A SURVEY ON GENERALIZED LINEAR MODELS (GLMS)

AND THEIR DIAGNOSTIC TOOLS

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

FAHAD ABDULAZIZ ALSIDRANI

DR. MUNNI BEGUM - ADVISOR

BALL STATE UNIVERSITY
MUNCIE, INDIANA
MAY 2017

A SURVEY ON GENERALIZED LINEAR MODELS (GLMS)

AND THEIR DIAGNOSTIC TOOLS

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

FAHAD ABDULAZIZ ALSIDRANI

Committee Approval:

Committee Chairperson	Date
Committee Member	Date
Committee Member	Date

Departmental Approval:

Departmental Chairperson	Date
Dean of Graduate School	Date

BALL STATE UNIVERSITY
MUNCIE, INDIANA
MAY 2017

ACKNOWLEDGMENTS

I would like to thank my supervisor Professor Munni Begum for her support and many thoughtful suggestions throughout this research. I also want to extend my gratitude to the other members of my committee, Dr. Rahmatullah Imon and Dr. Richard Stankewitz. I extend special thanks to all my professors in the Department of Mathematical Sciences. I am also grateful to the Department of Mathematical Sciences at Ball State University for the support during my study. I appreciate your patience, flexibility, wisdom, and kindness. Last but not the least, my deepest gratitude goes to my parents and brothers for their prayers and encouragement, and to my friends whom I have meet in Muncie for their love, patience, and sacrifice during my study. I am proud to be a product of the love and support that you have freely given to me over the years.

ABSTRACT

An important statistical development in the last four decades has been the advancement in the field of regression analysis for a variety of response variables under the umbrella of an exponential family of distributions. The basic idea is to use a generalized linear modeling scheme for the response variables distributed as an exponential family of distributions. Regression analysis for a single response is conceptually a simple method for investigating functional relationships among the response and a number of predictor variables. The general linear regression model requires that the response variable follows the normal distribution whereas the generalized linear regression models is an extension of the general linear model which allows the response variables to be non-normal within the exponential family of distributions.

An important aspect of any regression analysis is the diagnostic tools for evaluating whether and to what extent the assumption for the regression models are valid. Diagnostics tools are well studied and understood for linear regression models. In this research, a survey on the diagnostic tools for generalized linear regression models such as, Poisson regression, Logistic regression and Multinomial regression are conducted. Also, numerical examples are provided to illustrate these tools for Poisson regression, Logistic regression and Multinomial regression.

Table of Contents

List of Tables

CHAPTER 5

CHAPTER 1

Introduction

1.1 Background

The term Generalized Linear Models (GLMs) was first introduced by Nelder and Wedderburn in 1972. Generalized Linear Models have become a commonly used tool of data analysis. Regression analysis is probably the most popular and commonly used statistics in all branches of knowledge. Regression analysis can be traced back to Sir Francis Galton who observed that children's heights tend to revert to the average height of the population rather than diverting from it. It is a conceptually simple method for investigating functional relationships among variables. The user of regression analysis attempts to discern the relationship between a dependent (response) variable and one or more independent (explanatory/ predictor/ regressor) variables. It can be used to predict the value of a response variable from knowledge of the values of one or more explanatory variables. Most application of regression analysis involve the use of more than one explanatory variable. Regression models that employ more than one independent variables are called multiple regression models. When all parameters enter the equation linearly it is called multiple linear regression model. A simple linear regression model with a single explanatory variable, also called covariate, can be written as:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.1}$$

where y_i is the response variable and x_i is the predictor variable, β_0, β_1 are unknown parameters and ε_i is the error term. A simple linear regression model depicts the relationships between the two continuous variables y and x. The simple linear regression gets its adjective simple because it concerns the study of only one predictor variable. Note that for n observations the model (1.1) can be written as follows:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1.2}$$

for $i = 1,2, \dots, n$.

The multiple linear regression gets its adjective multiple because that the response value y is often influenced by two predictor variables or more. A multiple linear regression model with two or more covariates can be written as:

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{1.3}$$

where Y_i is the response variable, $\beta_1, \beta_2, \dots, \beta_k$ are a set of unknown parameters, x_i for $i = 1, 2, \dots, k$ is the i^{th} explanatory variable, and ε_i is the random error term. Usually, we assume that the random error term is distributed normally with mean zero and some variance, and hence, we have the expected value of the response Y_i, i.e., $E(Y_i)$ in the linear regression model written as:

$$E(Y_i) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} . \tag{1.4}$$

However, the generalized linear regression models of each response variable Y_i and a set of explanatory variables $x_{i1}, \dots, x_{ik}$ can be written as:

$$g[E(Y_i)] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} . \tag{1.5}$$

So then, for the responses $Y_1, \dots, Y_N$, and a set of explanatory variables $x_{i1}, \dots, x_{ik}$, the expression (1.5) can be written in a matrix notation as:

$$g[E(\boldsymbol{y})] = \boldsymbol{X\beta} \tag{1.6}$$

where $\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$ is a vector responses, $\boldsymbol{X}$ is a matrix whose elements are constants representing the set of explanatory variables, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$ is a vector of unknown parameters, and g is the link function. There are many other response variables which do not have a normal distribution. Examples of such variables are binary response (success/failure), count response, ordinal response,

and nominal response. The variables belong to a more general and larger family of distributions known as the exponential family. Suppose that we have the observations $y_1, \dots, y_n$, which represent the response values of a random sample of size n, and that follow an exponential family distribution. Then its probability density function $f(y_i, \theta_i)$ can be expressed in the form:

$$f(y_i, \theta_i) = exp[y_i\theta_i + b(\theta_i) + c(y_i)] \tag{1.7}$$

where b and c are some known functions.

We discuss four different response variables from the exponential family as follows:

1- Binary Response:

The binomial probability distribution can be written in the form of an exponential family as follows:

$$\begin{aligned} f(y_i, \theta_i) &= \binom{n_i}{y_i} {p_i}^{y_i}(1-p_i)^{n_i - y_i} \\ &= exp\left[ln\binom{n_i}{y_i} + ln({p_i}^{y_i}) + ln((1-p_i)^{n_i-y_i})\right] \\ &= exp\left[ln\binom{n_i}{y_i} + y_i\, ln(p_i) + (n_i - y_i) ln\,(1-p_i)\right] \\ &= exp\left[ln\binom{n_i}{y_i} + y_i\, ln(p_i) + n_i\, ln(1-p_i) - y_i ln\,(1-p_i)\right] \\ &= exp\left[y_i ln\left(\frac{p_i}{1-p_i}\right) + n_i\, ln(1-p_i) + ln\binom{n_i}{y_i}\right] \\ &= exp[y_i\theta_i + b(\theta_i) + c(y_i)] \end{aligned} \tag{1.8}$$

where $\theta_i = ln\left(\frac{p_i}{1-p_i}\right)$, $b(\theta_i) = n_i\, ln(1-p_i)$ and $c(y_i) = ln\binom{n_i}{y_i}$.

2- Count Response:

Let y_i be any random variables representing event counts, and let μ_i be the probability of the i^{th} event occurring in any given trial. The Poisson distribution can be written in the form of an exponential family as follows:

$$f(y_i, \mu_i) = \frac{{\mu_i}^{y_i}\, e^{-\mu_i}}{y_i!}$$

$$= exp\left[ln\left(\frac{{\mu_i}^{y_i}\, e^{-\mu_i}}{y_i!}\right)\right]$$

$$= exp[ln({\mu_i}^{y_i}\, e^{-\mu_i}) - ln\,(y_i!)]$$

$$= exp[ln({\mu_i}^{y_i}) + ln(e^{-\mu_i}) - ln\,(y_i!)]$$

$$= exp[y_i\, ln(\mu_i) - \mu_i - ln\,(y_i!)] \qquad (1.9)$$

$$= exp[y_i\theta_i + b(\theta_i) + c(y_i)]$$

where $\theta_i = ln(\mu_i)$, $b(\theta_i) = -\mu_i$ and $c(y_i) = -ln\,(y_i!)$.

3- Ordinal and Nominal Response:

Consider the $k+1$ cell multinomial distribution with cell probabilities $p_1, \dots, p_k, p_{k+1} = 1 - \sum_{i=1}^{k} p_i$. Let $Y = (y_1, \dots, y_k)$ be a collection of integer-valued random variables representing event counts, where y_k represents the count of number of times the k^{th} event occurs in a set of M independent trials. Let θ_i be the probability of the i^{th}event occurring in any given trial with $\sum_{i=1}^{k} y_i \leq M$. Then the multinomial distribution can be written in the form of exponential family as follows:

$$f(y_i, \theta_i) = \frac{M!}{(\prod_{i=1}^{k} y_i!)\,(M - \sum_{i=1}^{k} y_i)!} \prod_{i=1}^{k} {p_i}^{y_i}\ (1 - \sum_{i=1}^{k} p_i)^{M - \sum_{i=1}^{k} y_i}, \sum_{i=1}^{k} y_i \leq M$$

$$= \frac{M!}{(\prod_{i=1}^{k} y_i!)\,(M-\sum_{i=1}^{k} y_i)!}\; exp\Big[\sum_{i=1}^{k} y_i\, ln(p_i) - ln\big(1-\sum_{i=1}^{k} p_i\big)\big(\sum_{i=1}^{k} y_i\big) + M\; ln\,(1-\sum_{i=1}^{k} p_i)\Big]$$

$$= \frac{M!}{(\prod_{i=1}^{k} y_i!)\,(M-\sum_{i=1}^{k} y_i)!}\; exp\left[\sum_{i=1}^{k} y_i\; ln\left(\frac{p_i}{1-\sum_{i=1}^{k} p_i}\right) + M\; ln\,(1-\sum_{i=1}^{k} p_i)\right] \quad (1.10)$$

$$= exp[y_i\theta_i + b(\theta_i) + c(y_i)]$$

where $\theta_i = ln\left(\frac{p_i}{1-\sum_{i=1}^{k} p_i}\right)$, and $c(y_i) = M\; ln\,(1-\sum_{i=1}^{k} p_i)$.

4- Continuous Response:

Most continuous responses in practice follow the normal distribution. The normal distribution can be written in the form of an exponential family as follows:

$$f(y_i,\mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\; e^{-(y_i-\mu_i)^2/2\sigma^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\; e^{-({y_i}^2-2y_i\mu_i+{\mu_i}^2)/2\sigma^2}$$

$$= exp\left[ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\; e^{-({y_i}^2-2y_i\mu_i+{\mu_i}^2)/2\sigma^2}\right)\right]$$

$$= exp\left[-\frac{1}{2}ln(2\pi\sigma^2) - \frac{{y_i}^2}{2\sigma^2} + \frac{y_i\mu_i}{\sigma^2} - \frac{{\mu_i}^2}{2\sigma^2}\right]$$

$$= exp\left[\frac{y_i\mu_i}{\sigma^2} - \frac{{\mu_i}^2}{2\sigma^2} - \frac{{y_i}^2}{2\sigma^2} - \frac{1}{2}\; ln(2\pi\sigma^2)\right] \quad (1.11)$$

$$= exp[y_i\theta_i + b(\theta_i) + c(y_i)]$$

where $\theta_i = \frac{\mu_i}{\sigma^2}$, $b(\theta_i) = -\frac{{\mu_i}^2}{2\sigma^2}$ and $c(y_i) = -\frac{{y_i}^2}{2\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2)$.

Some models are used to fit regressions for univariate responses with Binomial, Poisson, and Multinomial distributions. The main idea of statistical modelling and the modelling process are involved in the following steps:

i- Estimating parameters used in the models and testing hypotheses.

ii- Checking how well the models fit the observed data.

1.2 Research Methodology

Diagnostic tools for regression analysis provide goodness of fit of the model to observed data. In addition, regression diagnostics for the linear model are well used to measure the influence of one or more observations on the regression analysis. The diagnostics use statistics which measure the effects of deleting a single point from the observed data. The regression diagnostics define the relationship between the least squares fit of the linear model to a complete set of n cases, and the fit to the $(n-1)$ cases remaining after the deletion of a single case (Williams, 1987). Linear regression analysis can be a very effective way to model the observed data, and there had been some works for diagnostics in generalized liner models. The objective of this research is to survey available diagnostics tools for GLM. In particular, we will consider the following regressions

- Diagnostics tools for Poisson regression will be explored. The Poisson regression is often used for modeling count data. The approach considered enables one to concentrate on describing the relation between a count response and a set of predictor variables through the regression model.
- Diagnostics tools for Logistic regression will be explored. The Logistic regression is used to model dichotomous outcome variables or with a data set that has a binary response. The relationship between response and the explanatory variable are described by logistic function.

- Diagnostics tools for Multinomial Logistic regression will be explored. The Multinomial Logistic regression is used to model nominal outcome variables, in which the logarithm of the log odds of the outcomes are modeled as a linear combination of the predictor variables.

We will explore the diagnostic methods for these generalized linear models mentioned above in our research.

1.3 The Generalized linear model

A linear model (LM) describes a continuous response variable as a function of one or more predictor variables, which fits models of the form:

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{1.12}$$

where Y_i is the response variables, $\beta_1, \dots, \beta_k$ are a set of unknown parameters, x_i for $i = 1,2 \dots, k$ is the i^{th} explanatory variable, and ε_i is the random error term. Usually, we assume that the random error term is distributed normally with mean zero and some variance. A generalized linear model is a flexible generalization of ordinary linear regression models, which allows the response variables to have error distribution other than the normal distribution. The generalized linear model is allowing us to us fit regression models for exponential families such as binomial, multinomial or Poisson distributions as described in the previous section. The generalized linear regression allows the linear model to be related to the response variable with a link function g. However, the generalized linear regression models of each response variable Y_i and a set of explanatory variables $x_{i1}, \dots, x_{ik}$ can be written as:

$$g[E(Y_i)] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}\,. \tag{1.13}$$

So then, for the responses $Y_1, \dots, Y_N$, and a set of explanatory variables $x_{i1}, \dots, x_{ik}$, the expression (1.13) can be written in a matrix notation as:

$$g[E(\boldsymbol{y})] = \boldsymbol{X\beta} \tag{1.14}$$

where $\boldsymbol{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}$ is a vector responses, $\boldsymbol{X}$ is a matrix whose elements are constants representing the set of explanatory variables, $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}$ is a vector of unknown parameters, and g is the link function (usually not normal). The most common GLM's are Poisson regression and logistic regression, which will be introduced later.

1.4 The Binomial Distribution

The binomial distribution is particularly useful for modeling count of successes or failures given a number of independent trials such as votes received given an electorate.

Suppose $$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, 3 \dots, n \\ 0 & elsewhere \end{cases} \quad (1.15)$$

where $0 < p < 1, and\ n = 1,2, \dots$, is the sum of the probabilities of these $\binom{n}{x}$ mutually exclusive events. If x successes occur, where $x = 0, 1, 2, \dots n$, then $(n - x)$ failures occur. So, the number of ways of selecting the x position for the x successes in the n trials is given by:

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!}. \quad (1.16)$$

A random variable X that has a probability mass function (pmf) of the form of $p(x)$ is said to have a binomial distribution, and any such $p(x)$ is called a binomial pmf. Note that if n is a positive integer,

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} y^k\, x^{n-k}\,. \quad (1.17)$$

The binomial distribution is denoted by the symbol $b(n, p)$, where the constants n and p are called the parameters of the binomial distribution. Also, the moment generating function (mgf) of a

binomial distribution is defined as follows:

$$M(t) = \sum_{x=0}^{n} e^{tx}\, p(x)$$

$$= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x\, (1-p)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x}\, (pe^t)^x\, (1-p)^{n-x}$$

$$= [(1-p) + pe^t]^n \tag{1.18}$$

for all real values t. Now the mean and the variance μ and σ^2 respectively, of X can be easily computed from $M(t)$ by finding its derivatives $M'(t)$ and $M''(t)$.

Since, $$M'(t) = n\,[(1-p) + pe^t]^{n-1}\,(pe^t), \tag{1.19}$$

and $$M''(t) = n\,[(1-p) + pe^t]^{n-1}\,(pe^t) + n\,(n-1)\,[(1-p) + pe^t]^{n-2}\,(pe^t)^2, \tag{1.20}$$

it follows that:

$$\mu = M'(0) = np \text{ and } \sigma^2 = M''(0) - \mu^2 = np\,(1-p). \tag{1.21}$$

1.5 The Multinomial Distribution

The multinomial distribution is a generalization form of the binomial distribution and it can be generalized as follows. Suppose we have a random experiment that repeated n independent times, say $C_1, C_2, \ldots, C_k$ are mutually exclusive and exhaustive ways of each repetition of the experiment results. Let p_i be the probability that the outcome is an element of C_i for $i = 1,2, \ldots, k$, and let p_i remain constant throughout the n independent repetitions $i = 1,2, \ldots, k$. Then define the random variable X_i to be equal to the number of outcomes that are elements of $C_i, i = 1,2, \ldots, k-1$.

Furthermore, let $x_1, x_2, \ldots, x_{k-1}$ be nonnegative integers, so that $x_1 + x_2 + \cdots + x_{k-1} \leq n$. Then the probability that exactly x_1 terminations of the experiment are in C_1, and the probability that exactly x_{k-1} terminations are in C_{k-1}, and hence exactly $n - (x_1 + x_2 + \cdots + x_{k-1})$ terminations are in C_k is:

$$\frac{n!}{x_1! \ldots x_{k-1}! x_k!} \, {p_1}^{x_1} \ldots {p_{k-1}}^{x_{k-1}} \, {p_k}^{x_k} \tag{1.22}$$

where x_k is merely an abbreviation for $n - (x_1 + x_2 + \cdots + x_{k-1})$. This is the multinomial probability mass function (pmf) of $k - 1$ random variables $X_1, X_2, \ldots, X_{k-1}$ of the discrete type (Wackerly and Scheaffer, 1996). Note that the number of distinguishable arrangements of $x_1 \, C_1, x_2 \, C_2, \ldots, x_k \, C_k$ is

$$\binom{n}{x_1}\binom{n-x_1}{x_2} \ldots \binom{n-x_1- \ldots -x_{k-2}}{x_{k-1}} = \frac{n!}{x_1! \ldots x_{k-1}! \, x_k!}. \tag{1.23}$$

And the probability of each of these distinguishable arrangements is

$${p_1}^{x_1} \ldots {p_{k-1}}^{x_{k-1}} \, {p_k}^{x_k}. \tag{1.24}$$

Suppose that $k = 3$ and let $X = X_1$ and $Y = X_2$, then $n - X - Y = X_3$. The joint probability mass function (pmf) of X and Y is given by:

$$p(x, y) = \begin{cases} \dfrac{n!}{x! \; y! \; (n - x - y)!} {p_1}^x \, {p_2}^y \, {p_3}^{n-x-y} \\ 0 \qquad\qquad\qquad\qquad elsewhere \end{cases}$$

where x and y are nonnegative integers with $x + y \leq n$, and p_1, p_2 and p_3 are positive proper fractions with $p_1 + p_2 + p_3 = 1$. Then we said that X and Y have a trinomial distribution. Now if n is a positive integer and p_1, p_2 and p_3 are fixed constants, then we would have that:

$$\sum_{x=0}^{n}\sum_{y=0}^{n-x}\frac{n!}{x!\,y!\,(n-x-y)!}\,{p_1}^x\,{p_2}^y\,{p_3}^{n-x-y}$$

$$=\sum_{x=0}^{n}\frac{n!}{x!\,(n-x)!}\,{p_1}^x\,\sum_{y=0}^{n-x}\frac{(n-x)!}{y!\,(n-x-y)!}\,{p_2}^y$$

$$=\sum_{x=0}^{n}\frac{n!}{x!(n-x)!}\,{p_1}^x\,(p_2+p_3)^{n-x}$$

$$=(p_1+p_2+p_3)^n\,. \tag{1.25}$$

Consequently, the moment generating function (mgf) of a trinomial distribution can be defined as follows:

$$M(t_1,t_2)=\sum_{x=0}^{n}\sum_{y=0}^{n-x}\frac{n!}{x!\;y!\;(n-x-y)!}\,(p_1\,e^{t_1})^x\,(p_2\,e^{t_2})^y\,{p_3}^{n-x-y}$$

$$=\;(p_1\,e^{t_1}+p_2\,e^{t_2}+p_3)^n \tag{1.26}$$

for all real values of t_1 and t_2.

The moment generating function of the marginal distributions of X and Y are respectively given by:

$$M(t_1,0)=(p_1\,e^{t_1}+p_2+p_3)^n$$

$$=[(1-p_1)+p_1\,e^{t_1}]^n, \tag{1.27}$$

and

$$M(0,t_2)=(p_1+p_2\,e^{t_2}+p_3)^n$$

$$=[(1-p_2)+p_2\,e^{t_2}]^n. \tag{1.28}$$

And hence, we see that X is $b(n,p_1)$ and Y is $b(n,p_2)$. Therefore, the mean and the variance of X and Y are, respectively.

$$\mu_1 = np_1, {\sigma_1}^2 = np_1(1-p_1), \tag{1.29}$$

and

$$\mu_2 = np_2,\ {\sigma_2}^2 = np_2(1-p_2). \tag{1.30}$$

In general, the moment generating function (mgf) of a multinomial distribution is given by:

$$M(t_1, \dots, t_{k-1}) = (p_1\, e^{t_1} + \cdots + p_{k-1}\, e^{t_{k-1}} + p_k)^n \tag{1.31}$$

for all real values $t_1, t_2, \dots, t_{k-1}$.

1.6 Poisson Distribution

The Poisson distribution is often used to model counts such as the number of arrivals, death, or failures, in given time period. We know that the series

$$1 + n + \frac{n^2}{2!} + \frac{n^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{n^x}{x!} \tag{1.32}$$

converges to e^n for all values of n.

Let $n > 0$ and consider the function $p(x)$ defined as follows:

$$p(x) = \begin{cases} \frac{n^x\, e^{-n}}{x!} & x = 0,1,2,\dots \\ 0 & elsewhere \end{cases}. \tag{1.33}$$

Since $n > 0$, then $p(x) > 0$ and

$$\sum_x p(x) = \sum_{x=0}^{\infty} \frac{n^x\, e^{-n}}{x!} = e^{-n} \sum_{x=0}^{\infty} \frac{n^x}{x!} = e^{-n}\, e^n = e^0 = 1.$$

A random variable that has a probability mass function (pmf) of the form $p(x)$ is said to have a Poisson distribution with parameter n, and such any $p(x)$ is called a Poisson **(**pmf**)** with n parameters. Also, the moment generating function (mgf) of a Poisson distribution is defined as

follows:

$$M(t) = \sum_x e^{tx}\, p(x) = \sum_{x=0}^{\infty} e^{tx}\, \frac{n^x\, e^{-n}}{x!}$$

$$= e^{-n} \sum_{x=0}^{\infty} \frac{(n\, e^t)^x}{x!} = e^{-n}\ e^{ne^t}$$

$$= e^{n(e^t-1)} \tag{1.34}$$

for all values of t. Now the mean and the variance μ and σ^2 respectively, of X can be easily computed from $M(t)$ by finding its derivatives $M'(t)$ and $M''(t)$.

Since, $$M'(t) = e^{n(e^t-1)}\ (ne^t), \tag{1.35}$$

and $$M''(t) = e^{n(e^t-1)}\ (ne^t) + e^{n(e^t-1)}\ (ne^t)^2, \tag{1.36}$$

it follows that: $$\mu = M'(0) = n, \tag{1.37}$$

and $$\sigma^2 = M''(0) - \mu^2 = n+n^2-n^2 = n. \tag{1.38}$$

So, Poisson distribution has $\mu = \sigma^2 = n > 0$. Hence a Poisson distribution is can be written as:

$$p(x) = \begin{cases} \frac{\mu^x\, e^{-\mu}}{x!} & x = 0,1,2,\ldots \\ 0 & elsewhere \end{cases}. \tag{1.39}$$

Thus, the parameter n in a Poisson (pmf) is the mean μ.

Remark

Suppose that $Y_1, Y_2, \ldots, Y_m$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_m$, respectively. Then we have that the $\sum_{k=1}^{m} Y_k$ also has a Poisson distributions with mean equal to μ, where $\mu = \sum_{k=1}^{m} \mu_k = \mu_1 + \mu_2 + \cdots + \mu_m$. Now assume that we are given that $m = 2$, i.e., assume

that Y_1 and Y_2 are independent with distributions Poisson (μ_1) and Poisson (μ_2), respectively. Given that $Y_1 + Y_2 = n$, then

$$P(Y_1 = y_1, Y_2 = y_2 | Y_1 + Y_2 = n) = \frac{P(Y_1 = y_1, Y_2 = n - y_1)}{P(Y_1 + Y_2 = n)}$$

$$= \frac{\frac{\mu_1^{y_1} e^{-\mu_1}}{y_1!} \times \frac{\mu_2^{n-y_1} e^{-\mu_2}}{(n-y_1)!}}{\frac{(\mu_1+\mu_2)^n \, e^{-(\mu_1+\mu_2)}}{n!}}$$

$$= \binom{n}{y_1} \left(\frac{\mu_1}{\mu_1 + \mu_2}\right)^{y_1} \left(1 - \frac{\mu_1}{\mu_1 + \mu_2}\right)^{n-y_1}$$

$$= \binom{n}{y_1} \left(\frac{\boldsymbol{\mu_1}}{\boldsymbol{\mu_1 + \mu_2}}\right)^{\boldsymbol{y_1}} \left(\frac{\boldsymbol{\mu_2}}{\boldsymbol{\mu_1 + \mu_2}}\right)^{\boldsymbol{n - y_1}} \tag{1.40}$$

which is the Binomial distribution with $P(\text{success}) = \frac{\mu_1}{\mu_1+\mu_2}$ and $P(\text{failure}) = \frac{\mu_2}{\mu_1+\mu_2}$. Now for the general case, if we have m counts, assume that we are given $Y_1, Y_2, \ldots, Y_m$ are independent Poisson random variables with means $\mu_1, \mu_2, \ldots, \mu_m$, respectively, and given that $\sum_{k=1}^{m} Y_k = Y_1 + Y_2 + \cdots + Y_m = n$, with mean $\mu = \sum_{k=1}^{m} \mu_k = \mu_1 + \mu_2 + \cdots + \mu_m$, then

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m | \sum_{k=1}^{m} Y_k = Y_1 + Y_2 + \cdots + Y_m = n)$$

$$= \frac{P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m)}{P(Y_1 + Y_2 + \cdots + Y_m = n)}$$

$$= \frac{n!}{Y_1! Y_2! \ldots Y_m!} \left(\frac{\mu_1}{\mu}\right)^{y_1} \left(\frac{\mu_2}{\mu}\right)^{y_2} \ldots \left(\frac{\mu_m}{\mu}\right)^{y_m} \tag{1.41}$$

which is a Multinomial distribution. This relationship between Poisson and multinomial distributions is exploited while fitting GLM for count data.

CHAPTER 2

Estimations

2.1 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation begins with writing a mathematical expression known as the likelihood function of the sample data. The likelihood of a set of data is the probability of obtaining that set of data, given the chosen probability distribution model. Often it is easier to work with the log-likelihood function than with the likelihood function itself (Dobson, 2008). Let $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ denote a random vector, and let the density function of the y_i's be $f(\boldsymbol{y}; \boldsymbol{\theta})$ joint probability, which depends on the vector of parameters $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$ for $p < n$. The likelihood function $L(\boldsymbol{\theta}; \boldsymbol{y})$ is algebraically the same as the joint probability density function $f(\boldsymbol{y}; \boldsymbol{\theta})$, but the change in notation reflects a shift of emphasis from the random variables $\boldsymbol{y}$, with $\boldsymbol{\theta}$ fixed, to the parameters $\boldsymbol{\theta}$ with $\boldsymbol{y}$ fixed. Since L is defined in terms of the random vector y, it is itself a random variable. Let Ω denote the set of all possible values of the parameter vector $\boldsymbol{\theta}$; Ω is called the parameter space. The maximum likelihood estimator of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ which maximizes the likelihood function, that is, $L\left(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}\right) \geq L(\boldsymbol{\theta}; \boldsymbol{y})$ for all $\boldsymbol{\theta}$ in Ω. Equivalently, $\widehat{\boldsymbol{\theta}}$ is the value which maximizes the log-likelihood function $l(\boldsymbol{\theta}; \boldsymbol{y}) = log\, L(\boldsymbol{\theta}; \boldsymbol{y})$, since the logarithm function is monotonic. Thus,

$$l(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) \geq l(\boldsymbol{\theta}; \boldsymbol{y}) \text{ for all } \boldsymbol{\theta} \text{ in } \Omega.$$

Example: (Poisson distribution)

Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables each with the Poisson distribution given as:

$$f(y_i; \theta) = \frac{e^{-\theta}\ \theta^{y_i}}{y_i!} \text{ for } y_i = 0,1,2,\ldots$$

with the same parameter θ. Their joint distribution is

$$f(y_1, \dots, y_n, \theta) = \prod_{i=1}^{n} f(y_i; \theta) = \frac{e^{-\theta}\ \theta^{y_1}}{y_1!} \times \dots \times \frac{e^{-\theta}\ \theta^{y_n}}{y_n!} = \frac{e^{-n\theta}\ \theta^{\sum y_i}}{y_1! \dots y_n!}. \quad (2.1)$$

This is also the likelihood function $L(\theta;\ y1, \dots, yn)$. It is easier to use the log likelihood function

$$l(\theta; y_1, y_2, \dots, y_n) = \ lnL(\theta; y_1, y_2, \dots, y_n)$$

$$= \ (\sum_{i=1}^{n} y_i\ )\ ln\ \theta - n\theta - \sum_{i=1}^{n}(ln\ y_i!). \quad (2.2)$$

To find the maximum likelihood estimate θ, use

$$\frac{dl(\theta; y_1, y_2, \dots, y_n)}{d\theta} = \frac{1}{\theta}\sum_{i=1}^{n} y_i - n\ . \quad (2.3)$$

Solving $\frac{dl(\theta; y_1, y_2, \dots, y_n)}{d\theta} = 0$, yields:

$$\tilde{\theta} = \sum_{i=1}^{n} \frac{y_i}{n} = \ \bar{y}. \quad (2.4)$$

Since $\frac{d^2 l(\theta; y_1, y_2, \dots, y_n)}{d^2\theta} \ = \ -\sum_{i=1}^{n} \frac{y_i}{\theta^2} \ \leq \ 0$, l has its maximum value when $\theta = \tilde{\theta}$, confirming that $\bar{y}$ is the maximum likelihood estimate.

Example: (Binomial distribution)

Suppose we have an experiment with binomial distribution containing n trials resulted in observations, say $y_1, y_2, \dots, y_n$ such that:

$$y_i = \begin{cases} 1 & if\ the\ i^{th}\ trail\ was\ success \\ 0 & Otherwise \end{cases}$$

Then the likelihood of the observed sample is the probability of the observing $y_i, i = 1, 2, \dots, n$. Hence,

$$L(p) = L(y_1, y_2, \dots, y_n | \ p) = p^y\ (1-p)^{n-y} \quad (2.5)$$

where $y = \sum_{i=1}^{n} y_i$. Now to find the value of p that maximizes $L(p)$. Let us write $L(p)$ as the following:

$$L(p) = \begin{cases} (1-p)^n & if\ y = 0, and\ then\ L(p)\ is\ maximized\ when\ p = 0 \\ p^n & if\ y = n, and\ then\ L(p)\ is\ maximized\ when\ p = 1 \\ p^y\ (1-p)^{n-y} & if\ y = 1,2, \dots, n-1 \end{cases}$$

Clearly, $L(p) = p^y\ (1-p)^{n-y}$ is zero at $p = 0$ and $p = 1$, and it is continuous for all the values of p between 0 and 1. So, we only need to find the value of p that maximizes $L(p)$ for $y = 1,2, \dots, n-1$. Note that $ln[L(p)]$ is a monotonically increasing function of $L(p)$. Hence both functions $ln[L(p)]$ and $L(p)$ are maximized for the same value p, [since $L(p)$ is a product of functions of p]. Hence, we would have that:

$$ln[L(p;y)] = ln\ [p^y\ (1-p)^{n-y}]$$

$$= ln(p^y) + ln[(1-p)^{n-y}]$$

$$= y\ ln(p) + (n-y)\ ln(1-p). \tag{2.6}$$

Thus, by computing the derivative of $ln[L(p;y)]$ with respect to p, for $y = 1,2, \dots, n-1$, we have that:

$$\frac{dln[L(p;y)]}{dp} = y\ \left(\frac{1}{p}\right) + (n-y)\ \left(\frac{-1}{1-p}\right). \tag{2.7}$$

Solving $\frac{dln[L(p;y)]}{dp} = 0$, yields:

$$\frac{d\ ln[L(p;y)]}{dp} = \frac{y}{p} - \frac{(n-y)}{(1-p)} = 0$$

$$\Longrightarrow y - y\hat{p} - n\hat{p} + y\hat{p} = 0$$

$$\Longrightarrow y - n\hat{p} = 0 \Rightarrow \hat{p} = \frac{y}{n}. \tag{2.8}$$

Hence, $L(p)$ is maximized at $\hat{p} = 0$ when $y = 0,$ at $\hat{p} = 1$ when $y = n,$ and at $\hat{p} = \frac{y}{n}$ when $y = 1, 2, \dots, n-1$. Therefore, the maximum likelihood estimator (MLE) $\hat{p} = \frac{y}{n}$, is the function of the successes in the total number of trials n. To confirm that $\hat{p}$ is indeed the MLE, we compute:

$$\frac{d^2 ln[L(p;y)]}{dp^2} = \frac{y}{p^2} - \frac{(n-y)}{(1-p)^2} < 0.$$

Example: (Poisson distribution)

Suppose that $Y_1, Y_2, \dots, Y_n$ denote a random sample from the Poisson distribution with mean λ. So, by definition we have that:

$$f(y_i) = \frac{e^{-\lambda}\ \lambda^{y_i}}{y_i!} \text{ for } i = 0,1,2, \dots, n.$$

Then their density function can be written as:

$$f(y_1, \dots, y_n) = f(y_1) \times f(y_2) \times \dots \times f(y_n)$$

$$= \frac{e^{-\lambda}\ \lambda^{y_1}}{y_1!} \times \frac{e^{-\lambda}\ \lambda^{y_2}}{y_2!} \times \dots \times \frac{e^{-\lambda}\ \lambda^{y_n}}{y_n!}$$

$$= \frac{e^{-n\lambda}\ \lambda^{\sum y_i}}{y_1! \dots y_n!}.$$

Now to find the value of λ that maximized $L(y)$, note that the log likelihood function can be computed as:

$$ln\ [L(\lambda; y)] = ln[f(y_1, \dots, y_n)]$$

$$= ln\left[\frac{e^{-n\lambda}\ \lambda^{\sum y_i}}{y_1! \dots y_n!}\right] = ln\left[\frac{e^{-n\lambda}\ \lambda^{\sum y_i}}{\prod_{i=1}^{n} y_i!}\right]$$

$$= ln\left[\lambda^{\sum y_i}\right] + ln\left[e^{-n\lambda}\right] + ln\left[\frac{1}{\prod_{i=1}^{n} y_i!}\right]$$

$$= n\overline{y}\, ln(\lambda) - \ ln(\prod_{i=1}^{n} y_i!) - n\lambda. \tag{2.9}$$

Thus, by computing the derivative of $ln[L(\lambda; y)]$ with respect to λ, we have that:

$$\frac{dL(\lambda;y)}{d\lambda} = \frac{d}{d\lambda}[n\overline{y}\ \ In(\lambda) - \ ln(\prod_{i=1}^{n} y_i!) - n\lambda]$$

$$= n\overline{y}\ \left(\frac{1}{\lambda}\right) - n. \tag{2.10}$$

Solving $\frac{dL(\lambda;y)}{d\lambda} = 0$, yields:

$$\frac{d\ L(\lambda;y)}{d\lambda} = n\overline{y}\ \left(\frac{1}{\lambda}\right) - n = 0$$

$$\Longrightarrow n\overline{y} - n\lambda = 0$$

$$\Longrightarrow \lambda \equiv \overline{y}. \tag{2.11}$$

Therefore, the maximum likelihood estimator (MLE) $\lambda \equiv \overline{y}$.

2.2 Iteratively Weighted Least Squares Estimations (IWLS)

2.2.1 Ordinary Least Squares Estimation (OLS)

The method of least squares is used to estimate the regression coefficients in a multiple linear regression model. Consider the following straight line models

$$\begin{matrix} y_1 = \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_n + \varepsilon_n \end{matrix} \Leftrightarrow\ \boldsymbol{y} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \boldsymbol{\varepsilon} \tag{2.12}$$

with $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ and $\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$. Therefore, we can write (2.12) as:

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - A\boldsymbol{\beta} \tag{2.13}$$

where $A = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$. Note that $\|\boldsymbol{\varepsilon}\| = \sqrt{{\varepsilon_1}^2 + \cdots + {\varepsilon_n}^2} = \|\boldsymbol{y} - A\boldsymbol{\beta}\|$, we need to minimize this function over the choice of β_0 and β_1. So, then

$$\min_{\beta_0,\beta_1} \|\boldsymbol{\varepsilon}\|^2 = \min_{\beta_0,\beta_1} \|\boldsymbol{y} - A\boldsymbol{\beta}\|^2, \tag{2.14}$$

and

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} {\varepsilon_i}^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2.15}$$

Note if $\sum_{i=1}^{n} x_i$ denotes all the points in our sample, then we know that:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \Longrightarrow \sum_{i=1}^{n} x_i = n\overline{x}. \tag{2.16}$$

Then the first order condition is

$$\frac{dE}{d\beta_0} = 0$$

$$\Rightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \Longrightarrow \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i = n\overline{y} - n\beta_0 - n\beta_1\overline{x} = \overline{y} - \beta_0 - \beta_1\overline{x} = 0$$

$$\Rightarrow \beta_0 = \overline{y} - \beta_1\overline{x}, \tag{2.17}$$

and similarly, the second first condition is

$$\frac{dE}{d\beta_1} = 0$$

$$\Rightarrow -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \Longrightarrow \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - n\beta_0 \overline{x} - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0. \tag{2.18}$$

From (2.17) and (2.18), we have that:

$$\sum_{i=1}^{n} y_i x_i - (\overline{y} - \beta_1 \overline{x})\, n\overline{x} - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i - n\overline{y}\,\overline{x} = \beta_1 \left(\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2\right). \tag{2.19}$$

From (2.19), we have that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{y}\,\overline{x}}{\sum_{i=1}^{n} {x_i}^2 - n\overline{x}^2}, \tag{2.20}$$

and then

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x}, \tag{2.21}$$

which are the last squares estimates of β_0 and β_1.

Remark

Suppose that $y_{ij} = x'_{ij}\beta + \varepsilon_{ij}$ for $i = 1,2,..,n$ and $j = 1,2,\ldots,k$. Let $g: \mathbb{R}^n \longrightarrow \mathbb{R}$ be a function and let $\theta = g(\beta)$ be the parameter. If $\theta = \varepsilon_{ij} = \sigma^2$ for all i, j then the customary estimator θ is the (OLS) is given by $\hat{\theta}_0 = g(\hat{\beta}_0)$,

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} x_{ij}} \tag{2.22}$$

where y_{ij} are the responses variables, x_{ij} are the values of an n dimensional covariate, x_{ij}' is the transpose of x_{ij}, β is a n vector of unknown parameter and ε_{ij} are random error. Note that the (OLS) can be improved by the weighted least square estimator (WLSE) $\hat{\theta}_w = g(\hat{\beta}_w)$ and

$$\hat{\beta}_w = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} w_i\, x_{ij}\, y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} w_i\, x_{ij}\, x_{ij}} \tag{2.23}$$

for some weighted $w_i > 0, i = 1,2,..,n.$

If ${\sigma_i}^2$ are known, then we need to estimate ${\sigma_i}^2$ by ${\hat{\sigma}_i}^2$, we use $w_i = {\sigma_i}^{-2}$, and then we have:

$$\hat{\beta}_w = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} {\sigma_i}^{-2}\, x_{ij}\, y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{k} {\sigma_i}^{-2}\, x_{ij}\, x_{ij}} \tag{2.24}$$

for some weighted $w_i > 0, i = 1,2,..,n$. Now we need to find the vector of least squares estimators b that minimizes

$$S(\beta) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

$$= y'y - \beta'X'y - y'X\beta + \beta'X'\,X\beta$$

$$= y'y - 2\beta'X'y + \beta'X'\,X\beta. \tag{2.25}$$

Hence the least squares estimation must satisfy:

$$\left.\frac{dS}{d\beta}\right|_b = -2X'y + 2X'\,Xb = 0$$

$$\Rightarrow X'\,Xb = X'y. \tag{2.26}$$

And we called $b = (X'\,X)^{-1}X'y$, the ordinary least squares estimator of β.

2.2.2 Least Squares Estimation

Consider the independent observations $y_1, \dots, y_n$ of random variables with means $\mu_1, \dots, \mu_n$ and variances ${\sigma_1}^2, \dots, {\sigma_n}^2$, respectively. Let $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ for $p < n$. The method of least squares is about

estimating the parameter $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ by minimizing the squared discrepancies between observed data $y_i's$, and their expected value $E(y_i) = \mu_i(\boldsymbol{\beta})$. So, we have that:

$$S = \sum_{i=1}^{n} y_i[y_i - \mu_i(\boldsymbol{\beta})]^2. \tag{2.27}$$

Hence to obtain $\hat{\beta}$ we have to differentiate S with respect to β_i for all $i = 1, \dots, p$, and then we set $\frac{dS}{d\beta_i} = 0$ to get:

$$S = \sum_{i=1}^{n} w_i[y_i - \mu_i(\boldsymbol{\beta})]^2 \tag{2.28}$$

which is the weighted sum of minimizing the squared of the differences between the observed data $y_i's$, and their expected value, with weights $w_i = (\sigma_i{}^2)^{-1}$. In general, let $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, and let $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$ with variance covariance matrix $\boldsymbol{V}$. Then the weighted least squares estimator is obtained by minimizing, and hence, we have:

$$S = (\boldsymbol{y} - \boldsymbol{\mu})\, \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{2.29}$$

Example: (Binomial distribution)

Suppose we have that

$$S = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)^2}{\sigma_i{}^2}\right] \tag{2.30}$$

where $\mu_i = n_i\pi_i$ and $\sigma_i{}^2$ is the binomial variance at the i^{th} point with

$$\sigma_i{}^2 = n_i\pi_i[1 - \pi_i] = \frac{e^{-x_i{}'\beta}}{\left(1+e^{-x_i{}'\beta}\right)^2}. \tag{2.31}$$

Then we have that:

$$\min S = min_{\beta} \sum_{i=1}^{n} \left[\frac{(y_i-\mu_i)^2}{{\sigma_i}^2}\right]. \tag{2.32}$$

So, for fixed ${\sigma_i}^2$, we have that:

$$\frac{d\ min\ S}{d\beta} = -2\left[\frac{\sum_{i=1}^{n}(y_i-\mu_i)}{{\sigma_i}^2}\right]\left(\frac{d\mu_i}{d\beta}\right) \tag{2.33}$$

and $\mu_i = n_i\pi_i[1-\pi_i]x_i = {\sigma_i}^2 x_i$. Now set $\frac{d\ min\ S}{d\beta} = 0$, then we have that:

$$-2\left[\frac{\sum_{i=1}^{n}(y_i-\mu_i)}{{\sigma_i}^2}\right]{\sigma_i}^2 x_i = 0$$

$$\Rightarrow \ \sum_{i=1}^{n}(y_i-\mu_i)\,x_i = 0 \tag{2.34}$$

which is equal to $X'(y\ -\ \mu) = 0$, where $\mu = (\mu_1, \dots, \mu_n)$ and $y = (y_1, \dots, y_n)$.

Let $b = (b_1, \dots, b_n)$ be the final estimation of the model parameters, then we have that $E(b) = \beta$ and $var(b) = (X'WX)^{-1}$, where W is $n\times n$ diagonal matrix and the i^{th} diagonal element of W is $w_{ii} = n_i\pi_i[1-\pi_i]$ and hence, the linear predictor is given by:

$$\eta_i = {x_i}'b. \tag{2.35}$$

Note that for the binomial distribution, we have:

$${\sigma_i}^2 = n_i\pi_i[1-\pi_i] = \frac{e^{-{x_i}'\beta}}{\left(1+e^{-{x_i}'\beta}\right)^2},$$

and for Poisson distribution, we have:

$${\sigma_i}^2 = e^{{x_i}'\beta}. \tag{2.36}$$

CHAPTER 3

Special Regression Models

In the following section, we present some special cases of generalized linear models (GLMs).

3.1 The Logistic Regression

The logistic regression model is one of the popular mathematical models for the analysis of binary data. Logistic regression for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous. Logistics regression methods use one of the three types of categorical response variables: binary, ordinary, nominal. A binary response has two categories with no natural order (for example, success-failure or yes-no). An ordinal response has three or more categories with a natural ordering (for example none, mild, and extra). A nominal response has three or more categories with no natural ordering (for example, blue, black, red, yellow or sunny, rainy, and cloudy).

3.2 Logistic Regression Model

Consider a simple k variable regression model where $k = n + 1$, whose general form for the general regression model is given by:

$$Y = f(\beta_0 + \sum_{i=1}^{n} \beta_i X_i) \tag{3.1}$$

where β_0 has a constant value, and the β_i for all $i = 1,2,\dots,n$ are the estimated weights of X_i, the transformed raw data.

3.2.1 Logistic Response Function

Generally, where the variable is binary, there is a considerable empirical evidence indicating that the shape of the respond function is nonlinear. Suppose that a binary random variable y has a Bernoulli distribution, i.e., y takes either the value 1 or the value 0 with probabilities $\pi(x)$ or $1 -$

$\pi(x)$ respectively, where $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$. Note that $\pi(x) = p(y = 1 \mid x)$, which is the conditional probability of $y = 1$, given x. So, the specific form of the logistic regression model with unknown parameter $\beta_1, \beta_2, \dots, \beta_n$ is given by:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x_1+ \dots +\beta_n x_n)}}{1+e^{(\beta_0+\beta_1 x_1+ \dots +\beta_n x_n)}}. \tag{3.2}$$

The function $\beta_0 + \beta_1 x_1 + \ \dots + \beta_n x_n$ is called the linear predictor. The transformation of $\pi(x)$ is called the logit transformation, and is given by:

$$\text{Logit } \pi(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right). \tag{3.3}$$

The binary response model violates a number of ordinary least square assumptions. So, we will use the maximum likelihood method to estimate the parameters of the logistic regression model assuming that Y is a Bernoulli random variable. The solution of the parameters of the logistic regression model is obtained by iteration. Suppose we have two groups defined as following:

	Diseased	Not diseased
Exposed	π_1	$1 - \pi_1$
Not exposed	π_2	$1 - \pi_2$

Note that for the simple logistic model $\pi_i = \frac{e^{\beta_i}}{1+e^{\beta_i}}$, then $\theta = 1$ if there is no difference between the two groups, the exposed and not exposed. To see this, suppose that there is no difference between the exposed and not exposed groups, i.e., $\beta_1 = \beta_2$ and write $\theta = e^{\beta_1-\beta_2} = e^{\beta_1-\beta_1} = e^0 = 1$.

3.3 The Multinomial Logistic Regression Model

The multinomial regression model is a generalization of the logistic model (the binary response), to the model for multiple (more than two) category response variables that have a multinomial

distribution as described in (1.5). To see this let us consider any categorical variable Y with three categories. Then we only need two logit models as logistic regression model uses the binary outcomes variable. Assume that there are n explanatory variables in this model, i.e., $x = (x_1, x_2, \dots, x_n)$. Then we have that:

$$ln\left(\frac{p(Y=1 \mid x)}{p(Y=3 \mid x)}\right) = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1n}x_n\,, \tag{3.4}$$

and

$$ln\left(\frac{p(Y=2 \mid x)}{p(Y=3 \mid x)}\right) = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2n}x_n. \tag{3.5}$$

Then the response probabilities can be written as the following, but note that $\pi(x) = p(y = k \mid x)$, which is the conditional probability of $y = k$, given x with $k = 1,2$ and 3.

$$p(Y=1 \mid x) = \frac{e^{(\beta_{10}+\beta_{11}x_1+\dots+\beta_{1n}x_n)}}{1 + e^{(\beta_{10}+\beta_{11}x_1+\dots+\beta_{1n}x_n)} + e^{(\beta_{20}+\beta_{21}x_1+\dots+\beta_{2n}x_n)}}$$

$$= \frac{e^{(\widehat{\beta}_1 x)}}{1 + e^{(\widehat{\beta}_1 x)} + e^{(\widehat{\beta}_2 x)}},$$

$$p(Y=2 \mid x) = \frac{e^{(\beta_{20}+\beta_{21}x_1+\dots+\beta_{2n}x_n)}}{1 + e^{(\beta_{10}+\beta_{11}x_1+\dots+\beta_{1n}x_n)} + e^{(\beta_{20}+\beta_{21}x_1+\dots+\beta_{2n}x_n)}}$$

$$= \frac{e^{(\widehat{\beta}_2 x)}}{1 + e^{(\widehat{\beta}_1 x)} + e^{(\widehat{\beta}_2 x)}},$$

$$p(Y=3 \mid x) = \frac{1}{1 + e^{(\beta_{10}+\beta_{11}x_1+\dots+\beta_{1n}x_n)} + e^{(\beta_{20}+\beta_{21}x_1+\dots+\beta_{2n}x_n)}}$$

$$= \frac{1}{1+e^{(\widehat{\beta}_1 x)} + e^{(\widehat{\beta}_2 x)}},$$

where $\hat{\beta}_1 x = \beta_{10} + \beta_{11}x_1 + \cdots + \beta_{1n}x_n$ and $\hat{\beta}_2 x = \beta_{20} + \beta_{21}x_1 + \cdots + \beta_{2n}x_n$ are unknown parameters. The conditional likelihood function given the covariates for sample of n independent observations is

$$L(\beta) = \prod_{i=1}^{n}\left[\left(p(Y=1 \mid x)\right)^{y_{1i}} \times \left(p(Y=2 \mid x)\right)^{y_{2i}} \times \left(p(Y=3 \mid x)\right)^{y_{3i}}\right]$$

$$= \prod_{i=1}^{n}\left[\left(\frac{e^{\hat{\beta}_1 x}}{1+e^{\hat{\beta}_1 x}+e^{\hat{\beta}_2 x}}\right)^{y_{1i}} \times \left(\frac{e^{\hat{\beta}_2 x}}{1+e^{\hat{\beta}_1 x}+e^{\hat{\beta}_2 x}}\right)^{y_{2i}} \times \left(\frac{1}{1+e^{\hat{\beta}_1 x}+e^{\hat{\beta}_2 x}}\right)^{y_{3i}}\right]. \quad (3.6)$$

Now by taking the log of both sides we have that:

$$ln\,(L(\beta)) = ln \prod_{i=1}^{n}\left[\left(\frac{e^{\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{1i}} \times \left(\frac{e^{\hat{\beta}_2 x_i}}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{2i}} \times \left(\frac{1}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{3i}}\right]$$

$$= \sum_{i=1}^{n} ln\left(\frac{e^{\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{1i}} \times \; ln\left(\frac{e^{\hat{\beta}_2 x_i}}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{2i}} \times \; ln\left(\frac{1}{1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i}}\right)^{y_{3i}}$$

$$= \sum_{i=1}^{n}\Big[\, y_{1i}\left(\hat{\beta}_1 x_i - ln(1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i})\right) + y_{2i}\left(\hat{\beta}_2 x_i - ln(1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i})\right) -$$

$$y_{3i}\left(ln(1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i})\right)\Big]$$

$$= \sum_{i=1}^{n}\left[y_{1i}\,\hat{\beta}_1 x_i + y_{2i}\,\hat{\beta}_2 x_i - (y_{1i}+y_{2i}+y_{3i})\; ln(1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i})\right]$$

$$= \sum_{i=1}^{n}\left[y_{1i}\,\hat{\beta}_1 x_i + y_{2i}\,\hat{\beta}_2 x_i - ln(1+e^{\hat{\beta}_1 x_i}+e^{\hat{\beta}_2 x_i})\right]. \quad (3.7)$$

Since $\sum_{i=1}^{n} y_{ji} = 1$ for each j, we have $y_{1i} + y_{2i} + y_{3i} = 1$.

3.4 The Poisson Regression

The Poisson regression model is derived from the Poisson distribution by parameterizing the relation between the mean parameter μ and covariates (regressor) x, and is usually known as a log-linear model. Suppose that each random variable $Y_i, i = 1, 2, \ldots, n$ represents a count variable where Y_i is denoting the number of events for n observations and means μ_i. The Poisson model assumes that the counts for the outcome variables have a Poisson distribution as described in (1.6). Note that the mean and variance of this distribution can be shown to be

$$E(Y_i) = \mu_i = Var(Y_i) \;\text{ for }\; i = 1, 2, \ldots, n.$$

Then we have that the linear predictor function is $\eta_i = x'_i\beta$, and the link function is $\log E(Y_i) = \log(\mu_i) = \eta_i$, for the log link usual with Poisson data. Then the log likelihood function for n observations can be written as following:

$$\begin{aligned} ln\big(L(\beta)\big) &= ln\left(\prod_{i=1}^{n} \frac{e^{-\mu_i}\,\mu_i{}^{y_i}}{y_i!}\right) \\ &= \sum_{i=1}^{n}(y_i \ln(\mu_i) - \mu_i - \ln(y_i!)) \\ &= \sum_{i=1}^{n}(y_i\,\eta_i - e^{\eta_i} - \ln(y_i!)) \\ &= \sum_{i=1}^{n}\big(y_i\,x'_i\beta - e^{x'_i\beta} - \ln(y_i!)\big). \end{aligned} \tag{3.8}$$

Now by computing the first derivative of $\ln\big(L(\beta)\big)$ with respect to β, and the set the result to zero we have that:

$$\begin{aligned} &\frac{d\ln\big(L(\beta)\big)}{d\beta} = 0 \\ &\Rightarrow y_i\,x'_i - x'_i\,e^{x'_i\beta} = 0 \\ &\Rightarrow y_i\,x'_i = x'_i\,e^{x'_i\beta} \\ &\Rightarrow y_i\,x'_i = x'_i\,\hat{\mu} \end{aligned} \tag{3.9}$$

where $\hat{\mu} = e^{x'_i\beta}$ is the vector of the estimated means.

CHAPTER 4

Diagnostic Tools for Generalized Linear Models

In this chapter, we introduce two types of diagnostic tools of goodness of fit, deviance test statistics and chi-square goodness of fit statistics. Diagnostics are designed to find problems with assumptions of any statistical procedure. In addition, it is found that the residual deviance provides better goodness of fit measure for generalized linear models (GLMs) than Pearson statistics. The higher numbers in deviance always indicate a bad fit, and imply poor fit to the observed data.

4.1 Deviance Test Statistics

The goodness of model fit in general, is referring to measuring how well do the observed data correspond to the fitted model. One method for goodness of fit is to use the deviance statistics D, which is a measure of discrepancy between observed and fitted values. Deviance is an important idea associated with a fitted GLM. It can be used to test the fit of the link function and linear predictor to the data, or to test the significance of a particular predictor variable (or variables) in the model. Hence, the deviance test statistic is given by:

$$D = 2\left(ln\left[L_s\left(\hat{\beta}\right)\right] - ln\left[L_m(\hat{\beta})\right]\right) \tag{4.1}$$

where $ln\left[L_s(\hat{\beta})\right]$ is the maximized log likelihood of the saturated model, and $ln\left[L_m(\hat{\beta})\right]$ is the maximized log likelihood of the fitted model. For large samples the distribution of the deviance is approximately a chi-squared (will be discussed later) with $n - p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, the deviance can be used directly to test the goodness of fit of the model. Under specific regularity conditions D converges to a χ^2 distribution with l degrees of freedom and l is the difference between the number parameters in the saturated model and the number of parameters in the model being considered, i.e.,

$$D \sim \chi^2_{(n-p)}. \tag{4.2}$$

Therefore, the deviance can be used to test the null hypothesis that any subset of the β's is equal to 0. We can test the null hypothesis

$$H_0: \beta_{n-p} = 0, \tag{4.3}$$

and then H_0 is can be rejected when

$$D \geq \chi^2_{(1-\alpha)}, \tag{4.4}$$

where α is a fixed level of significance.

4.2 Chi-square Goodness of Fit Statistics

The Chi-square test for goodness of fit is designed to test whether observed frequencies differ significantly from expected frequencies. An alternative measure of goodness of fit is Pearson's chi-squared statistic, we can use it to test the null hypothesis that the data comes from a specific parametric distribution for example, binomial and Poisson distribution. The Pearson chi-squared statistic is defined as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{4.5}$$

which has the alternative interpretation of:

$$\chi^2 = \sum_{i=1}^{n} \frac{(observed_i - expected_i)^2}{expected_i}. \tag{4.6}$$

However, if we have a generalized linear model with responses y_i, the sum of weights $w_i = \sum_{j=1}^{n_i} w_{ij}$, then the fitted mean $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i, and the variance function $v(\mu)$. Therefore, the Pearson's chi-squared statistic can be written as:

$$\chi^2 = \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{4.7}$$

The numerator is the squared difference between observed and fitted values, and the denominator is the variance of the observed value. If χ^2 has the distribution $\chi^2(n)$ with n degrees of freedom which is defined the sum of n independent random variables $y_1, \dots, y_n$, then its expected value is $E(\chi^2) = n$ and its variance is $var(\chi^2) = 2n$. The Pearson statistic has the same form for Poisson and binomial data.

4.3 Diagnostic Tools for Count Data

Suppose that $Y_1, Y_2, \dots, Y_n$ denote a random sample from the Passion distribution with mean λ. So, by definition we have that:

$$f(y_i) = \frac{e^{-\lambda}\ \lambda^{y_i}}{y_i!}\ , i = 0,1,2, \dots, n\ .$$

Then the ML model can be computed as follows:

$$ln\ [L(\lambda; y)] = ln[f(y_1, \dots, y_n)]$$

$$= ln\left[\frac{e^{-n\lambda}\ \lambda^{\Sigma y_i}}{y_1! \dots y_n!}\right] = ln\left[\frac{e^{-n\lambda}\ \lambda^{\Sigma y_i}}{\prod_{i=1}^{n} y_i!}\right]$$

$$= ln\left[\lambda^{\Sigma y_i}\right] + ln\left[e^{-n\lambda}\right] + ln\left[\frac{1}{\prod_{i=1}^{n} y_i!}\right]$$

$$= n\overline{y}\ ln(\lambda) -\ ln(\textstyle\prod_{i=1}^{n} y_i!) - n\lambda.$$

Thus,

$$\frac{d\ L(\lambda; y)}{d\lambda} = \frac{d}{d\lambda}\left[n\overline{y}\ In(\lambda) -\ ln\left(\prod_{i=1}^{n} y_i!\right) - n\lambda\right]$$

$$= n\overline{y}\ \left(\frac{1}{\lambda}\right) - n.$$

Solving $\frac{d\,L(\lambda;y)}{d\lambda} = 0$, yields:

$$\frac{d\,L(\lambda;y)}{d\lambda} = n\overline{y}\left(\frac{1}{\lambda}\right) - n = 0$$

$$\Longrightarrow n\overline{y} - n\lambda = 0$$

$$\Longrightarrow \lambda \equiv \overline{y}.$$

For the model of a count outcome variable (Poisson distribution), the deviance statistic is computed as:

$$D = 2\left(\log\left[L_s(\hat{\beta})\right] - \log\left[L_m(\hat{\beta})\right]\right)$$

$$= 2\left(\sum_{i=1}^{n}\left[y_i log\left(\frac{y_i}{e^{(X_i\hat{\beta})}}\right)\right] - \sum_{i=1}^{n}\left[\,y_i - e^{(X_i\hat{\beta})}\right]\right)$$

$$= 2\sum_{i=1}^{n}\left[y_i log\left(\frac{y_i}{\hat{\lambda}_i}\right) - \left(\,y_i - \hat{\lambda}_i\,\right)\right]$$

$$= 2\sum_{i=1}^{n}\left[y_i\left(log(y_i) - log\,(\hat{\lambda}_i)\right) - \left(\,y_i - \hat{\lambda}_i\,\right)\right] \qquad (4.8)$$

which does not contain unknown parameters, so can be calculated from the data. The first term is identical to the binomial deviance, representing twice a sum of observed times log of observed over fitted. The second term is a sum of the differences between observed and fitted values and is usually zero. Since the deviance is a measure of how well the model fits the data, then we say the model fits well the observed values y_i if they are close to their predicted means λ_i which is causing both terms in *D* to be small, and so the deviance to be small. Hence the Pearson's chi-squared statistic is given by:

$$\chi^2 = \sum_{i=1}^{n}\frac{(\,y_i - e^{(X_i\hat{\beta})})^2}{e^{(X_i\hat{\beta})}}$$

$$= \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \quad (4.9)$$

4.4 Diagnostic Tools for Binary Data

Suppose that we have $Y_i \sim$Binomial (m_i, π_i), where Y_i is the number of successes for the i^{th} combination of levels of the predictor variables $i = 1,2, \dots, n$. By definition we have that:

$$p(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & y = 0, 1, 2, 3 \dots, n \\ 0 & elsewhere \end{cases}.$$

Then the ML model can be computed as follows: Define n parameters $\theta_1, \theta_2, \dots, \theta_n$ where $\theta_i = \pi_i$. Then we have that:

$$L_s(\theta) = \prod_{i=1}^{n} \binom{m_i}{y_i} {\theta_i}^{y_i} (1-\theta_i)^{m_i - y_i}$$

$$l_s(\theta) = \sum_{i=1}^{n} ln \binom{m_i}{y_i} + \sum_{i=1}^{n} y_i \ln \theta_i + \sum_{i=1}^{n} (m_i - y_i) \ln(1-\theta_i).$$

Thus, by computing the derivative of $l_s(\theta)$ with respect to θ_i, we have that:

$$\frac{d\, l_s(\theta)}{d\theta_i} = y_i \left(\frac{1}{\theta_i}\right) + (m_i - y_i) \left(\frac{-1}{1-\theta_i}\right)$$

Solving $\frac{d\, l_s(\theta)}{d\theta_i} = 0$, yields:

$$\frac{d\, l_s(\theta)}{d\theta_i} = \frac{y_i}{\theta_i} - \frac{(m_i - y_i)}{(1-\theta_i)} = 0$$

$$\Longrightarrow y_i - y_i\hat{\theta}_i - m_i\hat{\theta}_i + y_i\hat{\theta}_i = 0$$

$$\Longrightarrow y_i - m_i\hat{\theta}_i = 0$$

$$\Rightarrow \hat{\theta}_i = \frac{y_i}{m_i}. \quad (4.10)$$

Let $p_i = \widehat{\theta_i}$, so that p_i is the observed proportion of successes for the i^{th} combination of levels of the predictor variables. Let $\hat{\pi}_i$ be the MLE of π_i under the proposed model, i.e., $\hat{\pi}_i = g^{-1}(\sum_{j=1}^{p} x_{ij}\hat{\beta}_j)$. We can then compute the deviance as:

$$D = 2\left(ln\left[L_s\left(\hat{\beta}\right)\right] - ln\left[L_m(\hat{\beta})\right]\right)$$

$$= 2[\sum_{i=1}^{n} y_i \, ln \, p_i + \sum_{i=1}^{n}(m_i - y_i) \, ln \, (1 - p_i) -$$

$$\sum_{i=1}^{n} y_i \, ln \, \hat{\pi}_i - \sum_{i=1}^{n}(m_i - y_i) \, ln(1 - \hat{\pi}_i)]$$

$$= 2\sum_{i=1}^{n}\left[y_i \, ln\left(\frac{p_i}{\hat{\pi}_i}\right) + (m_i - y_i) \, ln\left(\frac{1-p_i}{1-\hat{\pi}_i}\right)\right]. \quad (4.11)$$

For Y_i~Binomial(m_i, π_i) data, the Pearson chi-squared statistic is normally defined as:

$$\chi^2 \equiv \sum_{i=1}^{n} \frac{(y_i - m_i\pi_i)^2}{m_i\pi_i(1-\pi_i)} \quad (4.12)$$

where Y_i is the number of successes for the i^{th} combination of levels of the predictor variables $i = 1,2, \dots, n$.

4.5 Diagnostic Tools for Normal Data

Suppose we have that $Y_i \sim N(\mu_i, \sigma^2)$ for $i = 1, 2, \dots, n$ and consider the model

$$E(Y_i) = \mu_i = X_i^T \beta \quad (4.13)$$

where Y_i are independent. The log likelihood function can be computed as:

$$l(\beta; y) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2 - \frac{1}{2} n \log(2\pi\sigma^2). \quad (4.14)$$

For the saturated model, all μ_i can be different so β has n elements $\mu_1, \mu_2, \dots, \mu_n$. Then by differentiating $l(\beta;\ y)$ with respect to each μ_i, and solving the estimating equation we obtain that $\hat{\mu}_i = y_i$. Therefore, the maximum value of the log likelihood function for the saturated model is given by:

$$l(b_{max}; y) = -\frac{1}{2} n \log(2\pi\sigma^2). \quad (4.15)$$

For any other model with $p < n$, let

$$b = (X^T X)^{-1} X^T y \tag{4.16}$$

be the maximum likelihood estimator. The corresponding maximum value for log likelihood function is written as:

$$l(b; y) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - {X_i}^T b)^2 - \frac{1}{2} n \log(2\pi\sigma^2). \tag{4.17}$$

Therefore, the deviance can be computed as:

$$D = 2[l(b_{max}; y) - l(b; y)]$$

$$= 2\left(-\frac{1}{2} n \log(2\pi\sigma^2) - [-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - {X_i}^T b\right)^2 - \frac{1}{2} n \log(2\pi\sigma^2)]\right)$$

$$= \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i - {X_i}^T b\right)^2$$

$$= \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 \tag{4.18}$$

where $\hat{\mu}_i$ denoted the fitted value ${X_i}^T b$.

CHAPTER 5

Numerical Examples

In this chapter we consider three different data sets, Default, Wage and UScrime. The fist data set is assumed to come from the binomial family, the second one from multinomial family and the third one from the Poisson family. For each data set, first we fit the model to all the data set then we identify the predictor that is not a statistically significant, and we remove it from the model. Then we fit the model again to determine the most significant predictors, and we have a better fit to the data. The parameter estimates are obtained through the R program. These data set are taken from R Documentation.

5.1 Binomial Data

Usage: Default (Credit Card Default Data). Description: A simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. The data frame with 10,000 observations on the following four variables. Default: A factor with levels No and Yes indicating whether the customer defaulted on their debt. Student: A factor with levels No and Yes indicating whether the customer is a student. Balance: The average balance that the customer has remaining on their credit card after making their monthly payment. Income: Income of customer.

The five number summary is a set of descriptive statistics that provide information about a dataset. It consists of the five most important sample percentiles: Min: the sample minimum, 1st Qu: the first quartile, Median: the middle value, 3rd Qu.: the third quartile and Max: the sample maximum. In the following Table 5.1 we presented the result of the five number summary for Balance and Income in the Default data set.

Table 5.1: Summary Statistics for Balance and Income in the Default data set

	Min.	1st Qu.	Median	3rd Qu.	Max.
Balance	0.0	481.7	823.6	1166.3	43808
Income	772	21340	34553	43808	73554

Table 5.2: Summary Statistics for Default and Student in the Default data set

	No	Yes	%
Default	9667	333	0.0333
Student	7056	2944	0.2944

Table 5.2 shows that over 10,000 observations there are only 333 customers whom are defaulted to their debt, and 2,944 of the customers are students. Now we fit this data to the model then the logistic regression model can be written as:

$$log\frac{Pr[Default = Yes]}{Pr[Default = No]} = -10.87 - 0.6468x_1 + 0.00573x_2 + 0.000003x_3$$

The following Table 5.3 presents the parameter estimates and the standard error from the logistic regression model for all variables in the Default data set.

Table 5.3: Parameter Estimates from Full Logistic Regression Model

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	$-1.087e+01$	$4.923e-01$	-22.080	$2e-16$ ***
Student (Yes)	$-6.468e-01$	$2.363e-01$	-2.738	0.00619 **
Balance	$5.737e-03$	$2.319e-04$	24.738	$2e-16$ ***
Income	$3.033e-06$	$8.203e-06$	0.370	0.71152

Table 5.3 gives us that the null deviance is 2,920.6 on 9,999 degrees of freedom, and the residual deviance is 1,571.5 on 9,996 degrees of freedom. The significance codes are simply categorizations of the p-value. The coefficient that is marked $(***)$ has a p-value < 0.001, and the coefficient that is marked $(**)$ has p-value < 0.01. Also, we see that the income is not a statistically significant predictor, and so we remove it from the model. Then we fit the model again to determine most significant predictors. Now after we fit the model to the reduced data then the logistic regression model can be written as:

$$log\frac{Pr[Default = Yes]}{Pr[Default = No]} = -10.75 - 0.7149x_1 + 0.00538x_2$$

The following Table 5.4 presents the parameter estimates and the standard error from the logistic regression model for variables in the Default data set after removing the income predictor.

Table 5.4: Parameter Estimates from Reduced Logistic Regression Model

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	$-1.075e+01$	$3.692e-01$	-29.116	$2e-16$ ***
Student (Yes)	$-7.149e-01$	$1.475e-01$	-4.846	$1.26e-06$ ***
Balance	$5.738e-03$	$2.318e-04$	24.750	$2e-16$ ***

Here in Table 5.4 we have that the null deviance is 2,920.6 on 9,999 degrees of freedom, and the residual deviance is 1,571.7 on 9,997 degrees of freedom. The significance codes are simply categorizations of the p-value. The coefficient that is marked $(***)$ has a p-value < 0.001. We know that the large deviance or chi-square value implies poorly fitted observations with respect to the model. So, we see that there is no big change in the null deviance and the residual deviance after we have removed the income, and hence, we conclude that the model fits the data, and the reduced model with Student status and Balance predictors fit the data well.

5.2 Multinomial Data

Usage: Wage. Description: Wage and other data for a group of 3000 workers in the Mid-Atlantic region. The data frame with 3000 observations on the following 12 variables. Year: Year that wage information was recorded. Age: Age of worker. Sex: Gender. Maritl: A factor with levels 1. Never Married 2. Married 3. Widowed 4. Divorced and 5. Separated indicating marital status. Race: A factor with levels 1. White 2. Black 3. Asian and 4. Other indicating race. Education: A factor with levels 1. < HS Grad 2. HS Grad 3. Some College 4. College Grad and 5. Advanced Degree indicating education level. Region: Region of the country (mid-Atlantic only). Jobclass: A factor with levels 1. Industrial and 2. Information indicating type of job. Health: A factor with levels 1. Good and 2. Very Good indicating health level of worker. Health_ins: A factor with levels 1. Yes, and 2. No indicating whether worker has health insurance. Logwage: Log of workers' wage. Wage: Workers raw wage.

The five number summary is a set of descriptive statistics that provide information about a dataset. It consists of the five most important sample percentiles: Min: the sample minimum, 1st Qu: the first quartile, Median: the middle value, 3rd Qu.: the third quartile and Max: the sample maximum. In the following Table 5.5 we present the result of the five number summary for Year, Age, Logwage and Wage predictors in the Wage data set.

Table 5.5: Summary Statistics for some Variables in Wage Data Set

	Min.	1st Qu.	Median	3rd Qu.	Max.
Year	2003	2004	2006	2008	2009
Age	18.00	33.75	42.00	51.00	80.00
Logwage	3.000	4.447	4.653	4.857	5.763
Wage	20.09	85.38	104.92	128.68	318.34

The following Table 5.6 gives a summary statistics for Education, Maritl, Race, Sex, Jobclass, Health, Health_ins and Region predictors in the Wage data set.

Table 5.6 Summary Statistics for some Variables in Wage Data Set

	< HS Grad	HS Grad	Some College	College Grad	Advanced Degree
Education	268	971	650	685	426
	Separated	Divorced	Widowed	Married	Never Married
Maritl	55	204	19	2074	648
	White	Black	Asian		Other
Race	2480	293	190		37
	Male	Female			
Sex	3000	0			
	Industrial	Information			
Jobclass	1544	1456			
	Good	Very Good			
Health	858	2142			
	Yes	No			
Health_ins	2083	917			
	Middle Atlantic	Other			
Region	3000	0			

Table 5.6 shows that the group of 3,000 male workers in the Middle Atlantic region where 2,142 of them have very good health. Now we fit this data to the model, then the multinomial logistic regression model can be written as:

$$log[Wage] = -42.18179 - 0.18561x_1 - 0.01716x_2 - 1.46198x_3 - 3.49564x_4 -$$
$$1.27917x_5 - 2.50237x_6 - 0.53547x_7 - 0.41331x_8 + 0.68373x_9 -$$
$$1.53915x_{10} - 2.38319x_{11} - 0.77893x_{12} + 4.86794x_{13} +$$
$$0.64614x_{14} - 0.18688x_{15} + 4.41498x_{16} + 113.26508x_{17}$$

The following Table 5.7 presents the parameter estimates and the standard error from the multinomial logistic regression model for all variables in the Wage data set.

Table 5.7: Parameter Estimates from Full Multinomial Logistic Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-42.18179	229.31872	-0.184	0.8541
Year	-0.18561	0.11449	-1.621	0.1051
Age	-0.01716	0.02317	-0.740	0.4592
Maritl 2. Married	-1.46198	0.65235	-2.241	0.0251 *
Maritl 3.Widowed	-3.49564	2.96917	-1.177	0.2392
Maritl 4.Divorced	-1.27917	1.07126	-1.194	0.2325
Maritl 5.Separated	-2.50237	1.79938	-1.391	0.1644
Race 2. Black	-0.53547	0.79644	-0.672	0.5014
Race 3. Asian	-0.41331	0.96561	-0.428	0.6687
Race 4. Other	0.68373	2.10209	0.325	0.7450
Education 2.HS Grad	-1.53915	0.88136	-1.746	0.0809
Education 3.Some College	-2.38319	0.94682	-2.517	0.0119 *
Education 4. College Grad	-0.77893	0.97362	-0.800	0.4238
Education 5. Advanced Degree	4.86794	1.10266	4.415	$1.05e-05$ ***

Jobclass 2. Information	0.64614	0.49152	1.315	0.1887
Health 2. Very Good	-0.18688	0.52922	-0.353	0.7240
Health_ins 2. No	4.41498	0.54460	8.107	$7.51e-16$ ***
Logwage	113.26508	0.82815	136.769	$2e-16$ ***

Table 5.7 shows that the null deviance is 5,222,086 on 2,999 degrees of freedom, and the residual deviance is 474,161 on 2,982 degrees of freedom. The significance codes are simply categorizations of the p-value. The coefficient that is marked (***) has a p-value < 0.001, and the coefficient that is marked (*) has p-value < 0.05. Also, we see that the race is not a statistically significant predictor, and so we remove it from the model. Then we fit the model again to determine most significant predictors. Now after we fit the model to the reduced data then the multinomial logistic regression model can be written as:

$$log[Wage] = -36.39061 - 0.18858x_1 - 0.01696x_2 - 1.43532x_3 - 3.52124x_4 - 1.22065x_5 - 2.53542x_6 - 1.55223x_7 - 2.42072x_8 - 0.81486x_9 + 4.80972x_{10} + 0.60658x_{11} + 4.41400x_{12} + 113.25963x_{13}$$

The following Table 5.8 presents the parameter estimates and the standard error from the multinomial regression model for all variables in the Wage data set after removing the race predictor.

Table 5.8: Parameter Estimates from Reduced Multinomial Logistic Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-36.39061	229.02321	-0.159	0.8738
Year	-0.18858	0.11434	-1.649	0.0992
Age	-0.01696	0.02270	-0.747	0.4551
Maritl 2. Married	-1.43532	0.64770	-2.216	0.0268 *
Maritl 3. Widowed	-3.52124	2.96532	-1.187	0.2351
Maritl 4. Divorced	-1.22065	1.06829	-1.143	0.2533
Maritl 5. Separated	-2.53542	1.79748	-1.411	0.1585
Education 2. HS Grad	-1.55223	0.87877	-1.766	0.0774
Education 3. Some College	-2.42072	0.94363	-2.565	0.0104 *
Education 4. College Grad	-0.81486	0.96596	-0.844	0.3990
Education 5. Advanced Degree	4.80972	1.09217	4.404	$1.10e-05$ ***
Jobclass 2. Information	0.60658	0.48774	1.244	0.2137
Health_ins 2. No	4.41400	0.54379	8.117	$6.91e-16$ ***
Logwage	113.25963	0.82307	137.606	$2e-16$ ***

Here in Table 5.8 we have that the null deviance is 5,222,086 on 2,999 degrees of freedom, and the residual deviance is 474,297 on 2,986 degrees of freedom. If the null deviance is small, it means that the null model explains the data very good. Likewise, with the residual deviance. Since the null deviance shows how well the response variable is predicted by a model that includes only the intercept whereas residual with inclusion of independent variables. The significance codes are simply categorizations of the p-value. The coefficient that is marked (***) has a p-value <

0.001, and the coefficient that is marked $(*)$ has p-value < 0.05. So, we see that there is no big change in the null deviance, or the residual deviance after we have removed the race, and we know that the higher numbers in deviance always indicate a bad fit. Hence, we conclude that the values of deviance in the reduced model imply a poor fit.

5.3 Poisson Data

Usage: (UScrime the effect of punishment regimes on crime rates). Description: Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1,960 given in this data frame. The variables seem to have been re-scaled to convenient numbers. The data frame contains the following 16 variables.

M: Percentage of males aged 14–24. So: Indicator variable for a Southern state. Ed: Mean years of schooling. Po1: Police expenditure in 1960.Po2: Police expenditure in 1,959.LF: Labour force participation rate. M. F: Number of males per 1,000 females. Pop: State population. NW: Number of non-whites per 1,000 people. U1: Unemployment rate of urban males 14–24. U2: Unemployment rate of urban males 35–39.GDP: Gross domestic product per head. Ineq: Income inequality. Prob: Probability of imprisonment. Time: Average time served in state prisons. y: Rate of crimes in a particular category per head of population.

The five number summary is a set of descriptive statistics that provide information about a data set. It consists of the five most important sample percentiles: Min: the sample minimum, 1st Qu: the first quartile, Median: the middle value, 3rd Qu.: the third quartile and Max: the sample maximum. In the following Table 5.9 we present the result of the five number summary for all variables in the UScrime data set: the effect of punishment regimes on crime rates.

Table 5.9: Summary Statistics for all Variables in UScrime Data Set

Now we fit this data to the model, then the Poisson regression model can be written as:

	M	SO	ED	PO1	PO2	LF	M.F	Pop
Min:	119.0	0.000	87.0	45.0	41.00	480.0	934.0	3.0
1st Qu:	130.0	0.000	97.5	62.5	58.50	530.5	964.5	10.0
Median:	136.0	0.000	108.0	78.0	73.00	560.0	977.0	25.0
3rd Qu:	146.0	1.000	114.5	104.5	97.00	593.0	992.0	41.50
Max:	177.0	1.000	122.0	166.0	157.00	641.0	1071.0	168.0
	NW	U1	U2	GDP	Ineq	Prob	Time	y
Min:	2.0	70.0	20.0	288.0	126.0	0.00690	12.20	342.0
1st Qu:	24.0	80.5	27.5	459.5	165.5	0.03270	21.60	658.5
Median:	76.0	92.00	34.00	357.0	176.0	0.04210	25.80	831.0
3rd Qu:	132.5	104.00	38.50	591.5	227.5	0.05445	30.45	1057.5
Max:	423.0	142.00	58.00	689.0	276.0	0.11980	44.00	1993.0

$$\log(y) = 0.6242 + 0.008694x_1 + 0.07387x_2 + 0.01954x_3 + 0.01600x_4$$
$$- 0.008322x_5 - 0.00002354x_6 - 0.0003372x_7 + 0.001465x_8$$
$$+ 0.0004990x_9 - 0.004587x_{10} + 0.01526x_{11} + 0.002021x_{12}$$
$$+ 0.008889x_{13} - 6.333x_{14} - 0.003324x_{15}$$

In the following Table 5.10 we present the parameter estimates and the standard error from the Poisson regression model for all variables in the UScrime: the effect of punishment regimes on crime rates.

Table 5.10: Parameter Estimates from Full Poisson Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	$6.242e-01$	$2.424e-01$	2.576	$0.01001\ *$
M	$8.694e-03$	$6.702e-04$	12.973	$2e-16\ ***$
So	$7.387e-02$	$2.363e-02$	3.126	$0.00177\ **$
Ed	$1.954e-02$	$1.054e-03$	18.532	$2e-16\ ***$
Po1	$1.600e-02$	$1.641e-03$	9.746	$2e-16\ ***$
Po2	$-8.322e-03$	$1.855e-03$	-4.487	$7.24e-06\ ***$
LF	$-2.354e-05$	$2.334e-04$	-0.101	0.91967
M.F	$-3.372e-04$	$3.344e-04$	-1.008	0.31326
Pop	$-1.465e-03$	$1.872e-04$	-7.826	$5.05e-15\ ***$
NW	$4.990e-04$	$1.079e-04$	4.626	$3.73e-06\ ***$
U1	$-4.587e-03$	$6.762e-04$	-6.783	$1.18e-11\ ***$
U2	$1.526e-02$	$1.305e-03$	11.694	$2e-16\ ***$
GDP	$2.021e-03$	$1.650e-04$	12.251	$2e-16\ ***$
Ineq	$8.889e-03$	$3.717e-04$	23.916	$2e-16\ ***$
Prob	$-6.333e+00$	$3.814e-01$	-16.603	$2e-16\ ***$
Time	$-3.324e-03$	$1.178e-03$	-2.823	$0.00476\ **$

In the above Table 5.10 we have the null deviance is 7,071.1 on 46 degrees of freedom, and the residual deviance is 1,329.9 on 31 degrees of freedom. The significance codes are simply categorizations of the p-value. The coefficient that is marked $(***)$ has a p-value < 0.001, the coefficient that is marked $(**)$ has p-value < 0.01, and the coefficient that is marked $(*)$ has p-value < 0.05. Also, we see that the LF (Labour force participation rate) is not a statistically

significant predictor, and so we remove it from the model. Then we fit the model again to determine most significant predictors. Now after we fit the model to the reduced data then the Poisson regression model can be written as:

$$\log(y) = 0.6266730 + 0.0087032x_1 + 0.0750647x_2 + 0.0195118x_3 +$$
$$+ 0.0159639x_4 - 0.0082732x_5 - 0.0003543x_6 - 0.0014681x_7 +$$
$$+ 0.0004952x_8 - 0.0045564x_9 + 0.0152509x_{10} + 0.0020202x_{11} +$$
$$+ 0.0088851x_{12} - 6.3292214x_{13} - 0.0033125x_{14}$$

The following Table 5.11 we present the parameter estimates and the standard error from the Poisson regression model for all variables in UScrime data set after removing LF.

Table 5.11: Parameter Estimates from Reduced Poisson Regression Model

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	0.6266730	0.2411420	2.599	0.009356 **
M	0.0087032	0.0006638	13.111	$2e-16$ ***
So	0.0750647	0.0204509	3.670	0.000242 ***
Ed	0.0195118	0.0010196	19.136	$2e-16$ ***
Po1	0.0159639	0.0016091	9.921	$2e-16$ ***
Po2	-0.0082732	0.0017900	-4.622	$3.80e-06$ ***
M.F	-0.0003543	0.0002881	-1.230	0.218767
Pop	-0.0014681	0.0001843	-7.967	$1.63e-15$ ***
NW	0.0004952	0.0001011	4.899	$9.61e-07$ ***
U1	-0.0045564	0.0006068	-7.509	$5.97e-14$ ***
U2	0.0152509	0.0013018	11.715	$2e-16$ ***

GDP	0.0020202	0.0001646	12.276	$2e-16$ ***
Ineq	0.0088851	0.0003699	24.023	$2e-16$ ***
Prob	-6.3292214	0.3798968	-16.660	$2e-16$ ***
Time	-0.0033125	0.0011720	-2.826	0.004706 **

Table 5.11 shows that the null deviance is 7,071.1 on 46 degrees of freedom, and the residual deviance is 1,329.9 on 32 degrees of freedom. We know that if the null deviance is really small, then the null model explains the data pretty well. Likewise, with the residual deviance. The significance codes are simply categorizations of the p-value. The coefficient that is marked $(***)$ has a p-value < 0.001, and the coefficient that is marked $(**)$ has p-value < 0.01. Since the null deviance shows how well the response variable is predicted by a model that includes only the intercept whereas residual with inclusion of independent variables. So, we see that there is no big change in the null deviance, and the residual deviance after we have removed the LF, and we know that the higher numbers in deviance always indicate a bad fit. Hence, we conclude that the values of deviance in the reduced model imply a poor fit.

CHAPTER 6

Conclusion

The binomial credit card default was predicted using the logistic regression model. The income factor was found to be insignificant to predict default status, yet the other two factors (student status and balance) were highly significant. These variables have the opposite effect on Default status: the "Yes" student status subtracts the default status, whereas the positive balance adds to the status. The null deviance of the model with the reduced quantity of variables indicated that the student status and balance are reliable variables to predict default status.

The multinomial data presented the prediction of wage by 12 variables. The logistic regression showed that there were few factors that had influence on wage. These were marital status (married), college and advance degree, and absence of health insurance. When the analysis was run with only the significant variables, the null deviance indicated that the model fit data are not reliable. Therefore, the prediction of wage with the variables is not reliable, despite of the significant values of some coefficients.

The Poisson regression model for punishment effect on crime rates was studied on data set with 16 variables that covered 47 states. The first model approximation indicated that there are numerous significant predictors, including males' quantity (% aged 14-24), police expenditures, southern state population, etc. There were only two insignificant variables: labor force rage and number of males per 1000 of females. When the insignificant variables were removed, and the null deviance of the model was analyzed, the poor fit was indicated.

The prediction of binomial, multinomial, and Poisson data with the logistic regression model indicated that a good fit was obtained only for the binomial model. Possibly, the multinomial and

Poisson data models included too many predictor variables, which are the source of errors. The analysis of these data with smaller quantity of predictors might produce better results.

References

D. A. Williams. Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletion. *Journal of the Royal Statistical Society*, 36(2), c, (1987) pp. 181-191.

Dobson, A. J., & Barnett, A. G. (2008). An introduction to generalized linear models. Boca Raton: CRC Press.

Ehrlich, I. (1973) Participation in illegitimate activities: a theoretical and empirical investigation. *Journal of Political Economy*, 81, 521–565.

E. L. Frome. The Analysis of Rates Using Poisson Regression Models: Medical and Health Sciences Division. Oak Associated Universities, Oak Ridge, Tennessee.

Games, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) *An Introduction to Statistical Learning with applications in R*,www.StatLearning.com, Springer-Verlag, New York

Gill, J. (2001). Generalized linear models: A unified approach. Thousand Oaks, CA: Sage Publications.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2005). Introduction to mathematical statistics. Upper Saddle River, NJ: Pearson Education.

Myers, R. H., Montgomery, D. C., & Vining, G. G. (2002). Generalized linear models: With applications in engineering and the sciences. New York: J. Wiley.

Rencher, A. C. (2000). Linear Models in Statistics. New York: Wiley.

Ross, S. M. (1976). A first course in probability. New York: Macmillan.

Vandaele, W. (1978) Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation*, eds A. Blumstein, J. Cohen and D. Nagin, pp. 270–335. US National Academy of Sciences.

Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS.* Third Edition.

Springer.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (1996). Mathematical statistics with applications. Belmont, CA: Duxbury Press.

Appendix

```
[Workspace loaded from ~/.R Data]
> Library (MASS)
> Library (ISLR)
```

1- GLM for the binomial family

```
> Data ("Default")
> Head (Default)

  Default  Student   Balance     Income
1 No       No        729.5265    44361.625
2 No       Yes       817.1804    12106.135
3 No       No        1073.5492   31767.139
4 No       No        529.2506    35704.494
5 No       No        785.6559    38463.496
6 No       Yes       919.5885    7491.559

> Attach (Default)
> mod.bin1 = glm (default ~ student + balance + income, family = binomial ("logit"),
data = Default)
> Summary (mod.bin1)
Call:
glm (formula = default ~ student + balance + income, family = binomial ("logit"),
   Data = Default)

Deviance Residuals:
 Min        1Q        Median      3Q        Max
-2.4691    -0.1418   -0.0557     -0.0203   3.7383
```

Coefficients:

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-1.087e+01	4.923e-01	-22.080	2e-16 ***
Student Yes	-6.468e-01	2.363e-01	-2.738	0.00619 **
Balance	5.737e-03	2.319e-04	24.738	2e-16 ***
Income	3.033e-06	8.203e-06	0.370	0.71152

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1571.5 on 9996 degrees of freedom

AIC: 1579.5

Number of Fisher Scoring iterations: 8

> mod.bin2 = glm (default ~ student + balance, family = binomial ("logit"), data = Default)

> Summary (mod.bin2)

Call:

glm (formula = default ~ student + balance, family = binomial ("logit"), Data = Default)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.4578	-0.1422	-0.0559	-0.0203	3.7435

Coefficients:

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	-1.075e+01	3.692e-01	-29.116	2e-16 ***
Student Yes	-7.149e-01	1.475e-01	-4.846	1.26e-06 ***
Balance	5.738e-03	2.318e-04	24.750	2e-16 ***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2920.6 on 9999 degrees of freedom

Residual deviance: 1571.7 on 9997 degrees of freedom

AIC: 1577.7

Number of Fisher Scoring iterations: 8

2- GLM for the multinomial family

> Data ("Wage")

> Head (Wage)

Year age sex Maritl race education region Jobclass health Health_ins Logwage wage

231655 2006 18 1. Male 1. Never Married 1. White 1. < HS Grad 2. Middle Atlantic 1. Industrial 1. <=Good 2. No 4.318063 75.04315

86582 2004 24 1. Male 1. Never Married 1. White 4. College Grad 2. Middle Atlantic 2. Information 2. >=Very Good 2. No 4.255273 70.47602

161300 2003 45 1. Male 2. Married 1. White 3. Some College 2. Middle Atlantic 1. Industrial 1. <=Good 1. Yes 4.875061 130.98218

155159 2003 43 1. Male 2. Married 3. Asian 4. College Grad 2. Middle Atlantic 2. Information 2. >=Very Good 1. Yes 5.041393 154.68529

11443 2005 50 1. Male 4. Divorced 1. White 2. HS Grad 2. Middle Atlantic 2. Information 1. <=Good 1. Yes 4.318063 75.04315

376662 2008 54 1. Male 2. Married 1. White 4. College Grad 2. Middle Atlantic 2. Information 2. >=Very Good 1. Yes 4.845098 127.11574

> attach (Wage)

> lm (wage~ year + age, data = Wage)

Call:

lm (formula = wage ~ year + age, data = Wage)

Coefficients:

(Intercept) year age

-2318.5309 1.1968 0.6992

> mod.mult1 = glm (wage ~ year + age + maritl + race + education + jobclass + health + health_ins + logwage, data=Wage)

> Summary (mod.mult1)

Call:

glm (formula = wage ~ year + age + maritl + race + education + jobclass + health + health_ins + logwage, data = Wage)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-16.721	-5.359	-3.058	0.712	94.152

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-42.18179	229.31872	-0.184	0.8541
Year	-0.18561	0.11449	-1.621	0.1051
Age	-0.01716	0.02317	-0.740	0.4592
Maritl 2. Married	-1.46198	0.65235	-2.241	0.0251 *
Maritl 3. Widowed	-3.49564	2.96917	-1.177	0.2392
Maritl 4. Divorced	-1.27917	1.07126	-1.194	0.2325
Maritl 5. Separated	-2.50237	1.79938	-1.391	0.1644
Race 2. Black	-0.53547	0.79644	-0.672	0.5014
Race 3. Asian	-0.41331	0.96561	-0.428	0.6687
Race 4. Other	0.68373	2.10209	0.325	0.7450
Education 2. HS Grad	-1.53915	0.88136	-1.746	0.0809
Education 3. Some College	-2.38319	0.94682	-2.517	0.0119 *
Education 4. College Grad	-0.77893	0.97362	-0.800	0.4238
Education 5. Advanced Degree	4.86794	1.10266	4.415	1.05e-05 ***
Jobclass 2. Information	0.64614	0.49152	1.315	0.1887
Health 2. >=Very Good	-0.18688	0.52922	-0.353	0.7240
Health_ins 2. No	4.41498	0.54460	8.107	7.51e-16 ***
Logwage	113.26508	0.82815	136.769	2e-16 ***

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for Gaussian family taken to be 159.0076)

Null deviance: 5222086 on 2999 degrees of freedom

Residual deviance: 474161 on 2982 degrees of freedom

AIC: 23740

Number of Fisher Scoring iterations: 2

> mod.mult2 = glm (wage ~ year + age + maritl + education + jobclass + health_ins + logwage, data = Wage)

> Summary (mod.mult2)

Call:

glm (formula = wage ~ year + age + maritl + education + jobclass + health_ins + logwage, data = Wage)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-16.624	-5.369	-3.104	0.726	94.368

Coefficients:

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-36.39061	229.02321	-0.159	0.8738
Year	-0.18858	0.11434	-1.649	0.0992
Age	-0.01696	0.02270	-0.747	0.4551
Maritl 2. Married	-1.43532	0.64770	-2.216	0.0268 *
Maritl 3. Widowed	-3.52124	2.96532	-1.187	0.2351
Maritl 4. Divorced	-1.22065	1.06829	-1.143	0.2533
Maritl 5. Separated	-2.53542	1.79748	-1.411	0.1585
Education 2. HS Grad	-1.55223	0.87877	-1.766	0.0774
Education 3. Some College	-2.42072	0.94363	-2.565	0.0104 *
Education 4. College Grad	-0.81486	0.96596	-0.844	0.3990
Education 5. Advanced Degree	4.80972	1.09217	4.404	1.10e-05 ***
Jobclass 2. Information	0.60658	0.48774	1.244	0.2137
Health_ins 2. No	4.41400	0.54379	8.117	6.91e-16 ***
Logwage	113.25963	0.82307	137.606	2e-16 ***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1

(Dispersion parameter for Gaussian family taken to be 158.8401)

Null deviance: 5222086 on 2999 degrees of freedom

Residual deviance: 474297 on 2986 degrees of freedom

AIC: 23733

Number of Fisher Scoring iterations: 2

3- GLM for the Poisson family

> Data (UScrime)

> Attach (UScrime)

> Head (UScrime)

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time	y
1	151	1	91	58	56	510	950	33	301	108	41	394	261	0.084602	26.2011	791
2	143	0	113	103	95	583	1012	13	102	96	36	557	194	0.029599	25.2999	1635
3	142	1	89	45	44	533	969	18	219	94	33	318	250	0.083401	24.3006	578
4	136	0	121	149	141	577	994	157	80	102	39	673	167	0.015801	29.9012	1969
5	141	0	121	109	101	591	985	18	30	91	20	578	174	0.041399	21.2998	1234
6	121	0	110	118	115	547	964	25	44	84	29	689	126	0.034201	20.9995	682

> Summary (UScrime)

M	So	Ed	Po1	Po2	LF	M.F
Min:119.0	Min: 0.0000	Min:87.0	Min:45.0	Min:41.00	Min:480.0	Min:934.0
1Q:130.0	1Q:0.0000	1Q:97.5	1Q:62.5	1Q:58.50	1Q:530.5	1Q:964.5
Median:136.0	Median:0.0000	Median:108.0	Median:78.0	Median:73.00	Median:560.0	Median:977.0
3Q:146.0	3Q:1.0000	3Q:114.5	3Q:104.5	3Q: 97.00	3Q:593.0	3Qu: 992.0
Max:177.0	Max:1.0000	Max:122.0	Max:166.0	Max:157.00	Max:641.0	Max:1071.0

Pop	NW	U1	U2	GDP	Ineq	Prob
Min:3.00	Min:2.0	Min:70.00	Min:20.00	Min:288.0	Min:126.0	Min:0.00690
1Q: 10.00	1Q: 24.0	1Q: 80.50	1Q:27.50	1Q:459.5	1Q:165.5	1Q:0.03270
Median: 25.00	Median: 76.0	Median: 92.00	Median: 34.00	Median: 537.0	Median: 176.0	Median: 0.04210
3Q:41.50	3Q:132.5	3Q:104.00	3Q:38.50	3Q:591.5	3Q:227.5	3Q:0.05445
Max:168.00	Max:423.0	Max:142.00	Max:58.00	Max:689.0	Max:276.0	Max:0.11980

Time	y
Min:12.20	Min: 342.0
1Q:21.60	1Q: 658.5
Median:25.80	Median: 831.0
3Q:30.45	3Q:1057.5
Max:44.00	Max:1993.0

> mod.Pois1 = glm (formula = y ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + GDP + Ineq + Prob + Time, family = Poisson, data = UScrime)

> Summary (mod.Pois1)

Call:

glm (formula = y ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + GDP + Ineq + Prob + Time, family = Poisson, data = UScrime)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.2919	-3.4129	-0.0613	3.8498	13.6840

Coefficients:

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	6.242e-01	2.424e-01	2.576	0.01001 *
M	8.694e-03	6.702e-04	12.973	2e-16 ***
So	7.387e-02	2.363e-02	3.126	0.00177 **
Ed	1.954e-02	1.054e-03	18.532	2e-16 ***
Po1	1.600e-02	1.641e-03	9.746	2e-16 ***
Po2	-8.322e-03	1.855e-03	-4.487	7.24e-06 ***
LF	-2.354e-05	2.334e-04	-0.101	0.91967
M.F	-3.372e-04	3.344e-04	-1.008	0.31326
Pop	-1.465e-03	1.872e-04	-7.826	5.05e-15 ***
NW	4.990e-04	1.079e-04	4.626	3.73e-06 ***
U1	-4.587e-03	6.762e-04	-6.783	1.18e-11 ***
U2	1.526e-02	1.305e-03	11.694	2e-16 ***

GDP 2.021e-03 1.650e-04 12.251 2e-16 ***

Ineq 8.889e-03 3.717e-04 23.916 2e-16 ***

Prob -6.333e+00 3.814e-01 -16.603 2e-16 ***

Time -3.324e-03 1.178e-03 -2.823 0.00476 **

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for Poisson family taken to be 1)

Null deviance: 7071.1 on 46 degrees of freedom

Residual deviance: 1329.9 on 31 degrees of freedom

AIC: 1764.4

Number of Fisher Scoring iterations: 4

> mod.Pois2 = glm (formula = y ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + GDP + Ineq + Prob + Time, family = Poisson, data = UScrime)

> Summary (mod.Pois2)

Call:

glm (formula = y ~ M + So + Ed + Po1 + Po2 + M.F + Pop + NW + U1 + U2 + GDP + Ineq + Prob + Time, family = Poisson, data = UScrime)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-12.2954	-3.4120	-0.0398	3.8477	13.6846

Coefficients:

	Estimate	Std. Error	z value	Pr (>|z|)
(Intercept)	0.6266730	0.2411420	2.599	0.009356 **
M	0.0087032	0.0006638	13.111	2e-16 ***
So	0.0750647	0.0204509	3.670	0.000242 ***
Ed	0.0195118	0.0010196	19.136	2e-16 ***
Po1	0.0159639	0.0016091	9.921	2e-16 ***
Po2	-0.0082732	0.0017900	-4.622	3.80e-06 ***

M.F	-0.0003543	0.0002881	-1.230	0.218767
Pop	-0.0014681	0.0001843	-7.967	1.63e-15 ***
NW	0.0004952	0.0001011	4.899	9.61e-07 ***
U1	-0.0045564	0.0006068	-7.509	5.97e-14 ***
U2	0.0152509	0.0013018	11.715	2e-16 ***
GDP	0.0020202	0.0001646	12.276	2e-16 ***
Ineq	0.0088851	0.0003699	24.023	2e-16 ***
Prob	-6.3292214	0.3798968	-16.660	2e-16 ***
Time	-0.0033125	0.0011720	-2.826	0.004706 **

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

(Dispersion parameter for Poisson family taken to be 1)

Null deviance: 7071.1 on 46 degrees of freedom

Residual deviance: 1329.9 on 32 degrees of freedom

AIC: 1762.4

Number of Fisher Scoring iterations: 4