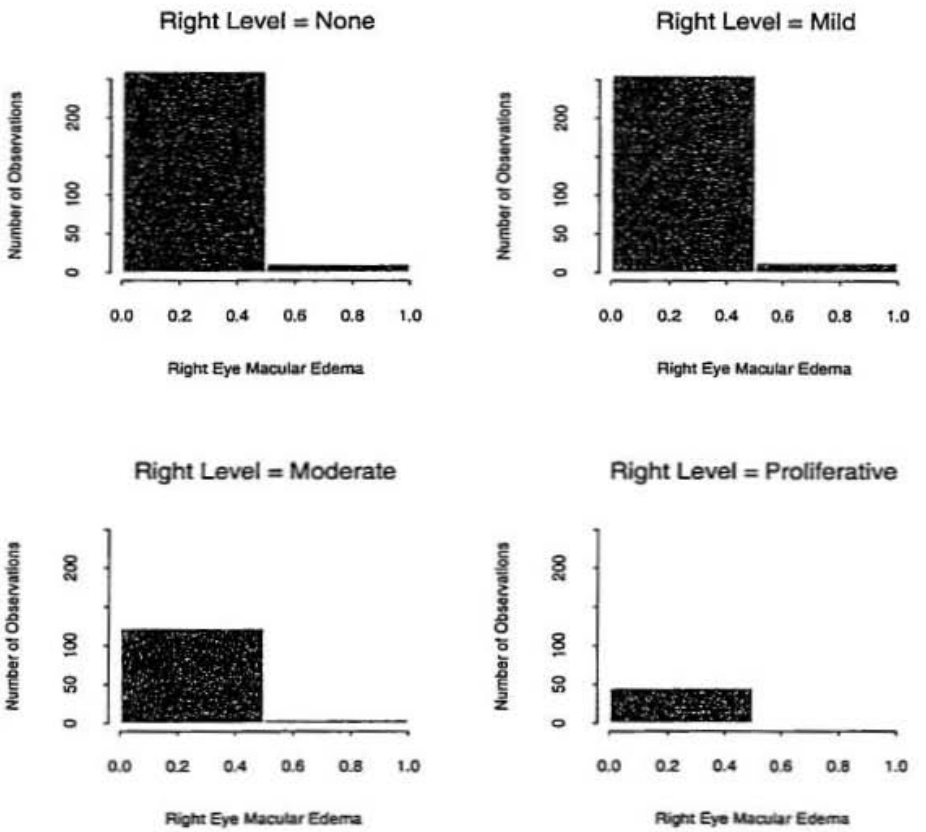

Figure A.20: Histogram of the Distribution of the Covariate Right Eye Macular Edema within each of the ordered categories