BAYESIAN MODEL CHECKING IN MULTIVARIATE DISCRETE
REGRESSION PROBLEMS

Fanglong Dong

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2008

Committee:

James H. Albert, Advisor

Madhu Rao,
Graduate Faculty Representative

Hanfeng Chen

John. T. Chen
ABSTRACT

James H. Albert, Advisor

Ordinal data are common in the academic area such as a student grade, A, B, C, D, or F, also ordinal data are common in other area such as customer satisfaction survey. It is straightforward to fit a regression model to reflect the relationship between the response and the predictors. Since the response in an ordinal data set is a vector, it is not clear how the traditional statistics define residuals and detect outliers because of the dimension of response. Since the introduction of latent variable, we can model the data using the latent variable and we have a new type of residual called latent residual. With the help of introduction of latent variable into the model, it is easy to define residuals and detect outliers. In practice there are usually more than one predictor in the data set and we need to decide to choose variable that should be included in the model. We look at from a frequentist’s perspective and a Bayesian perspective. Also when we fit a model to a data set, we care about how well this model fit the data set, and we look from both a frequentist’s perspective and a Bayesian perspective. Usually methods from a frequentist’s perspective rely on the asymptotic distribution to draw a conclusion and sometime this will become a problem especially when the sample size is small, on the contrary, methods from a Bayesian perspective use simulation and thus it removes the reliance on the asymptotic distribution. Chapter 3 talks about methods for outlier detection problems and Chapter 4 talks about goodness-of-fit and model selection problems, in Chapter 5 we apply the methods from Chapter 3 and Chapter 4 to the BGSU student data set. Chapter 6 summarized the whole dissertation and possible future research interest and applications.
ACKNOWLEDGMENTS

First of all, I would like to thank Dr. Jim Albert, my advisor, for his constant support and many suggestions throughout this research. I also wish to thank him for the friendship and all the advice and encouragement that he shared about life in general.

I also want to extend my gratitude to the other members of my committee, Dr. John Chen, Dr. Hanfeng Chen, and Dr. Madhu Rao, for their time and advice.

I am grateful to the department of Mathematics and Statistics for all the support and for providing a wonderful research environment. I especially wish to thank Marcia Seubert, Cyndi Patterson, and Mary Busdeker for all their help. The dissertation fellowship for the period 2007-2008 was very important to the completion of this work.

I wish to thank my colleagues and friends from Bowling Green, Laura, Jasmine, Ann-Gee, Ai, and all the other people for all the fun and supports.

Finally, I thank my beloved wife, Qiali, for her full support, encouragement, love, and patience, and our coming baby Catherine for bringing all the joy and happiness in our lives during our stay in Bowling Green. Without them this work could never have come to existence.

Fanglong Dong
Bowling Green, Ohio.
December, 2008
## Table of Contents

**CHAPTER 1: Overview**  
1.1 Ordinal Data and Ordinal Regression ................................................. 1  
1.2 Binary Data and Binary Regression Models ........................................... 4  
  1.2.1 Models for Binary Data ............................................................ 5  
1.3 Estimating Binary Regression Coefficient ............................................. 8  
  1.3.1 Likelihood for Binary Data ....................................................... 8  
  1.3.2 Maximum Likelihood Estimate for Binary Data ............................... 9  
1.4 Binary Regression Model Diagnostics .................................................. 9  
  1.4.1 Pearson Residual for Binary Data .............................................. 9  
  1.4.2 Deviance Residual for Binary Data .......................................... 10  
  1.4.3 Adjusted Deviance Residual for Binary Data ............................... 11  
1.5 Goodness-of-fit and Model Selection for Binary Data .......................... 11  
1.6 Models for Ordinal Data ....................................................................... 13  
1.7 Classical Fitting for Ordinal Regression Model ..................................... 18  
  1.7.1 The Likelihood Function for Ordinal Data .................................... 18  
  1.7.2 Maximum Likelihood Estimate for Ordinal Data ........................... 20  
1.8 Ordinal Regression Model Diagnostics ................................................ 20  
  1.8.1 The M Test for Ordinal Data to Detect Outliers ............................ 21  
  1.8.2 Locally Best Invariant Test to Detect the Existence of Outliers ....... 22  
  1.8.3 Using Outlier Score to Detect Outliers ...................................... 22
## CHAPTER 2: Bayesian Models for Ordinal Data

1. **Bayes's Rule**
   - **Prior Predictive Distribution**
   - **Posterior Predictive Distribution**
2. **Latent Variables**
3. **Likelihood Function Using Latent Variable**
4. **Prior Distribution**
   - **Noninformative Prior**
   - **Informative Prior**
5. **Review of Methods of Simulating from the Posterior Distribution**
   - **Metropolis-Hastings Algorithm**
   - **Independent Chain Algorithm**
   - **Random Walk Algorithm**
   - **Gibbs Sampling Algorithm**
6. **Parameter Estimation for Ordinal Data**
   - **Traditional Method for Parameter Estimation**
   - **MCMC Sampling from the Posterior Distribution**
7. **Model Selection from a Bayesian Perspective**
8. **Residuals from a Bayesian Perspective**
9. **Goodness-of-fit from a Bayesian Perspective**

## CHAPTER 3: Outlier Detection

1. **A Frequentist’s Method to Detect Outliers**
2. **A Bayesian Perspective to Detect Outliers**
   - **Latent Residuals for Ordinal Regression Model**
5.4 Goodness-of-fit for the Model

5.4.1 Goodness-of-fit from a Frequentist’s Perspective

5.4.2 Goodness-of-fit from a Bayesian Perspective

5.5 Summary

CHAPTER 6: Summary and Conclusions

Appendices

CHAPTER A: Part of R Code for Chapter 3

A.1 Ordinal.probit Program

A.2 Probitpost Program

A.3 Use Bayes Factor to Detect Outliers R Code

A.4 Individual Deviance to Detect Outliers R Code

A.5 Posterior Predictive Distribution of $D_i(y^\text{rep})$ to Detect Outliers R Code

A.6 Parameter-dependent $D_i(y^\text{rep}, \beta)$ to Detect Outliers R Code

A.7 Ordinal.probit.hierarchical R Code

CHAPTER B: Part of R Code for Chapter 4

B.1 Using Deviance as a Measure of GOF R Code

B.2 Using Deviance $D(y^\text{rep})$ as a Measure of GOF R Code

B.3 Using Parameter-dependent Deviance as a Measure of GOF R Code

B.4 Pearson Chi Square Statistic as a Measure of GOF R Code

B.5 Posterior Predictive Distribution of $X^2(y^\text{rep})$ as a Measure of GOF R Code

B.6 Using Parameter-dependent $X^2(y^\text{rep}, \beta)$ as a Measure of GOF R Code

B.7 Probitpost_model1_ch4.K.1 R Code

B.8 Probit_model1_compare_ch4.K.100 R program

B.9 Probitpost_model2_ch4 R program

B.10 Probit_model2_compare_ch4.K.1000 R program
List of Figures

1.1 The plot of SAT score vs grade .................................................. 3
1.2 Comparison of three regression models ....................................... 7
1.3 Graph of fitted probabilities of getting grade for three regression models . . . 16
2.1 Latent variable interpretation of ordinal data ................................. 29
3.1 The barplot of fitted probabilities for all observations ...................... 47
3.2 The plot of individual deviance statistic $D(y)$ against observation ID . . . 50
3.3 The posterior mean of latent residuals against observation ID ........... 54
3.4 Using theoretical distribution of latent residuals to detect outliers ....... 56
3.5 Apply $3\sigma$ rule to latent residuals to detect outliers .................... 57
3.6 QQ plot to detect outliers for ordinal data .................................. 61
3.7 Using Baves factor to detect outliers ........................................ 63
3.8 Using individual deviance $D(y^{rep})$ to detect outliers .................. 65
3.9 Using $D(y^{rep}, \theta)$ to detect outliers ....................................... 68
3.10 Using hierarchical model to detect outliers for $v=2$ and 4 case ......... 70
3.11 Using hierarchical model to detect outliers for $v=10$ and 20 case ....... 71
4.1 Using $D(y^{rep})$ as a measure of goodness-of-fit .......................... 78
4.2 Scatterplot of deviance $D(y^{rep}, \theta)$ against $D(y, \theta)$ ................... 80
5.1 Scatterplot of proportion of Math 126 getting A against grade .......... 96
5.2 Scatterplot of proportion of Math 126 getting A against ACT math score . . . 97
5.3 Scatterplot of proportion of Math 126 getting A against placement score . . 98
5.4 The probability of getting A against variable ACT while fixing pgrade at 4 . 100
5.5 The probability of getting C against variable ACT while fixing pgrade at 4 . 101
5.6 The probability of getting A against variable pgrade . . . . . . . . . . . . . . 102
5.7 The probability of getting C against variable pgrade . . . . . . . . . . . . . . 103
5.8 The plot of individual deviance statistic $D_i(y)$ for BGSU data . . . . . . . . 108
5.9 Using $D(y^{rep})$ to detect outliers for BGSU data . . . . . . . . . . . . . . 109
5.10 Using $D(y^{rep}, \beta)$ to detect outliers for BGSU data . . . . . . . . . . . . . 110
5.11 The histogram of simulated deviance . . . . . . . . . . . . . . . . . . . . . 112
5.12 Using $D(y^{rep})$ as a measure of goodness-of-fit for BGSU data . . . . . . 113
5.13 Using $D(y^{rep}, beta)$ as a measure of goodness-of-fit for BGSU data . . . . 114
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The grade example data set</td>
<td>2</td>
</tr>
<tr>
<td>1.2</td>
<td>The analysis of deviance table for the binary data</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>The result of probit regression for ordinal data using \textit{polr} function</td>
<td>44</td>
</tr>
<tr>
<td>3.2</td>
<td>The fitted probability table for all students</td>
<td>46</td>
</tr>
<tr>
<td>3.3</td>
<td>Using M test in Fuchs and Kenett (1980) to detect outliers</td>
<td>49</td>
</tr>
<tr>
<td>4.1</td>
<td>Analysis of Deviance table for model selection</td>
<td>84</td>
</tr>
<tr>
<td>4.2</td>
<td>The influence of choice of K in model M5 on the Bayes factor</td>
<td>88</td>
</tr>
<tr>
<td>4.3</td>
<td>The influence of choice of K on the Bayes factor for model M7</td>
<td>90</td>
</tr>
<tr>
<td>5.1</td>
<td>The student Math 126 data set at Bowling Green State University</td>
<td>95</td>
</tr>
<tr>
<td>5.2</td>
<td>Probit Regression Model for the BGSU students</td>
<td>99</td>
</tr>
</tbody>
</table>
CHAPTER 1

Overview

1.1 Ordinal Data and Ordinal Regression

Ordinal data are frequently encountered in social science. A particular question on a survey may have the responses like “Strongly Agree”, “Agree”, “Neutral”, “Disagree” and “Strongly Disagree”. Another example of ordinal data is the grade that a student in a class may get, which is F, D, C, B, or A. The defining property of ordinal data is that there exists a clear ordering of the response categories, but no underlying interval scale between them. For example, it is generally reasonable to assume that a grade has an ordering of the form

\[ F < D < C < B < A \]

but it usually does not make sense to assign integer values to those categories. Thus, statements of this type

\[ A - B = D - F \]

usually do not hold.

To illustrate the new method from a Bayesian perspective and some possible issues that arise in modeling ordinal data using traditional methods, consider the grade example in Table 1.1.
<table>
<thead>
<tr>
<th>Student number</th>
<th>grade</th>
<th>SAT-M score</th>
<th>Grade in prerequisite probability course</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D</td>
<td>525</td>
<td>B</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>533</td>
<td>C</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>545</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>582</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>581</td>
<td>C</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>576</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>572</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>609</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>559</td>
<td>C</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>543</td>
<td>D</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>576</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>525</td>
<td>A</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>574</td>
<td>F</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>582</td>
<td>D</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>574</td>
<td>C</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>471</td>
<td>B</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>595</td>
<td>B</td>
</tr>
<tr>
<td>18</td>
<td>D</td>
<td>557</td>
<td>C</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>557</td>
<td>A</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>584</td>
<td>A</td>
</tr>
<tr>
<td>21</td>
<td>A</td>
<td>599</td>
<td>B</td>
</tr>
<tr>
<td>22</td>
<td>D</td>
<td>517</td>
<td>C</td>
</tr>
<tr>
<td>23</td>
<td>A</td>
<td>649</td>
<td>A</td>
</tr>
<tr>
<td>24</td>
<td>B</td>
<td>584</td>
<td>C</td>
</tr>
<tr>
<td>25</td>
<td>F</td>
<td>463</td>
<td>D</td>
</tr>
<tr>
<td>26</td>
<td>C</td>
<td>591</td>
<td>B</td>
</tr>
<tr>
<td>27</td>
<td>D</td>
<td>488</td>
<td>C</td>
</tr>
<tr>
<td>28</td>
<td>B</td>
<td>563</td>
<td>B</td>
</tr>
<tr>
<td>29</td>
<td>B</td>
<td>553</td>
<td>B</td>
</tr>
<tr>
<td>30</td>
<td>A</td>
<td>549</td>
<td>A</td>
</tr>
</tbody>
</table>

Table 1.1: The grade example data set. In this data set, the first column is the student number, the second column is the current grade that each student has, the third one is the SAT-M score and the last column is the grade in the prerequisite probability course.
Table 1.1 represents a general form of ordinal data using student grades in a statistics class as an example. The first column is the student number, the second column is the current grade in the class, which is an ordinal variable, the third one is the SAT-M score, and the last column is the grade in the prerequisite probability course.

In statistics, before we make a statistical analysis of a data set, it is always a good idea to make an exploratory graph. In this SAT grade example data set, the SAT score ranges from 463 to 649, we manually divide them into 4 categories, which is from 400-500, 501-550, 551-600, 601-650, and count the proportions of A, B, C, D, and F’s in each SAT grade interval. I then plot a barplot of the proportions in each SAT grade interval. I get the Figure 1.1.

Figure 1.1: The plot of SAT score vs grade. The x-axis is the SAT score interval, where 475 represents grade interval 450-500, 525 represents grade interval 501-550, 575 represents grade interval 551-600, and 625 represents grade interval 601-650. The y-axis is the proportions of each grade in each SAT grade interval.
From Figure 1.1, one can see that if SAT grade are higher, then the chance that one student gets a better grade is also getting higher. SAT score seems to have effects on the grade. One natural question would be: Is there is a model to describe the relationship between the student’s SAT grade and the probability that he gets A, B, C, D, or F? Suppose that model exists, should we include the variable SAT in the model? If we decide to include SAT in the model, What is the probability of getting an “A” if one student’s SAT grade is 550? Also notice there is another variable Grade in prerequisite probability course, should we include this variable Grade in prerequisite probability course in the model too? In addition, there are students like observation number 19 who gets a high SAT score (SAT=577) but get a low grade (Observation 19 gets a F), are they consistent with the hypothetical model? How to identify observations like observation 19? To answer these questions, we have to give the model first.

1.2 Binary Data and Binary Regression Models

As a special case of ordinal data, binary data have received considerable attention during last 40 years. Binary data has only two kinds of responses, for example, we can define pass=“greater than or equal to grade C” and fail=“D or F” for the student grade example. By collapsing the grade into only 2 categories, we lose some information, but in some situations, what we focus on is just two categories (for example, whether a patient has some kind of disease or not), which explains binary data are so widely used in daily life.

Since binary data is a special case of ordinal data, in order to make the content more readable and complete, at the same time, in order to make the transition from binary data to ordinal data easier, we introduce the models for the binary data first.
1.2.1 Models for Binary Data

For the binary data, since there are only two categories, if we know the probability of individual $i$ of falling in one category $p$, then the probability of individual $i$ falling in another category will be $1 - p$, where $p$ is the probability of individual $i$ falling in one category. Because of the correlation structure, we can just focus on one category.

In order to make the content consistent, we use the same SAT grade example data set, the only change will be there will not be 5 categories (A, B, C, D, and F), instead there will be two categories, success and failure. We define success = “getting A, B or C” and failure = “getting D or F”. The most common model for binary data is the general link model, which is defined as

$$\theta_i = F(x_i'\beta)$$  \hspace{1cm} (1.1)

where $F$ is the cumulative density function (CDF), $\theta_i$ is the probability of individual $i$ getting a success, $x_i$ is the vector of explanatory variable for individual $i$, $\beta$ is the binary regression parameter. If we choose $F$ to be the standard logistic distribution CDF, which is given by equation (1.2)

$$F(x) = \frac{1}{1 + e^{-x}}, \quad -\infty < x < \infty$$  \hspace{1cm} (1.2)

then equation (1.1) becomes

$$\theta_i = \frac{1}{1 + e^{-x_i'\beta}}$$  \hspace{1cm} (1.3)

Rewrite this equation (1.3) into a more familiar form, which is called the logistic regression model, given by equation (1.4)

$$\log \frac{\theta_i}{1 - \theta_i} = x_i'\beta$$  \hspace{1cm} (1.4)

If we choose $F$ to be the standard normal distribution CDF, then equation (1.1) becomes the probit regression model

$$\theta_i = \Phi(x_i'\beta)$$  \hspace{1cm} (1.5)
If we choose to use the extreme distribution CDF

\[ F(x) = 1 - e^{-e^x}, \quad -\infty < x < \infty \]  

(1.6)

then equation (1.4) becomes

\[ \theta_i = 1 - e^{-e^{x_i'\beta}} \]  

(1.7)

Rewrite equation (1.7) into the more familiar form called complementary log-log regression model, given by equation (1.8)

\[ \log(1 - \log(1 - \theta_i)) = x_i'\beta \]  

(1.8)

There is no particular preference of choosing one model over the other. For illustration purpose, we use the revised data set, which has only two categories, failure (grade D or F) or success (grade A, B, C). We fit the logistic regression model, the probit regression model and the complementary log-log regression model on the same revised data with SAT as the only explanatory variable, then we can plot the fitted probability for success from these three models, which leads to Figure 1.2. From Figure 1.2 we can see that the predicted probabilities from both the probit link regression model and the logistic regression models are almost indistinguishable, however, the complementary log-log link gives substantively different fitted probabilities than the probit regression model or the logistic regression model. In order to facilitate the Bayesian simulation, which we will talk about in Chapter 2, I choose probit regression model.
Figure 1.2: The fitted probabilities for the success category using the logistic, the probit and the complementary log log regression models. The X-axis is the SAT grade, the Y-axis is the fitted probabilities for success.
1.3 Estimating Binary Regression Coefficient

1.3.1 Likelihood for Binary Data

To derive the likelihood for binary regression models, recall that for a binomial random variable \( Y \), which denotes the number of success in \( n \) trials each having success probability \( p \), the probability density function can be expressed as

\[
f(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}\quad (1.9)
\]

let \( y_1, \ldots, y_m \) denote the number of success in independent binary experiments with total trials \( n_1, \ldots, n_m \) and the success probabilities \( p_1, \ldots, p_m \) respectively. The likelihood function for \( y_1, \ldots, y_m \) is the product of the probability density functions but is reviewed as a function of \( p \):

\[
L(p_1, \ldots, p_m) = \prod_{i=1}^{m} p_i^{y_i} (1-p_i)^{n_i-y_i}, 0 \leq p_i \leq 1\quad (1.10)
\]

In the regression setting, we choose the probit link so that the success probability for individual \( i \) is given by \( p_i = \Phi(x'_i \beta) \), where \( x_i \) is the vector of explanatory variable for individual \( i \), \( \beta \) is the regression parameter, \( x'_i \beta \) is a shorthand notation of the combination of the explanatory variables. For predicting probabilities of success for individual \( i \) based on his SAT score,

\[
x'_i \beta = \beta_0 + \beta_1 \times SAT_i
\]

In more general problem with \( r \) explanatory variables,

\[
x'_i \beta = \beta_0 + \beta_1 \times SAT_i + \cdots + \beta_r \times SAT_r
\]

Replacing \( p_i \) in likelihood \( L \) we get

\[
L(\beta) = \prod_{i=1}^{m} \left( \Phi(x'_i \beta) \right)^{y_i} (1 - \Phi(x'_i \beta))^{n_i-y_i}\quad (1.11)
\]
1.3.2 Maximum Likelihood Estimate for Binary Data

As stated above, likelihood function given by equation (1.1). can be viewed as a function of the unknown parameter $\beta$, we have to maximize the likelihood in order to get the maximum likelihood estimator (MLE) $\hat{\beta}$ for the unknown parameter $\beta$. Since the publication of Nelder and Wedderburn’s paper in 1972, finding MLE can be easily solved by Newton-Rephson method using iteratively reweighted least square, and a byproduct of this procedure is the standard error of these estimator.

1.4 Binary Regression Model Diagnostics

Since binary data only has two responses and if we know the probability of subjects falling in one category, then we can know the probability of subjects falling in other category, we can just focus on just one category to do statistical analysis. In the SAT grade data example, we want to use the category of success. A residual is measurement of agreement between an observed response and its corresponding fitted value under a certain model, and is defined as the difference between the fitted values and the observed values for the univariate case. If the model is appropriate, then one would expect the fitted values should be “close” to the observed values. Anscombe and Tukey (1963) considered outlier to be “observation that have such large residuals, in comparison with most of others, as to suggest that they ought to be treated specially”. Since outliers sometimes can seriously affect the estimated model fitting parameters values thus it can affect the precision of prediction, we have to take care of outliers before we make any prediction.

1.4.1 Pearson Residual for Binary Data

Assume that we use a probit link,

$$p_i = \Phi(x_i'\beta)$$
for the revised SAT-grade data, which has only two categories (success and failure) and SAT as the only explanatory variable. Let $\hat{\beta}$ denote the MLE of the regression parameter using iteratively reweighted least square (IRLS), the fitted values for the $i$th individual may be rewritten as

$$\hat{y}_i = n_i \hat{p}_i, \text{ where } \hat{p}_i = \Phi(x_i'\hat{\beta})$$

The traditional residual is the Pearson residual, which is obtained by dividing the $i$th residuals, canonically defined as

$$y_i - \hat{y}_i$$

by its standard error, where $y_i$ is the observed count for individual $i$, $\hat{y}_i$ is the estimated counts for individual $i$. Because the variance of a binomial distribution observation is $n_i p_i (1 - p_i)$, the standard error of equation (1.12) is given by $\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}$, the Pearson residual for binary case is then given by

$$r_{i,p} = \frac{y_i - \hat{y}_i}{\sqrt{n_i \hat{p}_i (1 - \hat{p}_i)}}$$

(1.13)

Noted that for $n \geq 5$, Pearson residual given by equation (1.13) can be reasonably approximated by a standard normal distribution.

### 1.4.2 Deviance Residual for Binary Data

Deviance residuals are related to the deviance statistic, which is defined as twice of the difference of the log likelihood function evaluated at the observed proportions and the fitted proportions. The log likelihood function is given by

$$\text{Log Likelihood} = \sum_{i=1}^{n} y_i \log p_i + (n_i - y_i) \log (1 - p)$$

(1.14)

Then the deviance statistic is given by

$$D = \sum_{i=1}^{n} 2\{y_i \log \left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\}$$

(1.15)
The deviance residual, \( r_{i,D} \) is defined for the \( i \)th observation as the signed square root of the \( i \)th observation’s contribution to the total model deviance. The deviance residual is given by equation (1.16)

\[
 r_{i,D} = \text{sign}(y_i - \hat{y}_i)\{2[y_i\log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i})]\} 
\]

1.4.3 Adjusted Deviance Residual for Binary Data

Pierce and Schaffer in 1986 proposed adjusted deviance residual based on the fact that there is some bias in the deviance residual, and it is defined as equation (1.17)

\[
 r_{i,ad} = r_{i,D} + \frac{1 - 2\hat{p}_i}{6\sqrt{n_i\hat{p}_i(1 - \hat{p}_i)}} 
\]

Note that among these three residuals, i.e., Pearson residual, deviance residual, and the adjusted deviance residual, the adjusted deviance residual is often most nearly normally distribution and yields accurate approximations even for small \( n_i \).

1.5 Goodness-of-fit and Model Selection for Binary Data

The first traditional method for testing the goodness-of-fit for a model utilizes the famous deviance statistic \( D = \sum_{i=1}^{n} 2[y_i\log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - \hat{y}_i})] \), where \( \hat{y}_i = n_i\hat{p}_i \) is the fitted value from the model. If the model is “true”, then this deviance statistic follows an asymptotic \( \chi^2 \) distribution with \( n - q \) degree of freedom, where \( n \) is the number of binomial observations and \( q \) is the dimension of the regression parameters.

It is natural that there are several explanatory variable in the data set and most of the time it is not wise to include all the variables in the model. To decide which variable should be included in the model is a matter of model selection. For the purpose of model selection, deviance statistic from several competing models are often reported and used as a tool to identify which model is better. The most common way to do model selection is using an
Analysis of Deviance table. Suppose that there are two models M1 and M2, if M2 contains all the parameters of M1 and other parameters, then we said that model M1 is nested in model M2. We can then fit the data with model M1, we can get a deviance statistic D1 with degree of freedom DF1, we can also fit the data with model M2, we then can get another deviance statistic D2 with degree of freedom DF2. The difference between these two deviance statistic D1 and D2 has an approximately $\chi^2$ with degree of freedom DF1-DF2. Note that model M2 has “extra” parameters than model M1. If we should really include those “extra” variable in the model, then the difference of deviance should be larger than the critical value of $\chi^2$ distribution with degree of freedom DF1-DF2, in other words, the difference of deviance should be “significant”.

For an illustration purpose, we use the SAT-grade binary data as the example, here we have two explanatory variable, one is SAT, the other one is the prerequisite probability class. We wonder if it is proper to include the variables both SAT and prerequisite math class in the model, so we fit three models:

- M1: There is just the constant parameter in the model, no variable SAT or prerequisite probability class
- M2: There is the constant part and variable SAT, no prerequisite probability class
- M3: There are three variables in the model, the constant part, the variable SAT and prerequisite probability class as a factor variable.

We fit these three models to the binary SAT grade example, and get the Analysis of Deviance Table 1.2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Covariates</th>
<th>Deviance</th>
<th>Degree of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Constant part</td>
<td>36.65</td>
<td>29</td>
</tr>
<tr>
<td>M2</td>
<td>Constant+ SAT</td>
<td>22.27</td>
<td>28</td>
</tr>
<tr>
<td>M3</td>
<td>Constant+SAT+prerequisite probability course</td>
<td>18.92</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 1.2: The analysis of deviance table for the binary SAT grade example
From Table 1.2, we can see that model M1 is nested in model M2 in the sense that all the variables in model M1 is contained in model M2, also notice that model M2 has the “extra” variable SAT included in the model, and the difference of deviance between model M1 and model M2 is $36.65 - 22.27 = 14.38$ and the difference of degree of freedom between model M1 and model M2 is 1. The critical value for $\chi^2$ with 1 degree of freedom is 3.84, since $14.38 > 3.84$, we can conclude that variable SAT should be included in the model. Now take a look at model M2 and model M3, the difference between model M2 and model M3 is that model M3 has the “extra” variable prerequisite probability course and now the difference of deviance between model M2 and model M3 is $22.27 - 18.92 = 3.35$ and the difference of degree of freedom is $28 - 24 = 4$. The critical value for $\chi^2$ distribution with 4 degrees of freedom is 9.49, since $3.35 < 9.49$, we conclude that the variable prerequisite probability course should not be included in the model.

1.6 Models for Ordinal Data

Models for binary case can be easily extend to the ordinal case with some minor modifications. A large class of models have been proposed for analyzing ordinal categorical data, for example, McCullagh in his classic paper (McCullagh, 1980) proposed the famous proportional odds model. Suppose that there are $k$ categories. Let $p_{ic}$ denote the probability of individual $i$ falling in a category $c$, and define $\theta_{ic} = p_{i1} + p_{i2} + \cdots + p_{ic}$ to be the cumulative probability of the individual $i$ falls in the category $c$ or below, then the regression component can be written in general as

$$\theta_{ic} = F (\gamma_c - x_i'\beta)$$

(1.18)

where $F$ is called the link function. The parameters $\gamma_c$ are generally of little interest but are usually referred to as “cutoff points”, $x_i'$ is the vector of explanatory variable for individual $i$, $\beta$ is the regression parameters, and $x_i'\beta$ is still the shorthand notation of the combination of explanatory variables as the same case in binary case. The usual choice for the link function
is the cumulative density function of some distribution, for example, if a logistic link function is assumed, which is $F(x) = \frac{1}{1+e^{-x}}$, $-\infty < x < \infty$, equation (1.18) becomes

$$\theta_{ic} = \frac{1}{1 + e^{-(\gamma_c - x'_i \beta)}}$$ (1.19)

Similar to the technique in binary case, we rewrite equation (1.19) into the more familiar form called logistic link model given by (1.20)

$$\log \left( \frac{\theta_{ic}}{1 - \theta_{ic}} \right) = \gamma_c - x'_i \beta$$ (1.20)

Notice for this logistic link model (1.20), the ratio of the odds for the event $y_1 \leq c$ to the odds of the $y_2 \leq c$ is

$$\frac{\theta_{1c}/ (1 - \theta_{1c})}{\theta_{2c}/ (1 - \theta_{2c})} = \exp (x'_2 \beta - x'_1 \beta)$$ (1.21)

independently of the category of response $c$. For this reason, model (1.21) is often called the proportional odds model (McCullagh, 1980).

Another common regression model for ordinal data assumes complementary log-log link

$$F(x) = 1 - e^{-e^x}, -\infty < x < \infty$$ (1.22)

in equation (1.18), which becomes

$$\theta_{ic} = 1 - e^{e^{\gamma_c - x'_i \beta}}$$ (1.23)

Again, rewrite equation (1.23) into a more familiar model called complementary log-log regression model, given by equation (1.24)

$$\log \left[ -\log (1 - \theta_{ic}) \right] = \gamma_c - x'_i \beta$$ (1.24)
If we interpret $1 - \theta_{ic}$ as the probability of survival time beyond category $c$ for the $i$th observation, this model may be considered as a discrete version of the proportional hazards model proposed by Cox (1972).

Another commonly used link function to model ordinal data is the CDF of standard normal distribution. With such a link, model (1.18) becomes

$$\theta_{ic} = \Phi(\gamma_c - x_i'\beta)$$

(1.25)

This model is referred to as the ordinal probit model. The ordinal probit model produces predicted probabilities similar to those obtained from the logistic link model and complementary log-log link model. In Figure 1.3, we fit logistic link regression model, probit link, and complementary log-log on the SAT grade example given in Table 1.1 and plotted the predicted cumulative probability for getting a grade of $f$ and less than $c$.

From Figure 1.3 we can see that the fitted cumulative probabilities from both the probit regression model and the logistic regression model are also the same, which is very similar to the case of binary case. However, the fitted cumulative probabilities from the complementary log-log regression model are substantively different from the ones from logistic and probit regression model. We will see that the ordinal probit regression model processes a property that makes sampling from its posterior distribution particularly efficient, which we will address in Chapter 2. For this purpose, we will focus on the ordinal probit model in this thesis.

However, there are some shortcoming for the logistic, probit link or complementary log-log link model in McCullagh, for example, Anderson raised the question of indistinguishability of dimension and claimed that general link function can not handle the dimension indistinguishability well, so Anderson proposed the stereotype model in Anderson (1984), as given by equation (1.26)

$$\frac{p(y = i|x)}{p(y = G|x)} = \exp(\gamma_i - x'\beta)$$

(1.26)
Figure 1.3: Graph of fitted probabilities of getting grade \( f \) and less than \( c \) using the logistic, probit and complementary log-log models for the statistics class data set. The solid line is the fitted probabilities using complementary log-log link, the dashed lines is the fitted probabilities using probit link, the dotted line is the fitted probabilities using logistic link.
where in this stereotype model, $p(y = i|x)$ is the probability of the observation $x$ falls in category $i$, $p(y = G|x)$ is the probability of the observation $x$ falls in reference category $G$.

There are other models for the ordinal data, for example, Feinberg in 1980 proposed an alternative model called continuation-ratio model to the proportional odds model for the analysis of ordinal data. Still, let $p_{ic}$ denote the probability of individual $i$ falling in a category $c$ and $\theta_{ic} = p_{i1} + p_{i2} + \cdots + p_{ic}$ be the cumulative probability of the individual $i$ falling in the category $c$ or below, then the continuation-ratio model is given by

$$\log\left(\frac{p_{i,c}}{1 - \theta_{ic}}\right) = \alpha_c - x'_i\beta$$

(1.27)

where $\alpha_c$ are the unknown intercepts parameters, which satisfy the condition that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_k$, $\beta$ is a vector of unknown regression parameter corresponding to the vector of explanatory covariate $x_i$. Continuation-ratio model given by equation (1.27) can essentially be viewed as the ratio of two conditional probabilities, one is the probability of individual $i$ falling in category $c$, the other is the probability of individual $i$ falling in category $c+1$ and above.

The last model is called the adjacent-category logistic model mentioned in (Agresti,1990). This model involves modeling of two adjacent category probabilities $p_{ic}$ and $p_{ic+1}$, which represents the probability of individual $i$ falling in category $c$ and $c+1$. The model has the following representation:

$$\log\left(\frac{p_{i,c}}{p_{i,c+1}}\right) = \alpha_c - x'_i\beta_c$$

(1.28)

where $\alpha_c$ are still the unknown intercepts parameters satisfying the condition that $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_k$, $\beta_c$ is a vector of unknown regression parameter corresponding to the vector of explanatory covariate $x_i$. 
1.7 Classical Fitting for Ordinal Regression Model

Like the binary case, in order to make any inference about the unknown parameters, we have to deal with the likelihood function first.

1.7.1 The Likelihood Function for Ordinal Data

To derive the likelihood for ordinal regression models, recall that for a multinomial random variable vector \( Y = (y_1, \cdots, y_n) \), which denotes the number of success in \( N \) trials each having success probability vector \( p = (p_1, \cdots, p_n) \), the probability density function can be expressed as

\[
f(y_1, \cdots, y_n, p_1, \cdots, p_n) = \frac{n!}{y_1! \cdots y_n!} p_1^{y_1} \cdots p_n^{y_n} \tag{1.29}
\]

For the ordinal probit model with the grade data set, suppose there are \( n \) (\( n=30 \) in the SAT grade example) observations and \( k \) (\( k=5 \) in the SAT grade example, which are F, D, C, B, and A) categories. Since the response variable \( y \) has \( k \) possible categories, it is convenient to introduce \( k-1 \) dummy variables \( y = (y_1, y_2, \cdots, y_{k-1}) \) with components

\[
y_{ij} = \begin{cases} 
1 & \text{if } y_i = j, j = 1, 2, \cdots, k - 1 \\
0 & \text{else.}
\end{cases}
\]

where \( y_i = j \) denotes the individual \( i \) falls in category \( j \), we then have

\[
y_i = j \text{ is equivalent to } y_{ij} = (0, \cdots, 1, \cdots, 0)
\]

The probability are simply connected by

\[
p(y_i = j) = P(y_{ij} = 1)
\]
We note that no matter what value \( y \) takes on, \( \sum_{j=1}^{k} y_{ij} = 1 \). Assuming that the observations are independent, then the likelihood function is in equation (1.30)

\[
L = \prod_{i=1}^{n} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} \cdots p_{ik}^{y_{ik}}
\]  

(1.30)

For the SAT grade example (we have 5 categories so \( k=5 \)), the likelihood is

\[
L = \prod_{i=1}^{n} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} \cdots p_{i5}^{y_{i5}}
\]  

(1.31)

According to the basic rule of calculus, if we want to maximize the likelihood function in order to get the maximum likelihood estimator (MLE), we have to take derivative with respect to every component in the regression parameter vector \( \beta \) and let every derivative equal 0. Since the likelihood function is in the product form, it is not convenient to deal with the likelihood directly, the usual method is to take the natural log of likelihood so that the product can be transformed into a summation. The log-likelihood function in this case becomes

\[
l = \sum_{i=1}^{n} y_{i1} \log p_{i1} + y_{i2} \log p_{i2} + \cdots + y_{ik} \log p_{ik}
\]  

(1.32)

The log likelihood for the SAT grade example (we have 5 categories so \( k=5 \)) is given by

\[
l = \sum_{i=1}^{n} y_{i1} \log p_{i1} + y_{i2} \log p_{i2} + \cdots + y_{i5} \log p_{i5}
\]  

(1.33)

If we choose probit link, which gives the fitted cumulative probability of individual \( i \) falling in category less than or equal \( c \) by

\[
\theta_{ic} = \Phi(\gamma_c - x_i'\beta)
\]  

(1.34)
then the log likelihood given by equation (1.32) becomes

\[ l = \sum_{i=1}^{n} y_{i1} \log \Phi(\gamma_{1} - x_{i}'\beta) + y_{i2} \log (\Phi(\gamma_{2} - x_{i}'\beta) - \Phi(\gamma_{1} - x_{i}'\beta)) + \cdots + y_{ik} \log (1 - \Phi(\gamma_{k-1} - x_{i}'\beta)) \]

(1.35)

1.7.2 Maximum Likelihood Estimate for Ordinal Data

The usual method of estimating parameters in ordinal regression model is using maximum likelihood method, which is taking partial derivative of log-likelihood function with respect to the unknown parameters and let each partial derivative to be zero. However, due to the complicated nonlinear structure of the log likelihood give by (1.35) the usual derivative method usually do not have analytical solution, we have to find the solution to the derivatives numerically. McCullagh (1980) talked about the details of using Newton-Raphson method with Fisher scoring to get the estimator of unknown parameters. Albert and Johnson (1999) use the iteratively reweighed least square (IRLS) algorithm to find both the maximum likelihood estimate and the asymptotic covariance matrix for the probit regression model. This process of finding the maximum likelihood estimator is complicated in theory but in practice it is easy, for example, It can be easily complemented with software such as PROC LOGISTIC in SAS program or `polr` function in the package `MASS` for software R.

1.8 Ordinal Regression Model Diagnostics

The diagnostics techniques for binary data usually do not hold for the ordinal data. In binary data case, we have two kinds of responses and if we know one, then we know the other, so we can focus on just one category. However, in ordinal data case, we usually have more than one category in the data set and we cannot focus on one category and ignore the rest categories. Moving from binary distribution to multivariate (n>2) means not only that there are more residuals to be examined, but also the residuals are correlated so that
it is hard to detect outliers in the traditional method, the Pearson residual for the binary case cannot be extended to the ordinal case. However, the deviance residual for binary case can be extended to ordinal case with minor modification. Recall that in binary case, the deviance statistic is given by

\[ r_{i,D} = \text{sign}(y_i - \hat{y}_i) \left\{ 2[y_i \log\left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log\left( \frac{n_i - y_i}{n_i - \hat{y}_i} \right)] \right\} \]

For the ordinal case, it is preferable to examine the value of the deviance contribution from each individual observation directly, given by

\[ d_i = \sum_{j=1}^{k} I(y_i = j) \log(I(y_i = j)/\hat{p}_{ij}) \]

here, \( \hat{p}_{ij} \) is the fitted probability of observation \( i \) falling in category \( j \) using the maximum likelihood estimator \( \hat{\beta} \) and it is given (if we pick the probit regression model) by

\[ \hat{p}_{ij} = \Phi(\gamma_{j-1} - x_i \hat{\beta}) \]

\( \gamma_{j-1} \) is the \( j \)-th cutoff point, \( I(y_i = j) \) is the indicator function to show that observation \( i \) falls in category \( j \). Observations that contribute disproportionately to the overall model deviance should be recognized with suspicion.

### 1.8.1 The M Test for Ordinal Data to Detect Outliers

Fuchs and Kenett (1980) proposed the M test based on the maximum adjusted residuals to detect outliers. In their context, they consider only one variable \( n \) from multinomial distribution with parameter \( N \) and \( p_i \), where \( N = \sum_{i=1}^{k} n_i \), and \( p_i \), \( i = 1, \cdots, k \). They define

\[
Z_i = \frac{n_i - N\hat{p}_i}{N\hat{p}_i(1 - \hat{p}_i)}, \quad i = 1, \cdots, k
\]
as the adjusted residuals, where \( \hat{p}_i \) is the fitted probability of the observation fall in category \( i \), also they found the distribution of the adjusted residuals to be multivariate normal distribution with mean 0 and covariate matrix \( R \). Then they picked the largest \( Z_i \) and compared it with the critical value \( M^* \). They also gave the computation method of the critical value of \( M^* \). They claimed that M-test asymptotically has more power than the Chi Square test in some cases especially when there is only one outlier. However, Fuchs and Kenett did not fully use the advantage of the residuals because they just used the maximum of the adjusted residuals, in this sense, they lost some information about the model.

1.8.2 Locally Best Invariant Test to Detect the Existence of Outliers

Schwager and Margolin (1982) found the critical region for the locally best invariant test of the null hypothesis \( H_0 \): there is no outliers vs \( H_\alpha \): some outliers are present. Unfortunately, Schwager and Margolin did not give the specific method on how to find the outliers if \( H_0 \): there is no outlier, is rejected.

1.8.3 Using Outlier Score to Detect Outliers

Zijlstra, Ark, and Sijtsma (2006) propose two definitions for outlier score, namely, item-based outlier score and item-pair based outlier score. No matter which method they use, the higher score, the more chance that this observation is an outlier. Tukey’s fences method is one method of checking these outlier scores. Zijlstra, Ark, and Sijtsma use the interquartile (25% quantile to 75% quantile, in a more familiar notation, Q1 and Q3) to select those observations that have outlier score greater than \( Q_3 + \frac{3}{2}(Q_3-Q_1) \). Another method to check those outlier scores is using the extreme studentized deviate (ESD) procedure. The generalized ESD tests \( H_0 \): these outlier scores to be normally distributed with mean \( \mu \) and variance \( \sigma^2 \) vs \( H_\alpha \): these outlier scores to be normally distributed with mean \( \mu + a \) and variance \( \sigma^2 \). A test
statistics is given and an asymptotic distribution of t distribution is also obtained.

1.8.4 Gupta et al.’s Residual for Ordinal Data

Gupta, Nguyen and Pardo (2006) propose a new family of residuals and influential measures based on the $\phi$-divergence test statistic, which is based on

$$D_\phi(\hat{p}, p(\theta)) = \sum_{j=1}^{M} p_j(\theta)\phi\left(\frac{\hat{p}_j}{p_j(\theta)}\right)$$

where $p_j(\theta)$ is the probability parameter of the multinomial distribution, $\hat{p}$ is the estimator of $p_j$. Gupta et al. also get asymptotic distribution for the standardized new family of residuals, which is a normal distribution without surprise.

1.9 Goodness-of-fit for Ordinal Regression Model

It is very common that there are more than one explanatory variable in the model, in order to make the prediction more accurate and more reliable, we have to decide which variable should be included in the model. The usual way of testing whether the estimated coefficient is significant depends on the t-test statistic. They usually test $H_0$: coefficient for one particular variable=0 vs $H_1$: coefficient for that particular variable$\neq 0$. By using the MLE method, one can get the estimate of the unknown parameters, and one also can get the estimated standard error of the known parameters. If $H_0$ is true, one can get the a t-test statistic, and one can compute the probability of greater than the t-test statistic that we just computed. In traditional statistics, that probability value is called p-value. If p-value is less than 0.05, then one rejected the null hypotheses and said that this coefficient is significant and the corresponding variable thus should be included in the model, otherwise, if the p-value is greater than 0.05, then this coefficient is not significant and the corresponding variable should be removed from the model.
Another way to decide which variable should be included in the model is using deviance statistic for binary cases given by (1.15). The deviance statistic can be transferred to ordinal data case without any major modification except for the increase of the number of categories. The deviance for ordinal case is defined by

$$D(y) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} \{ I(y_i = j) \log \left( \frac{I(y_i = j)}{\hat{p}_{ij}} \right) \}$$  \hspace{1cm} (1.36)

Like before, $I(y_i = j)$ is an indicator function to show that observation $i$ falling in category $j$, $\hat{p}_{ij}$ is the fitted probability of observation $i$ falls in category $j$ using the maximum likelihood estimator $\hat{\beta}$ and it is given (if we pick the probit regression model) by

$$\hat{p}_{ij} = \Phi(\gamma_{j-1} - x_i \hat{\beta})$$

$\gamma_{j-1}$ is the $j$-$1$th cutoff point, $k$ is the number of categories, $n$ is the number of the observations. As is the case with any fitted model, especially in the sensitive area such as medical study and the financial case, before using the model to make any prediction about “future” observation, we should assess its overall goodness-of-fit and examine the contribution to the fit of each observation. The most commonly used ones are the usual likelihood ratio deviance and Pearson Chi Square statistic, which means the fit of the given model versus the “saturated” model. The “saturated” model can have as many parameters as the number of observations so that it can fit the model 100% perfectly, however it itself is of little interest because you can not make any reliable based on “saturated” model.

The Pearson Chi Square statistic in general is given by

$$X^2 = \sum_{i=1}^{n} (y_i - \hat{u}_i)' \Sigma_i^{-1}(\hat{\beta}) (y_i - \hat{u}_i)$$

where $y_i$ is the vector of observed count in each category for observation $i$, $\hat{u}_i$ is the vector of expected count for the observation $i$ in each category, $\Sigma_i^{-1}(\hat{\beta})$ is the estimated variance-
covariance matrix for the observation \( i \).

The problem with usual likelihood ratio deviance and Pearson Chi Square statistic to test the goodness-of-fit for one model is that if the expected counts of each cell is less than 5, then the \( \chi^2 \) distribution is not a good approximation. Lapsitz, Fitzmaurice and Molenberghs (1996) propose a goodness-of-fit test statistic by assigning scores to response category \( k \), and this goodness-of-fit test statistic have approximately \( \chi^2 \)-distribution when the model has been correctly specified. Johnson and Albert (1999) use deviance statistic, defined as

\[
D = 2 \sum_{i=1}^{n} \sum_{j=1}^{C} I(y_i = j) \log(Iy_i = j)/\hat{p}_{ij}
\]

Johnson and Albert also point out that the deviance statistic follows a \( \chi^2 \) distribution only when observations are groups according to covariate values and the expected counts in each cell become large. When the number of observations observed at each covariate values is small, then the deviance statistic is not well approximated by a \( \chi^2 \) distribution.
CHAPTER 2

Bayesian Models for Ordinal Data

2.1 Bayes’s Rule

From a classical (or frequentist’s) framework, parameters are considered to be fixed quantities, we can either use maximum likelihood method or moment estimator method to get the estimated values for these parameters. However, from a Bayesian framework, parameters are often considered to be random variables following some distributions. Let $\theta$ denote the parameters. Usually a distribution is assigned to the unknown parameters before we fit a model to the data, and this distribution is called a prior distribution denoted by $p(\theta)$, which reflects our historical knowledge about the unknown parameters. After we observe the data, we can write down the likelihood function as described by equation (1.30). Let $L(y|\theta)$ denote the likelihood function, then the joint probability distribution of $y$ and $\theta$ is

$$p(\theta, y) = p(\theta) L(y|\theta)$$

Using the idea of conditional distribution of $\theta$ on $y$, we get the posterior distribution of $\theta$, which is given by equation (2.1)

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta) L(y|\theta)}{p(y)}$$

(2.1)
where \( p(y) \) is called marginal likelihood. It can be computed either 
\[
p(y) = \sum_{\theta} p(\theta) L(y|\theta)
\]
and the summation is over all the possible values of \( \theta \) if \( \theta \) is discrete, or 
\[
p(y) = \int_{\theta} p(\theta) L(y|\theta)
\]
if \( \theta \) is continuous.

Equation (2.1) is often referred as Bayes’s rule, named after the Reverend Thomas Bayes. The posterior distribution of \( \theta \) conditional on \( y \) given by equation (2.1) reflects the updated belief about the unknown parameter \( \theta \) after we observe the data. Before we observe the data \( y \), we consider the prior distribution \( p(\theta) \) as the distribution of the unknown parameter \( \theta \). After we observe the data \( y \), we consider the posterior distribution \( p(\theta|y) \) as the distribution of the unknown parameter \( \theta \). Since we are interested in learning about the unknown parameter \( \theta \), we often ignore the part of \( p(y) \) because it is considered to be a constant for the parameter \( \theta \). Usually the most frequently used form of posterior distribution of \( \theta \) is give in Equation (2.2)
\[
p(\theta|y) \propto p(\theta) L(y|\theta) \tag{2.2}
\]
Our task is to simulate values of \( \theta \) from its posterior distribution and summarize them so that we can learn about the unknown parameter \( \theta \). The simple expression in equation (2.2) encapsulates the technical core of Bayesian inference: The primary task of any specific application is to develop the model \( p(y, \theta) \) and perform necessary computations to summarize \( p(\theta|y) \) in appropriate ways.

### 2.1.1 Prior Predictive Distribution

To make an inference about an unknown but observable “future” \( \tilde{y} \) value, before the actual data \( y \) are observed, the distribution of “future” \( \tilde{y} \) has the distribution of
\[
p(\tilde{y}) = \int p(y, \theta) d\theta = \int p(\theta)p(y|\theta) d\theta \tag{2.3}
\]
This is often called the marginal distribution of \( y \), but more informatively is called the prior predictive distribution by Andrew Gelman et.al (2003): prior because it is before the time
that we observe the actual data and it is not conditional on the previous observation of the
process, predictive because it is the distribution for the observable “future” y value.

2.1.2 Posterior Predictive Distribution

Judging from the name, the posterior predictive is the prediction of the observable “future”
\( \tilde{y} \) value after we observed the actual data set: posterior because it is conditional on the
observed y value and predictive is because it is the prediction of the “future” value. The
distribution of \( \tilde{y} \) conditional on the observed y value is given by

\[
p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta \\
= \int p(\tilde{y}, |y, \theta)p(\theta|y)d\theta \\
= \int p(\tilde{y}|\theta)p(\theta|y)d\theta
\] (2.4)

2.2 Latent Variables

To model ordinal data, one common strategy is to introduce the idea of latent variable into
the regression model. Latent variable is an unobservable variable and often assumed to be
drawn from a continuous distribution centered on a mean value that varies from individual
to individual. Often the mean value is modeled as a linear function of the student’s covariate
vector. We define a latent variable by equation (2.5).

\[
Z_i = x_i'\beta + \varepsilon_i
\] (2.5)

where \( \varepsilon_i \) is from a standard normal distribution (if we choose the probit regression model)
or a standard logistic link function (if we choose the logistic regression model).

To illustrate the idea of the latent variable, we use the student’s grade data given by
Table 1.1. From a latent variable perspective, this model is equivalent to assuming that we
can associate a latent performance variable with each student, and the distribution of the
unknown variable has a normal distribution (if we choose probit regression model) centered
on a linear function of the student’s SAT score. For a class with five grades, a total of four
unknown additional grade cutoff points must be introduced, denoted by $\gamma_1, \cdots, \gamma_4$. Also,
because the response categories are ordered, we have to put some constraints on the values
of the unknown grade cutoff points. Let the upper grade cutoff for an F be denoted by $\gamma_1$,
the upper grade cutoff for a D be denoted by $\gamma_2$, and so on, the ordering constraint may be
stated mathematically as

$$-\infty < \gamma_1 \leq \gamma_2 \leq \gamma_3 \leq \gamma_4 < +\infty$$

For computation convenience, we add two more cutoff points, which are $\gamma_0 = -\infty$ and
$\gamma_5 = +\infty$. When $Z$ falls between the grade cutoff $\gamma_{c-1}$ and $\gamma_c$, this individual is classified
into category C. The latent variable $Z_i$ can be shown by Figure 2.1.

![Figure 2.1: Latent variable interpretation of ordinal data. In this plot, the latent variable
for the $i$th student is from a normal distribution with mean $x_i'\beta$ and standard deviation 1.
If this latent variable $Z_i$ falls between $\gamma_1$ and $\gamma_2$, then the $i$th observation will get a D.
To link the latent variable model with ordinal data, let $f$ denote the standard normal]}
probability density function (pdf) and $F$ denote the standard normal cumulative density function (cdf). Let $p_{ic}$ denote the probability that individual $i$ receives a grade of $C$. Then from Equation (2.5), it follows that

\[ p_{ic} = \text{prob} \left( \gamma_{c-1} < Z_i < \gamma_c \right) \]
\[ = \int_{\gamma_{c-1}}^{\gamma_c} f \left( Z_i - x_i'\beta \right) dZ_i \]
\[ = F \left( \gamma_c - x_i'\beta \right) - F \left( \gamma_{c-1} - x_i'\beta \right) \]

(2.6)

### 2.3 Likelihood Function Using Latent Variable

To obtain the likelihood function using the latent variable, let us assume that there are $k$ possible categories, denoted by $1, \cdots, k$. Also suppose that $n$ items are observed and that the categories assigned to these $n$ items are denoted by $y_1, \cdots, y_n$. Associated with the $i$th individual’s response, we define a latent variable

\[ Z_i = x_i'\beta + \varepsilon_i \]

where $x_i$ is the vector of covariates associated with the $i$th individual, $\beta$ is the vector of unknown parameter, $\varepsilon_i$ is distributed as standard normal distribution if we pick the probit regression model. We observe the category $y_i = C$ if the latent variable $Z_i$ falls between $\gamma_{c-1}$ and $\gamma_c$. Let $p_i$ denote the vector of the probability that the $i$th observation falls in each category, that is, $p_i = (p_{i1}, p_{i2}, \cdots, p_{ik})$, where $p_{ic}$ is the probability that the $i$th observation fall in category $C$. Let $y = (y_1, y_2, \cdots, y_n)$ denote the observed vector of response for all individuals. Then the likelihood is given by Equation (2.7)

\[ L(p_i|y) = \prod_{i=1}^{n} p_{iy_i} \]

(2.7)
Substituting the values of $p_{ic}$ from Equation (2.6) leads to the likelihood in Equation (2.8)

$$L(\beta, \gamma) = \prod_{i=1}^{n} \left[ F(\gamma_{yi} - x_i'\beta) - F(\gamma_{yi-1} - x_i'\beta) \right]$$

(2.8)

If we parameterize the model by using the idea of latent variable, then likelihood (2.8) becomes

$$L(\beta, \gamma, Z) = \prod_{i=1}^{n} f(Z_i - x_i'\beta) I(\gamma_{yi-1} \leq Z_i < \gamma_{yi})$$

(2.9)

### 2.4 Prior Distribution

#### 2.4.1 Noninformative Prior

In the situation where little prior information is available, the easiest way to construct a prior distribution for the unknown cutoff points (suppose that there are $k$ categories) $\gamma_1, \gamma_2, \cdots, \gamma_{k-1}$ and regression parameter vector $\beta$ is to fix one value of the unknown cutoff point, say $\gamma_1$, at 0. After fixing the value of $\gamma_1$, a uniform prior can then be assumed for the remaining cutoff points, subject to the constraints that

$$\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{k-1}$$

Also a uniform prior can be assumed on the unknown regression parameter vector $\beta$, we also assume the prior distributions for the cutoff points and the regression parameters are independent.

If we choose the a noninformative prior on the unknown cutoff points $\gamma_2, \cdots, \gamma_{k-1}$ and the regression parameter $\beta$, then the posterior distribution in this case is proportional to

$$p(\beta, \gamma, Z|y) \propto \prod_{i=1}^{n} f(Z_i - x_i'\beta) I(\gamma_{yi-1} \leq Z_i < \gamma_{yi})$$

(2.10)
The reason why we need to fix one cutoff point, for example, \( \gamma_1 \) to be 0 is because that if \( \gamma_1 \) does not equal to 0, then we can always subtract \( \gamma_1 \) from the regression parameter \( \beta \) to make the first cutoff points \( \gamma_1 \) equal to 0. Another important reason is that if we do not fix one cutoff point, then we can not get one unique estimate of the regression parameter vector and the rest cutoff points.

The maximum of a posterior estimator for the unknown cutoff points and regression parameter vector using the informative prior is identical to the maximum likelihood estimator using traditional method. In general, MLE provides satisfactory point estimator for the cell probabilities in multivariate distribution for each observation when moderate counts are observed in the \( k \) categories.

### 2.4.2 Informative Prior

Noninformative prior is for the case that we do not have much knowledge about the unknown parameters, however, in the real life, most of the time we have some knowledge about the unknown parameters even before we observe the actual data. In that case, we can put an informative prior on the unknown parameters to reflect our belief. Johnson and Albert (1999) pointed out a more attractive and easier method of constructing an informative prior is using the conditional means approach of Bedrick, Christensen and Johnson (1996) method. Suppose that there are \( k \) categories, so there will be \( k-1 \) unknown cutoff points \( \gamma_1, \gamma_2, \cdots, \gamma_{k-1} \), but since we fix the first cutoff point \( \gamma_1 \) to be 0, we actually have \( k-2 \) unknown cutoff points. Also we suppose that the dimension of regression parameter vector \( \beta \) is \( q \). To construct an informative prior for \( \{ \gamma_2, \cdots, \gamma_{k-1}, \beta \} \), we must specify \( M=q+k-2 \) values of the covariate vector \( x \)—call these covariate values \( x_1, \cdots, x_M \). For each covariate vector \( x_j \), we specify a prior probability estimate and number of hypothetical evidence to support our estimate of the corresponding cumulative probability \( \theta_j \). Thus, for each covariate value, two items are specified:

- A guess at the cumulative probability \( \theta_j \) corresponding to covariate \( x_j \)
• the number of hypothetical observations to support this guess. Denote this sample size \( K_j \).

Then the prior distribution can be constructed as following

\[
g(\theta_1, \ldots, \theta_M) \propto \prod_{j=1}^{M} \theta_j^{K_j g_j - 1}(1 - \theta_j)^{K_j (1 - g_j) - 1} \quad (2.11)
\]

Recall that if we choose the general link, then \( \theta_j = F(\gamma_j - 1 - x_j' \beta) \), where \( F \) is cumulative density function of some distribution. By using this relation, we can transform equation (2.11) back to the equation using \( \beta, \gamma_2, \ldots, \gamma_{k-1} \)

\[
g(\beta, \gamma_2, \ldots, \gamma_{k-1}) \propto \prod_{j=1}^{M} F(\gamma_j - 1 - x_j' \beta)^{K_j g_j - 1}(1 - F(\gamma_j - 1 - x_j' \beta))^{K_j (1 - g_j) - 1} f(\gamma_j - x_j' \beta) \quad (2.12)
\]

where \( f \) is the probability density function of the same distribution whose cdf is \( F \). For an illustration purpose, we choose the probit regression model, then the prior distribution is given by

\[
g(\beta, \gamma_2, \ldots, \gamma_{k-1}) \propto \prod_{j=1}^{M} \Phi(\gamma_j - 1 - x_j' \beta)^{K_j g_j - 1}(1 - \Phi(\gamma_j - 1 - x_j' \beta))^{K_j (1 - g_j) - 1} f(\gamma_j - x_j' \beta) \quad (2.13)
\]

So the posterior distribution using informative prior are proposal to the following distribution

\[
p(\beta, \gamma, Z_i | y) \propto \prod_{j=1}^{M} \Phi(\gamma_j - 1 - x_j' \beta)^{K_j g_j - 1}(1 - \Phi(\gamma_j - 1 - x_j' \beta))^{K_j (1 - g_j) - 1} f(\gamma_j - x_j' \beta) \\
\times \prod_{i=1}^{n} f(Z_i - x_i' \beta) I(\gamma_{y_i-1} \leq Z_i < \gamma_{y_i}) \quad (2.14)
\]
2.5 Review of Methods of Simulating from the Posterior Distribution

No matter whether we choose the informative or noninformative prior, we are facing a very complicate posterior distribution given by equation (2.14) or (2.10). So the traditional method of taking partial derivative with every component of regression parameter vector \( \beta \) and cutoff points \( \gamma_2, \ldots, \gamma_{k-1} \) and letting every partial derivative equal 0 usually does not have analytical solution. We have to resort on Markov Chain Monte Carlo (MCMC) method to get the simulated values of the unknown quantities of \( \gamma_2, \ldots, \gamma_{k-1}, \beta \) and \( Z \).

MCMC is a general technique for generating samples from a probability distribution in high dimension space. The key in MCMC is to construct an irreducible, aperiodic Markov chain for which the stationary distribution is the target function. A Markov chain is stochastic process \( \{x(t), t = 0, 1, 2, \ldots\} \) that takes on a finite or countable number of possible values and has the following property:

\[
p(x^{(t+1)} | x^{(t)}, \ldots, x^{(1)}) = p(x^{(t+1)} | x^{(t)})
\]  

(2.15)

where each \( x^{(t)} \) is called a state in the Markov Chain. In other word, equation (2.15) implies that the next state depends only on the present state, which is often called the Markov property. A Markov chain is called irreducible if, starting from any one of the states, it is possible to get to any other state (not necessarily in one jump). A Markov chain is periodic if it can visit certain portions of the state space only at certain regularly space intervals. State \( j \) has period \( d \) if the probability of going from state \( j \) to state \( j \) in \( n \) steps is 0 for all the \( n \) not divisible by \( d \). If every state in a Markov chain has the period \( d=1 \) then we call that Markov chain aperiodic.

There are two most frequently used algorithms in MCMC simulation, which are Metropolis-Hastings algorithm and Gibbs sampling. Two special cases of the Metropolis-Hastings algo-
rithm are the independence chain and random walk algorithms.

2.5.1 Metropolis-Hastings Algorithm

Metropolis-Hastings algorithm was first proposed by Metropolis et.al in Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) for investigating properties such as the equation of state for substances. After the introduction of Metropolis’s algorithm, Hastings (1970) generalized Metropolis’s algorithm and proposed relevant theory, techniques of application.

Let $f$ be the function that we are interested in, which is called the target function, and we want to simulate a Markov chain from the function $f$. The Metropolis-Hastings algorithm begins at $t=0$ with $x^{(0)}$ drawn from the proposal density $g(x)$. The essence of Metropolis-Hastings is as following:

1. Sample a candidate value $x^*$ from a proposal density $g(\cdot | x^{(t)})$

2. Compute the Metropolis-Hastings Ratio as

$$R(x^*, x^{(t)}) = \frac{f(x^*)g(x^{(t)} | x^*)}{f(x^{(t)})g(x^* | x^{(t)})}$$

3. Generate a random number from a uniform distribution on $(0,1)$—call this random number $u$. If the $R(x^*, x^{(t)}) > u$, then we accept $x^{(t+1)} = x^*$, otherwise, $x^{(t+1)} = x^{(t)}$, In other words, accept the candidate $x^*$ as the acceptance probability equals minimum between $R(x^*, x^{(t)})$ and 1.

Clearly Metropolis-Hastings algorithm guarantees that this simulated sequence of values $x^{(t)}, t = 0, 1, \cdots$ is a Markov chain because of the way of simulating $x^*$ in step 1. However, it does not assure that the simulated sequence $x^{(t)}, t = 0, 1, \cdots$ is irreducible and aperiodic.
2.5.2 Independent Chain Algorithm

It is usually difficult to compute the condition distribution \( g(x^{(t)}|x^*) \) and \( g(x^*|x^{(t)}) \) in the general Metropolis-Hastings algorithm. In order to simplify the computation, we choose the proposal function \( g \) such that \( g(x^*|x^{(t)}) = g(x^*) \), and this yields the independence chain algorithm. In this case, the Metropolis-Hastings Ratio becomes

\[
R(x^*, x^{(t)}) = \frac{f(x^*)g(x^{(t)})}{f(x^{(t)})g(x^*)}
\]

which makes the computation process much easier.

2.5.3 Random Walk Algorithm

The independence chain still needs to choose the proposal density \( g(\cdot) \), which sometimes is not an easy job especially when there are more than one parameter that we are interested in, so a revised Metropolis-Hastings algorithm called random walk algorithm is proposed. The idea of the random walk algorithm is generating a candidate \( x^* \) by \( x^* = x^{(t)} + \epsilon \), where \( \epsilon \) is a random variable from a distribution \( h(\epsilon) \). In this case, \( g(x^*|x^{(t)}) = h(x^* - x^{(t)}) \) and the common choices for \( h \) include uniform distribution over a ball centered at the origin, a scaled standard normal distribution and a scaled student t distribution. The reason why we want to choose a scaled distribution is to make sure that the simulated values are not highly correlated so that they can be “mixed” well.

2.5.4 Gibbs Sampling Algorithm

No matter independence chain algorithm or random walk algorithm, we did not look into the inside structure of the target function \( f \). When there are more than one parameter in the function \( f \), the conditional structure of one parameter on the rest parameters may have a nice distribution. Gibbs sampling takes advantage of this property and it turns out that sometimes it is a very efficient algorithm when the conditional distribution of one parameter
on the rest parameters has a nice distribution. The brief description of the Gibbs algorithm is as following:

1. choose an ordering the components of $x^{(t)}$

2. for each $i$ in the ordering chosen above, sample $x^*|x_{-i}^{(t)} \sim f(x_i|x_{-i}^{(t)})$, where $x_{-i}^{(t)}$ denote the rest components of $x^{(t)}$ without $x_{-i}^{(t)}$

3. once step 2 has been completed for each component of $x$ in the selected order, set $x^{(t+1)} = x^*$

4. repeat step 2 and step 3

2.6 Parameter Estimation for Ordinal Data

2.6.1 Traditional Method for Parameter Estimation

As discussed in Chapter 1, we can use iteratively reweighted least square (IRLS) algorithm to get the point estimator for the unknown cutoff points $\gamma_2, \cdots, \gamma_{k-1}$ and the regression parameter vector $\beta$. The polr function in package MASS for the software R can perform ordinal regression fitting using probit link, logistic link, complementary log log link and cauchit link, also the program PROC LOGISTIC in SAS can perform ordinal regression fitting using probit link, logistic link, complementary log log link.

2.6.2 MCMC Sampling from the Posterior Distribution

In order to use MCMC algorithms to simulate $\gamma_2, \cdots, \gamma_{k-1}, \beta$ and $Z_i$ from the posterior distribution for ordinal data given by equation (2.14) (using informative prior) or (2.10) (using noninformative prior), we will face two possible difficulties.

Firstly, in order to use Metropolis-Hastings algorithm, we have to find the proposal density. However, multivariate normal distributions are not well suited as the proposal
density for generating candidate vector for the cutoff points due to the constraints of $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_{k-1}$. Because of this constraint, every candidate drawn from a multivariate normal distribution that does not satisfy this constraint will be discarded. This can cause the simulating candidate points process inefficient so that a more complicated proposal density is expected to overcome this problem.

Secondly, if we increase the number of categories, then the probability of a candidate being accepted often decreases dramatically. To overcome this side effect, a combination of Metropolis-Hastings and Gibbs sampling is often used to sampling regression parameters from the complicated posterior distribution for ordinal data.

Several algorithms have been proposed to efficiently simulate the unknown regression parameter $\beta$ and the cutoff points $\gamma_2, \cdots, \gamma_{k-1}$ for the ordinal case, including Albert and Chib(1993), Cowles(1993), and Nandram and Chen(1996). Albert and Chib in 1993 applied the idea of latent variables to the ordinal data and used Gibbs sampling to sample regression parameter $\beta$, the unknown cutoff points $\gamma = (\gamma_2, \cdots, \gamma_{k-1})$ and latent variable $Z_i$, they also found the condition distribution of one parameter conditioned on the rest parameters as following:

1. $[\beta|y, Z_i, \gamma]$ is distributed as multivariate normal with mean $(X'X)^{-1}X'Z_i$ and variance-covariance matrix $(X'X)^{-1}$, where $X = (x'_1, \cdots, x'_n)$, $x'_i$ is the vector of the covariate for individual $i$.

2. $[Z_i|y_i, \beta, \gamma]$ is distributed as a normal distribution with mean $x'_i\beta$ and variance 1 with truncation by $\gamma_{y_{i-1}}$ and $\gamma_{y_i}$

3. $[\gamma|Z_i, y, \beta]$ is distributed as a uniform distribution on the interval $[\max\{\max\{Z_i : y_i = j\}, \gamma_j\}, \min\{\min\{Z_i : y_i = j + 1\}, \gamma_{j+1}\}]$

Albert and Chib’s algorithm is good in general, however, there is a problem with the simulation of $\gamma$ conditional on $[Z_i, y, \beta]$, sometimes the interval $[\max\{\max\{Z_i : y_i = j\}, \gamma_j\}, \min\{\min\{Z_i : y_i = j + 1\}, \gamma_{j+1}\}]$ is not wide so we get highly correlated simulated $\gamma$ values
and slow convergence. Having realized this problem, Cowles (1993) proposed an improved algorithm to simulate the cutoff points $\gamma$ using Metropolis-Hastings algorithm. Instead of simulating one variable at a time, Cowles used the following algorithm:

1. $[\beta|y, Z_i, \gamma]$ is distributed as multivariate normal with mean $(X'X)^{-1}X'Z_i$ and variance-covariance matrix $(X'X)^{-1}$, again $X = (x'_1, \cdots, x'_n)$, $x'_1$ is the vector of the covariate for individual $i$.

2. $[Z_i|y_i, \beta, \gamma]$ is distributed as a normal distribution with mean $x'_i\beta$ and variance 1 with truncation by $\gamma_{y_i - 1}$ and $\gamma_{y_i}$.

3. $[\gamma, Z_i|y, \beta]$ can be expressed as the product of $[\gamma|y, \beta]$ and $[Z_i|\gamma, y, \beta]$.

The key part in Cowles’s algorithm is the conditional distribution of $\gamma$ on $y$ and $\beta$, and Cowles (1993) gave the complicated distribution by equation (2.16).

\[
[\gamma|y, \beta] \propto \prod_{i:y_i=2} \{\Phi(\gamma_2 - x'_i\beta) - \Phi(-x'_i\beta)\} \prod_{i:y_i=3} \{\Phi(\gamma_3 - x'_i\beta) - \Phi(\gamma_2 - x'_i\beta)\} \\
\cdots \prod_{i:y_i=k-1} \{\Phi(\gamma_{k-1} - x'_i\beta) - \Phi(\gamma_{k-2} - x'_i\beta)\} \prod_{i:y_i=k} \{1 - \Phi(\gamma_{k-1} - x'_i\beta)\}
\] (2.16)

where $\Phi$ is the standard normal cumulative distribution function. Cowles used the Metropolis-Hasting algorithm to simulate the candidate $\gamma'_j$ from the following proposal distribution

\[
[\gamma'_j|\gamma_{j-1,new}, \gamma'_{j+1}^{\text{last iteration}}] \sim N(\gamma_{j+1}^{\text{last iteration}}, \sigma^2_{\gamma})
\] (2.17)

where $\gamma_{j-1,new} \leq \gamma'_j \leq \gamma_{j+1}^{\text{last iteration}}$, $\gamma_{j-1,new}$ is the simulated values for the $j$-th cutoff point, $\gamma_{j+1}^{\text{last iteration}}$ is the simulated value in the last iteration for $j+1$-th cutoff point.

However, as pointed out by Tierney (1994), the truncated normal distribution given by (2.17) might not serve as a good proposal distribution because it is not spread out enough. Cowles (1994) also pointed out that a good $\sigma^2_{\gamma}$ in equation (2.17) is hard to obtain even using a conventional updating scheme.
To solve the problem in the article of Cowles(1994), Nandram and Chen(1995) used a different strategy to accelerate Gibbs sampling convergence. Nandram and Chen reparameterized the model by

$$
\delta = \frac{1}{\gamma_{k-1}}, \gamma_j^* = \delta \gamma_j, j = 0, 1, \cdots, k, \beta^* = \delta \beta, Z^* = \delta Z
$$

where $k$ is the number of categories. Though it seems that Nandram and Chen introduced a new variable $\delta$ into the model, the conditional distribution becomes simpler after the reparameterization, which made the sampling process more efficient. The distribution is given by

1. $[\beta^* | y, z^*, \delta^2]$ is distributed as multivariate normal with mean $(X'X)^{-1}X'z^*$ and variance-covariance matrix $\delta^2(X'X)^{-1}$, again $X = (x'_1, \cdots, x'_n)$, $x'_i$ is the vector of the covariate for individual $i$.

2. $[z^*_i | y_i, \beta^*, \gamma^*]$ is distributed as a normal distribution with mean $x'_i \beta^*$ and variance $\delta^2$ with truncation by $\gamma^*_y$ and $\gamma^*_y$.

3. $[\delta^2 | z^*_i, y, \beta^*]$ has an inverse gamma distribution with the shape parameter $\frac{n+k+m}{2}$ and scale parameter $\frac{1}{2}(Z^*-X'\beta)'(z^*-X'\beta')$, in other words, $[\delta^2 | z^*_i, y, \beta^*] \sim IG(\frac{n+k+m}{2}, \frac{1}{2}(z^*-X'\beta)'(z^*-X'\beta'))$

4. The simulation of the rest cutoff points is based on Metropolis-Hastings algorithm using a Dirichlet distribution as the proposal distribution.

We are using the algorithm from Nandram and Chen (1996) to simulate the regression parameter $\beta$, the cutoff points $\gamma_2, \cdots, \gamma_{k-1}$ and the latent variables $Z$. The reason why we choose their algorithm is because Nandram and Chen’s algorithm provides faster convergence speed and smaller autocorrelation between the simulated values than Cowles’s algorithm.
2.7 Model Selection from a Bayesian Perspective

It is common that there are more than one variables in the data set and in most cases it is not wise to include all the variables in the model. We have to decide which variable should be included in the model before we can get the right information of the variables. Depending on the choice of variables included in the model, we usually have several competing models in hand, which leads to a model selection issue. One of the model selection from a Bayesian perspective is based on the concept of Bayes factor. Bayes factor is based on the ratio of marginal likelihood of two models, namely model M1 and model M2. Bayes factor is given by

$$\text{Bayes factor}(M_2, M_1) = \frac{p(y|M_2)}{p(y|M_2)} = \frac{\int p(y, \theta|M_2)d\theta}{\int p(y, \theta|M_1)d\theta} = \frac{\int p(\theta|M_2)p(y|\theta, M_2)d\theta}{\int p(\theta|M_1)p(y|\theta, M_1)d\theta}$$ (2.18)

If Bayes factor is significantly greater than 1, then we should choose model M2 over model M1. Some scholars preferred to take log base 10 of the Bayes factor value, and if they get a negative number, it means that we should choose model M2 over model M1.


2.8 Residuals from a Bayesian Perspective

Residual from a Bayesian perspective is much easier than a frequentist’s perspective. From a frequentist’s perspective, it is not clear to define the residual because the expected probability for each observation is a vector; however, with the help of introduction of latent variable, defining residuals from a Bayesian perspective becomes much easier because we get one single
response for each observation. The latent residual for observation $i$ is defined by

$$ r_{i,l} = Z_i - x_i'\beta $$

(2.19)

Theoretically, the latent residuals $r_{1,l,\ldots}, r_{n,l}$ are distributed as independently from distribution $F$. Deviation from the model structure are reflected by atypical values of these quantities drawn from $F$ and can be checked by traditional method such as Quantile-Quantile (QQ) plot. There are of course other kind of residuals, such as posterior predictive residual. Posterior predictive residual for ordinal data models are defined using the actual observed value and the posterior predicted $y$ value and it is defined as

$$ r_{i,pp} = y_i - y_i^* $$

(2.20)

where $y_i$ is an actual observed response for observation $i$, $y_i^*$ is the predicted response for observation $i$

### 2.9 Goodness-of-fit from a Bayesian Perspective

Another important aspect when we fit a model to a data set is to check the goodness-of-fit because goodness-of-fit measures how well a model fits the data set. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question, and the common choice of test statistic is deviance and Pearson Chi Square statistic. The goodness-of-fit from a Bayesian perspective is based on the deviance statistic given by

$$ D(y) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j) \log\left(\frac{I(y_i = j)}{p_{ij}}\right) $$

(2.21)

where $k$ is the number of categories, $n$ is the number of observations, $p_{ij}$ is the probability of observation $i$ falling in category $j$, $I(y_i = j)$ is the indicator function that observation $i$ falls
in category $j$. Instead of using the MLE of $p_{ij}$ in the traditional method, which depends on the asymptotic distribution of $\chi^2$ with certain degrees of freedom, we can use the simulated $\beta$ and cutoff points $\gamma_2, \ldots, \gamma_{k-1}$ to get the simulated values of $p_{ij}$ by equation (2.22).

\[ p_{ij} = \Phi(\gamma_{j-1} - x'_i \beta) \] (2.22)

where a probit regression model is assumed, then we can simulate the “future” observation $y^{rep}$ from a multinomial distribution with sample size equals 1 and probability vector $(p_{i1}, p_{i2}, \ldots, p_{ik})$. In this case, for each simulated values of $\beta$ and cutoff points $\gamma_2, \ldots, \gamma_{k-1}$, we can get one deviance statistic $D(y^{rep})$, as a result we can get $m$ simulated deviance $D(y^{rep})$, where $m$ is the number of simulations, call it $D_1(y^{rep}), D_2(y^{rep}), \ldots, D_m(y^{rep})$, then we compare the simulated deviance $D_1(y^{rep}), D_2(y^{rep}), \ldots, D_m(y^{rep})$ with the observed deviance $D(y)$. If the observed deviance $D(y)$ falls in the extreme end of simulated deviance values, then we conclude that the model for the ordinal data is not a suitable fit.
CHAPTER 3

Outlier Detection

3.1 A Frequentist’s Method to Detect Outliers

Fitting an ordinal probit regression model by maximum likelihood using the method of iterative reweighted least square (IRLS) is straightforward. In some software, for example, the function `polr` in R package `MASS` can fit a probit regression model to an ordinal data set and get the estimate of the vector of regression parameters $\beta$ and cutoff points $\gamma_1, \cdots, \gamma_{k-1}$. A probit regression model is fit on the grade data with covariate SAT in Table 1.1 in Chapter 1 and the result is given by Table 3.1.

Call:
`polr(formula = grade ~ sat, method = "probit")`

Coefficients:
sat
0.02382519

Intercepts:
1|2 2|3 3|4 4|5
11.22 12.67 13.49 14.78

Residual Deviance: 73.49
AIC: 83.49

Table 3.1: The result of probit regression for ordinal data using `polr` function in R package `MASS`.

Note that from Table 3.1 we have $k=5$ categories so we need four intercepts to separate
these five categories. To calculate the estimated probability of each observation falling in a specific category, for example, the estimated probability that observation $i$ falls in category $j$, we use the formula

$$\hat{p}_{ij} = \Phi(\hat{\gamma}_j - \text{SAT}_i \times \hat{\beta}) - \Phi(\hat{\gamma}_{j-1} - \text{SAT}_i \times \hat{\beta})$$ (3.1)

where the $\hat{\gamma}_j$ is the estimated $j$th cutoff point. Based on equation (3.1), we get the predicted probability of each observation falling in each category, which is given by Table 3.2.

From Table 3.2 we can see the probability of one observation falling in each category. However, information given by Table 3.2 is difficult for us to judge which observation is an outlier just looking at the probability vector of each observation falling in each category. A graphic way of looking for outliers is to look at the bar plot of fitted probabilities for each observation. We plot the fitted probability for each observation falling in each category against observation ID and get Figure 3.1.

From Figure 3.1 we can graphically see the fitted probability of falling in each category for all observations and this barplot can help us identify possible outliers based on the interpretation of a probability. As is known to all, a probability is the chance that an event happening in the repeated experiments. If the probability of a particular event happening is small, then we did not expect that event to happen in one experiment. Notice that there are two suspicious observations in Figure 3.1, one is observation number 19, who actually gets an “F” but the fitted probability of getting an “F” is very small; the other one is observation number 30, who actually gets an “A”, but the fitted probability of getting an “A” is also very small. Based on this, we can identify observation number 19 and number 30 as the possible outliers. We can not conclude those two observations are outliers due to the fact the cut-point of small probability varies by individual.

Since the fitted probabilities for each observation is a vector, it is not clear on how to define residuals and detect outliers from a frequentist’s perspective. Some research has been
<table>
<thead>
<tr>
<th>Obs ID</th>
<th>Grade</th>
<th>Prob of F</th>
<th>Prob of D</th>
<th>Prob of C</th>
<th>Prob of B</th>
<th>Prob of A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D</td>
<td>0.1</td>
<td>0.47</td>
<td>0.27</td>
<td>0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>0.07</td>
<td>0.42</td>
<td>0.3</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>0.04</td>
<td>0.34</td>
<td>0.32</td>
<td>0.27</td>
<td>0.04</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
<td>0</td>
<td>0.11</td>
<td>0.24</td>
<td>0.46</td>
<td>0.18</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>0</td>
<td>0.12</td>
<td>0.25</td>
<td>0.46</td>
<td>0.17</td>
</tr>
<tr>
<td>6</td>
<td>B</td>
<td>0.01</td>
<td>0.14</td>
<td>0.27</td>
<td>0.44</td>
<td>0.14</td>
</tr>
<tr>
<td>7</td>
<td>C</td>
<td>0.01</td>
<td>0.16</td>
<td>0.28</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>0</td>
<td>0.03</td>
<td>0.12</td>
<td>0.45</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>0.02</td>
<td>0.24</td>
<td>0.31</td>
<td>0.36</td>
<td>0.07</td>
</tr>
<tr>
<td>10</td>
<td>C</td>
<td>0.04</td>
<td>0.35</td>
<td>0.32</td>
<td>0.25</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>0.01</td>
<td>0.14</td>
<td>0.27</td>
<td>0.44</td>
<td>0.14</td>
</tr>
<tr>
<td>12</td>
<td>B</td>
<td>0.1</td>
<td>0.47</td>
<td>0.27</td>
<td>0.15</td>
<td>0.01</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>0.01</td>
<td>0.15</td>
<td>0.27</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>14</td>
<td>C</td>
<td>0</td>
<td>0.11</td>
<td>0.24</td>
<td>0.46</td>
<td>0.18</td>
</tr>
<tr>
<td>15</td>
<td>B</td>
<td>0.01</td>
<td>0.15</td>
<td>0.27</td>
<td>0.44</td>
<td>0.13</td>
</tr>
<tr>
<td>16</td>
<td>D</td>
<td>0.5</td>
<td>0.42</td>
<td>0.06</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>B</td>
<td>0</td>
<td>0.07</td>
<td>0.18</td>
<td>0.48</td>
<td>0.27</td>
</tr>
<tr>
<td>18</td>
<td>D</td>
<td>0.02</td>
<td>0.26</td>
<td>0.32</td>
<td>0.34</td>
<td>0.06</td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>0.02</td>
<td>0.26</td>
<td>0.32</td>
<td>0.34</td>
<td>0.06</td>
</tr>
<tr>
<td>20</td>
<td>B</td>
<td>0</td>
<td>0.1</td>
<td>0.23</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>21</td>
<td>A</td>
<td>0</td>
<td>0.05</td>
<td>0.17</td>
<td>0.48</td>
<td>0.3</td>
</tr>
<tr>
<td>22</td>
<td>D</td>
<td>0.14</td>
<td>0.5</td>
<td>0.24</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>23</td>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.23</td>
<td>0.75</td>
</tr>
<tr>
<td>24</td>
<td>B</td>
<td>0</td>
<td>0.1</td>
<td>0.23</td>
<td>0.47</td>
<td>0.19</td>
</tr>
<tr>
<td>25</td>
<td>F</td>
<td>0.58</td>
<td>0.37</td>
<td>0.04</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>C</td>
<td>0</td>
<td>0.08</td>
<td>0.2</td>
<td>0.48</td>
<td>0.24</td>
</tr>
<tr>
<td>27</td>
<td>D</td>
<td>0.35</td>
<td>0.51</td>
<td>0.12</td>
<td>0.03</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>B</td>
<td>0.01</td>
<td>0.22</td>
<td>0.3</td>
<td>0.38</td>
<td>0.08</td>
</tr>
<tr>
<td>29</td>
<td>B</td>
<td>0.03</td>
<td>0.28</td>
<td>0.32</td>
<td>0.32</td>
<td>0.05</td>
</tr>
<tr>
<td>30</td>
<td>A</td>
<td>0.03</td>
<td>0.31</td>
<td>0.32</td>
<td>0.29</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 3.2: The fitted probability table for all students. The first column is the observation ID, the second column is the actual grade each student gets, and the third column to the last column are the probabilities of each observation falling in category F, category D, category C, category B, and category A respectively.
Figure 3.1: The barplot of fitted probabilities for all observations falling in each category versus observation ID. The x-axis is the observation ID, the y-axis is the fitted probabilities for each observation falling in each category, and the marked grade is the actual grade that this observation obtains. For each observation, segments from bottom to top are “F”, “D”, “C”, “B”, and “A”
conducted on how to define residuals for ordinal data, however, as pointed out by McCullagh (1980), several factors complicate this situations of defining residuals thus it also complicates the process of finding outliers. “Firstly, the residuals, however standardized, must take one of a limited number of possible values which causes problems in rankit or related plots when the cell counts are small. Secondly, it is far from clear in general that cell residuals are the relevant quantities to examine”. McCullagh (1980) use the contribution to the likelihood ratio statistic from each multinomial sample as a residual. However, this kind of residual is always positive and does not reflect the direction of departure of the fitted value from the observed value.

Fuchs and Kenett (1980) propose the M test based on the maximum adjusted residuals to detect outliers. In their context, they consider only one variable \( n = (n_1, \cdots, n_k) \) from a multinomial distribution with parameter \( N \) and \( p_i \), where \( N = \sum_{i=1}^{k} n_i \), and \( p_i, i = 1, \cdots, k \). They define

\[
Z_i = \frac{n_i - N\hat{p}_i}{\sqrt{N\hat{p}_i(1 - \hat{p}_i)}} \quad i = 1, \cdots, k
\]
as the adjusted residuals, where \( \hat{p}_i \) is the fitted probability of the observation fall in category \( i \). In this case, Fuchs and Kenett get \( k \) \( Z_i' \)'s for each observation, and they pick the largest \( Z_i' \)'s and compare it with the approximated critical value of the M test, which is approximated by \( \Phi^{-1}\left[1+\frac{(1-\alpha)^\frac{1}{2}}{2}\right] \) for the two-sided test, where \( \alpha \) is the significance level. We replicate this process and get the result shown in Table 3.3.

From Table 3.3, we can see that observation number 4, 19 and 30 have unusually large \( Z_i \) values, which are 3.13, 9.96 and 5.71 respectively. We then compare these two values with the approximated critical value of the M test, which is \( \Phi^{-1}\left[1+\frac{(1-\alpha)^\frac{1}{2}}{2}\right]=2.57 \), so we conclude that observation number 4, 19 and 30 are outliers based on the M test method. One problem with the M test method is Fuchs and Kenett do not give a specific asymptotic distribution for the M test so they use the approximated critical value, which is inaccurate for a small sample data set.
Table 3.3: Using M test in Fuchs and Kenett (1980) to detect outliers. The first and third column is observation ID, the second and fourth column is the $Z_i$ value.

<table>
<thead>
<tr>
<th>Obs ID</th>
<th>$Z_i$</th>
<th>Obs ID</th>
<th>$Z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.99</td>
<td>16</td>
<td>1.07</td>
</tr>
<tr>
<td>2</td>
<td>1.11</td>
<td>17</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>1.70</td>
<td>18</td>
<td>1.75</td>
</tr>
<tr>
<td>4</td>
<td>3.13</td>
<td>19</td>
<td>9.96</td>
</tr>
<tr>
<td>5</td>
<td>1.71</td>
<td>20</td>
<td>0.99</td>
</tr>
<tr>
<td>6</td>
<td>1.05</td>
<td>21</td>
<td>1.58</td>
</tr>
<tr>
<td>7</td>
<td>1.53</td>
<td>22</td>
<td>0.90</td>
</tr>
<tr>
<td>8</td>
<td>1.27</td>
<td>23</td>
<td>0.54</td>
</tr>
<tr>
<td>9</td>
<td>1.38</td>
<td>24</td>
<td>0.99</td>
</tr>
<tr>
<td>10</td>
<td>1.37</td>
<td>25</td>
<td>0.88</td>
</tr>
<tr>
<td>11</td>
<td>1.05</td>
<td>26</td>
<td>1.99</td>
</tr>
<tr>
<td>12</td>
<td>2.61</td>
<td>27</td>
<td>0.87</td>
</tr>
<tr>
<td>13</td>
<td>1.56</td>
<td>28</td>
<td>1.24</td>
</tr>
<tr>
<td>14</td>
<td>1.73</td>
<td>29</td>
<td>1.36</td>
</tr>
<tr>
<td>15</td>
<td>1.08</td>
<td>30</td>
<td>5.71</td>
</tr>
</tbody>
</table>

Another way to detect outliers is using the individual deviance. Recall that the deviance is defined as

$$D(y) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j) \log \left( \frac{I(y_i = j)}{p_{ij}} \right)$$  \hspace{1cm} (3.2)

where $\hat{p}_{ij}$ is the estimated probability of observation $i$ falling in category $j$, $I(y_i = j)$ is the indicator function of observation $i$ falling in category $j$. The deviance statistic can generally be used as a diagnostic tool of goodness-of-fit and model selection. Related to this deviance statistic, we can look at the individual contribution to the overall deviance. The individual deviance is defined by

$$D_i(y) = 2 \sum_{j=1}^{k} I(y_i = j) \log \left( \frac{I(y_i = j)}{\hat{p}_{ij}} \right)$$  \hspace{1cm} (3.3)

The individual deviance can be used as a method of detecting outliers. As described by Johnson and Albert (1999), "observations that contribute disproportionably to the overall deviance statistic should be identified with suspicion". Following the formula given by equation (3.3), we get one individual deviance statistic for each observation, as a total we can get $n$ individual deviance statistic, we can plot these $n$ individual deviances against observation.
From Figure 3.2 we can see observation number 19 and number 30 have unusually large individual deviance statistic values, so we identify them out with suspicion, however, we can not conclude that they are really outliers just based on individual deviance statistic because we do not have a specific quantity to measure how observations contribute disproportionally to the overall deviance.

There are certainly other methods to detect outliers, no matter which method we choose, we are facing the common problem of relying on the asymptotic distribution of the test statistic to draw a conclusion. To overcome these problems, we turn to a Bayesian perspective to detect outliers.
3.2 A Bayesian Perspective to Detect Outliers

3.2.1 Latent Residuals for Ordinal Regression Model

Recall that the latent variables for ordinal regression model is introduced by

\[ Z_i = x'_i \beta + \epsilon_i \]  \hspace{1cm} (3.4)

where \( \epsilon_i \) has a distribution according to the “link” function F. After the introduction of latent variables into the ordinal regression model, we can then define a new type of residual called latent residuals by

\[ r_{iL} = Z_i - x'_i \beta \]  \hspace{1cm} (3.5)

One of the benefits of introducing latent residuals is that we have only a single residual for each observation, so we are not constrained by the dimension of response variables. Another benefit of introducing latent residuals is that we can have the well-defined distribution of latent residuals so it is easier to detect outliers.

The latent residual \( r_{1L}, \cdots, r_{nL} \) are a priori an independent sample distributed according to the “link” function F. For example, if we choose a probit regression model, then \( r_{iL}, \cdots, r_{nL} \) are a priori independently distributed as a standard normal distribution; if we pick the logistic regression model, then \( r_{iL}, \cdots, r_{nL} \) are a priori independently distributed as a standard logistic distribution. After we observe the data set, the distribution of latent residuals will change from the prior distribution to the posterior distribution. We can use information from the distribution of latent residuals to detect outliers.

To inspect the posterior distribution of latent residuals, we need to get the expression for the posterior distribution of the \( i \)th latent residuals \( [r_{iL}|y] \). Note that the posterior distribution of latent residual \( [r_{iL}|y] \) can be expressed as following

\[ f(r_{iL}|y) = \int f(r_{iL}, \beta|y) d\beta = \int f(r_{iL}|y, \beta) f(\beta|y) d\beta \]  \hspace{1cm} (3.6)
so in order to find the posterior distribution of latent residuals, we have to find the expressions for \( f(r_{iL}|y, \beta) \).

Notice that \( y_i \) can be redefined using latent variables as

\[
y_i = \begin{cases} 
1 & \text{if } Z_i < \gamma_1 \\
2 & \text{if } \gamma_1 < Z_i < \gamma_2 \\
3 & \text{if } \gamma_2 < Z_i < \gamma_3 \\
4 & \text{if } \gamma_3 < Z_i < \gamma_4 \\
5 & \text{if } \gamma_4 < Z_i 
\end{cases}
\]

where \( y_i = j, j = 1, \cdots, k \) denote that observation \( i \) falls in category \( j \). Albert and Chib (1994) give the distribution of latent residuals conditional on \( \beta \) for the binary response data case. Following their idea, we extend their result to the ordinal case and get the posterior distribution of \([r_{iL}|y, \beta]\) for the probit regression model, which is given by

\[
f(r_{iL}|y_i, \beta) = \begin{cases} 
\frac{\phi(r_{iL})}{\Phi(\gamma_1 - x'_i \beta)} I(r_{iL} < \gamma_1 - x'_i \beta) & \text{if } y=1 \\
\frac{\phi(r_{iL})}{\Phi(\gamma_2 - x'_i \beta) - \Phi(\gamma_1 - x'_i \beta)} I(\gamma_1 - x'_i \beta < r_{iL} < \gamma_2 - x'_i \beta) & \text{if } y=2 \\
\frac{\phi(r_{iL})}{\Phi(\gamma_3 - x'_i \beta) - \Phi(\gamma_2 - x'_i \beta)} I(\gamma_2 - x'_i \beta < r_{iL} < \gamma_3 - x'_i \beta) & \text{if } y=3 \\
\frac{\phi(r_{iL})}{\Phi(\gamma_4 - x'_i \beta) - \Phi(\gamma_3 - x'_i \beta)} I(\gamma_3 - x'_i \beta < r_{iL} < \gamma_4 - x'_i \beta) & \text{if } y=4 \\
\frac{\phi(r_{iL})}{1 - \Phi(\gamma_4 - x'_i \beta)} I(r_{iL} > \gamma_4 - x'_i \beta) & \text{if } y=5 
\end{cases}
\]

where \( \phi() \) is the probability density function of a standard normal distribution, \( \Phi() \) is the cumulative density function of a standard normal distribution, and \( I() \) is the indicator function. Note that the posterior distribution of \( r_{iL} \) conditional on \( y \) and \( \beta \) is a truncated normal density.

We compute the posterior mean of latent residual \( r_{iL} \) conditional on \( \beta \) for different case
of $y_i$ and get the result given by

$$E[r_{iL}|y_i, \beta] = \begin{cases} 
-\frac{\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_1-x'_i\beta)} & \text{if } y_i=1 \\
-\frac{\phi(\gamma_2-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_2-x'_i\beta)-\Phi(\gamma_1-x'_i\beta)} & \text{if } y_i=2 \\
-\frac{\phi(\gamma_3-x'_i\beta)\phi(\gamma_2-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_3-x'_i\beta)-\Phi(\gamma_2-x'_i\beta)} & \text{if } y_i=3 \\
-\frac{\phi(\gamma_4-x'_i\beta)\phi(\gamma_3-x'_i\beta)\phi(\gamma_2-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_4-x'_i\beta)-\Phi(\gamma_3-x'_i\beta)} & \text{if } y_i=4 \\
\frac{\phi(\gamma_4-x'_i\beta)}{1-\Phi(\gamma_4-x'_i\beta)} & \text{if } y_i=5
\end{cases}$$

Also we compute the posterior variance of latent residual $r_{iL}$ conditional on $\beta$ for different case of $y_i$ and get the result given by

$$\text{Var}[r_{iL}|y_i, \beta] = \begin{cases} 
1 - \frac{(\gamma_1-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_1-x'_i\beta)} - (\frac{\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_1-x'_i\beta)})^2 & \text{if } y_i=1 \\
1 - \frac{(\gamma_2-x'_i\beta)\phi(\gamma_2-x'_i\beta)-\phi(\gamma_1-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_2-x'_i\beta)-\Phi(\gamma_1-x'_i\beta)} - (\frac{\phi(\gamma_2-x'_i\beta)}{\Phi(\gamma_2-x'_i\beta)-\Phi(\gamma_1-x'_i\beta)})^2 & \text{if } y_i=2 \\
1 - \frac{(\gamma_3-x'_i\beta)\phi(\gamma_3-x'_i\beta)-\phi(\gamma_2-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_3-x'_i\beta)-\Phi(\gamma_2-x'_i\beta)} - (\frac{\phi(\gamma_3-x'_i\beta)}{\Phi(\gamma_3-x'_i\beta)-\Phi(\gamma_2-x'_i\beta)})^2 & \text{if } y_i=3 \\
1 - \frac{(\gamma_4-x'_i\beta)\phi(\gamma_4-x'_i\beta)-\phi(\gamma_3-x'_i\beta)\phi(\gamma_2-x'_i\beta)\phi(\gamma_1-x'_i\beta)}{\Phi(\gamma_4-x'_i\beta)-\Phi(\gamma_3-x'_i\beta)} - (\frac{\phi(\gamma_4-x'_i\beta)}{\Phi(\gamma_4-x'_i\beta)-\Phi(\gamma_3-x'_i\beta)})^2 & \text{if } y_i=4 \\
1 + \frac{(\gamma_4-x'_i\beta)\phi(\gamma_4-x'_i\beta)}{1-\Phi(\gamma_4-x'_i\beta)} - (\frac{\phi(\gamma_4-x'_i\beta)}{1-\Phi(\gamma_4-x'_i\beta)})^2 & \text{if } y_i=5
\end{cases}$$

The posterior mean of latent residuals $[r_{iL}|y]$ can be computed by

$$E(r_{iL}|y) = E[E(r_{iL}|y, \beta)] \quad (3.7)$$

using the simulated draws of $\beta$. The posterior mean and posterior variance of latent residuals $[r_{iL}|y]$ may be significantly different from the prior moment (the prior mean of latent residuals is 0 and the prior variance of latent residuals is 1) due to the truncated normal distribution.

So far we have two methods to compute the posterior mean of latent residuals, one method is that we can approximate the posterior mean of the latent residuals $[r_{iL}|y]$ using
equation (3.6) and the simulated draws of $\beta$, the other method is that we can compute the posterior mean of the latent residuals using the simulated draws of latent variable $Z_i$ and $\beta$ by equation (3.5). We wonder if there is an difference between these two methods. One way to compare these two methods is by plotting the posterior means of latent residuals from these two methods against observation ID and get Figure 3.3.

![Figure 3.3: The posterior mean of latent residuals against observation ID. The x-axis is the observation ID, the y-axis are the posterior mean of latent residuals from two methods: the + sign is the posterior mean of latent residuals using equation (3.6), and the circle is the posterior mean of the latent residuals using equation (3.5).](image)

From Figure 3.3, we can see that posterior mean of latent residuals using equation (3.6) agrees with that using equation (3.5), we can also see that the posterior mean of latent residuals are far from 0 for some observations. The result in Figure 3.3 proves that our formula of the posterior mean of latent residuals is right so we can use equation (3.5) to get the simulated draws of latent residuals and compute the mean of the simulated latent residuals instead of using complicated equation (3.6). From a practical viewpoints, using equation (3.5) to compute the posterior mean of latent residuals is more attractive because
one of the byproducts of Nandram and Chen (1995) algorithm is that we can simulate latent variable $Z_i$.

To utilize the posterior distribution of latent residuals to detect outliers, one way is to consider the probability that the absolute value of latent residuals are greater than a pre-specified value $SK$, i.e., $p(|r_{iL}| > SK|y_i = j)$. Outlying observations correspond to residual densities that have locations away from zero thus we can look for large probability values of $p(|r_{iL}| > SK|y_i = j)$ for some constant value $SK$. Here we choose the pre-specified value $SK=3$. If the probability of $p(|r_{iL}| > SK|y_i = j)$ is large, then we can conclude the corresponding observation is an outlier. Note that $p(|r_{iL}| > SK|y_i = j)$ can be expressed as

$$p(|r_{iL}| > SK|y_i = j) = E[p(|r_{iL}| > SK|y_i = j, \beta)]$$  \hspace{1cm} (3.8)

For an illustration purpose, we only show the computation $p(|r_{iL}| > SK|y_i = j)$ for $y_i = 2$ case, the rest cases can be solved similarly. For $y_i = 2$ case, the distribution of latent residuals conditional on $\beta$ is given by

$$[r_{iL}|y_i, \beta] = \frac{\phi(r_{iL})}{\Phi(\gamma_2 - x_i'\beta) - \Phi(\gamma_1 - x_i'\beta)} I(\gamma_1 - x_i'\beta < r_{iL} < \gamma_2 - x_i'\beta)$$  \hspace{1cm} (3.9)

then the $p(|r_{iL}| > SK|y_i = j, \beta)$ is given by

$$p(|r_{iL}| > SK|y_i = 2, \beta) = \begin{cases} 1 & \text{SK} < \gamma_1 - x_i'\beta \\ 0 & \text{SK} > \gamma_2 - x_i'\beta \\ \frac{\Phi(\gamma_2 - x_i'\beta) - \Phi(SK)}{\Phi(\gamma_2 - x_i'\beta) - \Phi(\gamma_1 - x_i'\beta)} & \gamma_1 - x_i'\beta \leq \text{SK} \leq \gamma_2 - x_i'\beta \\ 0 & -\text{SK} < \gamma_1 - x_i'\beta \\ 1 & -\text{SK} > \gamma_2 - x_i'\beta \\ \frac{\Phi(-SK) - \Phi(\gamma_1 - x_i'\beta)}{\Phi(\gamma_2 - x_i'\beta) - \Phi(\gamma_1 - x_i'\beta)} & \gamma_1 - x_i'\beta \leq -\text{SK} \leq \gamma_2 - x_i'\beta \end{cases}$$

then the probability $p(|r_{iL}| > SK|y_i = 2)$ can be obtained by averaging the values of $p(|r_{iL}| > SK|y_i = 2, \beta)$ over simulated draws of $\beta$. We use the above expression for the case of
\( y_i = 1, \ldots, 5 \) and plot the probability against observation ID, which is given by Figure 3.4.

Figure 3.4: The probability of absolute values of latent residuals are greater than or equals to \( SK=3 \) against observation ID. The x-axis is the observation ID, the y-axis is the probability of \( p(|r_i| > SK|y_i = j) \). Larger probabilities corresponds to outliers.

From Figure 3.4 we can see that observation number 19 and 30 have unusually large probabilities of \( p(|r_i| > SK|y_i = j) \), which are 0.17 and 0.09 respectively, so we conclude that these two observations are outliers.

### 3.2.2 Apply 3\( \sigma \) Rule to Latent Residuals to Detect Outliers

As stated before, the latent residual \( r_{iL}, \ldots, r_{nL} \) are nominally an independent sample from a standard normal distribution if we choose the probit regression model, so we can apply the 3\( \sigma \) rule for the normal distribution to the simulated latent residuals to detect outliers. The 3\( \sigma \) rule for a normal distribution with mean \( \mu \) and standard deviation \( \sigma \) states that 68.27% of the data fall in the \( \sigma \) area of \( \mu \), namely, \( (\mu - \sigma, \mu + \sigma) \), 95.45% of the data fall in the
2σ area of μ, and 99.73 % of the data fall in the 3σ area of μ. If one observation is not an outlier, then we expect that 99.73 % of latent residuals by equation (3.5) using simulated draws of Zi and β should fall in the interval (-3,3). If the probability of the simulated latent residuals falling out (-3,3) is large, then we conclude the corresponding observation is an outlier.

The detailed process goes as following: Using the simulation algorithms in Nandram and Chen (1996), for each iteration, we can simulate one regression parameter vector β, one vector of cutoff points γ = (γ2, · · ·, γk−1) and one latent variable Zi for observation i, then we can compute latent residuals for observation i by \( r_{iL} = Z_i - x_i' \beta \). As a result of this, we can get m latent residuals for each observation, where m is the number of simulations. We can then compute the proportion of the simulated latent residuals \( r_{iL} \) falling out the interval (-3,3) for observation i. We plot the proportion of simulated latent residuals falling out the interval (-3,3) against observation ID and get Figure 3.5.

![Figure 3.5](image)

Figure 3.5: Apply 3σ rule to latent residuals to detect outliers. The x-axis is the observation ID, the y-axis is the proportions of simulated latent residuals falling out the interval (-3,3) for each observation. Large proportion of latent residuals falling out the interval (-3,3) indicates the corresponding observation is an outlier.
From Figure 3.5 we can see that observation number 19 and observation number 30 has unusually large probabilities of simulated latent residuals falling out the interval (-3,3), which are 55.86% and 24.94% respectively, so we conclude that observation number 19 and observation number 30 are outliers.

### 3.2.3 Quantile-Quantile Plot for Latent Residuals to Detect Outliers

In a linear regression model, the Quantile-Quantile (abbreviated as QQ) plot is a common tool for model assumption checking and outlier detection. In theory, a QQ plot can be used to check whether one sample really comes from a specific target population. For the given sample we can compute the quantile and we compute the theoretical quantile of that target distribution. We can make a scatterplot the quantile from given sample data against quantile from the theoretical target distribution. If the given sample data indeed come from the target distribution, then these two quantiles should be approximately the same, so the scatterplot should look like a 45 degree straight line. In fact, when we fit a linear regression model on a data set, it is a good idea to check the QQ plot for the residuals to see whether these residuals are normally distributed and look for outliers.

Following the QQ plot for a linear regression idea, we want to make a QQ plot for the latent residuals by equation (3.5) to check for outliers. One difficulty about extending the QQ plot for a linear regression model to the latent residuals case is that we have many simulated latent residuals for each observation. In order to make a QQ plot, we need to summarize the simulated latent residuals so that we get only one value for each observation. One way to summarize the simulated latent residuals is to sort the simulated latent residuals for each iteration, then we can compute the mean of the sorted latent residuals and count the proportion of the most appearing observation. The detailed summarizing process is as following:

In order to deal with this large amount of data more easily, we put the $n \times m$ latent
residuals into a matrix form with dimension $n \times m$, where $n$ is the number of observations and $m$ is the number of iterations. In this case, each column in the latent residual matrix corresponds with one iteration, and each row corresponds with one observation. To understand the process better, we present the $n \times m$ latent residual matrix by

$$
\begin{bmatrix}
\text{iter}_1 & \cdots & \text{iter}_m \\
\text{obs}_1 & - & \cdots & - \\
\vdots & \vdots & \vdots & \vdots \\
\text{obs}_n & - & \cdots & -
\end{bmatrix}
$$

Then for each iteration, we sort the latent residuals from the smallest to the largest, also we write down the corresponding observation number for the sorted latent residuals for this iteration, and put them into another matrix. As a result, we will get two different matrices, one matrix is the sorted latent residuals with dimension $n \times m$, where the first row corresponds to the smallest latent residuals in all the iterations, the last row corresponds to the largest latent residual in all the iterations; the second matrix is the observation number matrix with dimension $n \times m$, where the first column is the observation number corresponding with the sorted latent residuals in the first column of the first matrix. A graph show of two $n \times m$ matrices containing the sorted latent residual and corresponding observation ID is given by

$$
\begin{bmatrix}
\text{iter}_1 & \cdots & \text{iter}_m \\
\text{row}_1 & \min\{r_{iL}^{(1)}\} & \cdots & \min\{r_{iL}^{(m)}\} \\
\vdots & \vdots & \vdots & \vdots \\
\text{row}_n & \max\{r_{iL}^{(1)}\} & \cdots & \max\{r_{iL}^{(m)}\}
\end{bmatrix}
\begin{bmatrix}
\text{iter}_1 & \cdots & \text{iter}_m \\
\text{row}_1 & \text{obs ID}_1^{(1)} & \cdots & \text{obs ID}_1^{(m)} \\
\vdots & \vdots & \vdots & \vdots \\
\text{row}_n & \text{obs ID}_n^{(1)} & \cdots & \text{obs ID}_n^{(m)}
\end{bmatrix}
$$

where $\min\{r_{iL}^{(1)}\}$ is the smallest latent residual for the first iteration, $\max\{r_{iL}^{(1)}\}$ is the largest latent residuals for the first iteration, $\min\{r_{iL}^{(m)}\}$ is the smallest latent residuals for the $m$th iteration, and $\max\{r_{iL}^{(m)}\}$ is the largest latent residuals for the $m$th iteration; $\text{obs ID}_1^{(1)}$ is the observation ID corresponding to the smallest latent residuals for the first iteration,
obs ID\(_{(n)}^{(1)}\) is the observation ID corresponding to the largest latent residuals for the first iteration, obs ID\(_{1}^{(m)}\) is the observation ID corresponding to the smallest latent residuals for the \(m\)th iteration, and obs ID\(_{n}^{(m)}\) is the observation ID corresponding to the largest latent residuals for the \(m\)th iteration.

The remaining process is straightforward: For the matrix containing the sorted latent residuals, compute the mean of the latent residuals for each row, as a result, we will get a \(n\) latent residuals. For the matrix containing the observation ID corresponding to the matrix containing sorted latent residuals, we count the most frequently appearing observation ID for each row and write down the proportion of appearing. Nominally, the mean of sorted latent residuals should be distributed as a standard normal distribution. To make a QQ plot, we can simulate \(n\) random draws from a standard normal distribution and compute the quantile of these \(n\) simulated random draws, we can then plot the mean of sorted latent residuals against the quantile of these \(n\) random variables, which leads to Figure 3.6.

In Figure 3.6 (#19, 92%) means that observation number 19 appearing 92 percent of the time of having the smallest latent residuals in all the iterations, (#30, 72.28%) means that observation number 30 appearing 72.28 percent of the time of having the largest latent residual in all the iterations. From Figure 3.6 we can see that observation number 19 and observation number 30 stay clearly from the 45-degree straight line, so we conclude that these two observations are outliers.

### 3.2.4 Use Bayes Factor to Detect Outliers

Another method that we can adopt is using Bayes factor to detect outliers. Suppose that the model is defined as

\[
y \sim f(y|\theta)
\]  

(3.10)
Figure 3.6: QQ plot for the mean of sorted latent residuals. The x-axis is the quantile of the random variables from a standard normal distribution, the y-axis is the row mean of sorted latent residuals using the matrix containing the sorted latent residuals. Observations clearly far from the straight line are considered to be outliers.
The parameter $\theta$ is assigned a prior distribution

$$\theta \sim g(\theta) \quad (3.11)$$

then the posterior distribution of $\theta$ is given by

$$g(\theta|y) = f(y|\theta)g(\theta) \quad (3.12)$$

the marginal likelihood is defined as

$$m(y) = \int f(y|\theta)g(\theta)d\theta \quad (3.13)$$

The Bayes factor is the ratio of marginal likelihood from two models and it is defined

$$BF_{21} = \frac{m(y|M_2)}{m(y|M_1)} \quad (3.14)$$

we write $m(y|M_2)$ only to emphasize that this is the marginal likelihood for model M2. If model M2 is a better fit for a given data set than model M1, then the Bayes factor $BF_{21}$ will be a large number to indicate that we should support model M2. In order to use Bayes factor to detect outliers, each time we propose a new model containing a new data set.

- Model $M_{-j}$: The data set is the original data set with observation $j$ deleted.
- Model $M_0$: The data is the original data set.

Assuming that we can use a probit regression model and we put vague uniform priors on the parameters, then the marginal likelihood for model $M_{-j}$ is given by

$$m(y|M_{-j}) = \int \prod_{i \neq j} [\Phi(\gamma_{yi} - x_i^\prime \beta) - \Phi(\gamma_{yi-1} - x_i^\prime \beta)]d\theta \quad (3.15)$$
We can get the estimated marginal likelihood for both models using the function \texttt{laplace}
in R package \textit{LearnBayes}, then we can compute the Bayes factor of these two models. If observation \( j \) is an outlier, then model \( M_{-j} \) with observation \( j \) removed will get a more accurate estimate than the original data set, so we expect the Bayes factor of Model \( M_{-j} \) over model \( M_0 \) to be unusually large, which indicates that we should support the data set with observation \( i \) removed. Based on this idea, we compute the Bayes factor of model \( M_{-j} \) over model \( M_0 \) for all the observations and get Figure 3.7.

![Figure 3.7](image.png)

**Figure 3.7**: Using Bayes factor to detect outliers. The x-axis is the observation ID, the y-axis is the logarithm (base 10) of Bayes factor of model \( M_{-j} \) over model \( M_0 \), where model \( M_{-j} \) is the original data set with observation \( j \) deleted, model \( M_0 \) is the original data set. Large Bayes factor value indicates that corresponding observation is an outlier.

From Figure 3.7 we can see the observation number 19 and number 30 have unusually large Bayes factors, which are 1.99 and 1.48 respectively on a logarithm base 10 scale, Based on this result, we conclude that observation number 19 and 30 are outliers.
3.2.5 Posterior Predictive Distribution of Individual Deviance

Using individual deviance $D_i(y)$ to detect outliers can only identify observations with suspicion due to the lack of a specific value to determine how large is (observations that) contributed disproportionately to the overall deviance statistic should be identified with suspicion. To solve this problem, we use the posterior predictive distribution of individual deviance using the future observation. Let $y^{rep}$ denote the replicated data that could have been observed, or, to think predictively, as the data we would see tomorrow if the experiment that produced $y$ today were replicated with the same model and the same parameter values $\theta$ that produced the observed data, as described by Gelman et al (2003). In this case, the individual deviance is defined as

$$D_i(y^{rep}) = 2 \sum_{j=1}^{k} I(y_i^{rep} = j) \log\left( \frac{I(y_i^{rep} = j)}{\hat{p}_{ij}} \right)$$

(3.16)

For each simulated values of “future” observation $y^{rep}$, we can get one individual deviance $D(y^{rep})$ using equation (3.16), in the end we can get $m$ individual deviances $D_i(y^{rep})$ and we can compare these individual deviances $D_i(y^{rep})$ with $D_i(y)$ for each observation. If the observation is not an outlier, then the individual deviance statistic $D(y)$ should not fall in the extreme end of the posterior predictive distribution of those $m$ individual deviance statistics $D_i(y^{rep})$. One way to check whether the individual deviance statistic $D(y)$ falls in the extreme end of the posterior predictive distribution of $D_i(y^{rep})$ is to compute the tail probability of $D_i(y) < D_i(y^{rep})$. If the tail probability is small, it means that $D(y)$ falls in the extreme end of posterior predictive distribution of $D_i(y^{rep})$ thus we conclude that the corresponding observation is an outlier. The process of using the posterior predictive distribution of individual deviance $D_i(y^{rep})$ to detect outliers goes as following:

1. simulate one regression parameter vector $\beta$, $k-2$ unknown cutoff points $(\gamma_2, \cdots, \gamma_{k-1})$, and one latent variable vector $Z = (Z_1, \cdots, Z_n)$ from the posterior distribution using the Nandram and Chen (1996) algorithm.
2. compute the probability of individual \( i \) falling in category \( j \) given by the formula

\[
p_{ij} = \Phi(\gamma_j - x'_i \beta) - \Phi(\gamma_{j-1} - x'_i \beta)
\]

3. simulate “future” observation \( y_{i\text{rep}} \) from a multinomial distribution with sample size=1 and probability vector\(=(p_{i1}, \cdot \cdot \cdot, p_{ik})\).

4. compute individual deviance given by equation (3.16).

After we run the simulation \( m \) times, we can get \( m \) individual deviances \( D(y_{\text{rep}}) \) for each observation, also we need to compute \( D(y) \) using equation (3.3) for each observation. Then we can compute the tail probability of \( D(y) < D(y_{\text{rep}}) \) for each observation and plot the probabilities against observation ID, which leads to Figure 3.8.

Figure 3.8: The plot of probability of individual deviance \( D(y_{\text{rep}}) \) are greater than the individual deviance \( D(y) \) for each observation against observation ID. The x-axis is the observation ID, the y-axis is probability of individual deviances \( D(y_{\text{rep}}) \) are greater than the individual deviance \( D(y) \). Smaller probabilities indicate that the corresponding observations are outliers.
From Figure 3.8, we can see that observation number 19 and number 30 have unusually small probabilities, which are 0.96% and 4.46% respectively, we can then conclude that observation number 19 and number 30 are outliers.

### 3.2.6 Parameter-Dependent Individual Deviance

Using the posterior predictive distribution of individual deviance to detect outliers treats individual deviance as a function of “future” observation $y^{rep}$. Another method of detecting outliers using individual deviance from a Bayesian perspective treats individual deviance as a function of two variables—the “future” observation $y^{rep}$ and the simulated regression vector $\theta = (\gamma, \beta)$. This parameter-dependent test statistics allows us to compare directly the discrepancy between the observed data and the posited model, instead of between the data and the best fit of the model. This type of parameter-dependent test statistics have been discussed by Tsui and Weerahandi (1989) and Meng (1994) for the case of testing parameters within a model. We want to apply the parameter-dependent individual deviance test statistic to detect outliers. The individual deviance statistic in this case can be defined as

$$D_i(y^{rep}, \theta) = 2 \sum_{j=1}^{k} I(y_i^{rep} = j) \log \left( \frac{I(y_i^{rep} = j)}{p_{ij}} \right)$$  \hspace{1cm} (3.17)

where $\theta = (\gamma, \beta)$ and $p_{ij} = \Phi(\gamma_{j-1} - x'_i\beta) - \Phi(\gamma_{j-2} - x'_i\beta)$. For each simulated values of “future” observation $y^{rep}$ and simulated regression vector $\theta$, we can get one individual deviance $D(y_i^{rep}, \theta)$, in the end we can get $m$ individual deviances $D(y_i^{rep}, \theta)$ for each observation. Also for the simulated $\beta$ and $\gamma$ values, we can compute the individual deviance $D(y, \theta)$. We can compare these individual deviances $D(y_i^{rep}, \theta)$ with $D(y, \theta)$. If the observation is not an outlier, then the individual deviance statistic $D(y_i, \theta)$ should not fall apart from those individual deviance $D(y_i^{rep}, \theta)$ for that observation. Similar to the method of using the posterior predictive distribution of individual deviance to detect outliers, we can take a look at the probability of $m$ individual deviance $D(y_i^{rep}, \theta)$ are greater than individual deviance statistic.
If the probability is small, then we conclude the corresponding observation is an outlier. The process of using parameter-dependent individual deviance to detect outliers goes as following:

1. Simulate one regression parameter vector $\beta$, the unknown cutoff points $(\gamma_2, \cdots, \gamma_{k-1})$, one latent variable vector $Z = (Z_1, \cdots, Z_n)$ from the posterior distribution using the Nandram and Chen (1996) algorithm.

2. Compute the probability of individual $i$ falling in category $j$ given by the formula $p_{ij} = \Phi(\gamma_j - x'_i\beta) - \Phi(\gamma_{j-1} - x'_i\beta)$ if we choose a probit regression model.

3. Simulate “future” observation $y_{i \text{ rep}}$ from a multinomial distribution with sample size=1 and probability vector=$\left(p_{i1}, \cdots, p_{ik}\right)$, where $k$ is the number of categories.

4. Compute individual deviance $D(y_{i \text{ rep}}, \theta)$ by equation (3.17)

5. Compute individual deviance $D(y, \theta)$ using the observed $y$ and the simulated $\beta$ and $\gamma$.

After we run the simulation $m$ times, we will get $m$ individual deviance statistic $D(y_{i \text{ rep}}, \theta)$ and $D(y_i, \theta)$ for each observation. We can compute the probability of these $m D(y_{i \text{ rep}}, \theta)$ are greater than $D(y_i, \theta)$ for each observation and plot these probabilities against observation ID, which is given by Figure 3.9.

From Figure 3.9, we can see that observation number 19 and observation 30 have very small probabilities of $D(y_{\text{rep}}, \beta)$ being greater than $D(y_i, \theta)$, which are 0.38% and 1.32% respectively, so we can conclude that observation number 19 and number 30 are outliers.

### 3.2.7 Hierarchical Model to Detect Outliers

In many cases, statistical applications involving multiple parameters can be regarded as related or connected in some way, hence hierarchical model is introduced into statistics to solve this correlated parameter problem and then becomes widely used in the complex data structure. Another reason why hierarchical model is popular is because nonhierarchical models...
Figure 3.9: The plot of tail probability of individual deviance $D(y_{i}^{rep}, \theta)$ are greater than $D(y_{i}, \theta)$ for each observation against observation ID. The x-axis is the observation ID, the y-axis is probability of $D(y_{i}^{rep}, \theta)$ are greater than $D(y_{i}, \theta)$. Unusually small or unusually large probabilities indicate the corresponding observation is an outlier.

are usually inappropriate for hierarchical data: If we use too few parameters, then we can not fit large data set accurately, on the other hand, if we fit too many parameters, then we tend to overfit the large data set and can not get reliable prediction values, while using hierarchical model, we can have enough parameters to fit the data well, and using a population distribution to describe the independence structure of the parameters, thereby we can avoid the problem of overfit.

A lot of applications using hierarchical model have been developed recently, for example, Fay and Herriot (1979) used hierarchical model for estimating income for small places through the combining information from neighborhood, Manton et al. (1989) fit a hierarchical Bayesian model to cancer mortality rates and compare maps of direct and adjusted mortality risks. Bock (1989) use hierarchical model method to study the effect of school policies, practices or climates on the student outcome. Lindsay and Smith (1972) give the theoretical result of posterior distribution of unknown parameter for the hierarchical model,
Albert and Chib (1993) use the hierarchical model for binary data. Inspired by Albert and Chib (1993) idea, here we want to introduce a hierarchical model for ordinal data set to detect outliers. Notice in the previous methods, such as using 3σ rule or individual deviance to detect outliers, we assume that all the latent variable $Z_i$ has the same variance value 1. On the contrary, the hierarchical model assumes that the latent variable $Z_i$ for each observation has different variance, Hierarchical model for the ordinal case makes more sense because if there are outliers in the data set, then we expect that outliers have a larger variance than the “normal” data. The hierarchical model for the ordinal data is given as following:

$$Z_i \sim N(x_i' \beta, \frac{1}{\lambda_i})$$ (3.18)

$$\lambda_i \sim \text{Gamma}(\frac{v}{2}, \frac{v}{2})$$ (3.19)

$$v \sim \text{flat prior}$$ (3.20)

Notice that $Z_i$ now has the variance $\frac{1}{\lambda_i}$, so if observation $i$ is an outliers, then we expect the value of $\lambda_i$ to be small, which will make the variance of $Z_i$ large. In order to use the Gibbs sampling algorithm to simulate draws of parameter, we have to find the distribution of one variable conditional on the rest variables. The conditional distribution is given by

$$[Z_i^*|\beta^*, \delta^2, \lambda_i] \sim \text{truncated } N(x_i' \beta^*, \frac{\delta^2}{\lambda_i})$$ (3.21)

$$[\beta^*|Z_i^*, \delta^2, \lambda_i] \sim N((x'W^{-1}x)^{-1}x'W^{-1}Z_i^*, (x'W^{-1}x)^{-1}\delta^2)$$ (3.22)

where

$$W = \text{diag}\left(\frac{1}{\lambda_i}\right)$$
\[
[\lambda_i | Z_i^*, \delta^2, \beta^*] \sim \text{Gamma}(\frac{v+1}{2}, \frac{v}{2} + \frac{(Z_i^* - x_i'\beta^*)^2}{2\delta^2})
\] (3.23)

\[
[\delta^2 | Z_i^*, \beta^*, \lambda_i] \sim \text{Inverse Gamma}(\frac{n + m + k}{2}, \frac{1}{2} \sum_{i=1}^{n} \lambda_i(Z_i^* - x_i'\beta^*)^2)
\] (3.24)

Using Gibbs Sampling with the conditional distribution given by equation (3.21), (3.22), (3.23), and (3.24), we can simulate the values of \( \lambda \) for each observation, which is the key parameter that we are interested in. We plot the mean of simulated \( \lambda \) values for each observation against observation ID and get Figure 3.10 and Figure 3.11.

Figure 3.10: Using hierarchical model to detect outliers. The left graph is the plot of mean of simulated \( \lambda \) values vs observation ID using degree of freedom \( v=2 \), the right graph is the plot of mean of simulated \( \lambda \) values vs observation ID using degree of freedom \( v=4 \). Lower values of mean of simulated \( \lambda \) values corresponds to outliers.

From both Figure 3.10 and Figure 3.11, we can see that observation number 19 and number 30 have smaller using the different degree of freedom for \( v \), namely, use the degree
of freedom 2, 4, 10, 20 respectively, which suggest that these two observations have large variance, so we conclude that observation number and observation number 30 are outliers.

Figure 3.11: Using hierarchical model to detect outliers. The left graph is the plot of mean of simulated λ values vs observation ID using degree of freedom v=10, the right graph is the plot of mean of simulated λ values vs observation ID using degree of freedom v=20. Lower values of mean of simulated λ values corresponds to outliers.
CHAPTER 4

Goodness-of-fit and Model Selection

It is straightforward to fit a model for a data set using software. Generally there are many possible models for one data set and we need to select the “best” model for the given data set. Also for the chosen model, we need to check whether this model really reflects the information given by the data set. Goodness-of-fit measures indicate how well a statistical model fits a set of observations and it typically summarizes the discrepancy between observed values and the fitted values under the model.

4.1 Testing Goodness-of-fit from a Frequentist’s Perspective

4.1.1 Deviance Statistic from a Frequentist’s Perspective

The traditional method of assessing goodness-of-fit in generalized linear models is based on the deviance statistic. Let $y$ denote the observed data set, $\beta$ denote the vector of parameters and $\hat{\beta}$ denote the maximum likelihood estimate of $\beta$. The deviance statistic is defined as

$$D(y) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} I(y_i = j) \log\left( \frac{I(y_i = j)}{\hat{p}_{ij}} \right)$$  \hspace{1cm} (4.1)
where \( \hat{p}_{ij} \) is the estimated probability of observation \( y_i \) falling in category \( j \), \( k \) is the number of categories and \( n \) is the number of observations. We can compute the deviance statistic \( D(y) \) and compare with its asymptotic distribution, which is \( \chi^2 \) with \( n(k - 1) - q \) degrees of freedom. If the deviance statistic \( D(y) \) is greater than the critical value of the \( \chi^2 \) distribution, then we conclude that the proposal model is not a suitable fit for the given data set.

To illustrate the usage of the deviance statistic as a measure of goodness-of-fit for a model, we consider the data set in Table 1.1 in Chapter 1. If we choose a probit regression model, then the estimated probability \( \hat{p}_{ij} \) can be computed by equation (4.2).

\[
\hat{p}_{ij} = \Phi(\hat{\gamma}_j - x_i'\hat{\beta}) - \Phi(\hat{\gamma}_{j-1} - x_i'\hat{\beta})
\]

Using the maximum likelihood method, we can fit a probit regression model to the ordinal data and get estimates of \( \hat{\beta} = (-12.37, 0.027) \) and \( (\hat{\gamma}_2, \hat{\gamma}_3, \hat{\gamma}_4) = (1.69, 2.61, 3.98) \). We can compute the estimated probability of observation \( i \) falling in category \( j \) by

\[
\hat{p}_{ij} = \Phi(\hat{\gamma}_j - (-12.37 + x_i \times 0.027)) - \Phi(\hat{\gamma}_{j-1} - (-12.37 + x_i \times 0.027))
\]

then we can compute the deviance statistic \( D(y) \) using equation (4.1) and obtain \( D(y) = 74.07 \). The asymptotic distribution of the deviance statistic is \( \chi^2 \) with \( 30(5 - 1) - 2 = 118 \) degrees of freedom and the 95% critical value for a \( \chi^2 \) distribution with 118 degrees of freedom is 144.35. Since the observed deviance statistic value 74.07 is less than the critical value 144.35, we conclude that the probit regression model for the ordinal data set is a suitable fit.

However, one problem about using deviance statistic as a measure of goodness-of-fit for a model is that we have to reply on the asymptotic distribution. As pointed out by Johnson and Albert (1999, p. 138), “Asymptotically, the deviance statistic for ordinal models has a \( \chi^2 \) distribution only when observations are grouped according to covariate values and the expected counts in each cell becomes large. When only one observation is observed at each
covariate value, the deviance statistic is not well approximated by a $\chi^2$ distribution. Since we have only get 30 observations from Table 1.1 in Chapter 1 and we did not group the data according to the covariate value, we did not reach the criterion of the $\chi^2$ approximation. To avoid the reliance on the asymptotic distribution, we use the deviance statistic from a Bayesian perspective.

### 4.1.2 Pearson Chi Square Statistic from a Frequentist’s Perspective

Another common and useful statistic for goodness-of-fit in generalized linear model is the Pearson Chi Square statistic. The general form of Pearson Chi Square statistic is defined as

$$X^2(y) = \sum_{i=1}^{g}(y_i - \hat{\mu}_i)\Sigma_i^{-1}(\hat{\beta})(y_i - \hat{\mu}_i) \quad (4.4)$$

where $g$ is the number of groups (if the data is grouped), $\Sigma_i$ is the variance-covariance matrix for $y_i$, $\hat{\mu}_i$ is the estimated mean vector for $y_i$. For the ordinal response data case, $X^2$ can be rewritten in the more familiar form

$$X^2(y) = \sum_{i=1}^{g}X^2_F(y_i, \hat{p}_{ij}) \quad (4.5)$$

where

$$X^2_F(y_i) = n_i \sum_{j=1}^{k} \frac{(y_{ij} - \hat{p}_{ij})^2}{\hat{p}_{ij}} \quad (4.6)$$

where $\hat{p}_{ij}$ is the estimated probability for observation $y_i$ falling in category $j$, and $n_i$ is the number of observations for the $i$th group. We can compare the observed Pearson Chi Square statistic $X^2(y)$ with its asymptotic distribution, which is $\chi^2$ distribution with $n(k - 1) - q$ degrees of freedom. If the observed Pearson Chi Square statistic $X^2(y)$ is greater than the critical value of asymptotic $\chi^2$ distribution, then we conclude that the proposal model for the given data set is not a suitable fit. However, one problem about this procedure is that
we need to group the data and rely on the asymptotic distribution. As pointed out by Fahrmeir and Tutz (1994), in order to use this Pearson Chi Square statistic for the purpose of goodness-of-fit, the data should be grouped as far as possible. When only one observation is observed for each covariate value, the Pearson Chi Square is not well approximated by a $\chi^2$ distribution with $n(k - 1) - q$ degrees of freedom, where $n$ is the number of observations, $k$ is the number of categories, and $q$ is the number of regression parameters.

To illustrate the use of Pearson Chi Square as a measure of goodness-of-fit for the ordinal regression model, we use the sample data set from Table 1.1 in Chapter 1. As pointed by Fahrmeir and Tutz (1994), we need to group the data as far as possible before we can apply Pearson Chi Square statistic as a measure of goodness-of-fit. Since we only have 30 observations in that data set and we do not have many common covariate values so there is no need of grouping the data. We compute the Pearson Chi Square statistic by equation (4.5) and (4.6), and get a value of 187.4832. Then we find the 95% critical value for the $\chi^2$ distribution with $30 \times (5 - 1) - 2 = 118$ degrees of freedom is 144.36. Since the observed Pearson Chi Square statistic 187.48 is greater than the critical value 144.36, we conclude that this model is not good for the ordinal data.

Previously when we use deviance statistic as a measure of goodness-of-fit, we conclude that the ordinal regression model for the sample data set is a suitable fit; now when we use Pearson Chi Square statistic as a measure of goodness-of-fit, we conclude that the ordinal regression model is not a suitable fit for the same data set. The reason accounting for the contradiction is the poor approximation by the asymptotic $\chi^2$ distribution for the Pearson Chi Square statistic. To remove the dependence on an asymptotic distribution, we turn to a Bayesian perspective to look at Pearson Chi Square statistic as a measure of goodness-of-fit.
4.2 Testing Goodness-of-fit from a Bayesian Perspective

4.2.1 Posterior Predictive Distribution of Deviance Statistic

The posterior predictive distribution considers the deviance statistic as one variable function of the “future” observation \(y^{rep}\). The posterior predictive distribution of “future” observation \(y^{rep}\) can be expressed as

\[
f(y^{rep}|y) = \int f(y_i|\theta)g(\theta|y)d\theta
\]  \hspace{1cm} (4.7)

where \(f(y|\theta)\) is the sampling density for the given data set, \(g(\theta|y)\) is the posterior distribution of regression parameter \(\theta\). Equation (4.7) is referred as the posterior predictive distribution, since it is averaging the sampling density over the posterior distribution of \(\theta\). The deviance statistic is defined as

\[
D(y^{rep}) = 2 \sum_{i=1}^{n} \sum_{j=1}^{k} I(y^{rep} = j)log\left(\frac{I(y^{rep} = j)}{\hat{p}_{ij}}\right)
\]  \hspace{1cm} (4.8)

Assume that we have the ability to simulate draws of \(y^{rep}\), we can get the posterior predictive distribution of deviance statistic \(D(y^{rep})\) and we can compare the simulated \(D(y^{rep})\) with \(D(y)\). If the model for the given data set is a suitable fit, then the simulated “future” observation \(y^{rep}\) should be similar to the observed value \(y\), so we expect that \(D(y)\) does not fall in the extreme end of the posterior predictive distribution of \(D(y^{rep})\). One way to check whether \(D(y)\) falls in the extreme end of the posterior predictive distribution of \(D(y^{rep})\) is to look at the histogram of \(D(y^{rep})\) and check the position of \(D(y)\) in the histogram. A more precise way is to compute the tail probability of \(D(y^{rep}) < D(y)\). If the tail probability is usually small or usually large, then we conclude that the model is not a suitable fit for the data.

To illustrate the use of posterior predictive distribution of deviance statistic \(D(y^{rep})\) as a
measure of goodness-of-fit, we assume a non-informative prior is assigned to the parameters. Using Nandram and Chen (1996) algorithm, we can simulate values of the regression parameter $\beta$ and the unknown cutoff points $(\gamma_2, \cdots, \gamma_{k-1})$ from the posterior distribution, then we can simulated “future” observation $y^{rep}$ and compute deviance $D(y^{rep})$ by equation (4.8). Thus we can get the posterior predictive distribution of deviance statistic $D(y^{rep})$ and use it to judge the goodness-of-fit of models for the data set. The detailed simulation algorithm goes as following:

1. Simulate one set of $\beta$ values and cutoff points $\gamma_i$ values from the posterior distribution $[\beta, \gamma, z|y]$.

2. For the simulated $\beta$ and $\gamma$ value, compute the probability of observation $i$ falling in category $j$ using the formula $p_{ij} = \Phi(\gamma_j - x_i^\prime \hat{\beta}) - \Phi(\gamma_{j-1} - x_i^\prime \hat{\beta})$

3. Simulate “future” value $y^{rep}$ from the multinomial distribution with sample size=1 and probability vector=$(p_{i1}, \cdots, p_{ik})$

4. Compute the deviance statistic $D(y^{rep})$ using the equation (4.8)

After this simulation process $m$ times, we will get $m$ simulated deviance values $D(y^{rep})$. Also we need to compute the deviance statistic $D(y)$ using equation (4.1). We can then compare the deviance statistic $D(y)$ with these $m$ simulated deviance values. As stated above, one way to compare these two statistic $D(y^{rep})$ and $D(y)$ is to make the histogram of these simulated values of $D(y^{rep})$ and mark the position of $D(y)$ in the histogram. We make a histogram of $D(y^{rep})$ and mark the position of $D(y)$ on the histogram, which leads to Figure 4.1.

From Figure 4.1 we can see that the deviance statistic $D(y)$ does not fall in the extreme end of $m$ simulated deviance statistic $D(y^{rep})$, so we can conclude that the probit regression model for the ordinal data is a suitable fit. A more precise way to look at whether the observed deviance $D(y)$ falling in the extreme end of those simulated deviance statistic
Figure 4.1: Histogram of the simulated deviance statistic $D(y_{rep})$. The dotted line is the observed deviance $D(y)$. A model is a suitable fit if $D(y)$ does not fall in the extreme end of the histogram of $D(y_{rep})$.

$D(y_{rep})$ is to compute the tail probability $D(y_{rep}) < D(y)$. We do the computation and get a tail probability is 0.21, which is an evidence that $D(y)$ does not fall in the extreme end of $D(y_{rep})$.

### 4.2.2 Parameter-dependent Deviance Statistic

The posterior predictive distribution of deviance method treats deviance $D(y_{rep})$ as function of only one variable $y_{rep}$. The Bayesian formulation not only handles deviance $D(y_{rep})$ but also naturally allows the use of the test “statistics” depending on the unknown parameter $\theta$. As pointed out by Gelman, Meng and Stern(1996), this generalization *allows us to compare directly the discrepancy between the observed data and the posited model, instead of between the data and the best fit of the model*. Parameter-dependent test statistics have been discussed by Tsui and Weerahandi(1989) and Meng (1994) for the case of testing parameters within a model. Inspired by the idea of test “statistics” containing unknown parameters, we treat
deviance as a function of two variables $D(y^{rep}, \theta)$. In this case, the deviance statistic is defined as

$$D(y^{rep}, \theta) = 2 \sum_{i=i}^{n} \sum_{j=1}^{k} I(y^{rep} = j) \log \left( \frac{I(y^{rep} = j)}{p_{ij}} \right)$$

(4.9)

where $\theta = (\beta, \gamma)$, $p_{ij}$ is the probability of observation $i$ falling in category $j$ and it is computed by $p_{ij} = \Phi(\gamma_j - x'_i \beta) - \Phi(\gamma_{j-1} - x'_i \beta)$ for the ordinal regression model. Notice that deviance $D(y^{rep}, \theta)$ given by equation (4.9) is a natural extension of deviance $D(y^{rep})$ given by equation (4.8), which is the marginal distribution of (4.9) in the sense that the integral of equation (4.9) over the parameter $\theta$. To use the deviance $D(y^{rep}, \theta)$ as a measure of goodness-of-fit, we compare the deviance $D(y^{rep}, \theta)$ with the realized deviance statistic $D(y, \theta)$. As stated before, if the model for the ordinal data set is a suitable fit, then we expected the “future” observation $y^{rep}$ to be similar to the observed value $y$, in other words, we expect that the deviance $D(y, \theta)$ does not fall in the extreme end of $D(y^{rep}, \theta)$. However, directly locating the realized deviance $D(y, \theta)$ within the distribution of $D(y^{rep}, \theta)$ is impossible because both deviance statistic depending on the unknown parameter $\beta$. One way to locate the deviance $D(y, \theta)$ in the distribution of $D(y^{rep}, \theta)$ is plotting the simulated $D(y, \theta)$ against $D(y^{rep}, \theta)$ on the same graph. A more precise way is to look at the tail probability of $D(y^{rep}, \theta) < D(y, \theta)$.

The simulation algorithm goes as following:

1. Simulate $\beta$ and cutoff points $\gamma_i$ from the posterior distribution $[\beta, \gamma, z|y]$.

2. Compute the probability of observation $i$ falling in category $j$ using the formula $p_{ij} = \Phi(\gamma_j - x'_i \beta) - \Phi(\gamma_{j-1} - x'_i \beta)$

3. Simulate “future” value $y$ from the multinomial distribution with sample size=1 and probability vector $p_i = (p_{i1}, \cdots, p_{ik})$

4. Compute the deviance statistic $D(y^{rep}, \theta)$ by equation (4.9).

5. Compute the realized deviance $D(y, \theta)$ using equation (4.9), only replacing $y^{rep}$ by the observed value $y$. 
If we run this simulation $m$ times, we will get $m$ simulated deviances $D(y^{rep}, \theta)$ and $m$ realized deviances $D(y, \theta)$, then we can compare these $m$ simulated deviances $D(y^{rep}, \theta)$ with these $m$ simulated deviances $D(y, \theta)$ by making a scatterplot. The scatterplot is given by Figure 4.2.

![Figure 4.2: Scatterplot of deviance $D(y^{rep}, \theta)$ against $D(y, \theta)$. The x-axis is the realized deviance $D(y, \theta)$, the y-axis is the deviance $D(y^{rep}, \theta)$. The tail probability of $D(y^{rep}, \theta) < D(y, \theta)$ is estimated by the proportion of points below the 45 degree straight line.](image)

From Figure 4.2, we can see that the proportion of points below the 45 degree straight line is not close to 1 or 0, so we conclude that the realized deviance $D(y, \theta)$ do not locate in the extreme end of the simulated deviance $D(y^{rep}, \theta)$. Based on the scatterplot of $D(y, \theta)$ and $D(y^{rep}, \theta)$, we conclude that the probit regression model for the ordinal data set is a suitable fit. To be more confident about our conclusion, we compute the tail probability of $p(D(y^{rep}, \theta) < D(y, \theta))$. We do the computation and get a probability 10.64%, which is an evidence that $D(y, \theta)$ did not fall in the extreme end of $m$ simulated deviance statistic $D(y^{rep})$. 
4.2.3 Posterior Predictive Distribution of Pearson Chi Square Statistic

Posterior predictive distribution of Pearson Chi Square statistic considers the Pearson Chi Square statistic as one variable function $X^2(y^{rep})$, which is defined as

$$X^2(y^{rep}) = \sum_{i=1}^{g} \chi^2_p(y^{rep})$$

(4.10)

where

$$X^2_p(y^{rep}) = n_i \sum_{j=1}^{k} \frac{(y_{ij}^{rep} - \hat{p}_{ij})^2}{\hat{p}_{ij}}$$

(4.11)

and

$$\hat{p}_{ij} = \Phi(\hat{\gamma}_j - x_i' \hat{\beta}) - \Phi(\hat{\gamma}_{j-1} - x_i' \hat{\beta})$$

(4.12)

where $\hat{p}_{ij}$ is the estimated probability of observation $i$ falling in category $j$, and $\hat{\beta}$ is the maximum likelihood estimate of the regression parameter $\beta$. Similar to the case of using posterior predictive distribution of deviance as a measure of goodness-of-fit, we can get simulated $X^2(y^{rep})$ for each simulated “future” observation $y^{rep}$ and compare the simulated $X^2(y^{rep})$ with $X^2(y)$. If the proposal model is a suitable fit, then we expect that $X^2(y)$ does not fall in the extreme end of the simulated $X^2(y^{rep})$. As usual, we can compute the tail probability $p(X^2(y^{rep}) < X^2(y))$ to see whether $X^2(y)$ falls in the extreme end of the simulated $X^2(y^{rep})$. The simulation process goes as following:

1. Simulate $\beta$ and cutoff points $\gamma_i$ from the posterior distribution $[\beta, \gamma|y]$.
2. Compute the probability using the formula $p_{ij} = \Phi(\gamma_i - x'_i \beta)$
3. Simulate “future” value $y$ from the multinomial distribution with sample size=1 and probability vector=$(p_{i1}, \cdots, p_{ik})$
4. Compute the Pearson Chi Square statistic $X^2(y^{rep})$ given by equation (4.10), (4.11) and (4.12).
After we run the simulation $m$ iterations, we will get $m$ Pearson Chi Square statistic $X^2(y^{rep})$. Also we can compute $X^2(y)$ using the observed data by equation (4.6) and compare the simulated $X^2(y^{rep})$ with $X^2(y)$ by computing the tail probability $p(X^2(y^{rep}) < X^2(y))$. We do the computation and get the tail probability value 19.6%, which is an evidence that the probit regression model model for the ordinal data set is a suitable fit.

Compared with deviance statistic, the Pearson Chi Square statistic works better with grouped data set. There are certainly other methods can be applies as a measure of goodness-of-fit for a model, and methods from a frequentist’s perspective usually rely on an asymptotic distribution to determine whether the test statistic is significant, which in some cases can be misleading (the Pearson Chi Square statistic from a frequentist’s perspective case). The general cure is to simulate the test statistic and compare the simulated test statistic with the test statistic from a frequentist’s perspective. If the test statistic from a frequentist’s perspective does not fall in the extreme end of the simulated test statistics, then we conclude the model for the data set is a suitable fit.

4.3 Model Selection

Usually there is more than one explanatory variables in the model and we are not sure if all the explanatory variables are needed in explaining the variation in the response variables. If we include the redundant variable in the model, then we introduce variability in future observations, if we include too few variables in the model, we introduce the bias in the estimated values, so we have to decide which variable should be included in the model before we can make any estimations about unknown parameters or prediction about the future observation. Model selection has been studied thoroughly both from a frequentist’s perspective and a Bayesian perspective. Classical methods of model selection includes background elimination, forward selection and stepwise regression, which depends on subsequently adding or deleting predictors by means of mean square error or modified mean square error. Bayesian
methods of model selection use the following criteria: Bayesian Information Criteria (BIC, Schwarz, 1978), Asymptotic Information Criteria (AIC, Akaike, 1974), Bayes factor and pseudo-Bayes factor. For an illustration purpose, we take the data in Table 1.1 in Chapter 1 as an example, and we notice that there are two explanatory variables in that data set, one variable is the student’s SAT grade and the other variable is the student’s prerequisite class grade. We want to use a statistical measure to decide which variable should be included in the model.

4.3.1 Deviance for Model Selection

The deviance statistic defined by equation (4.1) is a common measure of goodness-of-fit of a model. In addition, the deviance statistic can also be used as a tool of model selection. As stated before, the deviance statistic has an asymptotic \( \chi^2 \) distribution. Suppose the model M1 is nested in model M2 in the sense that all the parameters in model M1 are also included in model M2 and probably model M2 contains an “extra” variable compared with model M1, in theory, the difference of deviances between model M1 and model M2 has an asymptotic \( \chi^2 \) distribution with the degree of freedom equals to the difference between the degree of freedoms for two models. If the realized difference between two deviances are greater than the critical values of the asymptotic \( \chi^2 \) distribution, then we can conclude that the “extra” variable should not be included in the model. deviance statistic from several competing models are often summarized in a table called analysis of deviance table and we can decide which variable should be included in the model based on the analysis of deviance table. For the sample data set, in order to decide which variable should be included in the model, we propose three models for the data set in Table 1.1

- Model M1: Only the constant term should be included in the model.
- Model M2: Only the constant and variable SAT should be included in the model.
- Model M3: Constant, variable SAT and variable prerequisite course grade should be
Note that model M1 is nested in model M2 because model M2 contains not only all the variables in model M1, but also an “extra” variable SAT compares with model M1, the same case that model M2 is nested in model M3. We can use the function polr in program R package MASS to get the approximated deviance for each model, and obtain Table 4.1.

<table>
<thead>
<tr>
<th>model</th>
<th>Covariate</th>
<th>Deviance</th>
<th>Degree of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Constant</td>
<td>89.67</td>
<td>27</td>
</tr>
<tr>
<td>M2</td>
<td>Constant+SAT</td>
<td>73.49</td>
<td>26</td>
</tr>
<tr>
<td>M3</td>
<td>Constant+SAT+Prerequisite Course Grade</td>
<td>71.80</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 4.1: Analysis of Deviance table to decide which variable should be included in the model.

The difference between Model M1 and Model M2 is that Model M2 contains the “extra” variable SAT. The difference of the deviances between model M1 and model M2 is 89.67-73.49=16.17, and the difference of the degree of freedom between model M1 and model M2 is 1. The 95% critical value for the $\chi^2$ distribution with 1 degree of freedom is 3.84. Since the realized difference of deviance between Model M1 and Model M2 16.17 is greater than the critical value 3.84, we conclude that the variable SAT should be included in the model.

Similarly, to decide whether variable prerequisite course grade should be included in the model, we look at the difference of deviances between Model M2 and Model M3. The difference of deviances between Model M2 and Model M3 is 73.48-71.80=1.68, and the difference between the degree of freedom for Model M2 and Model M3 is 1. Again, the 95% critical value for the $\chi^2$ distribution with 1 degree of freedom is 3.84. Since the observed difference of deviance between Model M2 and Model M3 1.68 is less than the critical value 3.84, we conclude that the variable prerequisite course grade should not be included in the model.

The idea of using the formation given by the analysis of deviance table to decide which variable should be included in the model relies on the comparison of difference of deviance statistic between two models with the asymptotic $\chi^2$ distribution. When the difference of deviances between two competing model is close to the critical value of the $\chi^2$ distribution,
it is hard to say whether this variable should be included in the model. To avoid the issue of dependence on an asymptotic distribution, We consider a Bayesian perspective for the model selection problem.

### 4.3.2 Bayes Factor for Model Selection

The Bayesian method for model selection is based on the notation of a Bayes factor. In Chapter 3 we used the Bayes factor to decide which observation is an outlier, here we want to use the Bayes factor as a model selection tool. Recall that the Bayes factor is the ratio of marginal likelihood from model M1 and model M2. In general, if we let $f(y|\theta)$ denote the sampling density of y and let $g(\theta)$ denote the prior distribution on the unknown parameter $\theta$, then the joint distribution of y and $\theta$ is the product of $f(y|\theta)$ and $g(\theta)$. The marginal likelihood is the integral of the joint distribution over the parameter $\theta$ defined by equation (4.13).

$$f(y|M) = \int f(y|\beta)g(\beta)d\beta \quad (4.13)$$

The notation $f(y|M)$ is used to emphasize the dependence of marginal likelihood on the model M. We can then define the Bayes factor of Model M2 over Model M1 as

$$BF_{21} = \frac{f(y|M_2)}{f(y|M_1)} \quad (4.14)$$

If model M2 is more appropriate for the data set than model M1, then Bayes factor $BF_{21}$ will produce a large number to indicate that we should support model 2. The Bayes factor approach was espoused by Jeffreys(1935), who developed a methodology for quantifying the evidence in favor of a scientific theory. A lot of researches and applications have been conducted after the introduction of Bayes factor, for example, Raftery(1995) points out that in regression, when there are many candidate independent variables, standard variable selection procedures can give very misleading results. P-values and the tests based upon them give unsatisfactory results, especially in large samples and the quantitative sociologists
use Bayes factor to deal with model selection in social science. Berger and Pericchi (1996) introduces a new criterion called intrinsic Bayes factor, which is a better approximation to the real Bayes factor. Kass and Raftery (1995) give a review of application of Bayes factor in the field of genetics, sports, ecology, sociology, and psychology.

To illustrate the use of a Bayes factor to decide which variable should be included in the model, we return to the same model selection problem, that is, we want to decide whether we should include variable SAT and variable prerequisite course grade in the model. First we consider two models, one model assumes that SAT score does not affect the probability of falling in each category and the other model assumes that SAT score does affect the probability of falling in each category. To reflect our intention of including or excluding SAT in the model, we need to construct an informative prior for the regression vector β and the cutoff points γ under two models. Recall the informative prior can be constructed by the conditional means family of prior distribution, which has been discussed by Bedrick, Christensen and Johnson (1996). In the conditional means prior construction process, we construct our beliefs about priors by simply specifying two items as following:

• a guess at the cumulative probability \( \theta_j \) corresponding to covariate \( x_j \)

• the number of hypothetical observations of to support this guess. Denote this number of prior sample \( K_j \)

Then the prior distribution can be constructed as following

\[
g(\theta_1, \cdots, \theta_M) \sim \prod_{j=1}^{M} \frac{K_j^{g_j-1}(1 - \theta_j)^{K_j(1-g_j)-1}}{\theta_j^{K_j g_j-1}(1 - \theta_j)^{K_j(1-g_j)-1}}
\]  

(4.15)

We have to specify \( q+k-2 \) prior distributions for the regression vector \( \beta \) and unknown cutoff points \( \gamma_2, \cdots, \gamma_{k-1} \), where \( k \) is the number of categories, and \( q \) is the number of regression parameters. To specify the effect of SAT on the probability of observations falling in each category, we have to specify five SAT values and five probability guesses to reflect our opinion. For example, these beliefs can be quantified by supposing that the probability of getting a
F with SAT=500 is 0.2, and the probability of getting a F with SAT=600 is also 0.2, the probability of getting a grade less than B with SAT=525 is 0.75, and the probability of getting a grade less than or equal B is 0.75, the probability of getting a grade less than C with SAT=600 is 0.7. Also we need to state the number of observations that this information is worth.

If we believe that SAT does affect the probability of falling in each category, then the prior statement should have little influence on the corresponding posterior distribution, thus we assign only one single observation to this prior. On the other hand, If we believe that SAT does not affect the probability of falling in each category, then we want to assign a prior which can be influential and force the posterior distribution to remove this term from the model. To make the prior informative influential, we use the same SAT scores and same probability of getting corresponding grade as in the previous case, the only difference is that we use a large prior precision number, say 100 to reflect our belief that SAT should not be included in the model.

In summary, we construct the two models in the following way:

- Model M4 (SAT should be included in the model):
  - when SAT=500, prob(less than F)=0.2, K=1
  - when SAT=600 prob(less than F)=0.2, K=1
  - when SAT=525 prob(less than B)=0.75, K=1
  - when SAT=570 prob(less than B)=0.75, K=1
  - when SAT=600 prob(less than C)=0.7, K=1

- Model M5 (SAT should not be included in the model):
  - when SAT=500 prob(less than D)=0.2, K=100
  - when SAT=600 prob(less than D)=0.2, K=100
  - when SAT=525 prob(less than B)=0.75, K=100
  - when SAT=575 prob(less than B)=0.75, K=100
  - when SAT=600 prob(less than C)=0.7, K=100
In order to compute the Bayes factor, we have to compute the ratio of marginal likelihoods from Model M4 and Model M5. One easy way to compute the marginal likelihood is using \textit{laplace} in R package \textit{LearnBayes}. In order to use the function \textit{laplace}, all we need to do is to write the logarithm of the posterior function and give the guess of posterior mode of the parameters, then the \textit{laplace} function will compute posterior mode, the associated variance-covariance matrix, and approximation to the log marginal likelihood. We run the \textit{laplace} function and get the approximated natural logarithm of marginal likelihood for Model M4 is 

$$\log(f(y|M_4)) = -51.14$$

and the approximated natural log of marginal likelihood for Model M5 is 

$$\log(f(y|M_5)) = -56.20$$

so the Bayes factor of Model M5 over Model M4 is given by equation (4.16):

$$BF_{54} = \frac{f(y|M_5)}{f(y|M_4)} = 0.0063$$

(4.16)

On a log 10 scale, the Bayes factor is -2.2. This indicate that we should support Model M4, in other words, there is a significant evidence that SAT should be included in the model.

One concern in the computation of Bayes factor is the influence of choice of K on the Bayes factor. To study the effect of choice of K on the Bayes factor of M5 over M4, we choose different values of K for model M5 and fix the K value in model M4 at 1 all the time, then we use the function \textit{laplace} to compute the Bayes factor. We get the following result presented in Table 4.2.

<table>
<thead>
<tr>
<th>K</th>
<th>Log 10 of Bayes factor M5 over M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-0.396</td>
</tr>
<tr>
<td>100</td>
<td>-2.198</td>
</tr>
<tr>
<td>1000</td>
<td>-3.799</td>
</tr>
</tbody>
</table>

Table 4.2: The influence of choice of K in model M5 on the Bayes factor. The left column is the choice of K for model M5 while the K value for model M4 is fixed at 1 all the time, and the right column is the base 10 log of Bayes factor of Model M5 over Model M4, where Model M5 supports that variable SAT should not be include in the model, and Model M4 supports that variable SAT should be included in the model. The computation is using the function \textit{laplace} in R package \textit{LearnBayes}.
From Table 4.2, we can see the Bayes factor of model M5 over model M4 is stable and gives us the same conclusion that we should include variable SAT in the model.

Similarly, to decide whether we should include variable *prerequisite course grade* in the model, we apply the same idea of using informative prior to reflect our belief of including or excluding variable *prerequisite course grade* in the model. Again, we need \( q+k-2 \) prior distributions for the regression vector \( \beta \) and unknown cutoff points \( \gamma_2, \cdots, \gamma_{k-1} \), which turns out that in this case we need \( 3+5-2=6 \) values of SAT, prerequisite course grade, probabilities of falling in specific categories and number of observations supporting our belief. Here we can propose two models too.

- **Model M6** (variable *prerequisite course grade* should be included in the model):
  - when SAT=500 and pregrade=D, prob(less than D)=0.6, \( K=1 \)
  - when SAT=500 and pregrade=A, prob(less than D)=0.6, \( K=1 \)
  - when SAT=550 and pregrade=D, prob(less than C)=0.7, \( K=1 \)
  - when SAT=550 and pregrade=A, prob(less than C)=0.7, \( K=1 \)
  - when SAT=600 and pregrade=D, prob(less than B)=0.8, \( K=1 \)
  - when SAT=600 and pregrade=A, prob(less than B)=0.8, \( K=1 \)

- **Model M7** (variable *prerequisite course grade* should not be included in the model):
  - when SAT=500 and pregrade=D, prob(less than D)=0.6, \( K=100 \)
  - when SAT=500 and pregrade=A, prob(less than D)=0.6, \( K=100 \)
  - when SAT=550 and pregrade=D, prob(less than C)=0.7, \( K=100 \)
  - when SAT=550 and pregrade=A, prob(less than C)=0.7, \( K=100 \)
  - when SAT=600 and pregrade=D, prob(less than B)=0.8, \( K=100 \)
  - when SAT=600 and pregrade=A, prob(less than B)=0.8, \( K=100 \)

After constructing Model M6 and Model M7, we can use the function *laplace* to compute the natural logarithm of marginal likelihood for Model M6 and Model M7, which are \( f(y|M6) = -46 \) and \( f(y|M7) = -51.28 \), then the Bayes factor of Model M7 over M6 is given by equation
Take the log base 10 of the $BF_{76}$, we get a value of 2.39, which is a strong evidence to support Model M7, in other words, we should not include variable \textit{prerequisite course grade} in the model.

To study the influence of supporting observation number K on the Bayes factor, we try different values of K for model M7 in the prior distributions while keeping the K value for model M6 at 1 all the time and compute the Bayes factor. We obtain Table 4.3.

<table>
<thead>
<tr>
<th>K</th>
<th>Log 10 of Bayes factor M7 over M6</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>2.39</td>
</tr>
<tr>
<td>100</td>
<td>3.53</td>
</tr>
<tr>
<td>1000</td>
<td>4.51</td>
</tr>
</tbody>
</table>

Table 4.3: The influence of choice of K on the Bayes factor. The left column is the choice of K for model M7 while the choice of K for model M6 is fixed at 1 all the time, and the right column is the base 10 log of Bayes factor of model M7 over Model M6, where model M7 supports that variable \textit{prerequisite course grade} should not be include in the model, Model M6 supports that variable \textit{prerequisite course grade} should be included in the model. The computation is done by the function \textit{laplace} in R package \textit{LearnBayes}.

From Table 4.3 we can see that the Bayes factor is stable for different choice of K and gives us the same conclusion that we should not include variable \textit{prerequisite course grade} in the model.

### 4.3.3 Hierarchical Analysis for Model Selection

The last method for model selection is the hierarchical model structure motivated by Lindley and Smith(1972). Lindley and Smith(1972) reanalyze the usual linear regression model using Bayesian methods and propose the hierarchical model. Lindley and Smith also give the general form of hierarchical model structure and detailed proof of the conditional distribution which can be used for Gibbs sampling. Albert and Chib (1993) adopt this idea and apply it to the model selection problem for a binary data set case. We extend Albert and
Chib(1993)’s idea to the ordinal data case.

The idea of using a hierarchical model to select between competing models goes as following: From a matrix point of view, given a particular probit regression model with regression parameter $\beta$ of dimension $q$, if there are some redundant predictors in the model, then we can reflect the redundancy by assuming that $\beta$ lies on a linear space $A\beta^0$, where $\beta^0$ is a $p$ dimension vector and $p < q$. Thus, a prior belief can be modeled by a hierarchical model for the purpose of model selection in the following form:

1. $Z$ is distributed $N(X'\beta, I)$
2. $\beta$ is distributed $N(A\beta^0, \sigma^2)$
3. $(\beta^0, \sigma^2)$ is distributed according to the prior density $\pi(\beta^0, \sigma^2)$

where $Z$ is the vector of latent variables, $X$ is the coefficient matrix, $\beta$ is the regression parameter vector with dimension $q$, $\beta^0$ is the reduced regression parameter with dimension $p$, which we believe that the “true” model should be, $\sigma^2$ is the variance which controls the variance of the reduced regression vector $\beta^0$ and it reflects the precision of prior belief that $\beta$ lies on the linear space of $A\beta^0$. If $\sigma^2$ is small enough, then we believe that $\beta$ should be reduced to $\beta^0$. In other words, we believe that only variables corresponding with $\beta^0$ should be included in the model, so the focus here is the values of $\beta^0$ and $\sigma^2$, however, the expression for $\beta^0$ and $\sigma^2$ does not have a nice form so we have to reply on the simulated draws of $\beta^0$ and $\sigma^2$ by Gibbs sampling algorithm.

In the usual practice, the hyper parameter $\beta^0$ and $\sigma^2$ are assumed to be independent with $\beta^0$ assigned a uniform prior and $\sigma^2$ assigned a noninformative prior. For an illustration purpose, we use hierarchical structure to decide whether we should include variable SAT or prerequisite course grade in the model using the same data set in Table 1.1 from Chapter 1. Albert and Chib(1993) using Gibbs sampling to simulate draws of $\beta^0$ and $\sigma^2$ based the conditional distribution of each parameter on the rest parameters for the binary case. We
extend their sampling algorithm to ordinal data case with some modification. Specifically, the conditional distribution needed for Gibbs sampling is given by

1. \( Z_i | y, \beta \) is distributed as a truncated normal distribution \( N(x_i' \beta, 1) \) with truncation by \( \gamma_{y_i-1} \) and \( \gamma_{y_i} \)

2. \( \beta | Z, \sigma^2 \) is distributed as multivariate normal \( N_q(\mu, V) \), where

\[
\mu = W_1 \hat{\theta}_1 + (I_2 - W_1) A \hat{\theta}_2 \\
\hat{\theta}_1 = (X'X)^{-1} X'Z \\
\hat{\theta}_2 = [A'X'(I_1 + XX'\sigma^2)^{-1}XA]^{-1} A'X'(I_1 + XX'\sigma^2)^{-1}Z \\
W_1 = [X'X + I_2/\sigma^2]^{-1}X'X \\
V = ((I_2 - W_1)A)[A'X'(I_1 + XX'\sigma^2)^{-1}XA]^{-1} + ((I_2 - W_1)A)' + [X'X + I_2/\sigma^2]^{-1}
\]

3. \( \sigma^2 | Z \) is distributed according to the density proportional to

\[
c(Z) \frac{|(I_1 + XX'\sigma^2)^{-1}|^{1/2}}{|A'X'(I_1 + XX'\sigma^2)^{-1}XA|^{1/2}} \exp\{-\frac{1}{2} Q(Z, XA\hat{\theta}_2, (I_1 + XX'\sigma^2))\} \pi(\sigma^2)
\]

where \( Q(Z, \mu, \Sigma) = (Z - \mu)^{-1} \Sigma^{-1} (Z - \mu) \), \( c(Z) \) the proportionality constant.

Based on the conditional distribution, we run the simulation and focus on the posterior mean of simulated draws of \( \sigma^2 \). If the posterior mean of simulated values of \( \sigma^2 \) is small, then we conclude that variables corresponding with \( \beta \) should be reduced to variables corresponding with \( \beta^0 \), otherwise, we think that we should keep variables corresponding with \( \beta \) in the model.

To decide whether we should include variable SAT in the model, we apply the idea of hierarchical analysis to the situation. The hierarchical model for this case is given by,

1. \( Z \) is distributed \( N(X \beta, I_1) \), where \( \beta \) is a \( 2 \times 1 \) vector.
2. \( \beta \) is distributed \( N(A\beta^0, \sigma^2 I_2) \), where \( A \) is a \( 2 \times 1 \) matrix and \( \beta^0 \) is a \( 1 \times 1 \) matrix.

3. \((\beta^0, \sigma^2)\) is distributed according to the prior density \( \pi(\beta^0, \sigma^2) \)

Using the above conditional distribution with the proper replacement, we simulate draws of \( \sigma^2 \) and compute the posterior mean of simulated values of \( \sigma^2 = 157.566 \), which is a large number for variance for the given data set. Based on the posterior mean of simulated \( \sigma^2 \), we conclude that variable SAT should be included in the model.

To decide whether we should include variable *prerequisite course grade* in the model, we apply the hierarchical structure to the same data set. To be specific, the hierarchical model for this case is given by

1. \( Z \) is distributed \( N(X\beta, I_1) \), where \( \beta \) is a \( 3 \times 1 \) vector.

2. \( \beta \) is distributed \( N(A\beta^0, \sigma^2 I_2) \), where \( A \) is a \( 3 \times 2 \) matrix and \( \beta^0 \) is a \( 2 \times 1 \) matrix.

3. \((\beta^0, \sigma^2)\) is distributed according to the prior density \( \pi(\beta^0, \sigma^2) \)

After using the conditional distribution in the Gibbs Sample and running the simulation, we get the posterior mean of simulated values of \( \sigma^2 = 0.069 \), which is a smaller number for the variance for the given data set. Based on the posterior mean of simulated \( \sigma^2 \), we conclude that variable *prerequisite course grade* should be removed from the model.
CHAPTER 5

Application to a Real Life Data Set

We proposed methods of detecting outliers and measures of goodness-of-fit in Chapter 3 and Chapter 4 and illustrated those methods using the sample data in Table 1.1 from Chapter 1. In this chapter we will apply the methods from Chapter 3 and Chapter 4 to the data set of Bowling Green State University students. Every year the Department of Mathematics and Statistics at Bowling Green State University offers a course named Math 126, Differential and Integral Calculus. In order to take Math 126, students have to get a grade of C or higher in a prerequisite course Math 120, Math 122, Math 128, or Math 130 or two years of high school algebra and one of geometry and a satisfactory placement exam score. Here we have a data set of 200 students with variables grade, pgrade, place, and ACT, where variable grade is the grade in Math 126, pgrade is the grade in the prerequisite course, placement is the grade in the placement score, and ACT is the student’s ACT math score. The purpose of this study has several aspects:

- Given a certain value of some predictors, we wish to learn about the probability of that students obtaining A, B, C, D or F.

- We wish to learn about the useful predictors in the model

- We wish to assess the goodness-of-fit of the model
• We are interested in identifying those students that have grades that are inconsistent with the fitted model.

The data set is given by Table 5.1.

<table>
<thead>
<tr>
<th>Student ID</th>
<th>Math 126 grade</th>
<th>ACT</th>
<th>placement test</th>
<th>pgrade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>22</td>
<td>41</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>19</td>
<td>41</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>20</td>
<td>32</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>23</td>
<td>42</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>25</td>
<td>42</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>15</td>
<td>20</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>B</td>
<td>18</td>
<td>32</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>24</td>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>193</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>194</td>
<td>B</td>
<td>19</td>
<td>32</td>
<td>C</td>
</tr>
<tr>
<td>195</td>
<td>C</td>
<td>22</td>
<td>32</td>
<td>C</td>
</tr>
<tr>
<td>196</td>
<td>C</td>
<td>19</td>
<td>20</td>
<td>C</td>
</tr>
<tr>
<td>197</td>
<td>A</td>
<td>24</td>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>198</td>
<td>B</td>
<td>21</td>
<td>41</td>
<td>B</td>
</tr>
<tr>
<td>199</td>
<td>B</td>
<td>21</td>
<td>41</td>
<td>A</td>
</tr>
<tr>
<td>200</td>
<td>C</td>
<td>22</td>
<td>32</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 5.1: The student Math 126 data set at Bowling Green State University. There are 200 observations, this table only shows the first 8 and the last 8 observations due to the limit of pages. The first column is the student ID, the second column is student’s Math 126 grade, the third column is student’s ACT math score, the fourth column is student’s placement test grade, and the last column is the student’s grade in the prerequisite course.

We begin with an exploratory graph to reveal the relationship between students’ Math 126 grade and variable ACT math score, placement score or prerequisite course grade. To illustrate, we take a look at the relationship between a student’s Math 126 grade with variable pgrade, ACT math score, placement test score respectively. In Table 5.1, we notice that the range of pgrade is 3, 4 and 5 and we count the proportion of getting A for each category of pgrade. Also we notice that the range of ACT math score is from 20-30, so we manually divide them into 5 ACT math score intervals, which are 20-22, 22-24, 24-26, 26-28 and 28-30 respectively, and we count the proportion of getting A in each ACT math score.
We also notice that the range of placement score is from 11 to 61 so we divide them into 10 categories, which are 11-16, 16-21, · · · , 56-61 and we count the proportion of getting A in each interval. We make scatterplots of proportion of students’ Math 126 getting A against students’ pgrade, ACT math score and obtain Figure 5.1, Figure 5.2 and Figure 5.3 respectively.

Figure 5.1: Scatterplot of proportions of Math 126 getting A against pgrade. The x-axis are the pgrade intervals, which are 3, 4, and 5 plotted on the x-axis; the proportions of getting A in Math 126 for each pgrade interval is plotted on the y-axis.

From Figure 5.1, 5.2 and 5.3 we can see that the chance of getting A in Math 126 is an increasing function of students’ prerequisite math score, so variable pgrade should be included in the model. We also see that the chance of getting A in Math 126 is an increasing function of students’ ACT math score, so variable ACT math score should be included in the model. Moreover, we can see that the chance of getting A in Math 126 is an overall increasing function of students’ placement score, so it may be proper to include variable
Figure 5.2: Scatterplot of proportions of Math 126 getting A against ACT math score. The x-axis are the ACT math score intervals, which are 20-22, 22-24, 24-26, 26-28 and 28-30 plotted on the x-axis, and we use the middle value of each interval to represent the corresponding interval; the proportions of getting A in Math 126 for each ACT interval is plotted on the y-axis.
Figure 5.3: Scatterplot of proportions of Math 126 getting A against placement. The x-axis are the placement score intervals, which are 11-16, 16-21, \ldots, and 56-61 plotted on the x-axis, and we use the middle value of each interval to represent the corresponding interval; the proportions of getting A in Math 126 for each placement score interval is plotted on the y-axis.
polr(as.factor(grade) ~ act + pgrade, method = "probit")

Call:
polr(formula = as.factor(grade) ~ act + pgrade, method = "probit")

Coefficients:
  act  pgrade
   0.059   0.75

Intercepts:
   1|2  2|3  3|4  4|5
   2.88 3.23 4.27 5.62

Residual Deviance: 516.26
  AIC: 528.26

Table 5.2: Probit Regression Model for the BGSU students using R.

placement in the model, which leads to a model selection issue. Moreover we notice that there are some students who do not get a good ACT math score, good placement score or good prerequisite course grade, but they get a good grade on Math 126, and it is interesting to identify those students, which leads to the outlier detection issue. Before we perform any further statistical analysis on the data set, we describe the regression model.

To model the probability of one student getting A, B, C, D or F, we propose a probit regression model given by

\[ \theta_{ic} = \Phi(\gamma_c - x_i'\beta) \] (5.1)

where \( \theta_{ic} \) is the cumulative probability of individual \( i \) falling in category \( c \) or below, \( \Phi \) is the cumulative density function (CDF) for a standard normal distribution, \( \gamma_c \) is the vector of unknown cutoff points, and \( x_i'\beta \) is the linear combination of predictors.

5.1 Model Interpretation

Suppose that we include variable pgrade and ACT in the model, and we fit the probit regression model to the BGSU student dataset, we obtain the following result given by Table 5.2.

From Table 5.2 we can easily compute the probability of one student getting a specific
grade such as A by the equation (5.1). One important issue when we fit this model is the interpretation of the coefficients of the parameters. Since we use the probit regression model and we have two variables in this model, there is no easy explanation of the meaning of the coefficients. One possible explanation is fixing one variable at its mean value and look at the relationship between the response and the other variable. To illustrate, we first take a look at the relationship between the probability of one student getting A in Math 126 against variable ACT. In order to reflect this relationship, we fix the value of pgrade at 4 and consider the probability of getting A in Math 126 against the variable ACT. We obtain Figure 5.4.

![Figure 5.4: The probability of getting A against variable ACT while fixing pgrade at 4. The x-axis is the ACT value, which is from 11 to 30, the y-axis is the probability of getting A in Math 126.](image)

From Figure 5.4, we can both see that the variable ACT and pgrade have a positive impact on the probability of getting A in Math 126. However, when we consider the probability of getting C in Math 126 against variable ACT while fixing pgrade at value 4, we obtain
a totally different relationship between the probability of getting C against variable ACT, which is given by Figure 5.5.

Figure 5.5: The probability of getting C against variable ACT while fixing pgrade at 4. The x-axis is the ACT value, which is from 11 to 30, the y-axis is the probability of getting C in Math 126.

Similarly, to consider the relationship of the probability of getting A in Math 126 against variable pgrade, we fix the value of ACT at its mean value and we obtain Figure 5.6. It seems that the probability of getting A in Math 126 is an increasing function of pgrade.

When we consider the probability of getting C in Math 126, we use the same technique of fixing ACT at its mean value and obtain Figure 5.7.

From Figure 5.7, we can see that the probability of obtaining C in Math 126 is a decreasing function of pgrade when we fix the value of ACT at its mean value.

In summary, there is no easy explanation of the meaning of coefficient unless for a particular category. A good method of explaining the meaning of one variable is plotting the probability of falling in a particular category against one variable.
Figure 5.6: The probability of getting A against variable pgrade while fixing variable ACT at its mean value. The x-axis is the pgrade value, which are 3, 4 and 5, the y-axis is the probability of getting A in Math 126.
Figure 5.7: The probability of getting C against variable pgrade while fixing variable ACT at its mean value. The x-axis is the pgrade value, which are 3, 4 and 5, the y-axis is the probability of getting C in Math 126.
5.2 Model Selection for the Application Data Set

5.2.1 Model Selection from a Frequentist’s Perspective

To look at the question of which variable should be included in the model, we first consider a frequentist’s perspective. The method of model selection from a frequentist’s perspective is based on the deviance statistic. In order to use deviance statistic to decide which variable should be included in the model, we propose three models:

- Model M0: only the constant should be included in the model
- Model M1: variable pgrade should be included in the model
- Model M2: variable pgrade and ACT should be included in the model
- Model M3: variable pgrade, ACT and placement should be included in the model

Then we fit four probit regression models for model M0, M1, M2 and M3 separately, as a result we get three deviance statistic $D_0$, $D_1$, $D_2$, and $D_3$ from model M0, M1, M2 and M3 respectively. In our data set, we get the deviance statistic $D_0 = 580.11$ from model M0, the deviance statistic $D_1 = 520.94$ from model M1 and the deviance statistic $D_2 = 516.25$ from model M2. The difference between $D_0$ and $D_1$ is $580.11 - 520.94 = 59.17$. Since the difference between deviance statistic $D_0$ and $D_1$ 59.17 is greater than the asymptotic $\chi^2$ 95% critical value 3.84, we conclude that variable pgrade should be included in the model. To decide whether we should include variable ACT in the model, we notice that the difference of deviance between $D_1$ and $D_2$ is $520.94 - 516.25 = 4.69$ and we compare this value with the 95% critical value of asymptotic $\chi^2$ distribution with 1 degrees of freedom, which is 3.84. Since the observed difference of deviance statistic between $D_1$ and $D_2$ 4.69 is greater the 95% critical values 3.84, we conclude that variable ACT should be included in the model. Similarly, to decide whether we should include variable placement in the model, we compute the difference of deviance statistic between model M2 and model M3 and we get $516.25 - 506 = 10.29$. The
95% critical value for the asymptotic $\chi^2$ distribution with 9 degrees of freedom is 16.92. Since the observed difference of deviance statistic between model M2 and M3 10.25 is less than the critical value 16.92, we conclude that variable placement should not be included in the model.

### 5.2.2 A Bayesian Perspective for Model Selection

A Bayesian perspective to decide which variable should be included in the model is based on the notation of Bayes factor. For an illustration purpose, we want to decide whether it is proper to include variable pgrade in the model, we reflect our beliefs about whether we should include variable pgrade in the model by using different informative priors, thus we propose two models:

- **Model M4** (variable pgrade should not be included in the model):
  - when pgrade=A, prob($\leq D$)=0.2, K=1000
  - when pgrade=B, prob($\leq D$)=0.2, K=1000
  - when pgrade=B, prob($\leq C$)=0.6, K=1000
  - when pgrade=C, prob($\leq C$)=0.6, K=1000
  - when pgrade=A, prob($\leq A$)=0.5, K=1000.

- **Model M5** (variable pgrade should be included in the model):
  - when pgrade=A, prob($\leq D$)=0.2, K=1
  - when pgrade=B, prob($\leq D$)=0.2, K=1
  - when pgrade=B, prob($\leq C$)=0.6, K=1
  - when pgrade=C, prob($\leq C$)=0.6, K=1
  - when pgrade=A, prob($\leq A$)=0.5, K=1

After combining the prior information into the posterior distribution, we can compute the Bayes factor based on the marginal likelihood from model M4 and model M5. We do the computation and get a Bayes factor $BF_{45} = 0.094$, which means that model M5 is more
likely to happen than model M4, so we conclude that variable pgrade should be included in the model.

So far we know that it is proper to include variable pgrade in the model, we want to decide whether we should include variable ACT in the model, so we assigning informative priors to reflect our beliefs about whether we should include variable ACT in the model, thus we propose another two models:

- **Model M6** (variable ACT should not be included in the model):
  - when pgrade=A, ACT=24, \( \text{prob}(\leq D) = 0.6, K=1000 \)
  - when pgrade=A, ACT=27, \( \text{prob}(\leq D) = 0.6, K=1000 \)
  - when pgrade=B, ACT=23, \( \text{prob}(\leq C) = 0.7, K=1000 \)
  - when pgrade=B, ACT=28, \( \text{prob}(\leq C) = 0.7, K=1000 \)
  - when pgrade=3, ACT=23, \( \text{prob}(\leq B) = 0.8, K=1000 \)

- **Model M7** (variable ACT should be included in the model):
  - when pgrade=A, ACT=24, \( \text{prob}(\leq D) = 0.6, K=1 \)
  - when pgrade=A, ACT=27, \( \text{prob}(\leq D) = 0.6, K=1 \)
  - when pgrade=B, ACT=23, \( \text{prob}(\leq C) = 0.7, K=1 \)
  - when pgrade=B, ACT=28, \( \text{prob}(\leq C) = 0.7, K=1 \)
  - when pgrade=3, ACT=23, \( \text{prob}(\leq B) = 0.8, K=1 \)

After we combine the prior information with the data, we can compute the Bayes factor based on the marginal likelihood from model M6 and model M7. We do the computation and get a Bayes factor \( BF_{67} = 0.38 \), which means that model M7 is more likely to happen than model M6, so we conclude that we should include variable ACT in the model.

To decide whether we should include variable placement in the model, we use the same trick of assigning informative priors, thus we propose two models:

- **Model M8** (variable placement should not be included in the model):
  - when pgrade=A, ACT=24, placement=20, \( \text{prob}(\leq D) = 0.6, K=1000 \)
when pgrade=A, ACT=24, placement=27, prob(≤D)=0.6, K=1000
when pgrade=B, ACT=27, placement=24, prob(≤C)=0.7, K=1000
when pgrade=B, ACT=27, placement=41, prob(≤C)=0.7, K=1000
when pgrade=C, ACT=23, placement=27, prob(≤B)=0.8, K=1000
when pgrade=C, ACT=23, placement=61, prob(≤B)=0.8, K=1000

- Model M9 (variable placement should be included in the model):
  when pgrade=A, ACT=24, placement=20, prob(≤D)=0.6, K=1
  when pgrade=A, ACT=24, placement=27, prob(≤D)=0.6, K=1
  when pgrade=B, ACT=27, placement=24, prob(≤C)=0.7, K=1
  when pgrade=B, ACT=27, placement=41, prob(≤C)=0.7, K=1
  when pgrade=C, ACT=23, placement=27, prob(≤B)=0.8, K=1
  when pgrade=C, ACT=23, placement=61, prob(≤B)=0.8, K=1

After we combining the prior information with the data, we compute the Bayes factor of model M8 over model M9 based on the marginal likelihood from Model M8 and model M9 and get a value $BF_{89} = 23.87$, which means that model M8 is more likelihood to happen than model M9. Based on this Bayes factor values, we conclude that variable placement should not be included in the model.

5.3 Outlier Detection

5.3.1 Outlier Detection from a Frequentist’s Perspective

As stated before, there are some students who do not have good grades or ACT math scores but those students do well in Math 126 and we are interested in identifying those students out. Including outlying observations in the model can affect the estimated coefficients for the variable thus it will cause inaccurate prediction of future value or inaccurate inference about parameters. One method of outlier detection from a frequentist’s perspective is based on the
individual deviance statistic, denoted by \( D_i(y) \) and defined by equation (3.3). Observations that contribute disproportionally to the whole deviance statistic should be identified with suspicion. We compute individual deviance statistic for those 200 observations and plot them against observation ID, which leads to Figure 5.8.

![Figure 5.8: The plot of individual deviance statistic \( D_i(y) \) against observation ID. The x-axis is the observation ID, the y-axis is the individual deviance statistic \( D_i(y) \). Unusually large values of individual deviance indicate outliers.](image)

From Figure 5.8 we can see that there are two observations having unusually large individual deviance statistic, which are observation number 17 and observation number 103, so we identify them out with suspicion. However, we cannot conclude whether those two observations identified by Figure 5.8 are outliers, because we do not have a specific value to determine how large is (the individual deviance) disproportionally contributed to the overall deviance.
5.3.2 Outlier Detection from a Bayesian Perspective

In order to use individual deviance statistic to detect outliers from a Bayesian perspective, we consider the new individual deviance statistic $D_i(y^{rep})$ defined by equation (3.16) using the replicated “future” observation. One way to compare the individual deviance $D_i(y)$ and individual deviance $D_i(y^{rep})$ is to compute the probability of deviance statistic $D_i(y^{rep})$ is greater than the individual deviance $D_i(y)$ for each observation. If the probability is small, it means that the observed individual deviance $D_i(y)$ are unlikely to happen in large number of repeated experiments and we can conclude the corresponding observation is an outlier.

We computes the probabilities of $D_i(y^{rep}) > D_i(y)$ and plot them against the observation ID, which leads to Figure 5.9.

![Figure 5.9](image)

Figure 5.9: The probability of $D_i(y^{rep})$ are greater than the individual deviance $D_i(y)$ against observation ID for each observation. The x-axis is the observation ID, the y-axis is the probability of $D_i(y^{rep})$ are greater than the individual deviance $D_i(y)$. Small probabilities indicate the corresponding observations are outliers.

From Figure 5.9 we can see that there are 2 observations with unusually smaller probabilities, which are 0.007 for observation number 17 and 0.005 for observation number 103, thus we conclude these two observations as outliers.
Another Bayesian method of using the individual deviance to detect outliers is adopting the idea of parameter-dependent test statistic $D_i(y^{rep}, \beta)$, which is defined in equation (3.17) and discussed in Chapter 3. Using parameter-dependent test statistic $D_i(y^{rep}, \beta)$ to detect outliers is to compute the probability of the individual deviance statistic $D_i(y^{rep}, \beta)$ is greater than the individual deviance $D_i(y, \beta)$ for each observation. If the probabilities are small, it means that the observed individual deviances $D_i(y, \beta)$ are unusually large and we can conclude the corresponding observations are outliers. We compute the probabilities and plot them against the observation ID, which leads to Figure 5.10.

![Figure 5.10](image)

Figure 5.10: The probability of $D(y_i^{rep}, \beta)$ are greater than the individual deviance $D_i(y, \beta)$ against observation ID for each observation. The x-axis is the observation ID, the y-axis is the probability of probability of $D(y_i^{rep}, \beta)$ is greater than the individual deviance $D_i(y, \beta)$. Small probabilities indicate the corresponding observations are outliers.

From Figure 5.10, we can see that there are also 2 observations with extremely smaller probabilities, which are 0.001 for observation number 17 and 0.0005 for observation number 103, thus we identify these two observations as outliers.

To be more confident about our conclusion that observation 17 and 103 are outliers, we take a further look at observation 17 and observation 103, we notice that observation 17
obtains a F while observation 17 has the ACT Math score 19 and an A on the prerequisite course. Observation 103 get a F in Math 126 while observation 103 has a ACT Math score 26 and an A on the prerequisite course. We compute the probability of obtaining F in Math 126 when ACT=19 and 26 and pgrade=A, and we get 0.02 and 0.008 respectively. These two probabilities are so small that they hardly happen in one experiment, so we conclude that the fact that observation 17 and 103 get a F in Math 126 is inconsistent with the probit regression model thus these two observations are outliers.

There are other methods of detecting outliers from both a frequentist’s perspective and Bayesian perspective. Generally speaking, the frequentist’s method depends on the asymptotic distribution to conclude whether one observation is an outlier. In contrast, the Bayesian method depends on the simulated values of “future” observation $y^{rep}$ to conclude whether an observation is an outlier and does not need to rely on an asymptotic distribution.

5.4 Goodness-of-fit for the Model

5.4.1 Goodness-of-fit from a Frequentist’s Perspective

To check the goodness-of-fit for the probit regression model for Math 126 grade example, we first consider this goodness-of-fit question from a frequentist’s perspective. The test statistics that we use is the deviance statistic $D(y)$, which is defined by (4.1). We do the computation for the probit regression model using variable pgrade and ACT as the predictors and get the value of deviance statistic $D(y)$, which is 516.26. The 95% critical value of its asymptotic distribution $\chi^2$ with $200 \times (5 - 1) - 2 = 798$ degrees of freedom is 864.83. Since the observed deviance statistic 516.26 is smaller than 864.83, we conclude that the probit regression model using ACT and pgrade for the Math 126 grade data set is a suitable model.

One concern about the frequentist’s method is how good the asymptotic $\chi^2$ distribution is. To study this, we fix the value of parameters and simulate the deviance value. We make a histogram of the simulated deviance value and obtain Figure 5.11.
Figure 5.11: The histogram of simulated deviance. The dotted line is the mean of the simulated deviance statistic $D(y)$ from a frequentist’s perspective and the dashed line is the density estimate.

From Figure 5.11 we can see the deviance statistic does not well approximate its asymptotic $\chi^2$ distribution with 797 degrees of freedom because the center of the histogram is far from the mean of theoretical $\chi^2$ distribution, which causes the problem of the validity in the frequentist’s perspective.

5.4.2 Goodness-of-fit from a Bayesian Perspective

The Bayesian perspective to measure the goodness-of-fit of the probit regression model for the Math 126 data set is also based on the posterior predictive distribution of deviance statistic $D(y^{rep})$, which is defined in equation (4.8). We use the posterior predictive observation $y^{rep}$ to compute the deviance statistic $D(y^{rep})$, then we compare the deviance statistic $D(y^{rep})$ with the deviance statistic $D(y)$. If the probit regression model is a suitable fit, then we expect that $D(y)$ not to fall in the extreme end of $D(y^{rep})$. We compare $D(y)$ with $D(y^{rep})$ by making a histogram of $D(y^{rep})$ and marking the position of $D(y)$, which is given by Figure
Figure 5.12: The histogram of deviance statistic $D(y^{rep})$ using the posterior predicted value $y^{rep}$. The dotted line is the deviance statistic $D(y)$ from a frequentist’s perspective. A model is a suitable fit if the deviance $D(y)$ does not fall in the extreme end of $D(y^{rep})$.

From Figure 5.12, we can see the deviance statistic $D(y)$ does not fall in the extreme end of the deviance statistic $D(y^{rep})$, so we conclude the probit regression for the Math 126 data set with variable ACT and pgrade is a suitable fit. To consolidate our conclusion, we look at the tail probability of $D(y^{rep}) > D(y)$. We do the computation and get a value of 24.92%, which verifies our conclusion that the probit regression model for the Math 126 data set is a suitable fit.

Another Bayesian method to measure the goodness-of-fit is based on the parameter-dependent deviance statistic, which is defined by equation (4.9) and applied in Chapter 4. We use the simulated “future” observation $y^{rep}$ and simulated $\beta$ values to compute the deviance $D(y^{rep}, \beta)$, and $D(y, \beta)$. We compare $D(y, \beta)$ with $D(y^{rep}, \beta)$ by making a scatterplot, which is given by Figure 5.13.

From Figure 5.13, we can see that the deviance $D(y^{rep}, \beta)$ is not significantly greater
Figure 5.13: The scatterplot of deviance $D(y^{rep}, \beta)$ against $D(y^{rep}, \beta)$. The x-axis is the deviance $D(y^{rep}, \beta)$, the y-axis is the deviance $D(y, \beta)$ and the dotted line is the 45-degree straight line. A model is a suitable fit if the deviance $D(y^{rep}, \beta)$ is not significantly greater than $D(y, \beta)$. 
than $D(y, \beta)$. We compute the probability of $D(y^{rep}, \beta) > D(y, \beta)$ and we get 16.28\%, which confirms that the probit regression model using ACT and pgrade as the predictors is a suitable fit.

## 5.5 Summary

In Chapter 5 we apply our method of detecting outliers and measuring the goodness-of-fit to a BGSU students data set. In this data set we have some observations that are outlying. We treat both from a frequentist’s perspective and a Bayesian perspective. The frequentist’s perspective usually depends on the asymptotic distribution, which most of times works out fine, but there are some cases when the frequentist’s method does not work such as the obtained statistic value is close to the critical value of its asymptotic distribution. On the contrary, the Bayesian perspective usually use the replicated “future” observation $y^{rep}$ and compare the statistic using $y^{rep}$ with statistic using the observed value $y$. One advantage of Bayesian perspective over the frequentist’s perspective is the removal of asymptotic distribution, which is especially useful under the circumstance that the test statistic does not have a nice form of asymptotic distribution.
CHAPTER 6

Summary and Conclusions

This dissertation started with the introduction of ordinal data and ordinal regression. It was followed by a sample data set that we will use till Chapter 4 to illustrate the methods both from a frequentist’s perspective and a Bayesian perspective. To be more prepared for the ordinal data and ordinal regression, we reviewed the binary data and binary regression case first. In the binary data and binary regression section, we introduce three regression models for binary data case, which are probit regression model, logistic regression and complimentary log-log regression model respectively. We also compare these three models and conclude that there is no preference of one model over others by plotting the fitted probabilities from these three models on one graph. We also introduce the technique of estimating parameters, detecting outliers and measuring the goodness-of-fit for the binary regression model both from a frequentist’s perspective.

Chapter 2 starts with the introduction of the Bayes rule, which is the basis of posterior distribution of the parameter. Then we introduce the important concept of latent variables, which is very useful in the simulation process. The introduction of latent variable can greatly change the expression of likelihood function. After that, we introduce two types of priors, one is the uninformative prior, and the other is the informative priors, which can be used to reflect the belief of which variable should be included in the model. Since Bayesian
method requires a lot of simulations, we then move to the part of simulations algorithms, which include Metropolis-Hastings algorithm, random walk algorithm and Gibbs Sampling algorithm. We also review the method of parameter estimation from both a frequentist’s perspective and a Bayesian perspective.

In Chapter 3 we focus on outlier detection problem for ordinal data. First, we want to detect outliers from a Frequentist’s perspective, and there are some difficulties in front, and the biggest one is that the fitted probability for each observation is a vector and this leads to a multivariate distribution problem, so far it is not clear how to define and detect outliers from a classic perspective because of the dimension of the response. A lot of researches have been done to define residual and detect outliers such as Fuchs and Kenett (1980), we adopt their method and apply the sample data set to detect outliers. However, these classical methods depend on the asymptotic distribution to detect outliers, which may not be accurate especially when the sample size is small. On the other hand, with the introduction of latent variables into the regression model, we are able to define latent residual for each observation. One advantage of latent residual over the classic residual is that we only one latent residual for each observation thus it is convenient to apply the traditional idea to detect outliers. Based on the latent residual, we propose the $3\sigma$ rule and QQ plot to detect outliers. Also we review the idea of Bayes factor and use it to detect outliers. The last method from a Bayesian perspective is that we use the hierarchical model structure to detect outliers.

We begin Chapter 4 with the reason why we need the measure of goodness-of-fit and we first look at the goodness-of-fit from a Frequentist’s perspective. The test statistic is the deviance and based on the deviance statistic we can conclude whether a model is suitable fit when we compare the observed deviance with its asymptotic distribution. The potential problem with the frequentist’s method is the reliance on the asymptotic distribution, which could cause a problem especially when the data sample is small or when the observed deviance are close the critical value of its asymptotic distribution. To avoid the problem of reliance on the asymptotic distribution, we use the deviance statistic from a Bayesian perspective; we
take a look at the posterior predictive distribution of deviance and the parameter-dependent deviance. By doing this we no longer need the asymptotic distribution. Next we look at the model selection problem. Firstly we use the difference of deviance between two competing models and compare it with the asymptotic distribution, and then we use the idea of Bayes factor to decide which variable should be included in the model. Lastly, we use the hierarchical model structure to decide which variable should be included in the model.

In Chapter 5 we apply the methods from both Chapter 3 and Chapter 4 to the BGSU student. We choose 200 data samples and we detect outliers, choose between competing models and measure goodness-of-fit from both a frequentist’s perspective and a Bayesian perspective. We use some old programs from Chapter 3 and Chapter 4 and revise it to fit the situation in Chapter 5.

We can find a lot of applications of ordinal data not only in academic area, but also in other area such as social science and medical study. The contribution of this dissertation is that we systematically study the possible problem of ordinal data, such as outlier detection problem, model selection problem and measure of goodness-of-fit problem. One possible future research is to extend our methodology to other area and we are sure that a Bayesian method will find its applications not only in the area of ordinal data analysis, but also in other area. One possible difficulties is that acknowledgement of simulation methods is required because Bayesian use a lot of simulation.
Appendix A

Part of R Code for Chapter 3

A.1 Ordinal.probit Program

#### This ordinal.probit program is used to simulate the beta,gamma and the latent variable Z from the posterior distribution

#### This program is using Nandram and Chen(1996)'s algorithm

```r
ordinal.probit=function (data,K,m)
{
    # oprobit fits the ordinal model (probit link)
    # P(y_i \leq k) = \Phi(\gamma_k - x_i \beta)
    # on parameters (delta,\gamma^*,\beta^*,z^*)
    # command: ordinal.probit(data,K,m)
    # input: data - [y x], where y is ordinal response (1,2,...) and x is regression matrix
    # K - number of levels (at least 3)
    # m - number of iterations of simulation algorithm
    # output: Mbeta - matrix of simulated values of beta*
    # (each row is a simulated vector)
```

### Mdelta2 - column vector of simulated values of \( \delta^2 \)

### Mgam - matrix of simulated values of gamma*

### Mz - matrix of simulated values of z*

```r
y = data[,1]; x = data[,-1]
c = table(y)
nk = dim(x); n = nk[1]; k = nk[2]
a = chol(solve(t(x) %% x))
gam = (1:(K-3))/(K-2)
cpoints = c(-Inf,0,gam,1,Inf)
lo = cpoints[y]; hi = cpoints[y+1]
z = c(-.5,0,gam,1)
beta = solve(t(x) ** x) ** (t(x) ** z)[y])
delta2 = .1
Mbeta = array(0,c(m,k))
Mdelta2 = array(0,c(m,1))
Mgam = array(0,c(m,length(gam)))
Mz = array(0,c(length(y),m))
alpha = .2
for (i in 1:m)
{
  mu = x ** beta; sigma = sqrt(delta2)
  pr = pnorm(cbind((lo-mu)/sigma,(hi-mu)/sigma))
  Mz[,i] = mu + sigma * qnorm(runif(length(lo)) * (pr[,2]-pr[,1])+pr[,1])
  if (K>3)
  {
    If = sum(log(pr[,2]-pr[,1]))
    p = diff(c(0,gam,1))
  }
}
```
lp = sum((alpha*ct[2:(K-1)]-1)*log(p))

p1 = rgamma(K-2, shape = alpha*ct[2:(K-1)], rate = 1)

p1 = p1 / sum(p1)

lp1 = sum((alpha*ct[2:(K-1)]-1)*log(p1))

gam1 = cumsum(p1)

cpoints1 = c(-Inf, 0, gam1, Inf)

lo1 = cpoints1[y]; hi1 = cpoints1[y+1]

pr1 = pnorm(cbind((lo1-mu)/sigma, (hi1-mu)/sigma))

lf1 = sum(log(pr1[,2]-pr1[,1]))

prob = exp((lf1-lp1)-(lf-lp))

if (runif(1) < prob)
{

gam = gam1[1:(length(gam1)-1)]

lo = lo1; hi = hi1
}

if (K >= 4)

Mgam[i,] = gam

a1 = (n+k+K)/2

b1 = .5*sum((Mz[i]-x%*%beta)^2)

delta2 = b1/rgamma(1, shape = a1, rate = 1)

mn = solve(t(x)%*%x)%*%t(x)%*%Mz[i]

beta = mn + sqrt(delta2)*t(a)%*%array(rnorm(k),c(k,1))

Mbeta[i,] = t(beta)

Mdelta2[i] = delta2

Mb = Mbeta/(sqrt(Mdelta2)%*%array(1,c(1,k)))
return(list(Mbeta=Mbeta,Mdelta2=Mdelta2,Mgam=Mgam,Mz=Mz))
}

## A.2 Probitpost Program

### This probitpost program is the posterior distribution for a probit regression model

```r
probitpost=function (beta,data)
{
### This program using the flat prior for the parameters
x = data[, 1]
n = data[, 2]
y = data[, 3]
beta0 = beta[, 1]
beta1 = beta[, 2]

omatrix=array(0,c(length(y),5))
for( i in 1:length(y))
{
  omatrix[i,]=c(1,0,0,0,0)*(y[i]==1)+c(0,1,0,0,0)*(y[i]==2)+c(0,0,1,0,0)*(y[i]==3)+
  c(0,0,0,1,0)*(y[i]==4)+c(0,0,0,0,1)*(y[i]==5)
}

z=0
for( i in 1:length(y))
{
  lp = beta0 + beta1 * x[i]
  p1=pnorm(estimated.cutpoints[1]-lp,0,1)
  p2=pnorm(estimated.cutpoints[2]-lp,0,1)-pnorm(estimated.cutpoints[1]-lp,0,1)
  p3=pnorm(estimated.cutpoints[3]-lp,0,1)-pnorm(estimated.cutpoints[2]-lp,0,1)
```

```
p4 = pnorm(estimated.cutpoints[4]-lp,0,1) - pnorm(estimated.cutpoints[3]-lp,0,1)

p5 = 1 - pnorm(estimated.cutpoints[4]-lp,0,1)

part1 = omatrix[i,][1]*log(p1)

part2 = omatrix[i,][2]*log(p2)

part3 = omatrix[i,][3]*log(p3)

part4 = omatrix[i,][4]*log(p4)

part5 = omatrix[i,][5]*log(p5)

z = z + part1 + part2 + part3 + part4 + part5
}
return(z)

### A.3 Use Bayes Factor to Detect Outliers R Code

```r
#### We want to use Bayes factor to detect outliers
#### Bayes factor is the ratio of marginal likelihood from two competing models
#### We use the idea of "leaving-one-out"
#### every time we leave one observation out and use the laplace function to find marginal likelihood for the new data set
#### We need to source in the probitpost program
#### We need to source in the ordinal.probit program
#### We need to source in the LearnBayes package
#### We need to read in the data set, which is named "example 4.5 data.txt"
source("/Users/nickwestlake/Desktop/Research/Book Ordinal Data Modeling/Chapter 4 R stuff/probitpost.R")

source("/Users/nickwestlake/Desktop/Research/Book Ordinal Data Modeling/Chapter 4 R stuff/ordinal.probit.R")
```
require(LearnBayes)
data=read.table("example 4.5 data.txt",header=T)
attach(data)
data=cbind(grade,constant,sat)
o-data=cbind(sat,constant,grade)
fit.original=laplace(probitpost,post.mean.beta,100,odata)
logm.original=fit.original$int
logm=rep(0,length(grade))
for (j in 1:length(grade))
{
  fit=laplace(probitpost,post.mean.beta,100,odata[-j,])
  logm[j]=fit$int
}
BF=array(0,c(length(grade),1))
BF=exp(logm-logm.original)
plot(1:length(grade),log(BF,10),xlab="Observation ID",ylab="Log (base 10) Bayes factor")
abline(v=c(1:length(grade)),lty=3)

A.4 Individual Deviance to Detect Outliers R Code

### We use the individual deviance from a frequentist’s perspective to detect outliers
### We compute $D_i(y)$ for each observation and pay attention to those observation who
### have an unusually large individual deviance
### This method can only identify observation with suspicion, it cannot conclude outliers.
### mat.grade is the grade matrix whose dimension is 30*5
mat.grade=array(0,c(length(grade),5))
for( i in 1:length(grade))
{
mat.grade[i,]=(grade[i]==1)*c(1,0,0,0,0)+(grade[i]==2)*c(0,1,0,0,0)+
(grade[i]==3)*c(0,0,1,0,0)+(grade[i]==4)*c(0,0,0,1,0)+(grade[i]==5)*c(0,0,0,0,1)
} post.p1=pnorm(estimated.cutpoints[1]-x%*%post.mean.beta)
post.p2=pnorm(estimated.cutpoints[2]-x%*%post.mean.beta)-post.p1
post.p3=pnorm(estimated.cutpoints[3]-x%*%post.mean.beta)-post.p1-post.p2
post.p4=pnorm(estimated.cutpoints[4]-x%*%post.mean.beta)-post.p1-post.p2-post.p3
post.p5=1-post.p1-post.p2-post.p3-post.p4
post.column.prob=array(c(post.p1,post.p2,post.p3,post.p4,post.p5),c(length(grade),5))
device1=array(0,c(length(grade),1))
for(i in 1:length(grade))
{
ColNum=which(mat.grade[i,]==1)
device1[i]=2*log(1/post.column.prob[i,ColNum])
}
plot(device1,xlab="Observation ID",ylab="Individual Deviance from a Frequentist’s Perspective")
abline(v=c(1:30),lty="dotted",col="black")

A.5 Posterior Predictive Distribution of $D_i(y^{rep})$ to Detect Outliers Code

### This method uses the posterior predictive distribution of individual deviance statistic
### the individual deviance in this case can be defined as $D_i(y^{rep})$
### we can get many $D_i(y^{rep})$s for each observation using the simulated $y^{rep}$ values.
we want to compare $D_i(y^{rep})$ and $D_i(y)$ for each observation

we compute the $P(D_i(y^{rep}) > D_i(y))$ for each observation and pay attention to small probabilities.

those observations which have unusually small probabilities are outliers

iter=5000

deviance3=array(0,c(length(grade),iter))

length(grade)=30 is the number of observations that we have

simulated.category=array(0,c(length(grade),5))

post.p1=pnorm(estimated.cutpoints[1]-x*x%*%post.mean.beta)
post.p2=pnorm(estimated.cutpoints[2]-x*x%*%post.mean.beta)-post.p1
post.p3=pnorm(estimated.cutpoints[3]-x*x%*%post.mean.beta)-post.p1-post.p2
post.p4=pnorm(estimated.cutpoints[4]-x*x%*%post.mean.beta)-post.p1-post.p2-post.p3
post.p5=1-post.p1-post.p2-post.p3-post.p4
post.column.prob=array(c(post.p1,post.p2,post.p3,post.p4,post.p5),c(length(grade),5))

for( i in 1:length(grade))

{for(m in 1:iter)

{simulated.category[i,]=rmultinom(1,1,post.column.prob[i,])
ColNum=which(simulated.category[i,]==1)
deviance3[i,m]=2*log(simulated.category[i,ColNum]/post.column.prob[i,ColNum])
}
}
tail.prob3=array(0,c(length(grade),1))

for(i in 1:length(grade))

{tail.prob3[i]=sum(deviance3[i,]¿deviance1[i])/iter}
```r
tail.prob3
plot(tail.prob3,xlab="Observation ID",ylim=c(0,1),ylab="Probability")
abline(v=c(1:30),lty="dotted",col="black")
####method 2 ~ using the simulated Y value and the simulated value of beta D(yrep,beta)

### A.6 Parameter-dependent $D_i(y_{rep}, \beta)$ to Detect Outliers

#### R Code

```
p4=pnorm(estimated.cutpoints[4]-x%*%beta[m,])-p1-p2-p3
p5=1-p1-p2-p3-p4
column.prob=array(c(p1,p2,p3,p4,p5),c(length(grade),5))
for( i in 1:length(grade))
{
simulated.Y[i,]=rmultinom(1,1,column.prob[i,])
ColNum41=which(simulated.Y[i,]==1)
ColNum42=which(mat.grade[i,]==1)
D41[i,m]=2*log(1/column.prob[i,ColNum41])
D42[i,m]=2*log(1/column.prob[i,ColNum42])
}
compare.prob=array(0,c(length(grade),1))
for( i in 1:length(grade))
{
compare.prob[i]=sum(D41[i,]>D42[i,])/iter
}
plot((1:length(grade)),compare.prob,ylim=c(0,1),xlab="Observation ID",ylab="Prob of D(yrep,beta)>D(y,beta)")
abline(v=c(1:length(grade)),lty=3,col="black")

A.7 Ordinal.probit.hierarchical R Code

### this program using the hierarchical structure to detect outliers
### previously we assume that the latent variable \( Z_i \) has the same variance, which is 1
### the hierarchical structure put a parameter on the variance of the distribution of latent variable \( Z_i \)
large variance of latent variables indicates the corresponding observation is an outlier.

```R
ordinal.probit.hierarchical=function (data,K,iter,v)
{
    #oprobit fits the ordinal model (probit link)
    ### P(y_i<=k) = Phi(gamma_k - x_i beta)
    ### using Nandram and Chen(1996)'s algorithm
    ### on parameters (delta, gamma*, beta*, z*)
    ### v is the degree of freedom
    ### input: data - [y x], where y is ordinal
    ### response (1,2,...) and x is regression matrix
    ### K - number of levels (at least 3)
    ### iter - number of iterations of simulation algorithm
    ### output: Mbeta - matrix of simulated values of beta*
    ### (each column is a simulated vector)
    ### Mdelta2 - column vector of simulated values of delta^2
    ### Mz-matrix of simulated values of z*, dimension of n*iter
    ### Mlambda-matrix of the simulated values of lambda
    y=data[,1]; x=data[,-1]
    ct=table(y)
    nk=dim(x); n=nk[1]; k=nk[2]
    gam=(1:(K-3))/(K-2)
    cpoints=c(-Inf,0,gam,1,Inf)
    lo=cpoints[y]; hi=cpoints[y+1]
    z=c(-0.5,0,gam,1.1)
    beta=solve(t(x)%*%x)%*%t(x)%*%z[y]
    delta2=0.0625
    Mbeta=array(beta,c(k,iter))
}
```
Mdelta2 = array(delta2, c(iter, 1));
Mz = array(z, c(n, iter))
Mlambda = array(rep(1, n), c(n, iter))
Mgam = array(gam, c(iter, length(gam)))
alpha = 0.2
for (i in 1:iter)
{
    ### simulate z from truncated normal distribution with mean xbeta
    mu = x %*% Mbeta[,i]; sigma = sqrt(Mdelta2[i]/Mlambda[,i])
    prob = pnorm(cbind((lo - mu)/sigma, (hi - mu)/sigma))
    Mz[,i] = mu + sigma * qnorm(runif(length(lo))*(prob[,2]-prob[,1])+prob[,1])
    ### this one will produce a diagonal matrix with diagonal w=diag(Mlambda[,i])
    ### simulate beta values from a normal distribution
    mn = solve(t(x) %*% solve(w) %*% t(x) %*% solve(w) %*% Mz[,i])
    varcov = solve(t(x) %*% solve(w) %*% t(x) %*% Mdelta2[i])
    Mbeta[,i] = rmnorm(1, mn, varcov)
    ### simulate lambda values from a gamma distribution
    shape1 = (v+1)/2
    rate1 = v/2 + (Mz[,i]-x %*% Mbeta[,i])^2/(2*Mdelta2[i])
    Mlambda[,i] = rgamma(n, shape1, rate=rate1)
    ### simulate delta2 values
    rate2 = sum(Mlambda[,i]*(Mz[,i]-x %*% Mbeta[,i])^2)/2
    Mdelta2[i] = 1/rgamma(1, shape2, rate=rate2)
    ### simulated values of gamma(cutoff points)
    if (K > 3)
    {
}
lf = sum(log(prob[,2]-prob[,1]))
p = diff(c(0,gam,1))
lp = sum((alpha*ct[2:(K-1)]-1)*log(p))
p1 = rgamma(K-2,shape=alpha*ct[2:(K-1)],rate=1)
p1 = p1/sum(p1)
lp1 = sum((alpha*ct[2:(K-1)]-1)*log(p1))
gam1 = cumsum(p1)
cpoints1 = c(-Inf,0,gam1,Inf)
lo1 = cpoints1[y]; hi1 = cpoints1[y+1]
prob1 = pnorm(cbind((lo1-mu)/sigma,(hi1-mu)/sigma))
lf1 = sum(log(prob1[,2]-prob1[,1]))
test.prob = exp((lf1-lp1)-(lf-lp))
if (runif(1)<test.prob)
{
gam = gam1[1:(length(gam1)-1)]
lo = lo1; hi = hi1
}
}
if (K>=4)
Mgam[i,] = gam
return(list(Mbeta=Mbeta,Mdelta2=Mdelta2,Mz=Mz,Mlambda=Mlambda,Mgam=Mgam))
Appendix B

Part of R Code for Chapter 4

B.1 Using Deviance as a Measure of GOF R Code

```r
### this method using deviance statistic as a measure of goodness-of-fit
### the asymptotic distribution of deviance statistic is Chi Square with n(k-1)-q
### here n is the number of observations;
### q is the number of the regression parameters;
### k is the number of the categories
### we compare the deviance $D(y)$ with the critical value of its asymptotic distribution
### if $D(y)$ is greater than the critical value, then we conclude the model is not a suitable fit.
### mat.grade is the grade matrix which is 30*5, where 30 is the number of observations and 5 is the number of categories
### post.column.prob is the matrix using the posterior mean of beta
post.p1=pnorm(estimated.cutpoints[1]-x%*%post.mean.beta)
post.p2=pnorm(estimated.cutpoints[2]-x%*%post.mean.beta)-post.p1
post.p3=pnorm(estimated.cutpoints[3]-x%*%post.mean.beta)-post.p1-post.p2
post.p4=pnorm(estimated.cutpoints[4]-x%*%post.mean.beta)-post.p1-post.p2-post.p3
```
post.p5=1-post.p1-post.p2-post.p3-post.p4
defiance1=0
for(i in 1:length(grade))
{
  ColNum=which(mat.grade[,]==1)
  defiance1=deviance1+2*log(mat.grade[i,ColNum]/post.column.prob[i,ColNum])
}

deviance1
### we expect the observed deviance are smaller than the critical value
### so that we can conclude that the model is a suitable.
qchisq(0.95,30*(5-1)-2)

### B.2 Using Deviance $D(y^{rep})$ as a Measure of GOF

#### Code

#### This is using posterior predictive distribution of deviance statistic $D(y^{rep})$ as a measure of goodness-of-fit

#### $D(y^{rep})$ is the deviance statistic using the “future” observation $y^{rep}$

#### We can get many $D(y^{rep})$s and we can compare them with $D(y)$

#### We can compute the probability $P(D(y^{rep}) > D(y)$

#### if the probability is small, then the model is not a suitable fit. iter=5000
deviance3=array(0,c(iter,1))

#### length(grade)=30 is the number of observations that we have

simulated.category=array(0,c(length(grade),5))

post.p1=pnorm(estimated.cutpoints[1]-x%*%post.mean.beta)

post.p2=pnorm(estimated.cutpoints[2]-x%*%post.mean.beta)-post.p1
post.p5 = 1 - post.p1 - post.p2 - post.p3 - post.p4
post.column.prob = array(c(post.p1, post.p2, post.p3, post.p4, post.p5), c(length(post.p1), 5))
for(m in 1:iter) {
p1 = pnorm(estimated.cutpoints[1] - x%*%beta[m,])
p2 = pnorm(estimated.cutpoints[2] - x%*%beta[m,]) - p1
p3 = pnorm(estimated.cutpoints[3] - x%*%beta[m,]) - p1 - p2
p5 = 1 - p1 - p2 - p3 - p4
column.prob = array(c(p1, p2, p3, p4, p5), c(length(p1), 5))
for( i in 1:length(grade)) {
simulated.category[i,] = rmultinom(1, 1, column.prob[i,])
ColNum = which(simulated.category[i,] == 1)
deviance3[m,] = deviance3[m,] + 2*log(simulated.category[i, ColNum] / post.column.prob[i, ColNum])
}

hist(deviance3, xlab = "Deviance Using Simulated Y", main = NULL)
abline(v = deviance1, lty = "dotted", lwd = 3)
tail.prob3 = sum(deviance3 > deviance1) / iter; tail.prob3

B.3 Using Parameter-dependent Deviance as a Measure of GOF R Code

### this method compares two deviance \( D(y, \beta) \) with \( D(y^{rep}, \beta) \)
### \( D(y, \beta) \) using the simulated \( \beta \) values and the observed \( y \)
### $D(y_{rep}, \beta)$ using the simulated $\beta$ values and the simulated $y_{rep}$

### we can compute the probability $P(D(y_{rep}, \beta) > D(y, \beta))$

### if the probability is too large or too small, then we conclude that the model is not a suitable fit

iter = 5000

deviance2 = array(0, c(iter, 1))
deviance4 = array(0, c(iter, 1))

### length(grade) = 30 is the number of observations that we have

simulated.category = array(0, c(length(grade), 5))

for (m in 1:iter)
{
    p1 = pnorm(estimated.cutpoints[1] - x %*% beta[m,])
    p2 = pnorm(estimated.cutpoints[2] - x %*% beta[m,]) - p1
    p3 = pnorm(estimated.cutpoints[3] - x %*% beta[m,]) - p1 - p2
    p5 = 1 - p1 - p2 - p3 - p4
    column.prob = array(c(p1, p2, p3, p4, p5), c(length(grade), 5))
    for (i in 1:length(grade))
    {
        ColNum = which(mat.grade[i,] == 1)
        deviance4[m,] = deviance4[m,] + 2 * log(1 / column.prob[i, ColNum])
        simulated.category[i,] = rmultinom(1, 1, column.prob[i,])
        ColNum2 = which(simulated.category[i,] == 1)
        deviance2[m,] = deviance2[m,] + 2 * log(simulated.category[i, ColNum2] / column.prob[i, ColNum2])
    }
}

plot(deviance4, deviance2, xlab = "Deviance $D(y, \beta)$", ylab = "Deviance $D(y_{rep}, \beta)$")
abline(0,1)

compare_prob=sum(deviance2>deviance4)/iter

B.4 Pearson Chi Square Statistic as a Measure of GOF

R Code

###this method computes the Pearson Chi Square statistic using the observed y values
###the asymptotic distribution of Pearson Chi Square statistic is Chi Square distribution
###with n(k-1)-q degrees of freedoms
### n is the number of observations, k is the number of categories, and q is the number
###of regression parameters
###If the obtained Pearson Chi Square statistic is greater than the critical values, then
###we conclude that the proposed model is a suitable fit.
###post.column.prob is the matrix using the posterior mean of beta
post.p1=pnorm(estimated.cutpoints[1]-x%*%post.mean.beta)
post.p2=pnorm(estimated.cutpoints[2]-x%*%post.mean.beta)-post.p1
post.p3=pnorm(estimated.cutpoints[3]-x%*%post.mean.beta)-post.p1-post.p2
post.p4=pnorm(estimated.cutpoints[4]-x%*%post.mean.beta)-post.p1-post.p2-post.p3
post.p5=1-post.p1-post.p2-post.p3-post.p4
post.column.prob=array(c(post.p1,post.p2,post.p3,post.p4,post.p5),c(length(grade),5))

pearson1=0
for(i in 1:length(grade))
{
  for(j in 1:k)
  {
    pearson1=pearson1+(mat.grade[i,j]-post.column.prob[i,j])^2/post.column.prob[i,j]
  }
}
B.5 Posterior Predictive Distribution of $X^2(y^{rep})$ as a Measure of GOF R Code

```r
iter=5000

pearson3=array(0,c(iter,1))

length(grade)=30 is the number of observations that we have

simulated.category=array(0,c(length(grade),5))

post.p1=pnorm(estimated.cutpoints[1]-x*%post.mean.beta)

post.p2=pnorm(estimated.cutpoints[2]-x*%post.mean.beta)-post.p1

post.p3=pnorm(estimated.cutpoints[3]-x*%post.mean.beta)-post.p1-post.p2

post.p4=pnorm(estimated.cutpoints[4]-x*%post.mean.beta)-post.p1-post.p2-post.p3

post.p5=1-post.p1-post.p2-post.p3-post.p4

post.column.prob=array(c(post.p1,post.p2,post.p3,post.p4,post.p5),c(length(post.p1),5))

for(m in 1:iter)
{
  p1=pnorm(estimated.cutpoints[1]-x*%beta[m,])
  p2=pnorm(estimated.cutpoints[2]-x*%beta[m,])-p1
  p3=pnorm(estimated.cutpoints[3]-x*%beta[m,])-p1-p2
```
p4=pnorm(estimated.cutpoints[4]-x%*%beta[m,])-p1-p2-p3
p5=1-p1-p2-p3-p4
column.prob=array(c(p1,p2,p3,p4,p5),c(length(p1),k))
for( i in 1:length(grade))
{
  simulated.category[i,]=rmultinom(1,1,column.prob[i,])
  pearson3[m,]=pearson3[m,]+sum((simulated.category[i,]-post.column.prob[i,])^2/post.column.prob[i,])
}

t.pearson3=pearson3[pearson3<=1000]
hist(t.pearson3,xlab="Pearson Chi Square Statistic Using Simulated Y and Posterior Mean of Simulated Beta",main=NULL)
abline(v=pearson1,lty="dotted",lwd=3)
tail.prob3=sum(pearson3>pearson1)/iter;tail.prob3

B.6 Using Parameter-dependent $X^2(y^{rep}, \beta)$ as a Measure of GOF R Code

###this method compare two distributions $X^2(y^{rep}, \beta)$ with $X^2(y, \beta)$

###$X^2(y^{rep}, \beta)$ using the simulated $\beta$ values and the simulated “future” observation $y^{rep}$

###$X^2(y, \beta)$ using the simulated $\beta$ values and the observed y values

###we can compare these two distribution by computing $P(X^2(y^{rep}, \beta) > X^2(y, \beta))$

###if the probability is small, then we conclude the proposed model is not a suitable fit

iter=5000

pearson2=array(0,c(iter,1));pearson4=pearson2
length(grade) = 30 is the number of observations that we have.

simulated.category = array(0, c(length(grade), 5))

for (m in 1:iter)
{
  p1 = pnorm(estimated.cutpoints[1] - x * beta[m,])
  p2 = pnorm(estimated.cutpoints[2] - x * beta[m,]) - p1
  p3 = pnorm(estimated.cutpoints[3] - x * beta[m,]) - p1 - p2
  p4 = pnorm(estimated.cutpoints[4] - x * beta[m,]) - p1 - p2 - p3
  p5 = 1 - pnorm(estimated.cutpoints[4] - x * beta[m,])

  column.prob = array(c(p1, p2, p3, p4, p5), c(length(p1), 5))

  for (i in 1:length(grade))
  {
    for (j in 1:5)
    {
      simulated.category[i,] = rmultinom(1, 1, column.prob[i,])
      
      pearson2[m] = pearson2[m] + (simulated.category[i, j] - column.prob[i, j])^2 / column.prob[i, j]
      pearson4[m] = pearson2[m] + (mat.grade[i, j] - column.prob[i, j])^2 / column.prob[i, j]
    }
  }
}

plot(pearson2, pearson4, xlab = "Pearson Chi Square Statistic Using Simulated Y and Simulated Beta", ylab = "Pearson Chi Square Statistic Using Observed Y and Simulated Beta", main = NULL)

tail.prob2 = sum(pearson2 > pearson4) / iter; tail.prob2
### B.7 Probitpost_model1_ch4_K_1 R Code

#### this program only inlcude SAT as a explanatory variable

#### this model support that sat should be included in the model.

#### instead, this one uses the informative prior to reflect the belief that we should include variable SAT

#### we totally need q+C-2=2+5-2=5 values of beta

#### use sat=520,sat=500,sat=540,sat=570 and sat=600

#### also we need to specify the prob to each sat grade

#### sat=500 prob(less than F)=0.2,k=1

#### sat=600 prob(less than F)=0.2,k=1

#### sat=525 prob(less than B)=0.75,k=1

#### sat=570 prob(less than B)=0.75,k=1

#### sat=600 prob(less than C)=0.7,k=1

#### number of support=1

`probitpost_model1_ch4=function (par,data)
{
  x = data[,1]
  n = data[,2]
  y = data[,3]
  beta0 = par[,1]
  beta1 = par[,2]
  gamma1=0
  gamma2=par[,3]
  gamma3=par[,4]
  gamma4=par[,5]
  omatrix=array(0,c(30,5))
  for( i in 1:length(y))
    omatrix=}
\begin{verbatim}
{
  omatrix[i,] = c(1,0,0,0)*(y[i]==1) + c(0,1,0,0)*(y[i]==2) + c(0,0,1,0)*(y[i]==3) +
  c(0,0,0,1)*(y[i]==4) + c(0,0,0,0)*(y[i]==5)
}

z = 0
for( i in 1:length(x))
{
  lp = beta0 + beta1 * x[i]
  p1 = pnorm(gamma1-lp,0,1)
  p2 = pnorm(gamma2-lp,0,1)-pnorm(gamma1-lp,0,1)
  p3 = pnorm(gamma3-lp,0,1)-pnorm(gamma2-lp,0,1)
  p4 = pnorm(gamma4-lp,0,1)-pnorm(gamma3-lp,0,1)
  p5 = 1-pnorm(gamma4-lp,0,1)
  part1 = omatrix[i,][1]*log(p1)
  part2 = omatrix[i,][2]*log(p2)
  part3 = omatrix[i,][3]*log(p3)
  part4 = omatrix[i,][4]*log(p4)
  part5 = omatrix[i,][5]*log(p5)
  z = z + part1 + part2 + part3 + part4 + part5
}

logprior.part1 = (1*0.2-1)*log(pnorm(gamma2-beta0-beta1*500,0,1)) +
  (1*(1-0.2)-1)*log(1-pnorm(gamma2-beta0-beta1*500,0,1)) +
  log(dnorm(gamma2-beta0-beta1*500,0,1))-log(beta(1*0.2,1*(1-0.2)))
logprior.part2 = (1*0.2-1)*log(pnorm(gamma2-beta0-beta1*600,0,1)) +
  (1*(1-0.2)-1)*log(1-pnorm(gamma2-beta0-beta1*600,0,1)) +
  log(dnorm(gamma2-beta0-beta1*600,0,1))-log(beta(1*0.2,1*(1-0.2)))
logprior.part3 = (1*0.75-1)*log(pnorm(gamma4-beta0-beta1*525,0,1)) +
  (1*(1-0.75)-1)*log(1-pnorm(gamma4-beta0-beta1*525,0,1)) +
  log(dnorm(gamma4-beta0-beta1*525,0,1))-log(beta(1*0.75,1*(1-0.75)))
logprior.part4 = (1*0.5-1)*log(pnorm(gamma5-beta0-beta1*750,0,1)) +
  (1*(1-0.5)-1)*log(1-pnorm(gamma5-beta0-beta1*750,0,1)) +
  log(dnorm(gamma5-beta0-beta1*750,0,1))-log(beta(1*0.5,1*(1-0.5)))
logprior.part5 = (1*0.75-1)*log(pnorm(gamma6-beta0-beta1*550,0,1)) +
  (1*(1-0.75)-1)*log(1-pnorm(gamma6-beta0-beta1*550,0,1)) +
  log(dnorm(gamma6-beta0-beta1*550,0,1))-log(beta(1*0.75,1*(1-0.75)))
\end{verbatim}
\[
(1*(1-0.75)-1)*\log(1-\text{pnorm}(\gamma_4-\beta_0 -\beta_1*525,0,1))+ \\
\log(\text{dnorm}(\gamma_4-\beta_0 -\beta_1*525,0,1))-\log(\beta(1*0.75,1*(1-0.75))) \\
\log\text{prior}.part4=(1*0.75-1)*\log(\text{pnorm}(\gamma_4-\beta_0 -\beta_1*570,0,1))+ \\
(1*(1-0.75)-1)*\log(1-\text{pnorm}(\gamma_4-\beta_0 -\beta_1*570,0,1))+ \\
\log(\text{dnorm}(\gamma_4-\beta_0 -\beta_1*570,0,1))-\log(\beta(1*0.75,1*(1-0.75))) \\
\log\text{prior}.part5=(1*0.7-1)*\log(\text{pnorm}(\gamma_3-\beta_0 -\beta_1*600,0,1))+ \\
(1*(1-0.7)-1)*\log(1-\text{pnorm}(\gamma_3-\beta_0 -\beta_1*600,0,1))+ \\
\log(\text{dnorm}(\gamma_3-\beta_0 -\beta_1*600,0,1))-\log(\beta(1*0.7,1*(1-0.7))) \\
\log\text{prior}=\log\text{prior}.part1+\log\text{prior}.part2+\log\text{prior}.part3+\log\text{prior}.part4+\log\text{prior}.part5 \\
\text{post}=\log\text{prior}+z \\
\text{return}(\text{post})
\]

### B.8 Probit_model1_compare_ch4_K_100 R program

#### this program only include sat as a explanatory variable
#### this model support that sat should be removed from the model
#### this one is not using the flat prior
#### instead, this one use the informative prior
#### we totally need q+C-2=2+5-2=5 values of beta
#### use sat=520,sat=500,sat=540,sat=570 and sat=600,
#### also we need to specify the prob to each sat grade
#### sat=500 prob(less than D)=0.2,k=100
#### sat=600 prob(less than D)=0.2,k=100
#### sat=525 prob(less than B)=0.75,k=100
#### sat=575 prob(less than B)=0.75,k=100
#### sat=600 prob(less than C)=0.7,k=100
probit_modell_compare_ch4_K_100=function(par,data)
{
  x = data[, 1]
  n = data[, 2]
  y = data[, 3]
  beta0 =par[, 1]
  beta1 =par[, 2]
  gamma1=0
  gamma2=par[,3]
  gamma3=par[,4]
  gamma4=par[,5]
  omatrix=array(0,c(30,5))
  for( i in 1:length(y))
  {
    omatrix[i,]=c(1,0,0,0,0)*(y[i]==1)+c(0,1,0,0,0)*(y[i]==2)+c(0,0,1,0,0)*(y[i]==3)+
    +c(0,0,0,1,0)*(y[i]==4)+c(0,0,0,0,1)*(y[i]==5)
  }
  z=0
  for( i in 1:length(x))
  {
    lp = beta0 + beta1 * x[i]
    p1=pnorm(gamma1-lp,0,1)
    p2=pnorm(gamma2-lp,0,1)-pnorm(gamma1-lp,0,1)
    p3=pnorm(gamma3-lp,0,1)-pnorm(gamma2-lp,0,1)
    p4=pnorm(gamma4-lp,0,1)-pnorm(gamma3-lp,0,1)
    p5=1-pnorm(gamma4-lp,0,1)
    part1=omatrix[i,][1]*log(p1)
part2=omatrix[i,][2]*log(p2)
part3=omatrix[i,][3]*log(p3)
part4=omatrix[i,][4]*log(p4)
part5=omatrix[i,][5]*log(p5)

z=z+part1+part2+part3+part4+part5

}\nlogprior.part1=(100*0.2-1)*log(pnorm(gamma2-beta0-beta1*500,0,1))+
(100*(1-0.2)-1)*log(1-pnorm(gamma2-beta0-beta1*500,0,1))+
log(dnorm(gamma2-beta0-beta1*500,0,1)))-log(beta(100*0.2,100*(1-0.2)))

logprior.part2=(100*0.2-1)*log(pnorm(gamma2-beta0-beta1*600,0,1))+
(100*(1-0.2)-1)*log(1-pnorm(gamma2-beta0-beta1*600,0,1))+
log(dnorm(gamma2-beta0-beta1*600,0,1)))-log(beta(100*0.2,100*(1-0.2)))

logprior.part3=(100*0.75-1)*log(pnorm(gamma4-beta0-beta1*525,0,1))+
(100*(1-0.75)-1)*log(1-pnorm(gamma4-beta0-beta1*525,0,1))+
log(dnorm(gamma4-beta0-beta1*525,0,1)))-log(beta(100*0.75,100*(1-0.75)))

logprior.part4=(100*0.75-1)*log(pnorm(gamma4-beta0-beta1*570,0,1))+
(100*(1-0.75)-1)*log(1-pnorm(gamma4-beta0-beta1*570,0,1))+
log(dnorm(gamma4-beta0-beta1*570,0,1)))-log(beta(100*0.75,100*(1-0.75)))

logprior.part5=(100*0.7-1)*log(pnorm(gamma3-beta0-beta1*600,0,1))+
(100*(1-0.7)-1)*log(1-pnorm(gamma3-beta0-beta1*600,0,1))+
log(dnorm(gamma3-beta0-beta1*600,0,1)))-log(beta(100*0.7,100*(1-0.7)))

logprior=logprior.part1+logprior.part2+logprior.part3+logprior.part4+logprior.part5

post=logprior+z

return(post)
### B.9 Probitpost_model2_ch4 R program

```r
#### this model use sat and pregrade as variable
#### this model support that pregrade should be included from the model
#### this one is not using the flat prior
#### instead, this one use the informative prior
#### we totally need q+C-2=3+5-2=6 values of beta
#### use sat=450,sat=520,sat=500,sat=540,sat=570 and sat=600
#### also we need to specify the prob to each sat grade
#### sat=500, pregrade=D(2), prob(less than D)=0.6, k=1
#### sat=500, pregrade=A(5), prob(less than D)=0.6, k=1
#### sat=550, pregrade=D(2), prob(less than C)=0.7, k=1
#### sat=550, pregrade=A(5), prob(less than C)=0.7, k=1
#### sat=600, pregrade=D(2), prob(less than B)=0.8, k=1
#### sat=600, pregrade=A(5), prob(less than B)=0.8, k=1
#### par should be a vector of 6*1
#### data should include Sat, Constant, Grade, Pregrade and keep in that format.

```R
probitpost_model2_ch4=function (par, data)
{
    x = data[, 1]
    n = data[, 2]
    y = data[, 3]
    pregrade=data[,4]
    beta0 = par[, 1]
    beta1 = par[, 2]
    beta2=par[,3]
    gamma1=0
    gamma2=par[,4]
```
gamma3=par[,5]
gamma4=par[,6]

omatrix=array(0,c(30,5))

for( i in 1:length(y))
{
  omatrix[i,]=c(1,0,0,0,0)*(y[i]==1)+c(0,1,0,0,0)*(y[i]==2)+c(0,0,1,0,0)*(y[i]==3)+
  +c(0,0,0,1,0)*(y[i]==4)+c(0,0,0,0,1)*(y[i]==5)
}
z=0

for( i in 1:length(x))
{
  lp = beta0 + beta1 * x[i]+beta2*pregrade[i]
  p1=pnorm(gamma1-lp,0,1)
  p2=pnorm(gamma2-lp,0,1)-pnorm(gamma1-lp,0,1)
  p3=pnorm(gamma3-lp,0,1)-pnorm(gamma2-lp,0,1)
  p4=pnorm(gamma4-lp,0,1)-pnorm(gamma3-lp,0,1)
  p5=1-pnorm(gamma4-lp,0,1)
  part1=omatrix[i,][1]*log(p1)
  part2=omatrix[i,][2]*log(p2)
  part3=omatrix[i,][3]*log(p3)
  part4=omatrix[i,][4]*log(p4)
  part5=omatrix[i,][5]*log(p5)
  z=z+part1+part2+part3+part4+part5
}

logprior.part1=(1*0.6-1)*log(pnorm(gamma2-beta0 -beta1*500-2*beta2,0,1)) +
  (1*(1-0.6)-1)*log(1-pnorm(gamma2-beta0 -beta1*500-2*beta2,0,1)) +
  log(dnorm(gamma2-beta0 -beta1*500-2*beta2,0,1))-log(beta(1*0.6,1*(1-0.6)))

logprior.part2=(1*0.6-1)*log(pnorm(gamma2-beta0 -beta1*500-5*beta2,0,1)) +
B.10 Probit_model2_compare_ch4_K_1000 R program

### this model use sat and pregrade as variable
### this one support that pregrade should NOT be include in the model
### this one is not using the flat prior
### instead, this one use the informative prior
### we totally need q+C-2=3+5-2=6 values of beta
### use sat=500, sat=550, sat=550, sat=540, sat=500 and sat=600
### also we need to specify the prob to each sat grade
### sat=500, pregrade=D(2), prob(less than D)=0.6, K=1000
### sat=500, pregrade=A(4), prob(less than D)=0.6, k=1000
### sat=550, pregrade=D(2), prob(less than C)=0.7, k=1000
### sat=550, pregrade=A(5), prob(less than C)=0.7, k=1000
### sat=600, pregrade=D(2), prob(less than B)=0.8, k=1000
### sat=600, pregrade=A(5), prob(less than B)=0.8, k=1000
### par should be a vector of 6*1
### data should include Sat, Constant, Grade, Pregrade and keep in that format.

```r
probit_model2_compare_ch4_K_1000=function (par, data)
{
  x = data[, 1]
  n = data[, 2]
  y = data[, 3]
  pregrade=data[,4]
  beta0 = par[, 1]
  beta1 = par[, 2]
  beta2=par[,3]
  gamma1=0
  gamma2=par[,4]
  gamma3=par[,5]
  gamma4=par[,6]
  omatrix=array(0,c(30,5))
  for( i in 1:length(y))
  {
    omatrix[i,]=c(1,0,0,0,0)*(y[i]==1)+c(0,1,0,0,0)*(y[i]==2)+c(0,0,1,0,0)*(y[i]==3)+
  ```
\( +c(0,0,0,1,0) \cdot (y[i] == 4) + c(0,0,0,0,1) \cdot (y[i] == 5) \)

\[
\]
\( z = 0 \)

\[
\text{for( i in 1:length(x))}
\]

\[
\}
\]
\( l p = \beta_0 + \beta_1 \cdot x[i] + \beta_2 \cdot \text{pregrade}[i] \)
\( p1 = \text{pnorm}(\gamma_1 - lp, 0, 1) \)
\( p2 = \text{pnorm}(\gamma_2 - lp, 0, 1) - \text{pnorm}(\gamma_1 - lp, 0, 1) \)
\( p3 = \text{pnorm}(\gamma_3 - lp, 0, 1) - \text{pnorm}(\gamma_2 - lp, 0, 1) \)
\( p4 = \text{pnorm}(\gamma_4 - lp, 0, 1) - \text{pnorm}(\gamma_3 - lp, 0, 1) \)
\( p5 = 1 - \text{pnorm}(\gamma_4 - lp, 0, 1) \)
\( \text{part1} = \text{omatrix}[i,][1] \cdot \log(p1) \)
\( \text{part2} = \text{omatrix}[i,][2] \cdot \log(p2) \)
\( \text{part3} = \text{omatrix}[i,][3] \cdot \log(p3) \)
\( \text{part4} = \text{omatrix}[i,][4] \cdot \log(p4) \)
\( \text{part5} = \text{omatrix}[i,][5] \cdot \log(p5) \)
\( z = z + \text{part1 + part2 + part3 + part4 + part5} \)

logprior.\text{part1} = (1000*0.6-1) \cdot \log(\text{pnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 2 \cdot \beta_2, 0, 1)) +
(1000*(1-0.6)-1) \cdot \log(1-\text{pnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 2 \cdot \beta_2, 0, 1)) +
\log(\text{dnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 2 \cdot \beta_2, 0, 1)) - \log(\text{beta}(1000*0.6,1000*(1-0.6)))

logprior.\text{part2} = (1000*0.6-1) \cdot \log(\text{pnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 5 \cdot \beta_2, 0, 1)) +
(1000*(1-0.6)-1) \cdot \log(1-\text{pnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 5 \cdot \beta_2, 0, 1)) +
\log(\text{dnorm}(\gamma_2 - \beta_0 - \beta_1 \cdot 500 - 5 \cdot \beta_2, 0, 1)) - \log(\text{beta}(1000*0.6,1000*(1-0.6)))

logprior.\text{part3} = (1000*0.7-1) \cdot \log(\text{pnorm}(\gamma_3 - \beta_0 - \beta_1 \cdot 550 - 2 \cdot \beta_2, 0, 1)) +
(1000*(1-0.7)-1) \cdot \log(1-\text{pnorm}(\gamma_3 - \beta_0 - \beta_1 \cdot 550 - 2 \cdot \beta_2, 0, 1)) +
\log(\text{dnorm}(\gamma_3 - \beta_0 - \beta_1 \cdot 550 - 2 \cdot \beta_2, 0, 1)) - \log(\text{beta}(1000*0.7,1000*(1-0.7)))
logprior.part4=(1000*0.7-1)*log(pnorm(gamma3-beta0-b*550-b*5*beta2,0,1))+
(1000*(1-0.7)-1)*log(1-pnorm(gamma3-beta0-b*550-b*5*beta2,0,1))+
log(dnorm(gamma3-beta0-b*550-b*5*beta2,0,1))-log(beta(1000*0.7,1000*(1-0.7)))
logprior.part5=(1000*0.8-1)*log(pnorm(gamma4-beta0-b*600-b*2*beta2,0,1))+
(1000*(1-0.8)-1)*log(1-pnorm(gamma4-beta0-b*600-b*2*beta2,0,1))+
log(dnorm(gamma4-beta0-b*600-b*2*beta2,0,1))-log(beta(1000*0.8,1000*(1-0.8)))
logprior.part6=(1000*0.8-1)*log(pnorm(gamma4-beta0-b*600-b*5*beta2,0,1))+
(1000*(1-0.8)-1)*log(1-pnorm(gamma4-beta0-b*600-b*5*beta2,0,1))+
log(dnorm(gamma4-beta0-b*600-b*5*beta2,0,1))-log(beta(1000*0.8,1000*(1-0.8)))
logprior=logprior.part1+logprior.part2+logprior.part3+logprior.part4+logprior.part5+
logprior.part6
post=logprior+z
return(post)

**B.11 Subspace_pgrade R Code**

subspace_pgrade=function(data,iter,c,K)
{

####this one use the projection of pgrade to decide whether variable pgrade should be
####included or not
####data is the original data set
####iter is the number of iterations
####probability that obs [i] falls in category [j] is given by p[i][j]=pnorm(gamma[j+1]-
x[i]*beta)
####in order to simulate from the posterior distribution of delta2, we have to use random walk

```r
```
### so I introduce the new variable \( \eta \), which is the log of \( \delta^2 \).

### \( c \) is the number that I use in the random walk.

### we make the transformation \( Z^* = \gamma z \), \( Beta^* = \gamma \beta \), and \( \gamma = 1 / \text{last cutoff point} \)

\[
\beta = \begin{pmatrix} -2.37204570, & 0.04509203, & 0.67195786 \end{pmatrix}
\]

\[
y = \text{data[,1]}
\]

\[
x = \text{data[-1]}
\]

\[
nk = \text{dim(x)}
\]

\[
n = nk[1]; k = nk[2]
\]

### \( n \) is the number of observations, \( k \) is the number of \( \beta \)'s. \( ct = \text{table(y)} \)

\[
\alpha = 0.2
\]

\[
\gamma = \begin{pmatrix} 0.2, & 0.6 \end{pmatrix}
\]

\[
cpoints = \begin{pmatrix} -\infty, & 0, & \gamma, & 1, & \infty \end{pmatrix}
\]

\[
lo = cpoints[y]; hi = cpoints[y+1]
\]

\[
z = \begin{pmatrix} -0.5, & 0, & \gamma, & 1 \end{pmatrix}
\]

\[
\delta^2 = 0.156
\]

### \( \delta \) is 1 over last cutoff point

\[
M\delta^2 = \text{array}(\delta^2, \text{c}(1, \text{iter}+1))
\]

\[
M\beta = \text{array}(\beta, \text{c}(\text{length}(\beta), \text{iter}+1))
\]

\[
Mz = \text{array}(1, \text{c}(n, \text{iter}+1))
\]

### the variance for \( \beta \) on the second stage.

\[
M\sigma^2 = \text{array}(0.01, \text{c}(1, \text{iter}+1))
\]

### \( M\gamma \) is the values of the two unknown cutoff points

\[
M\gamma = \text{array}(\gamma, \text{c}(\text{iter}+1, \text{length}(\gamma)))
\]

\[
\text{accept.count} = 0
\]

\[
\text{total.count} = 0
\]

for( i in 1:iter)
I1.mat = diag(rep(1,n))
I1.mat.star = Mdelta2[i]*I1.mat
A0 = array(c(1,0,0,0,1,0),c(3,2))
A0.star = sqrt(1)*array(c(1,0,0,0,1,0),c(3,2))
Msigma2.star = Msigma2[i]*Mdelta2[i]
eta = log(Msigma2.star)

I2.mat = diag(rep(1,3))
theta1.hat = solve(t(x)*x)*Mz[,i]
theta2.hat.part1 = t(x*A0.star)*solve(I1.mat.star + x*t(x)*Msigma2.star)*x*A0.star
theta2.hat.part2 = t(x*A0.star)*solve(I1.mat.star + x*t(x)*Msigma2.star)
theta2.hat = solve(theta2.hat.part1)*theta2.hat.part2*Mz[,i]
w1 = solve(t(x)**x + 1/(Msigma2.star)*I2.mat)**x
v.part1 = solve(t(x*A0.star)**x + Msigma2.star)**x*A0.star
v.part2 = solve(t(x)**x + 1/Msigma2.star)*I2.mat
v = ((I2.mat-w1)**A0.star)**v.part1**t((I2.mat-w1)**A0.star) + v.part2
u = w1**theta1.hat + (I2.mat-w1)**A0.star**theta2.hat
Mbeta[,i+1] = rmnorm(1,u,v)
### end of simulation of beta* values

#### this part simulate Z from the truncated normal with mean xb* and variance 1.

```r
mu = x %*% Mbeta[, i]; sigma = sqrt(Mdelta2[i])
prob = pnorm(cbind((lo - mu)/sigma, (hi - mu)/sigma))
Mz[, i + 1] = mu + sigma * qnorm(runif(length(lo)) * (prob[, 2] - prob[, 1]) + prob[, 1])
```

#### simulation of the cutoff points

```r
if (K > 3)
{
  lf = sum(log(prob[, 2] - prob[, 1]))
  p = diff(c(0, gam, 1))
  lp = sum((alpha * ct[2:(K - 1)] - 1) * log(p))
  p1 = rgamma(K - 2, shape = alpha * ct[2:(K - 1)], rate = 1)
  p1 = p1 / sum(p1)
  lp1 = sum((alpha * ct[2:(K - 1)] - 1) * log(p1))
  gam1 = cumsum(p1)
  cpoints1 = c(-Inf, 0, gam1, Inf)
  lo1 = cpoints1[y]; hi1 = cpoints1[y + 1]
  prob1 = pnorm(cbind((lo1 - mu)/sigma, (hi1 - mu)/sigma))
  lf1 = sum(log(prob1[, 2] - prob1[, 1]))
  test.prob = exp((lf1 - lp1) - (lf - lp))
  if (runif(1) < test.prob)
  {
    gam = gam1[1:(length(gam1) - 1)]
    lo = lo1; hi = hi1
  }
}
```

if (K >= 4)
Mgam[i+1,] = gam

### this part simulate delta2 using random walk algorithm.

### here I define a function of L1, which is the likelihood function.

### notice that beta is related to sigma2, Z is related to sigma2, u is also related to sigma2.

### So I have to include all of them into this likelihood function.

### start of L1 function

likelihood = function(eta)
{
  theta2.hat.part1 = t(x%^%A0.star).solve(I1.mat.star + x%^%t(x)*exp(eta))
  theta2.hat.part2 = t(x%^%A0.star).solve(I1.mat.star + x%^%t(x)*exp(eta))
  theta2.hat = solve(theta2.hat.part1) %*% theta2.hat.part2 %*% Mz[,i]
  part1 = det(solve(I1.mat.star + x%^%t(x)*exp(eta)))
  part2 = det(t(x%^%A0.star).solve(I1.mat.star + x%^%t(x)*exp(eta)))
  part3 = exp(-1/2*t(Mz[,i] - x%^%A0.star %*% theta2.hat).solve(I1.mat.star + x%^%t(x)*exp(eta)) %*% (Mz[,i] - x%^%A0.star %*% theta2.hat))
  return(sqrt(part1*part2)*part3)
}

### end of L1 function###

eta.sim = eta + c*rnorm(1,0,1)

ratio = likelihood(eta.sim)/likelihood(eta)

t = runif(1)

eta.new = eta.sim*(t < ratio) + eta*(t >= ratio)

accept.count = accept.count + (eta.new == eta.sim)

total.count = total.count + 1

Msigma2[i+1] = exp(eta.new)/Mdelta2[i]
### the simulation of delta square

```r
rate2 = sum((Mz[i] - x * Mbeta[i])^2) / 2
Mdelta2[i+1] = rate2/rgamma(1, shape=shape2, rate=1)
```

accept_rate = accept.count / total.count

return(list(Mdelta2 = Mdelta2, Mbeta = Mbeta, Mz = Mz, Mgam = Mgam,
accept_rate = accept_rate, Msigma2 = Msigma2))

```
Appendix C

Chapter 5 Data Set

C.1 Chapter 5 Data Set
<table>
<thead>
<tr>
<th>studentID</th>
<th>grade</th>
<th>act</th>
<th>placement</th>
<th>pgrade</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>22</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>19</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>20</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>23</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>25</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>15</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>18</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>24</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>24</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>21</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>19</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>25</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>18</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>25</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>20</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
<td>24</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>19</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
<td>25</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>22</td>
<td>31</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>17</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>19</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>17</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>17</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>21</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>26</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>22</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>24</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>29</td>
<td>4</td>
<td>21</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>22</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>20</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>20</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>21</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>5</td>
<td>23</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>19</td>
<td>32</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>studentID</th>
<th>grade</th>
<th>act</th>
<th>placement</th>
<th>pgrade</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>4</td>
<td>21</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>37</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>22</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>4</td>
<td>26</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>22</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
<td>19</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>42</td>
<td>3</td>
<td>19</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>43</td>
<td>5</td>
<td>17</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>44</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>20</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>46</td>
<td>1</td>
<td>22</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>48</td>
<td>3</td>
<td>17</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>49</td>
<td>3</td>
<td>19</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>23</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>51</td>
<td>2</td>
<td>17</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>52</td>
<td>1</td>
<td>16</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>53</td>
<td>5</td>
<td>26</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>54</td>
<td>4</td>
<td>22</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>55</td>
<td>3</td>
<td>23</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>56</td>
<td>4</td>
<td>24</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>57</td>
<td>1</td>
<td>18</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>58</td>
<td>1</td>
<td>20</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>59</td>
<td>5</td>
<td>25</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>60</td>
<td>4</td>
<td>19</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>61</td>
<td>4</td>
<td>24</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>62</td>
<td>4</td>
<td>24</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>63</td>
<td>3</td>
<td>25</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>21</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>65</td>
<td>5</td>
<td>23</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>66</td>
<td>4</td>
<td>20</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>67</td>
<td>2</td>
<td>24</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>68</td>
<td>4</td>
<td>24</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>69</td>
<td>3</td>
<td>15</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>70</td>
<td>5</td>
<td>30</td>
<td>61</td>
<td>5</td>
</tr>
<tr>
<td>studentID</td>
<td>grade</td>
<td>act</td>
<td>placement</td>
<td>pgrade</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-----</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>71</td>
<td>4</td>
<td>25</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>72</td>
<td>4</td>
<td>20</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>73</td>
<td>3</td>
<td>22</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>74</td>
<td>3</td>
<td>24</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
<td>19</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>76</td>
<td>4</td>
<td>21</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>77</td>
<td>4</td>
<td>21</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>78</td>
<td>4</td>
<td>16</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>79</td>
<td>5</td>
<td>19</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>16</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>81</td>
<td>4</td>
<td>20</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>82</td>
<td>5</td>
<td>20</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>83</td>
<td>4</td>
<td>21</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>84</td>
<td>4</td>
<td>19</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>85</td>
<td>4</td>
<td>28</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>86</td>
<td>4</td>
<td>21</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>87</td>
<td>4</td>
<td>21</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>88</td>
<td>1</td>
<td>19</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>89</td>
<td>5</td>
<td>21</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>90</td>
<td>5</td>
<td>23</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>91</td>
<td>3</td>
<td>16</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>92</td>
<td>4</td>
<td>17</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>93</td>
<td>4</td>
<td>21</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>94</td>
<td>4</td>
<td>16</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>95</td>
<td>3</td>
<td>20</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>96</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>97</td>
<td>3</td>
<td>19</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>98</td>
<td>4</td>
<td>16</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>99</td>
<td>5</td>
<td>23</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
<td>23</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>101</td>
<td>1</td>
<td>22</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>102</td>
<td>4</td>
<td>18</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>103</td>
<td>1</td>
<td>26</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>104</td>
<td>4</td>
<td>21</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>105</td>
<td>3</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>studentID</td>
<td>grade</td>
<td>act</td>
<td>placement</td>
<td>pgrade</td>
</tr>
<tr>
<td>-----------</td>
<td>-------</td>
<td>-----</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>141</td>
<td>2</td>
<td>21</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>142</td>
<td>3</td>
<td>20</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>143</td>
<td>4</td>
<td>20</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>144</td>
<td>3</td>
<td>18</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>145</td>
<td>5</td>
<td>18</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>146</td>
<td>5</td>
<td>18</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>147</td>
<td>4</td>
<td>20</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>148</td>
<td>3</td>
<td>15</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>149</td>
<td>4</td>
<td>24</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>150</td>
<td>1</td>
<td>19</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>151</td>
<td>3</td>
<td>17</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>152</td>
<td>1</td>
<td>17</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>153</td>
<td>4</td>
<td>20</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>154</td>
<td>3</td>
<td>18</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>155</td>
<td>1</td>
<td>22</td>
<td>42</td>
<td>3</td>
</tr>
<tr>
<td>156</td>
<td>3</td>
<td>18</td>
<td>27</td>
<td>3</td>
</tr>
<tr>
<td>157</td>
<td>3</td>
<td>21</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>158</td>
<td>2</td>
<td>17</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>159</td>
<td>3</td>
<td>19</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>160</td>
<td>4</td>
<td>24</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>161</td>
<td>4</td>
<td>20</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>162</td>
<td>4</td>
<td>24</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>163</td>
<td>4</td>
<td>26</td>
<td>42</td>
<td>4</td>
</tr>
<tr>
<td>164</td>
<td>4</td>
<td>17</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>165</td>
<td>4</td>
<td>16</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>166</td>
<td>4</td>
<td>26</td>
<td>41</td>
<td>4</td>
</tr>
<tr>
<td>167</td>
<td>3</td>
<td>18</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>168</td>
<td>1</td>
<td>16</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>169</td>
<td>2</td>
<td>17</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>170</td>
<td>4</td>
<td>20</td>
<td>41</td>
<td>5</td>
</tr>
<tr>
<td>171</td>
<td>1</td>
<td>20</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>172</td>
<td>5</td>
<td>23</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>173</td>
<td>4</td>
<td>24</td>
<td>32</td>
<td>4</td>
</tr>
<tr>
<td>174</td>
<td>4</td>
<td>21</td>
<td>27</td>
<td>3</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY


