Generalized Linear Models and Extensions

Lang Wu

Department of Statistics
University of British Columbia, Vancouver

2019

Contents

1 Introduction

Regression models are perhaps the most widely used statistical models. In most studies, data are usually collected on more than one variable. Depending on the objectives of the study, one variable may be chosen as a *response* (or *dependent variable*) and some other variables may be treated as possible *covariates*, which are also called *explanatory variables* or *independent variables* or *predictors*. A main goal is to understand the variation in the response variable, which is of primary interest. Much of the variation in the response may be explained (or predicted) by the covariates (or predictors). Sometimes we may simply wish to study the approximation relationship between the response and covariates. Note that, the *true relationship* between a response and covariates is typically unknown and it can be highly complicated. In most regression analyses, we only try to *approximate* this relationship. In this section, we provide a brief overview of the essential ideas of regression models.

In a regression model, covariates are used to partially explain the *systematic variation* in the response. The remaining unexplained variation in the response is treated as random and is often assumed to follow a probability distribution – it reflects the uncertainty of the response aside from its systematic parts. This probability distribution is usually chosen based on the *type of the response variable*. In the systematic part of a regression model, the mean response is often assumed to link the covariates through a linear combination of the covariates, called a *linear predictor*. The resulting model is called a generalized linear model. The use of a linear predictor in a regression model is usually based on its simplicity and easy interpretation. In a nonlinear regression model or in a nonparametric or semiparametric regression model, however, the mean response may be linked to covariates in any nonlinear form. In a survival regression model, we may link the covariates to the hazard rather than the mean response.

The type of a regression model is usually determined by the type of the response variable. For example, if the response is a *continuous* variable, we may consider a *linear regression model* with normally distributed random errors. If the response is a *binary* variable, we may consider a *logistic regression model* with the random error following a binomial distribution. If the response is a *count*, we may consider a *Poisson regression model* with a Poisson random error. If the response is the *time to an event* of interest (e.g., time to death), we may consider a *survival regression model* with the random error following a Weibull distribution. In summary, the following types of regression models are commonly used in practice:

- linear models,

- generalized linear models,
- nonlinear models,
- survival regression models.

All these regression models can be extended to the analysis of longitudinal or clustered data.

In regression models, a main objective is to understand the dependence of the response on the covariates. Basic ideas and approaches of regression techniques apply to both cross-sectional data and longitudinal data. However, special considerations are needed for longitudinal or clustered data in order to incorporate the correlations within clusters.

In practice the *true* or exact relationship between a response and covariates may be very complicated. In a regression model, we often attempt to *approximate* this relationship using a simple and easy-to-interpret model, such as linear regression models or some generalized linear models. These simple models usually do not represent the true relationship between the responses and the covariates, even if the models fit the observed data well, and they are sometimes called empirical models. Prediction outside the observed-data range based on empirical models is often dangerous. These empirical models have been popular mainly because of their simplicity, which is especially important before modern computers become available.

In some cases, based on subject-area knowledge we may have a good understanding of the underlying mechanisms which generate the data. In these cases we may be able to derive the (approximate) true relationship between a response and covariates. The resulting models are often nonlinear, leading to *nonlinear regression models.* Therefore, nonlinear regression models are typically *mechanistic*, and they may represent the true (or approximately true) relationship between the response and the covariates. For this reason, nonlinear regression models often provide more reliable predictions than linear regression models. With the availability of modern computers, which greatly reduce computational burden, nonlinear models should be preferred if available.

In a broader sense, there are two general approaches for analyzing multivariate data. One approach is called *supervised learning*, where we treat one variable as a response and other variables as possible covariates. Regression models are examples of supervised learning. If more than one variables are treated as responses, we have multivariate regression models. The other approach is called *unsupervised learning*, where we treat all variables equally or symmetrically, i.e., no variables are treated as responses and no variables are treated as covariates. In unsupervised learning, the goal is to understand the underlying structures in

all the variables. Examples of unsupervised learning include *principal components analysis, cluster analysis*, *factor analysis*, *contingency tables*, and *graphical models*. In the analysis of longitudinal data, regression models receive more attention and are thus the focus of this book.

In the following subsections we provide a brief review of commonly used regression models for cross-sectional data and for longitudinal data.

2 Regression Models: An Overview

2.1 Linear Models

Regression models for cross-sectional data have been well developed. The most widely used regression models are probably *linear regression models*, where the relationship between the mean response and covariates is assumed to be linear and the random error is usually assumed to be normal. A comprehensive discussion of linear regression models can be found in Draper and Smith (1998) and Weisberg (2005), among others. Faraway (2004) discussed linear models using statistical software R, while Littell et al. (2002) illustrated statistical software SAS for linear models. In this section, we provide a brief overview of common regression models for cross-sectional data, with a focus on general concepts and approaches without technical details.

Suppose that there are n individuals in the sample, and the data on individual i are $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, where y_i is the response and x_{ij}'s are covariates. A general (multiple) *linear regression model* can be written as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, 2, \cdots, n, \\ \epsilon_i &\text{ i.i.d. } \sim N(0, \sigma^2), \end{aligned} \tag{1}$$

where β_j's are unknown regression parameters linking the covariates to the response, and ϵ_i is a random error representing residual variation in the data. The standard assumptions for linear model (1) are:

- the random errors $\{\epsilon_1, \epsilon_2, \cdots, \epsilon_n\}$ or the response $\{y_1, y_2, \cdots, y_n\}$ are independent;
- the random errors ϵ_i's have a zero mean and a constant variance;
- the random errors ϵ_i's are normally distributed.

The *independence* of the observations $\{y_1, y_2, \cdots, y_n\}$ is an important assumption for classical linear regression models.

Unknown parameters in model (1) can be estimated by

- the *least-square method*, or
- the *maximum likelihood method*.

The least-square method finds parameters to minimize the sum of squares of differences between the expected responses under the model and the observed responses. The likelihood method finds parameters to maximize the likelihood. For linear model (1), the two methods produce the same estimates.

Once the unknown parameters are estimated, model checking or *model diagnostics* should be performed to check the reasonability of the model and the assumptions, which can informally be based on *residual plots* and other graphical techniques. Variable transformations may be used to improve model fitting. *Outliers and influential observations* should also be checked since they may greatly affect the resulting estimates and may lead to misleading inference.

Model selection or variable selection can be based on standard statistical methods, such as the stepwise method and AIC/BIC criteria, as well as on scientific considerations. In general, *parsimonious models* are preferred since they may avoid potential collinearity in the predictors and may improve precision of the main parameter estimates.

In linear model (1), when all the covariates are *categorical* or *discrete*, the model is equivalent to an *analysis of variance (ANOVA)* model, which allows a specific decomposition of total variation into systematic part and random part. ANOVA models are often used in designed experiments. When some covariates are categorical and some covariates are continuous, model (1) is sometimes called an *analysis of covariance* model.

2.2 Extensions of Linear Models

Linear regression models have been widely used due to their simplicity, which is important in the pre-computer era since closed-form or analytic expressions of parameter estimates can be derived. However, linear models require strong assumptions such as linearity, and they may not be appropriate when the response is (say) categorical. Moreover, unlike nonlinear models, linear models usually do not describe the data-generating mechanisms and they often do not provide reliable prediction outside the observed data range. Therefore, extensions of linear models have received great attention in the last few decades, due partially to the developments of modern computers and computational tools. Linear models may be extended in two directions:

- *non-normal distributions* for the random errors;
- *nonlinear* relationships between the response and covariates.

These two extensions are briefly described below.

The first extension is to allow *non-normal distributions* for the responses or random errors. This is necessary for some responses whose distributions are clearly non-normal (even after transformations), such as binary responses taking only two possible values (say 0 or 1). A natural family of candidate distributions is the *exponential family*, which includes normal distributions, binomial distributions, Poisson distributions, and other distributions. For example, if the response of interest is an indicator of whether an individual has cancer or not, the response is a binary variable with only two possible values (say, 0 or 1). In this case, linear model (1) cannot be used since the covariates can take any real values and the response will not follow a normal distribution. However, we may assume that the response y_i follows a binomial (Bernoulli) distribution and consider the following special nonlinear regression model, called a *logistic regression model*

$$\log\left\{\frac{P(y_i=1)}{1-P(y_i=1)}\right\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \cdots, n, \tag{2}$$

in which the mean response $E(y_i) = P(y_i = 1)$ and the covariates are linked through a monotone function

$$g(y) = \log(y/(1-y)),$$

called a *logit link* function, and the response y_i is assumed to follow a binomial (Bernoulli) distribution.

More generally, we may assume that the response follows a distribution in the exponential family and then we link the mean response to the covariates via a linear predictor. The resulting model is called a *generalized linear model (GLM)*. Specifically, a GLM can be written as follows

$$g(E(y_i)) = \eta_i \equiv \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \qquad i = 1, 2, \cdots, n, \tag{3}$$

where $g(\cdot)$ is a known monotone *link function* and η_i is a *linear predictor*. Note that a GLM is a special nonlinear model.

When $g(y) = y$ (the identity link) and y follows a normal distribution, GLM (3) reduces to the standard normal linear model (1). When $g(y) = \log(y/(1-y))$ (the logit link) and y follows a binomial distribution, GLM (3) reduces to a logistic regression model. When $g(y) = \log(y)$ and y follows a Poisson distribution, GLM (3) reduces to a Poisson regression model for count response. For comprehensive discussions of GLMs, see McCullagh and Nelder (1989), Fahrmeir et al. (2001), and McCulloch et al. (2008). Faraway (2005) describes GLM using software R.

Another extension of linear regression models is to allow the response to link the covariates in any nonlinear forms, leading to *nonlinear regression models*. We focus on the

common class of nonlinear models in which the response and covariates may be linked in a nonlinear forms but the response or random error are assumed to be normal. A nonlinear regression model is often *mechanistic* in the sense that it usually describes or approximately describes the data-generating mechanism, i.e., the underlying mechanism which generates the observed data. Thus, nonlinear regression models often provide better predictions than linear regression models, and the parameters in nonlinear models often have natural physical interpretations.

A nonlinear regression model can be written as

$$y_i = h(x_{i1}, \cdots, x_{ip}, \boldsymbol{\beta}) + \epsilon_i, \qquad i = 1, 2, \cdots, n,$$

where $h(\cdot)$ is a nonlinear function and ϵ_i follows a normal distribution. Statistical inference for nonlinear regression models is more complex than that of linear models because closed-form expressions of parameter estimates are typically unavailable. Moreover, good choices of starting values are needed and are important for nonlinear models since the likelihoods may have multiple modes. For comprehensive discussions of nonlinear models, see Seber and Wild (2003) and Bates and Watts (2007).

2.3 Statistical Inference

The most common approach for statistical inference of regression models is the *maximum likelihood method*. The likelihood method is widely used because it is generally applicable to a wide variety of models and it offers nice asymptotic properties – the maximum likelihood estimates are consistent, asymptotically normal, and asymptotically most efficient under some regularity conditions. A drawback of the likelihood method is that it requires distributional assumptions. So likelihood inference may be sensitive to departures from the assumed distributions and sensitive to outliers.

For linear and nonlinear regression models with normal responses, another commonly used method is the *least square method*. The least square method does not require distributional assumption for the data. However, the least square estimates may not be optimal. Moreover, distributional assumptions may still be required for inference, such as hypothesis testing and confidence intervals.

Another robust approach is the so-called *quasi-likelihood method* and the closely related *generalized estimating equations (GEE)* method in which one only needs to specify the first two moments without distributional assumptions. However, the GEE estimates are less efficient than the likelihood estimates if the distributional assumptions hold.

3 Linear and Nonlinear Regression Models

3.1 Introduction

In the analysis of multivariate data, there are two general approaches. One approach is to treat each variable *equally*, and the goal is to understand the correlation structure between the variables or to reduce the dimension of the data space. Examples include principal component analysis, factor analysis, and cluster analysis. Another approach is to treat one variable as a *response* and the other variables as *predictors*, and the goal is to understand the variation in the response that can be partially explained by predictors. Such models are called *regression models*. Regression analysis is an important component of multivariate analysis, since it allows researchers to focus on the effects of predictors on the response. Regression analysis also is widely applied in actuarial science and finance. It is a required educational component of the two main actuarial bodies in the U.S. and Canada, the Society of Actuaries and the Casualty Actuarial Society.

Regression models attempt to partially explain the variation in the response by the predictors. In other words, regression models attempt to find the approximate relationship between the response and predictors. For example, we may wish to find the approximate relationship between income (response) and education, age, experience, gender, etc (predictors). Or we may wish to find the approximate relationship between success (response) and age, education, gender, IQ score, attitude, etc (predictors). A regression model is a useful statistical tool to determine such an approximate relationship.

In practice, the true relationship between the response and predictors may be highly complicated and may not be known exactly, but an approximation to the true relationship is possible, based on the observed data. When the response variable is continuous, the simplest approximation is a linear approximation, which assumes that the response and the predictors have an approximate linear relationship. The resulting model is called a *linear regression model*. Then, we use observed data to estimate the linear relationship and hope that such a linear approximation will be satisfactory. As such, linear regression models are empirical models which only describe the observed data, without a true understanding of the underlying mechanism which generate the data. In many practice situations, linear models may provide reasonable approximations to the true relationship, even though the true relationships are unknown and complicated. When necessary, some variables may be transformed to make the linear approximations more reasonable. Therefore, linear regression models are the simplest but are also the most widely used regression models.

Nonlinear regression models, on the other hand, are typically based on the underlying mechanisms which generate the data, so derivations of nonlinear models require good understanding of the scientific problems. Thus, nonlinear models are usually closer to the true relationships than linear models, and predictions based on nonlinear models are more reliable than linear models. However, for many practical problems, it may be difficult to derive nonlinear models since the underlying data generation mechanisms may be highly complicated.

In this chapter, we briefly review both linear and nonlinear regression models, with a focus on linear models. References for linear regression models are extensive, including Draper and Smith (1998) and Weisberg (2005), among others. Interested readers can find more detailed discussions of linear models in these references.

3.2 Linear Regression Models

Suppose that data are available on $p+1$ variables $(y, x_1, x_2, \cdots, x_p)$, where variable y is chosen as a *response* based on scientific interest and the other variables are treated as *predictors* or *covariates*. If the data is a sample of size n, then the data may be denoted as $\{(y_i, x_{i1}, x_{i2}, \cdots, x_{ip}),\ i = 1, 2, \cdots, n\}$. We wish to build a regression model which describes the approximate relationship between the response and the predictors.

The simplest regression model is a linear regression model, where the response and predictors are assumed to have a linear relationship. A linear regression model can be written as follows

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{4}$$

or

$$E(y_i) = \mathbf{x}_i^T \boldsymbol{\beta},$$

where y_i is the response for individual i, β_j's are unknown parameters, x_{ij} is the j-th predictor for individual i, $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ip})^T$ is a collection of all predictors, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$ is a vector of regression parameters, and ϵ_i's are random errors with mean zeros, $i = 1, 2, \cdots, n$. In linear model (4), the response and the predictors are assumed to have a linear relationship, with the unexplained variation accounted by the random error ϵ_i.

Linear regression models are widely used in practice because they are simple and are easy to interpret, even though they may not exactly represent the true relationship between the response and covariates. For example, regression parameter β_j may be interpreted as the effect of predictor x_j on the response y: one unit change in x_j is associated with β_j units

change in y. Such a simple linear form also allows us to derive the distribution of y based on the assumed distribution of ϵ and study the properties of parameter estimates. Moreover, in practice, such a linear relationship assumption may be reasonable and useful, especially when some predictors are transformed.

The linear model (4) may be written in a more compact matrix form. Let

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Model (4) can then be written in a matrix form as follows

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{5}$$

Common assumptions for models (4) or (5) are

- the errors ϵ_i's are *independent*,
- the errors ϵ_i's have mean zero, i.e., $E(\boldsymbol{\epsilon}) = \mathbf{0}$, as well as a *constant variance* σ^2, i.e., $Var(\boldsymbol{\epsilon}) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix,
- the errors ϵ_i's are normally distributed, i.e., $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I_n)$.

These assumptions are required for statistical inference. In data analysis, these assumptions must be checked to see if they are valid or not. Based on these assumptions, the (marginal) distribution of the response $\mathbf{y}$ is given by

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n). \tag{6}$$

Note that, in a standard regression model, the predictors $\mathbf{x}_i$ are assumed to be fixed (i.e., they are not random variables), while the response y is assumed to be a random variable.

Once a linear model is assumed for the variables, the next step is to estimate the unknown parameters and make statistical inference, based on the observed data. The *least squares method* for estimating parameters $\boldsymbol{\beta}$ is to minimize

$$Q(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}).$$

The resulting parameter estimates, called the *least square estimates*, are given by

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}, \qquad \hat{\sigma}^2 = \frac{RSS}{n-p-1} = \frac{\mathbf{r}^T \mathbf{r}}{n-p-1}, \tag{7}$$

where

$$\mathbf{r} = \mathbf{y} - X\hat{\boldsymbol{\beta}} = (y_1 - \hat{y}_1, \cdots, y_n - \hat{y}_n)^T$$

is a vector of *residuals*, and

$$RSS = \mathbf{r}^T\mathbf{r} = \sum_i (y_i - \hat{y}_i)^2$$

is called the *residual sum of squares*. Residuals represent the differences between the fitted values based on the assumed model and the observed values of the response, so they can be used to check if the assumed model fits the data well or not. They play an important role in model checking or model diagnostics. It can be shown that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta},\ \sigma^2(X^TX)^{-1}). \tag{8}$$

This result can be used to construct confidence intervals and hypothesis testing for $\boldsymbol{\beta}$.

A more general method for parameter estimation in regression models is the maximum likelihood method. For linear model (4) or (5), however, the least square estimates are identical to the maximum likelihood estimates (MLEs). This may not be true for other regression models.

The *coefficient of determination* is defined as

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2},$$

where $\bar{y} = \sum_i y_i/n$. It is the proportion of variation in the response that is explained by the regression or by the predictors, so it indicates the usefulness of the regression model.

3.3 Model Selection and Model Diagnostics

In regression models (4) or (5), some predictors may not have significant effects on the response, so they should be removed from the models. This is called model selection or variable selection. In regression analysis, parsimonious or simple models are preferred over complex models. In other words, unimportant or non-significant predictors should be removed from the models in order to reduce the number of unknown parameters and increase the accuracies of the estimates for the remaining parameters in the models. In data analysis, we should avoid models with too many predictors or avoid large models.

In model selection or variable selection, we need to check the significance of each predictor in the model. This is equivalent to comparing a smaller model to a larger model with more predictors to see if the two models differ significantly. For example, we may compare model

I with predictors x_1, x_2 to a larger model II with predictors x_1, x_2, x_3. Let Ω be a larger model and ω be a smaller model so that model ω is nested within model Ω. We can perform a hypothesis test to compare the two nested models. The null hypothesis is that the two models are not significantly different, while the alternative hypothesis is that the larger model is significantly better. A commonly used test statistic is

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)},$$

where p and q are the numbers of parameters in model Ω) and model ω respectively, and RSS_Ω and RSS_ω denote the residual sum of squares of model Ω) and model ω respectively. We reject the null hypothesis if $F > F_\alpha(p-q, n-p)$, i.e., the larger model is significantly better, so we should choose the larger model Ω. If we fail to reject the null hypothesis, i.e., the two models are not significantly different, we should choose the smaller model ω.

The above F-test can also be used to test the significance of a single predictor. To test a continuous predictor x_j, an alternative approach is to use the usual t-test, with the test statistic given by

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)},$$

where $s.e.(\hat{\beta}_j)$ is the standard error of the parameter estimate $\hat{\beta}_j$ associated with covariate x_j. The result is equivalent to the F-test.

A more general test for comparing nested models is the likelihood ratio test. The likelihood ratio test statistic has an asymptotic χ^2 distribution. It is a very general testing procedure and can be used to compare other nested regression models such as nonlinear and generalized linear models, while the above F-test is often used for linear regression models.

In model selection or variable selection, sometimes we need to compare several models which may or may not be nested. That is, the models to be compared may contain different sets of predictors or may have different functional forms, not necessarily that one model is nested within the other model. For example, we may compare a model with predictors x_1 and x_2 to a model with predictors x_3 and x_4. In this case, we may use some general and more commonly used model selection criteria, such as AIC and BIC.

Note that model selection should not completely rely on statistical criteria. We should also consider scientific issues for the problems under consideration. For example, if our main goal is to evaluate a treatment effect for a disease, we should keep the treatment variable in the model, whether it is significant or not. Moreover, the final model should also make scientific sense. Therefore, model selection is usually a compromise between statistical

criteria and scientific consideration. If several models are similar based on statistical criteria, such as similar AIC values, we should choose the model which is scientifically reasonable or the model which makes sense in practice.

Regression models are assumed, with certain assumptions, so they may not reflect the true data-generation mechanisms. Thus, in data analysis, the assumed models must be checked to see if the models fit the observed data well and if the model assumptions are satisfied. This is called model diagnostics. In model diagnostics, residuals play an important role since they measure how far away the fitted response values to the observed response values. A residual plot usually plots the residuals against the fitted response values. Other plots may also be used. These graphical methods are the basic tools for model diagnostics.

More specifically, for model diagnostics we should check the following features of the assumed model

- goodness of fit, i.e., whether the model fits the observed data well. This can be checked based on residual plots. If the model fits well, the residuals should all be close to zero without clear patterns.

- constant variance assumption. If the residual plot shows some clear patterns, such as increasing or decreasing patterns, the variance may not be constant. In this case, we may try to make a transformation on the response, such as a log-transformation, or other ways to improve the model.

- normality assumption. We can use a normal quantile-quantile (QQ) plot. If the normality assumption holds, the QQ plot should show a roughly straight line. When the normality does not hold, sometimes a transformation on the response may be a good idea.

- outliers. This can be seen from the residual plot: observations with unusually large or small residuals may be outliers.

- influential observations, i.e., observations which may have big impacts on parameter estimates but are not necessary outliers. They can be checked based on the Cook's distances: observations with unusually large Cook's distances may be influential.

When outliers or influential observations are identified, they should be removed and studied separately.

In summary, model diagnostics can generally be done informally based on graphical tools such as residual plots, QQ plots, and Cook's distance plots, but some more formal methods

are also available and may be used in some cases. Note that there is no perfect model. Model diagnostics are important, but we should also consider the interpretation and simplicity of the models. In other words, we can accept a model that is reasonable, simple, and easy to interpret.

3.4 Examples with R

Example 1. The Current Population Survey (CPS) is used to supplement census information between census years. These data consist of a random sample of 534 persons from the CPS, with information on wages and other characteristics of the workers, including sex, number of years of education, years of work experience, occupational status, region of residence and union membership. We wish to determine whether wages are related to these characteristics. The data file contains 534 observations on 11 variables sampled from the CPS. Variable names are:

```
EDUCATION: Number of years of education.
SOUTH: Indicator variable for Southern Region
       (1=Person lives in South, 0=Person lives elsewhere).
SEX: Indicator variable for sex (1=Female, 0=Male).
EXPERIENCE: Number of years of work experience.
UNION: Indicator variable for union membership (1=Union member, 0=Not union member).
WAGE: Wage (dollars per hour).
AGE: Age (years).
RACE: Race (1=Other, 2=Hispanic, 3=White).
OCCUPATION: Occupational category (1=Management, 2=Sales, 3=Clerical,
            4=Service, 5=Professional, 6=Other).
SECTOR: Sector (0=Other, 1=Manufacturing, 2=Construction).
MARR: Marital Status (0=Unmarried,  1=Married)
```

In the following, we analyze the data in R using linear regression models. Note that factors (or categorical variables) such as "race" need to be made clear in R. Otherwise, the computer cannot recognize them and may treat them as numerical variables, which may lead to misleading results.

```
wage.dat <- read.table("wage.data", head=T)
attach(wage.dat)
# Categorical variables must be declared as follows
race <- factor(race, labels=c("other", "hispanic", "white"))   # categorical variable
occupation <- factor(occupation,  labels=c("management","sale","clerical",
                     "service","professional","other"))   # categorical variable
sector <- factor(sector, labels=c("other","manufacturing",
                 "construction"))   # categorical variable

# Fit a full linear regression model
```

```
fit1 <- lm(wage ~ education+south+sex+experience+union+age+
                  race+occupation+sector+marr)
summary(fit1)
....
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.2781     6.6976   0.340  0.73390
education                0.8128     1.0869   0.748  0.45491
south                   -0.5627     0.4198  -1.340  0.18070
sex                     -1.9425     0.4194  -4.631 4.60e-06 ***
experience               0.2448     1.0818   0.226  0.82103
union                    1.6017     0.5127   3.124  0.00188 **
age                     -0.1580     1.0809  -0.146  0.88382
racehispanic             0.2314     0.9915   0.233  0.81559
racewhite                0.8379     0.5745   1.458  0.14532
occupationsale          -4.0638     0.9159  -4.437 1.12e-05 ***
occupationclerical      -3.2682     0.7626  -4.286 2.17e-05 ***
occupationservice       -3.9754     0.8108  -4.903 1.26e-06 ***
occupationprofessional  -1.3336     0.7289  -1.829  0.06791 .
occupationother         -3.2905     0.8005  -4.111 4.59e-05 ***
sectormanufacturing      1.0409     0.5492   1.895  0.05863 .
sectorconstruction       0.4774     0.9661   0.494  0.62141
marr                     0.3005     0.4112   0.731  0.46523
Residual standard error: 4.282 on 517 degrees of freedom
Multiple R-squared: 0.3265,     Adjusted R-squared: 0.3056
F-statistic: 15.66 on 16 and 517 DF,  p-value: < 2.2e-16
```

The above full model, with 10 predictors, explains about 31% variation in wage. An F-test of the full model against the null model (i.e., the model without any predictors) gives a very small p-value, indicating the full model is useful. However, we see that many predictors in the full model are not significant, so some predictors may be removed from the model. In other words, we need to do model selection or variable selection. We consider a stepwise method for variable selection using R function *step()*. The stepwise method is a combination of the forward method, which adds one predictor at each step, and the backward method, which deletes one insignificant predictor at each step.

```
fit2 <- step(fit1, direction="both")
summary(fit2)
.......
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.97952    1.71053   1.157 0.247696
education               0.67229    0.09904   6.788 3.10e-11 ***
south                  -0.68858    0.41504  -1.659 0.097701 .
sex                    -1.84527    0.41523  -4.444 1.08e-05 ***
experience              0.09370    0.01656   5.657 2.54e-08 ***
union                   1.51738    0.50836   2.985 0.002970 **
occupationsale         -3.97544    0.91420  -4.349 1.65e-05 ***
occupationclerical     -3.34712    0.76002  -4.404 1.29e-05 ***
occupationservice      -4.14818    0.80534  -5.151 3.68e-07 ***
```

```
occupationprofessional -1.26791    0.72703  -1.744 0.081754 .
occupationother        -2.79902    0.75655  -3.700 0.000239 ***
Residual standard error: 4.284 on 523 degrees of freedom
Multiple R-squared: 0.3181,     Adjusted R-squared: 0.305
F-statistic: 24.39 on 10 and 523 DF,  p-value: < 2.2e-16
```

The above model (called Model 2) is smaller than the full model and contains only 6 predictors, but it can still explain about 31% variation in wage. In other words, the above smaller model is as good as the full model, so it is preferred since it contains less predictors and is thus simpler.

Note that, for categorical variables, we can test their significances using the R function *drop1()*, as shown below.

```
drop1(fit1, test="Chi")
Single term deletions
           Df Sum of Sq     RSS    AIC   Pr(Chi)
<none>                  9480.8 1570.1
education   1     10.26  9491.0 1568.7  0.447366
south       1     32.95  9513.7 1570.0  0.173481
sex         1    393.34  9874.1 1589.8 3.176e-06 ***
experience  1      0.94  9481.7 1568.2  0.818076
union       1    178.95  9659.7 1578.1  0.001578 **
age         1      0.39  9481.2 1568.1  0.881884
race        2     44.55  9525.3 1568.6  0.286020
occupation  5    641.71 10122.5 1595.1 1.523e-06 ***
sector      2     65.90  9546.7 1569.8  0.157309
marr        1      9.79  9490.6 1568.7  0.457771
```

It seems that only sex, union, and occupation are significant. So let's try to fit the linear model with these three covariates, and then we can use the R function *anova* to compare nested models using the F-test.

```
fit3 <- lm(wage ~ sex+union+occupation)
# Compare three nested models
anova(fit3, fit2, fit1)
Analysis of Variance Table
Model 1: wage ~ sex + union + occupation
Model 2: wage ~ education + south + sex + experience + union + occupation
Model 3: wage ~ education + south + sex + experience + union + age + race +
                occupation + sector + marr
  Res.Df     RSS Df Sum of Sq       F    Pr(>F)
1    526 10753.1
2    523  9599.4  3   1153.69 20.9708 7.795e-13 ***
3    517  9480.8  6    118.59  1.0778    0.3746
```

We see that Model 2 is almost as good as the full Model 3 since the p-value from the F-test is 0.374 (i.e., the two models are not significantly different), but Model 1 is significantly worse

than Model 2 (very small p-value). Thus, we should choose the larger Model 2, since the additional covariates in Model 2 explain substantial extra variation than Model 1.

Finally, we should do model diagnostics for the final Model 2, using graphical tools as described earlier.

```
par(mfrow=c(2,1))
plot(fitted(fit2), resid(fit2), main="Residual Plot",
     xlab="fitted value", ylab="residuals")  # residual plots
abline(a=0,b=0)
qqnorm(resid(fit2))  # Normal QQ  plot
```

From these diagnostic plots (see Figure 1), we see that Model 2 does not fit the data well, since the residual plot shows some pattern (increasing trend), indicating possibly non-constant variance, and the QQ plot also shows possible non-normality since some points deviate from a straightline.

To improve model fitting, we can try to make a log transformation on the response and then re-fit the model with the new response. The results are shown below.

```
wage2 <- log(wage)  # a log-transformation of "wage" as the new response
fit4 <- lm(wage2 ~ education+south+sex+experience+union+age+
                   race+occupation+sector+marr)   # re-fit the model with new response
fit5<- step(fit4, direction="both")  # variable selection
summary(fit5)  # Model 5
........
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.224289   0.172070   7.115 3.73e-12 ***
education                0.068838   0.009912   6.945 1.14e-11 ***
south                   -0.102588   0.041668  -2.462 0.014139 *
sex                     -0.213602   0.041842  -5.105 4.65e-07 ***
experience               0.009494   0.001723   5.510 5.65e-08 ***
union                    0.202720   0.051009   3.974 8.06e-05 ***
occupationsale          -0.355381   0.091448  -3.886 0.000115 ***
occupationclerical      -0.209820   0.076149  -2.755 0.006068 **
occupationservice       -0.385680   0.080855  -4.770 2.40e-06 ***
occupationprofessional  -0.047694   0.072746  -0.656 0.512351
occupationother         -0.254277   0.079781  -3.187 0.001523 **
sectormanufacturing      0.111458   0.054845   2.032 0.042636 *
sectorconstruction       0.099777   0.096481   1.034 0.301541
marr                     0.065464   0.041036   1.595 0.111257
Residual standard error: 0.4283 on 520 degrees of freedom
Multiple R-squared: 0.3573,     Adjusted R-squared: 0.3412
F-statistic: 22.24 on 13 and 520 DF,  p-value: < 2.2e-16
```

The above model (called Model 5) explains about 36% variation in wage (in log-scale), so it seems better than model 2. Let's check to see if it indeed provides a better fit than model 2 for the observed data, based on model diagnostics.

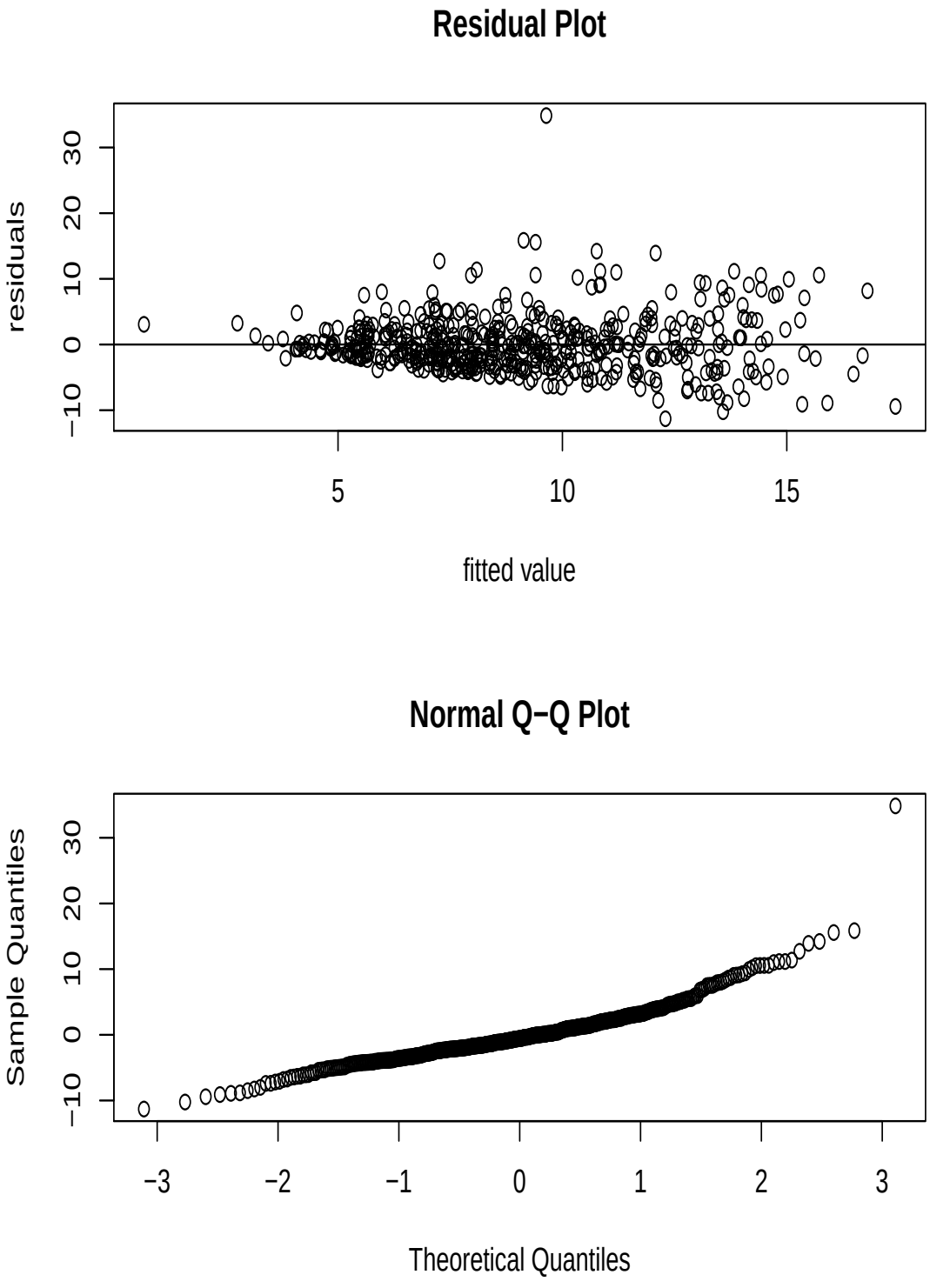

Figure 1: Residual plot and QQ plot for Model 2.

```
par(mfrow=c(2,1))
plot(fitted(fit5), resid(fit5), main="Residual Plot",
     xlab="fitted value", ylab="residuals")
abline(a=0,b=0)
qqnorm(resid(fit5))
```

The diagnostic plots are shown in Figure 2. We see that Model 5 fits the observed data better than Model 2. Thus, we can choose Model 5 as our final model. Note that better models may exist, such as models with interaction terms or transformed predictors, but it is also reasonable to simply choose Model 5, due to its simplicity. In other words, Model 5 may not be the best model, but it is a reasonable one. From Model 5, we see that the significant predictors are education, south, sex, experience, union, occupation, sector, and marr. A smaller model may be also possible. For example, the predictor "marr" may be removed from the model.

3.5 Nonlinear Regression Models

Linear regression models have been widely used because of their simplicity, which is an important advantage before modern computers become available. However, linear models usually only provide description of observed data, rather than trying to understand data, since they are usually chosen based on goodness-of-fit of the observed data. In other words, linear models usually provide little understanding of the data-generation mechanism. Nonlinear regression models, on the other hand, attempt to understand the mechanics of data generation, so they are often called mechanistic models or scientific models. There are some advantages of nonlinear models. First, nonlinear models may provide better predictions outside the range of observed data than that of linear models. Second, parameters in nonlinear models often have natural physical interpretations. Third, nonlinear models may require few parameters than the corresponding linear models that fit the data equally well. Note that, however, in many practical situations we do not know the data-generating mechanisms. In these cases, linear models would be good choices.

Unlike linear models, for nonlinear models there are typically no analytic or closed-form expressions for parameter estimates, so iterative algorithms are generally required obtain parameter estimates. Moreover, in fitting nonlinear models it is important to choose good *starting values* for the iterative algorithms since some likelihoods may have multiple modes.

Let y_i and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ be the response and predictors for individual i respectively,

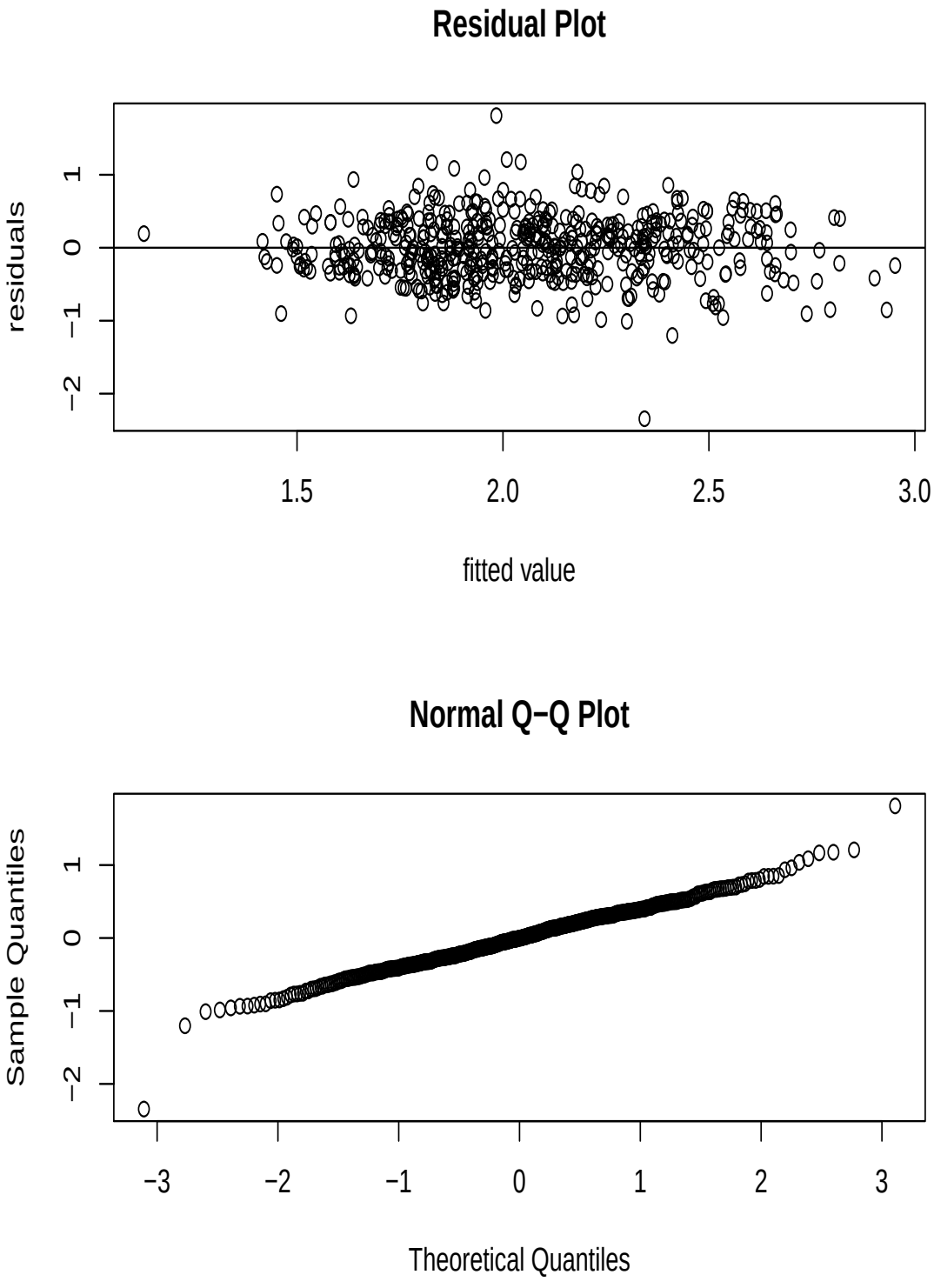

Figure 2: Residual plot and QQ plot for Model 5

$i = 1, 2, \cdots, n$. A general nonlinear regression model can be written as

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{9}$$

where h is a known nonlinear function, $\boldsymbol{\beta}$ is a vector of regression parameters, and ϵ_i is the random error. Assumptions for a standard nonlinear regression model are the same to those for a standard linear model, i.e., (i) the errors ϵ_i's are independent, (ii) the errors ϵ_i's have mean zero and constant variance σ^2, and (iii) the errors ϵ_i's are normally distributed.

Statistical inference for a nonlinear regression model can be based on the least squares method or the likelihood method. The ordinary least-squares estimator for parameter $\boldsymbol{\beta}$ is to minimize the sum of squares $\sum_{i=1}^{n}(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2$. This can be achieved by solving the following estimating equation

$$\sum_{i=1}^{n} \frac{\partial g(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left[y_i - g(\mathbf{x}_i, \boldsymbol{\beta})\right] = 0. \tag{10}$$

An iterative algorithm such as the Newton-Raphson method is often needed to solve the above equation.

Alternatively, under the normality assumption for the errors, i.e., ϵ_i i.i.d. $\sim N(0, \sigma^2)$, the MLE of $\boldsymbol{\beta}$ can be obtained by maximizing the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - g(\mathbf{x}_i, \boldsymbol{\beta}))^2}{2\sigma^2}\right].$$

So the MLE of $\boldsymbol{\beta}$ satisfies the following likelihood equation

$$\frac{\partial \log L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y})}{\partial \boldsymbol{\beta}} = 0,$$

which is identical to the least-squares equation (10). Therefore, for nonlinear regression models the ordinary least-squares estimator of $\boldsymbol{\beta}$ is also the same as the MLE of $\boldsymbol{\beta}$, and estimation for a nonlinear regression model is analogous to that for a linear regression model.

For nonlinear regression models, analytic or closed-form expressions for parameter estimates are unavailable. However, statistical inference can still be carried out based on the standard asymptotic results of likelihood methods under the usual regularity conditions. That is, under some regularity conditions, MLEs of the model parameters are consistent, asymptotically normal, and asymptotically most efficient. Confidence intervals and hypothesis testing can be based on the asymptotic normality of the MLEs. Therefore, with the availability of modern computers and software, statistical inference for nonlinear models does not offer much more difficulties than that for linear models.

In many cases, nonlinear models can be derived from a set of differential equations based on the understanding of the underlying data-generation mechanisms, as shown in the examples below. The developments of nonlinear models require close collaboration between statisticians and subject-area scientists, but such models may not be always available since the true data-generation mechanisms can be highly complex. Note that, in principle, any smooth nonlinear functions can be approximated by a high-order polynomial based on Taylor series expansions, if the functions are sufficiently smooth. However, high order polynomials are often unstable in replications of the data so they are generally not recommended.

Nonlinear models have been widely used in practice, such as HIV viral dynamics, pharmacokinetics, pharmacodynamics, molecular genetics, and growth or decay. More detailed discussions of nonlinear models can be found in Bates and Watts (1988), Seber and Wild (2003), and Wu (2009).

Example 1 (Growth curve models)

In the analysis of *growth curves*, nonlinear models are usually necessary. There are various growth curve models. Here we consider a simple monomolecular growth function. Let $y(t)$ be the size at time t (e.g., size of an animal), and let $\mu(t) = E(y(t))$. Suppose that the growth rate is proportional to the remaining size. Then $\mu(t)$ satisfies the following differential equation:

$$\frac{d\mu(t)}{dt} = \beta_1(\beta_0 - \mu(t)), \qquad \beta_1 > 0,$$

which can be solved analytically, with solution

$$\mu(t) = \beta_0 + \beta_2 e^{-\beta_1 t}.$$

Thus, given an observed sample, we can consider the following nonlinear regression model for estimating the parameters

$$y_{ij} = \beta_0 + \beta_2 e^{-\beta_1 t_{ij}} + e_{ij}, \qquad i = 1, \cdots, n,\ j = 1, \cdots, n_i, \tag{11}$$

where y_{ij} is the size for individual i at measurement time t_{ij} and e_{ij} is the corresponding measurement error. Note that, when $\beta_0 + \beta_2 = 0$ (i.e., when the initial size is 0), the above model is called the *von Bertalanffy growth curve*, which is often used in ecology to describe animal growth.

Example 2 (Pharmacokinetics)

Studies of *pharmacokinetics* are important in drug developments. Pharmacokinetics studies the course of absorption, distribution, metabolism, and elimination of some substance

in the body over time, given drug dose, i.e., how the drug moves through the body. Suppose that a substance enters the body via ingestion. Let $y(t)$ be the concentration of the substance in the body at time t (usually measured in the blood), and let $\mu(t) = E(y(t))$. Let $\mu_0(t)$ be the amount at the absorption site (e.g., stomach). A commonly used one-compartment model is based on the following differential equations

$$\begin{aligned} \frac{d\mu(t)}{dt} &= \beta_1\mu_0(t) - \beta_2\mu(t), \\ \frac{d\mu_0(t)}{dt} &= -\beta_1\mu_0(t), \end{aligned}$$

where β_1 is the absorption rate and β_2 is the elimination rate. The above differential equations have an analytic solution given by

$$\mu(t) = \frac{\beta_1 x}{(\beta_1 - \beta_2)\beta_3}\left(e^{-\beta_2 t} - e^{-\beta_1 t}\right),$$

where x is the dose of the substance and β_3 is the volume of distribution. Therefore, given an observed sample, we can consider the following nonlinear regression model for estimating the parameters

$$\begin{aligned} y_{ij} &= \frac{\beta_1 x_i}{(\beta_1 - \beta_2)\beta_3}(e^{-\beta_2 t_{ij}} - e^{-\beta_1 t_{ij}}) + e_{ij}, \\ & \qquad i = 1, \cdots, n, \quad j = 1, \cdots, n_i, \end{aligned} \tag{12}$$

where y_{ij} is the concentration for individual i at time t_{ij} and e_{ij} is the corresponding random error. This nonlinear model is widely used in pharmacokinetics.

4 Generalized Linear Models

4.1 Introduction

Both linear and nonlinear regression models typically assume that the response variable is continuous and follows a normal distribution. Nonlinear regression models extend linear regression models by allowing nonlinear relationships between the response and predictors, but the response is still assumed to be normally distributed as in linear models. In practice, however, there are different types of response variables, and many of them are not continuous variables and are unlikely to follow normal distributions, even after variable transformations. For example, if the response is a binary variable taking only two possible values (say, male or female, pass or fail, success or failure, etc), then the response variable cannot follow a normal distribution, no matter what transformation is used. In this case, linear or nonlinear regression models cannot be used. In this section, we describe a class of regression models for which the response variables can be binary, count, continuous but skewed, and more. This class of regression models is called generalized linear models.

Generalized linear models (GLMs) extend linear models by allowing the response variable to follow distributions in the *exponential family*, which includes a wide range of commonly used distributions such as normal, binomial, and Poisson distributions. In other words, in a GLM, the response variable can be continuous, discrete, and count. The covariates or predictors still enter the model in a linear fashion, but the response and predictors are linked by a nonlinear link function. A main advantage of a GLM is that it can be used to build a regression model when the response is discrete such as "pass/fail" and "success/failure". Moreover, GLMs include linear regression models as a special case. Therefore, GLMs greatly extend the applicability and popularity of regression models.

In this chapter, we first describe GLMs in a general form. Then, we discuss the two most popular GLMs, the logistic regression models and the Poisson regression models, in greater details.

4.2 The Exponential Family

In a GLM, the response variable is assumed to follow a distribution from the exponential family. The exponential family consists of many parametric distributions which share some common characteristics, including many of the most well known distributions. It is the most popular class of parametric distributions, although it is a small subset of all parametric distributions. We briefly describe the exponential family as follows.

A distribution in the *exponential family* has the following general form for its probability density function (pdf)

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{13}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions, θ is called the canonical parameter representing the location, and ϕ is called the *dispersion parameter* representing the scale. It can be shown that the mean and variance of a distribution from the exponential family are given by

$$E(y) = \mu = \frac{\partial b(\theta)}{\partial \theta}, \qquad Var(y) = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2}.$$

The following common distributions are in the exponential family: normal distribution, binomial distribution, multinomial distribution, Poisson distribution, gamma distribution, Dirichlet distribution, and some other distributions. However, some common distributions are not in the exponential family, such as the t-distribution and uniform distribution.

We focus on the three most important ones: normal, binomial, and Poisson. For the normal distribution $N(\mu, \sigma^2)$ with mean μ, variance σ^2, and pdf

$$f(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right],$$

we have

$$\begin{aligned}
&\theta = \mu, \qquad \phi = \sigma^2, \\
&a(\phi) = \phi, \qquad b(\theta) = \theta^2/2, \qquad c(y,\phi) = -(y^2/\phi + \log(2\pi\phi))/2.
\end{aligned}$$

For the binomial distribution with probability distribution

$$P(y=k) = \left(\begin{array}{c} n \\ k \end{array}\right)\mu^k(1-\mu)^{n-k},\ k = 0, 1, \cdots, n,$$

where $0 < \mu < 1$, we have

$$\begin{aligned}
&\theta = \log(\mu/(1-\mu)), \qquad \phi = 1, \\
&b(\theta) = -n\log(1-\mu), \qquad c(y,\phi) = \log\left(\begin{array}{c} n \\ y \end{array}\right), \\
&E(y) = n\mu, \qquad Var(y) = n\mu(1-\mu).
\end{aligned}$$

Note that the binomial distribution reduces to the Bernoulli distribution when $n = 1$, i.e.,

$$P(y=k) = \mu^k(1-\mu)^{1-k}, \qquad k = 0, 1,$$

with

$$E(y) = \mu = P(y = 1), \qquad Var(y) = \mu(1 - \mu).$$

For the Poisson distribution with probability distribution

$$P(Y = k) = (k!)^{-1} e^{-\mu} \mu^k, \qquad k = 0, 1, 2, \cdots,$$

where $\mu > 0$, we have

$$\begin{aligned} &\theta = \log(\mu), \qquad \phi = 1, \\ &a(\phi) = 1, \qquad b(\theta) = e^{\theta}, \qquad c(y, \phi) = -\log(y!), \end{aligned}$$

with $E(y) = Var(y) = \mu$.

The above three distributions are the most well known distributions in the exponential family. They are also most commonly used in GLMs. They can be used to model continuous, discrete, and count response data. Thus, GLMs can cover a wide variety of practical situations where regression analysis is needed. Note that, in a regression model, the predictors or covariates are viewed as fixed (i.e., not viewed as random variables) and thus can have any types. In other words, the type of regression model is determined by the type of the response variable, not the predictors.

4.3 The General Form of a GLM

A regression model links the mean of the response to a set of predictors or covariates. The set of predictors or covariates are usually combined in a linear way, which is called the linear predictor. Specifically, let

$$\mu_i = E(y_i)$$

be the mean response. Let $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})^T$ be a vector of predictors or covariates, $i = 1, \cdots, n$. We call the following linear combination

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$$

the *linear predictor*, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a $p \times 1$ vector of unknown parameters. In other words, the linear predictor η_i combines the predictors (covariates) in a linear form, which is the simplest and most common way to combine predictors. In a linear regression model, we link the mean response and the linear predictors as

$$E(y_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \qquad i.e., \qquad \mu_i = \eta_i.$$

That is, for linear regression models we use the identity function $g(x) = x$ to link the mean response μ_i and the linear predictor η_i, i.e., $g(\mu_i) = \eta_i$.

The identity link function $g(x) = x$ for linear regressions, however, may not be appropriate for other types of response. For example, if y_i is a binary variable, then the mean response $\mu_i = P(y_i = 1)$, which is a number between 0 and 1, while the value of the linear predictor η_i can take any value from $-\infty$ to ∞, so we cannot use the identity link to link the mean response to the linear predictor. In other words, for binary responses we should use other link functions to link the mean response to the linear predictor in regression modelling. For example, we may consider the following link function for binary response

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

Then, a regression model for the binary response y can be written as $g(\mu_i) = \eta_i$, i.e.,

$$\log\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \mathbf{x}_i^T\boldsymbol{\beta}, \qquad i = 1, 2, \cdots, n,$$

or

$$\text{odds} = \exp(\mathbf{x}_i^T\boldsymbol{\beta}),$$

or

$$\mu_i = P(y_i = 1) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta})}, \qquad i = 1, 2, \cdots, n,$$

which is the well known logistic regression model – the most popular generalized linear model. As an example, suppose that $y_i = 1$ if a person gets a cancer and $y_i = 0$ otherwise. Let x_i be the smoking status. Then, the above logistic regression model can be used to study the relationship between the probability (or odds) of getting a cancer if a person smokes.

A *generalized linear model (GLM)* can be written as

$$g(E(y_i)) = \mathbf{x}_i^T\boldsymbol{\beta}, \qquad i = 1, 2, \cdots, n, \tag{14}$$

where $g(\cdot)$ is a monotone and differentiable function, called the *link function*. Thus, a GLM has two components:

- the response y_i follows a distribution in the exponential family;
- the link function $g(\cdot)$ describes how the mean response $E(y_i) = \mu_i$ is related to the linear predictor η_i, i.e., $g(\mu_i) = \eta_i$.

The choice of the link function $g(\cdot)$ depends on the type of the response y_i. If the response is a *binary* variable, taking values 0 or 1, the most common choice of the link function is the following link, called *logit link*,

$$g(x) = \log\left(\frac{x}{1-x}\right).$$

The resulting model is called a *logistic regression model*. The logit link function offers nice interpretation: $P(y_i = 1)/(1 - P(y_i = 1))$ may be interpreted as "odds of success" (success is defined as "$y_i = 1$" here). Other link functions for binary response are also available. For example, we may choose $g(x) = \Phi^{-1}(x)$, where Φ is the standard normal cumulative distribution function (cdf). The resulting model is called a *probit model*. Another choice is the *complementary log-log link* function: $g(x) = \log(-\log(1 - x))$.

If the response is count and is assumed to follow the Poisson distribution, the most common choice of the link function is the log link:

$$g(x) = \log(x).$$

Both the logit link and the log link functions are called *canonical link functions*, since they can be naturally obtained from the assumed distributions written in the standard form of an exponential family (see, e.g., McCullagh and Nelder, 1989).

Since the link function in a GLM is often a nonlinear function, GLMs are special nonlinear regression models. The linear predictor is in a linear form, like in linear regression models. The nonlinear part is the link function. Note that, for nonlinear regression models, the predictors are not combined in a linear form but can be in any nonlinear forms in the models. So GLMs are essentially like linear models. In other words, GLMs are essentially "empirical models" rather than "scientific models". On the other hand, true nonlinear models are "scientific models", which are derived based on the data-generation mechanisms, and true nonlinear models usually do not contain linear predictors. Thus, GLMs are different from usual nonlinear models. Moreover, the response variable in a GLM is assumed to follow a distribution in the exponential family, while the response variable in a nonlinear model is assumed to follow the normal distribution. GLMs include linear regression models as a special case when the response distribution is normal and the link function is the identity function. In summary, GLMs are still restrictive in that they involve essentially linear models and only cover distributions from the exponential family.

4.4 Inference for GLM

A common approach for parameter estimation and inference for GLMs is to use the likelihood method. For a general GLM, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}.$$

Note that the regression parameters $\boldsymbol{\beta}$ is implicit in the loglikelihood function $l(\boldsymbol{\beta}, \phi)$ since

$$g(E(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta}, \qquad E(y_i) = \partial b(\theta_i)/\partial \theta_i.$$

The likelihood equation for $\boldsymbol{\beta}$ is given by

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = 0.$$

The resulting solution is a candidate for the MLE of $\boldsymbol{\beta}$. Since the loglikelihood $l(\boldsymbol{\beta}, \phi)$ is nonlinear in the parameters $\boldsymbol{\beta}$ and ϕ, MLEs are obtained using an iterative algorithm such as the *Newton-Raphson method* or the *iteratively reweighted least squares method* described in McCullagh and Nelder (1989). Since an iterative algorithm is used, sometimes convergence of the algorithm can be an issue, such as non-convergence, especially when the observed data are poor or the model is too complex.

The MLEs of the model parameters in a GLM share the usual asymptotic properties of MLEs: they are consistent, efficient, and asymptotically normally distributed. However, for finite samples in practice, the performance of the MLEs may depend on the sample size. For example, the standard errors of parameter estimates obtained based on the asymptotic Fisher information matrix may not always be reliable for small samples. Similarly, the asymptotic normality and the likelihood ratio test (deviance test) for small samples may not always perform well. For small samples, a better approach to obtain standard errors may be the bootstrap method. In summary, we should be careful in interpreting computer outputs of the parameter estimates and their standard errors for GLMs since these results are often based on asymptotic theory which may not hold for small samples.

Statistical inference for GLMs is often based on the *deviance*, which can be defined as the difference between the log-likelihoods for the full model and for the fitted model. Consider a model A. The deviance for model A is defined as

$$D(\mathbf{y}) = -2 \left[\log(f(\mathbf{y}|\hat{\theta}_A)) - \log(f(\mathbf{y}|\hat{\theta}_F)) \right],$$

where $\hat{\theta}_A$ is the parameter estimate under model A and $\hat{\theta}_F$ is the parameter estimate under the full model. The *full model* (or the *saturated model*) is the most complex model where

the data is explained exactly (i.e., it represents the data as being entirely systematic such as having n parameters for n data points). The *null model* is the smallest model where there is no relationship between the predictors and the response (i.e., it represents the data as being entirely random).

Note that the deviance is simply –2 times the log-likelihood ratio statistic of model A compared to the full model, so the deviance measures how close model A is to the full model (the perfect model). For linear regression models, the deviance is simply the residual sum of squares $RSS = \sum_i (y_i - \hat{y}_i)^2$, which usually measures the goodness-of-fit of the model or the discrepancy between observed data and fitted values. An alternative measure of discrepancy is the so-called *Pearson*'s χ^2 statistic defined as

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{Var(\hat{\mu}_i)}.$$

4.5 Model Selection and Model Diagnostics

Similar to linear regression models, for GLMs we should also do model selection or variable selection in data analysis. The basic ideas and methods for model selection of GLMs are similar to those for linear regression models, such as the criteria of AIC, BIC, and LRT. Models with small values of AIC or BIC are preferred, but AIC or BIC do not test the significance of a model over an alternative. Hypothesis tests can be used to compare models with significance. Two types of hypothesis tests are often considered:

- *Goodness of fit test*: test if the current model fits the observed data well by comparing the current model with the full model;
- *Compare two nested models*: compare a smaller model with a larger model.

These two types of test are described in more details below.

For goodness of fit tests, under some regularity conditions, the scaled deviance $D(\mathbf{y})/\phi$ and the Pearson's χ^2 statistics are both asymptotically distributed as the $\chi^2(d)$ distribution, where d is the number of parameters in model A. Note that, for binary data, this χ^2 approximation is poor. When comparing two *nested* models, under the null hypothesis of no difference between the two models, the difference in the deviances of the two models asymptotically follows the $\chi^2(d)$ distribution, with degrees of freedom d being the difference of the number of parameters in the two models being compared, i.e.,

$$D_S - D_L \to \chi^2(d), \qquad \text{as } n \to \infty,$$

where D_S and D_L are the deviances for the small model and the large model respectively. This result can be used for model comparison and model selection. For example, if the p-value is large (say larger than 0.05), we prefer the smaller model since there is no significant difference between the two models. When the p-value is small (say, less than 0.05), we prefer the larger model since the larger model fits significantly better than the smaller model. Note that the above χ^2 approximations are more accurate when comparing nested models than for the goodness of fit statistic. However, in both cases, the results are only approximate, and model selection should be combined with scientific interpretation of the selected model.

For testing the significances of individual predictors in a model, we can also consider a Wald-type test. Suppose that we wish to test the significance of predictor x_j, i.e., testing the hypotheses $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. The test statistic is given by

$$z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1), \qquad \text{asymptotically under } H_0.$$

So H_0 is rejected (i.e., predictor x_j is significant) if the p-value is small (say, less than 0.05). Since the χ^2 tests and the Wald-type tests are approximate, it is generally desirable to consider both types of test and compare the results to see if they agree or not, rather than relying on a single test. Generally, the deviance test is preferred.

Model Diagnostics

In model diagnostics for regression models, residuals play an important role. In linear regression models or Gaussian models, the residuals are defined as $r_i = y_i - \hat{\mu}_i,\ i = 1, 2, \cdots, n$, which are called *response residuals* in GLMs. For most GLMs, however, these residuals are not appropriate since the variance of the response is not constant so any patterns in the residual plots do not necessarily indicate lack of fit of the models. In other words, for some common distributions in the exponential family, such as the binomial distribution and the Poisson distribution, the variances depend on the mean parameters, while this is not the case for normal or Gaussian distributions in which the variance parameter σ is independent of the mean parameter μ. Thus, for most GLMs, the usual response residuals may not be appropriate. We need some modifications when defining the residuals for GLMs.

For GLM model diagnostics, we can use the following *Pearson residuals*

$$r_{ip} = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}}, \qquad i = 1, 2, \cdots, n,$$

which are usual residuals scaled by the standard deviations. Note that $\sum_i r_{ip}^2$ is simply the familiar Pearson statistic. Another type of residuals, called the *deviance residuals* r_{iD}, are

defined such that the deviance may be written as the sum of r_{iD}^2, i.e.,

$$\text{Deviance} = \sum_{i=1}^{n} r_{iD}^2 \equiv \sum_{i=1}^{n} d_i.$$

Thus, we can write

$$r_{iD} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}, \qquad i = 1, 2, \cdots, n.$$

Residuals allow us to check the goodness-of-fit of the models. Influential observations, on the other hand, are observations which have big effects on the resulting estimates of model parameters. In other words, removing a few influential observations may lead to big changes in parameter estimates. Thus, in data analysis influential observations should be removed and studied separately. In GLMs, influential observations may be checked using the *Cook's distances*, which are defined in a similar way to that in linear regression models. They are essentially the changes in model parameter estimates when the corresponding observations are removed in model fitting. Thus, observations with large values of Cook's distances may be influential.

When fitting a GLM to a dataset, model diagnostics often include the following: (i) check if there are any outliers or influential observations, and if so compare analysis without these observations to analysis with these observations; (ii) check whether the structure form of the model is reasonable, which includes choices of predictors and possible transformations of the predictors; (iii) check whether the stochastic part of the model is reasonable, such as the distributional assumptions or the nature of the variance. Common diagnostic methods include

- residual plots: we can plot the deviance residuals against the estimated linear predictor $\hat{\eta}_i$. This may help the choice or transformations of the predictors, but for a binary response the residual plot is not very useful due to the nature of the response data;
- transformation of the predictors: polynomial terms or interaction terms or other transformations may be considered;
- Cook's distance plot: check influential observations.

For data analysis using GLMs, an important component of model diagnostics is to check the so-called *over-dispersion problem* described below. This problem does not exist for linear regression models but is common for GLMs, due to possible relationship between mean and variance in GLMs. For Gaussian models such as linear or nonlinear regression models, there is no over-dispersion problem because the mean and variance are independent in normal distributions.

4.6 Over-Dispersion Problem

Regression models specify how the *mean structures* of the response y_i may depend on the predictors $\mathbf{x}_i$. For example, GLM assumes that

$$g(E(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta}.$$

For linear regressions, we introduce an additional parameter σ to account for the *variance* (or variation) of the data. However, this is not always possible. For some most common distributions in the exponential family, such as the binomial distribution and the Poisson distribution, the variance is determined by the mean. For example, if y_i follows a Bernoulli distribution, we have

$$Var(y_i) = E(y_i)(1 - E(y_i)),$$

and if y_i follows a Poisson distribution, we have

$$Var(y_i) = E(y_i).$$

This is very different from a linear regression model where the response is assumed to a normal distribution in which the variance is unrelated to the mean. That is, if y_i follows a normal distribution $N(\mu, \sigma^2)$, the mean μ and the variance σ^2 are independent and both can vary freely, which allows great flexibility in modelling real data. On the other hand, for Binomial and Poisson distributions, the strong relationship between the mean and variance is very restrictive in practice since the variation in the data may *not* agree with the theoretical variance which is determined by the mean. In other words, in a GLM we assume a mean structure $g(E(y_i)) = \mathbf{x}_i^T \boldsymbol{\beta}$ and assume that y_i follows a distribution in the exponential distribution, but the observed variation in the data may be different from the theoretical variance obtained from the assumed distribution. For example, for the Poisson regression model, the theoretical variance is the same as the mean, which is $E(y_i) = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$, but the variation in the data may be much larger or much smaller than the theoretical variance, so the assumed model is inappropriate.

If the variation in the data is larger (or smaller) than the theoretical variance determined by the assumed distribution, the problem is called an *over-dispersion* (or a *under-dispersion*) problem. When over-dispersion or under-dispersion problem arises in data analysis, the assumed distribution for the GLM does not hold, so this problem must be addressed for correct inference. Usually over-dispersion problems are more common than under-dispersion problems. Overdispersion problems can arise in longitudinal or clustered data if the correlation within clusters are not incorporated in the models.

One way to address the over-dispersion problem is to specify the mean and variance functions separately in a GLM, without a distributional assumption. This approach is called the *quasi-likelihood* method. For example, for a logistic regression model, we may assume that

$$Var(y_i) = \tau E(y_i)(1 - E(y_i)),$$

where τ is called a *dispersion parameter*. If $\tau = 1$, then y_i follows a binomial distribution. If $\tau \neq 1$, then y_i does not follows a binomial distribution. In this case, the model is called *quasibinomial*, which is not a parametric distribution, so the modified likelihood is called a *quasi-likelihood*. Similarly, for Poisson model, we may assume that

$$Var(y_i) = \tau E(y_i).$$

When $\tau \neq 1$, y_i does not follow a Poisson distribution, and the resulting quasi-likelihood model is called *quasi-Poisson*.

In data analysis using a statistical software such as R, we may consider the "quasi-likelihood" option and compare the results with that from a standard GLM fit to see if the results agree. If the results differ substantially, the result based on the "quasi-likelihood" option may be more reliable. Alternatively, we can obtain an estimate of the dispersion parameter τ and see if it is close to 1 or not. If $\hat{\tau} \approx 1$, the standard GLM may hold. Otherwise, we should consider the "quasi-likelihood" method.

4.7 More on Model Selection

In regression analysis, model selection or variable selection is an important step. In practice, there are often many potential predictors or covariates which may be included in a regression model. Too many predictors in a model may cause many potential problems such as multi-collinearity and poor parameter estimates. Generally, simple or parsimonious models are preferred, but simple models may not fit the data as well as complex models. A good model selection method should achieve a *balance* between simplicity and goodness of fit. Moreover, a model should make scientific sense. That is, a model should be meaningful in the application under consideration and should be easy to interpret.

There are many methods or criteria for model selection. Here we focus on the following commonly used methods: Akaike information criterion (AIC), Bayesian information criterion (BIC), the likelihood ratio test (LRT), and least absolute shrinkage and selection operator (LASSO). For illustration purpose, we focus on linear regression model selections, but the

methods may also be used in other regression models such as nonlinear and generalized linear models.

Consider the following linear regression model with p predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, 2, \cdots, n, \tag{15}$$

where y_i is the response for individual i, β_j's are unknown parameters, x_{ij} is the j-th covariate (predictor) for individual i, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$ is a vector of regression parameters, and ϵ_i's are random errors. We assume that ϵ_i i.i.d. $\sim N(0, \sigma^2)$. Then, $y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2)$, and the likelihood function $L(\boldsymbol{\beta}|\mathbf{y})$ for the y_i's can be easily obtained. Let $L_n(p)$ be the maximized value of the likelihood function

$$L_n(p) = \max_{\boldsymbol{\beta}} L(\boldsymbol{\beta}|\mathbf{y}).$$

The value of *Akaike information criterion (AIC)* is given by

$$AIC = -2\log(L_n(p)) + 2p,$$

where the first term measures the goodness of fit and the second term is a penalty for the number of parameters in the model. Thus, AIC describes a tradeoff between accuracy and complexity of the model or between bias and variance. Given a set of candidate models, the model with the smallest AIC value is preferred. In other words, we can start with several plausible models, find the models' corresponding AIC values, and then choose the model with smallest AIC value.

Another commonly used and closely related criteria is called the *Bayesian information criterion (BIC)* or the *Schwarz criterion*, which is derived using Bayesian arguments. The value of BIC is given by

$$BIC = -2\log(L_n(p)) + p\log(n).$$

Given a set of candidate models, the model with the smallest BIC value is preferred. Note that BIC penalizes the number of parameters more strongly than does the AIC. It has been shown that AIC is asymptotically optimal in selecting the model with the least mean square error, while BIC is not asymptotically optimal.

Note that AIC and BIC do not provide significance tests of models. That is, the best model selected by AIC/BIC may be better than other models, but not necessarily significantly better. For *nested* models, the likelihood ratio test (LRT) can be used to compare models with a given significance level. Suppose that we want to compare model (15) with a *smaller* model given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{iq} + \epsilon_i \qquad i = 1, 2, \cdots, n, \tag{16}$$

where $q < p$. Let $L_n(p)$ and $L_n(q)$ be the maximized values of the likelihood functions of models (15) and (16) respectively. Then, the LRT statistic is given by

$$T = 2(\log(L_n(p)) - \log(L_n(q))),$$

which asymptotically follows a $\chi^2(p-q)$ distribution. The LRT allows us to compare two nested model to see if one is significantly better than the other.

There are other model selection methods, such as Mallows' C_p, false discovery rate, and cross-validation. In data analysis, a common used strategy is to use the *stepwise* method: it is a combination of a forward selection and a backward elimination procedure. In each step, the selection may be based AIC, BIC, LRT, and other model selection criteria.

Note that the foregoing model selection or variable selection methods select subsets of all predictors, so they are discrete processes, i.e., predictors are either retained or dropped from the models (in other words, regression coefficients are either zeros or non-zeros). Thus, these methods can be quite unstable. For example, small changes in the data can result in very different models being selected. Moreover, a model which fits the data well may not be good for prediction. An important criterion for evaluating a model is its prediction error. The *prediction error (PE)* for a data point $\mathbf{x}_0$ for linear model (15) is given by

$$\begin{aligned} \text{PE} &= E((\mathbf{y} - \hat{\mathbf{y}})^2 \mid \mathbf{x} = \mathbf{x}_0) \\ &= \sigma^2 + \text{bias}^2(\hat{\mathbf{y}}) + \text{Var}(\hat{\mathbf{x}}|\mathbf{x}_0), \end{aligned}$$

where $\hat{\mathbf{y}}$ is the fitted value. The above decomposition is known as the *bias-variance tradeoff*. As a model becomes more complex (i.e., more terms are added in the model), the coefficient estimates suffer from higher variance.

The idea of ridge regression is to regularize the coefficients (i.e., control how large the coefficients grow). That is, instead of minimizing the residuals sum of squares as for the least-square estimates, we minimize

$$\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 \qquad \text{such that} \qquad \sum_{j=1}^{p}\beta_j^2 \leq t,$$

where the predictors $\mathbf{x}_i$ are standardized and t is a tuning parameter. This is equivalent to minimize the following *penalized residual sum of squares (PRSS)*:

$$PRSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2.$$

The resulting ridge estimate is given by

$$\tilde{\boldsymbol{\beta}} = (X^TX + \lambda I_p)^{-1}X^T\mathbf{y},$$

where X is the standardized design matrix and $\mathbf{y}$ is centered. For models with large p and small n, matrix X^TX may not be invertible (or matrix X^TX may be singular), while introducing parameter λ makes the problem non-singular. This is the original motivation for ridge regression. The tuning parameter may be selected using the cross-validation method.

Note that, in a linear regression model, when there are too many predictors, multicollinearity may happen and parameter estimates may become unstable. A ridge regression circumvents this problem. It makes the parameter estimates somewhat biased, but the variances of the estimates are smaller than that of the least square estimates, and their mean square errors may also be smaller than that of the least square estimates. Thus, the idea behind ridge regression is about bias-variance tradeoff. Ridge regression is a continuous process that *shrinks* regression coefficients and are thus more stable than subset selection methods (such as AIC/BIC/LRT), but it usually does not set regression coefficients to zeros (i.e., it does not select predictors).

In recent years, the *least absolute shrinkage and selection operator (LASSO)* has become increasingly popular, especially for high-dimensional data (i.e., large p, small n). The LASSO combines shrinkage and selection methods for linear regression. Unlike ridge regression, small value of t in LASSO will set some coefficients exactly to 0, which performs variable selection. In other words, it shrinks some regression coefficients and sets others to 0, so it retains the good features of both subset selection and ridge regression.

Specifically, consider the linear model (15). Suppose that the predictors x_{ij}'s are standardized so that they have mean 0 and variance 1. The LASSO estimate $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to} \quad \sum_j |\beta_j| \le t,$$

where $t \ge 0$ is a tuning parameter. This is equivalent to minimize the loss function

$$PRSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=0}^{p} |\beta_j|.$$

The tuning parameter t (or λ) controls the amount of shrinkage applied to the estimates. When the tuning parameter t is sufficiently large, the constraint has no effect and the LASSO solutions are just the usual least squares estimates. On the other hand, small values of t will cause shrinkage towards 0 (some regression coefficients may be exactly 0). That is, for small values of t, the solutions are shrunken versions of the least squares estimates. Therefore, choosing the tuning parameter is like choosing the number of predictors. Cross validation methods can be used to estimate the best value of t. Standard errors of the LASSO estimates can be obtained using bootstrap methods.

Unlike ridge regression, LASSO estimates have no analytic expressions. The LASSO solutions can be obtained using a quadratic programming method based on standard numerical analysis algorithms. But a better approach is the least angle regression method, which provides an efficient way to compute the LASSO solutions simultaneously for all values of t. The R package **lars** implements the LASSO. For problems with more predictors than observations, the least square method may not have solutions, but both ridge regression and the LASSO have solutions.

Finally, in model selection or variable selection, several models may have similar performances. In this case, the decision should be based on scientific considerations and simplicity or interpretations of the models.

4.8 Logistic Regression Models

Logistic regression models are perhaps the most widely used models in the GLM family. Logistic regression is also an important *machine learning* algorithm. A logistic regression model is used when the response y is a binary variable taking only two possible values (say, 0 or 1), which is very common in practice such as success/failure, cancer/health, death/alive, etc. In this case, it may be reasonable to assume that y follows a binomial or Bernoulli distribution.

Recall that the main idea of a regression model is to link the mean response to a set of predictors. When the response y is a binary variable, the mean response is $E(y) = P(y = 1)$, which is a value between 0 and 1, while the predictors may take any real values. Thus, a link function making the data ranges on both sides consistent is needed. Moreover, the link function should make interpretation easy. There are several choices for the link function. The most popular choice of the *logit link*

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right),$$

where $\mu = E(y) = P(y = 1)$. With the logit link, $g(\mu)$ can take any real value. The resulting GLM is the following *logistic regression model*

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p, \qquad i = 1, \cdots, n, \tag{17}$$

where $\mu_i = E(y_i) = P(y_i = 1)$, and y_i is assumed to follow a Binomial distribution with mean μ_i.

A main advantage of the logistic regression model is its attractive interpretation: $\mu_i/(1-\mu_i)$ can be interpreted as the *odds* of event "$y_i = 1$", so the parameter (say) β_j may be

interpreted as the change of odds in log-scale when predictor x_j is changed by 1 unit. Other link functions for binary responses include probit link and complementary log-log link, but they do not have the attractive interpretation as the logit link.

A logistic regression model is used when the response is a binary variable, which is very different from a continuous response variable as in a linear regression model. For a binary variable, the two values of the response variable represent two categories rather than real values. Thus, residuals are not well defined for logistic regression models. In other words, residual plots are not very useful for checking the goodness-of-fit of logistic regression models.

Model selection or variable selection for logistic regression models can be based on the χ^2 test of deviances, which is similar to the likelihood ratio test. Note that, for binary data, the deviance does not assess goodness of fit and it is not approximately χ^2 distributed. Thus, bootstrap methods may be preferred. For comparing two nested models, however, the χ^2 test based on the difference of the two deviances is still reasonable for binary data. For testing the significance of a single continuous predictor, say testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, the usual Wald-type z-test based on

$$z = \hat{\beta}_j / se(\hat{\beta}_j)$$

can still be used, although the deviance test is preferred. Other common model selection methods, such as AIC or BIC criteria, can also be used for model or variable selections. That is, models with small values of AIC or BIC are preferred. In data analysis, we may also plot $\hat{P}(y = 1)$ versus (say) x to help interpret the results (we may use boxplots if x is categorical).

For logistic regression models, the MLEs of model parameters are usually obtained based on an iterative algorithm such as the Newton's method. In some cases, the algorithm may not convergence. The reasons of *non-convergence* of a model may be as follows: (i) too many 0's or too many 1's in the response values; and (ii) multicollinearity problem. In this case, a possible solution is to simplify the model. One may also perform two-sample t-tests or bootstrap test to select potentially important predictors.

Example 1. Consider the well-known "wage data" based on the Current Population Survey (CPS). Suppose that wages below \$7.78 are considered to be "low" (defined as below median) and wages above \$7.78 are considered to be "high" (defined as above median). We wish to determine whether low/high wages are related to the variables in the dataset. This is an alternative way to study the relationship between wage and other variables, and the results may be compared with the linear regression models in the previous chapter. Note that, by

converting continuous data to binary data, some information may be lost, but binary data sometimes have more attractive and simpler interpretation such as high or low wages rather than the actual wage values.

```
wage.dat <- read.table("wage.data", head=T)
attach(wage.dat)
## For categorical variables, the function "factor" should be used to identify them
race <- factor(race, labels=c("other", "hispanic", "white"))   # categorical variable
occupation <- factor(occupation,  labels=c("management","sale","clerical",
                     "service","professional","other"))   # categorical variable
sector <- factor(sector, labels=c("other","manufacturing",
                 "construction"))   # categorical variable
# We convert the original response "wage" into a binary variable
# based on whether "wage>7.78".
wage3 <- as.numeric(wage>7.78)
```

Note that categorical variables "race, occupation, sector" need to be declared in R using the **factor()** function so that they will not be treated as numerical variables by computer. When a continuous variable (such as "wage") is converted into a binary variable, some information will be lost, but it sometimes gains easier interpretation in practice (such as high or low wage).

```
# Fit a logistic regression model with all predictors
fit1.glm <- glm(wage3 ~ education+south+sex+experience+union+age+
                        race+occupation+sector+marr, family=binomial)
summary(fit1.glm)
.......
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             16.8137   803.1175   0.021 0.983297
education                3.7579   133.8528   0.028 0.977603
south                   -0.5260     0.2323  -2.265 0.023528 *
sex                     -0.8326     0.2350  -3.543 0.000395 ***
experience               3.5065   133.8528   0.026 0.979101
union                    1.2282     0.2911   4.219 2.45e-05 ***
age                     -3.4642   133.8528  -0.026 0.979352
racehispanic            -0.6049     0.5864  -1.032 0.302259
racewhite                0.1849     0.3202   0.577 0.563776
occupationsale          -1.4599     0.5110  -2.857 0.004277 **
occupationclerical      -0.5582     0.4173  -1.338 0.181032
occupationservice       -1.3227     0.4532  -2.918 0.003519 **
occupationprofessional  -0.3700     0.4201  -0.881 0.378434
occupationother         -1.0871     0.4442  -2.447 0.014395 *
sectormanufacturing      0.5849     0.3028   1.931 0.053438 .
sectorconstruction       0.8012     0.5295   1.513 0.130261
marr                     0.3883     0.2251   1.725 0.084518 .
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 740.27  on 533  degrees of freedom
Residual deviance: 577.33  on 517  degrees of freedom
AIC: 611.33
```

Note that, for a categorical variable with k categories, there are $k-1$ estimates, with each estimate being the contract between that category and the baseline category (the R default is the first category). In this case, we can test the significance of whole categorical variable using R function drop1(), which compares nested models with one variable being dropped at each time.

```
drop1(fit1.glm, test="Chi")
Single term deletions
           Df Deviance    AIC     LRT   Pr(Chi)
<none>          577.33 611.33
education   1   580.57 612.57  3.2365 0.0720146 .
south       1   582.52 614.52  5.1828 0.0228115 *
sex         1   590.21 622.21 12.8727 0.0003334 ***
experience  1   579.19 611.19  1.8565 0.1730255
union       1   596.54 628.54 19.2022 1.176e-05 ***
age         1   579.00 611.00  1.6642 0.1970358
race        2   579.99 609.99  2.6526 0.2654604
occupation  5   592.81 616.81 15.4761 0.0085106 **
sector      2   582.39 612.39  5.0559 0.0798228 .
marr        1   580.32 612.32  2.9812 0.0842359 .
```

Thus, variables "south, sex, union, and occupation" seem to be highly predictive for "low/high wage" when they are tested *individually*. Next, let's do a variable selection based on AIC values using a stepwise method to see which variables will be selected simultaneously.

```
# Stepwise method for variable selection
fit2.glm <- step(fit1.glm)
 summary(fit2.glm)
......
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            -3.945607   0.984340  -4.008 6.11e-05 ***
education               0.301113   0.057976   5.194 2.06e-07 ***
south                  -0.570361   0.229312  -2.487 0.012873 *
sex                    -0.816617   0.233073  -3.504 0.000459 ***
experience              0.042208   0.009787   4.313 1.61e-05 ***
union                   1.190237   0.287789   4.136 3.54e-05 ***
occupationsale         -1.419285   0.506888  -2.800 0.005110 **
occupationclerical     -0.559147   0.413983  -1.351 0.176807
occupationservice      -1.321571   0.450408  -2.934 0.003344 **
occupationprofessional -0.344214   0.416351  -0.827 0.408385
occupationother        -1.043318   0.440786  -2.367 0.017936 *
sectormanufacturing     0.581945   0.301277   1.932 0.053409 .
sectorconstruction      0.844982   0.529516   1.596 0.110542
marr                    0.389090   0.223787   1.739 0.082095 .
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 740.27  on 533  degrees of freedom
Residual deviance: 581.72  on 520  degrees of freedom
AIC: 609.72
```

We see that the stepwise method based on AIC values only removes variables "age" and "race" from the original model. Let's compare the two nested models using the deviance test to see if they are significantly different.

```
# Compare the two nested models using deviance test
anova(fit1.glm, fit2.glm)
Analysis of Deviance Table
Model  Resid. Df Resid. Dev Df  Deviance
1       517       577.33
2       520       581.72     3    4.3823
```

We see that the difference in deviances from the two models is 4.3823, and the two models differ by 3 parameters (3 degrees of freedom). Compared with a $\chi^2(3)$-distribution 5th-percentile critical value, we see that models 1 and 2 are not significantly different (at 5% level), and the p-value of the χ^2 test is 0.22. Thus, we should choose the smaller model, which is model 2.

Note that model selections or variable selections often involve comparing nested models. R function anova() can also be used to test each covariate sequentially, i.e., it can compare models by dropping/adding one covariate at a time using a χ^2 test. This approach is especially helpful for categorical covariates since the Wald-type z-tests may not be good choice for categorical variables.

```
anova(fit2.glm, test="Chi")
Analysis of Deviance Table
Terms added sequentially (first to last)
           Df  Deviance  Resid. Df Resid. Dev P(>|Chi|)
NULL                         533     740.27
education   1   49.794       532     690.48 1.707e-12 ***
south       1   10.001       531     680.48  0.001564 **
sex         1   16.838       530     663.64 4.072e-05 ***
experience  1   38.594       529     625.05 5.218e-10 ***
union       1   18.576       528     606.47 1.632e-05 ***
occupation  5   16.282       523     590.19  0.006083 **
sector      2    5.442       521     584.75  0.065803 .
marr        1    3.029       520     581.72  0.081773 .
```

From the above results, we see that covariates education, south, sex, experience, union, and occupation seem significant when tested sequentially, while covariates sector and marr are not.

Model diagnostics for logistic regression models can be challenging since the response variable is binary, representing two (unordered) categories. One important approach is to check if the assumed binomial distribution holds or if there is an over-dispersion problem. We can use the *quasi-likelihood* method or the *quasibinomial* model and compare the results:

```
fit3.glm <- glm(wage3 ~ education+south+sex+experience+union+
                          occupation+sector+marr, family=quasibinomial)
......
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -3.945607   0.997052  -3.957 8.63e-05 ***
education               0.301113   0.058725   5.127 4.15e-07 ***
south                  -0.570361   0.232274  -2.456 0.014393 *
sex                    -0.816617   0.236083  -3.459 0.000587 ***
experience              0.042208   0.009914   4.258 2.45e-05 ***
union                   1.190237   0.291506   4.083 5.14e-05 ***
occupationsale         -1.419285   0.513435  -2.764 0.005907 **
occupationclerical     -0.559147   0.419330  -1.333 0.182975
occupationservice      -1.321571   0.456225  -2.897 0.003929 **
occupationprofessional -0.344214   0.421728  -0.816 0.414760
occupationother        -1.043318   0.446479  -2.337 0.019830 *
sectormanufacturing     0.581945   0.305168   1.907 0.057076 .
sectorconstruction      0.844982   0.536355   1.575 0.115768
marr                    0.389090   0.226678   1.716 0.086667 .

(Dispersion parameter for quasibinomial family taken to be 1.025996)
```

The estimated dispersion parameter is $\hat{\tau} = 1.02$, very close to 1. Thus, the assumed binomial distribution may hold or there may be no over-dispersion problem, suggesting that the results from *fit2.glm* may be reliable.

Influential observations can be checked using R function cooks.distance().

```
cooks.distance(fit2.glm)
          1           2           3           4           5           6
3.082100e-04 1.651538e-03 8.009031e-04 3.612468e-04 1.119035e-03 1.380721e-03
.....

plot(cooks.distance(fit2.glm), ylab="Cook's distance", main="Cook's Distance")
```

There is one observation which may be influential. We performed an analysis with this observation removed, but we obtained similar results as that based on all observations. Thus, there seem no obvious influential observations for this dataset.

Recall that this dataset was also analyzed based on linear regression models in previous chapter, where a similar set of significant covariates was selected, but the interpretations of the regression coefficients are different for linear regressions and logistic regressions. These results indicate that the selected covariates may be related to wage, whether wage is viewed as a continuous variable or a discrete variable.

The final selected model should not just be based on statistical criteria such as AIC values or χ^2 tests. We should also take into account scientific considerations and interpretations of the results, especially when choosing several models with similar AIC values.

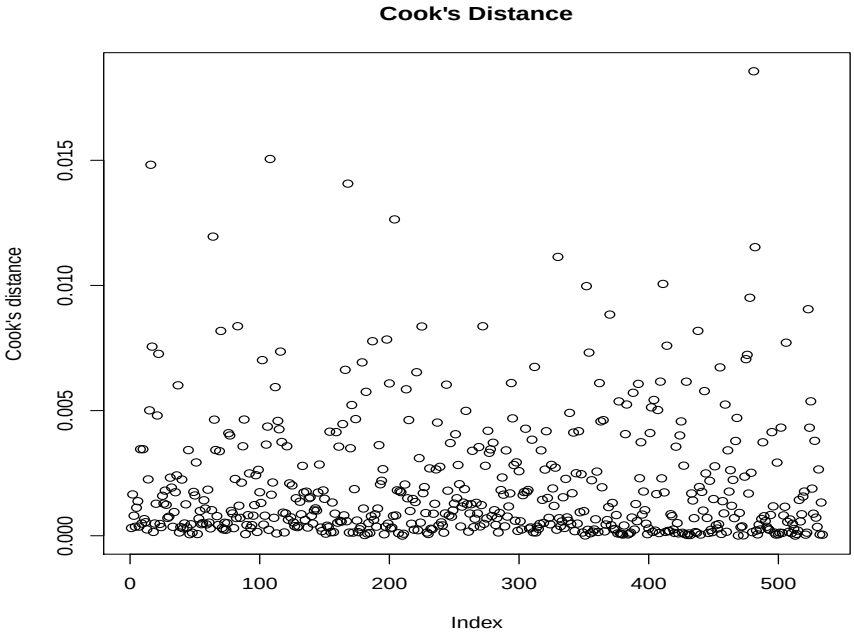

Figure 3: Cook's distance plot

4.9 Poisson Regression Models

If the response y is a count, it may be reasonable to assume that y follows a Poisson distribution. Then, the Poisson GLM is a natural choice. For the Poisson GLM, the standard link function is the following log-link

$$g(\mu) = \log(\mu),$$

where $\mu = E(y)$. The resulting GLM is called a *Poisson GLM* and it can be written as follows

$$\log(\mu_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p, \tag{18}$$

where $\mu_i = E(y_i)$, and y_i is assumed to follow a Poisson distribution with mean μ_i. The regression parameters in the Poisson GLM have attractive interpretation: parameter (say) β_j represents the change in the mean response (in log-scale) when covariate x_j is changed by 1 unit, which is similar to that for linear models but in log-scale.

When the response is a count, it may be reasonable to assume that the count follows a Poisson distribution and thus the Poisson GLM is a natural choice. However, in practice,

the observed count data do not necessarily follow a Poisson distribution. In other words, the Poisson distributional assumption may not hold in practice. This can be seen when the variation in the observed data is much larger or smaller than the mean value, since the variance and the mean should be the same if the distribution is Poisson. That is, when over-dispersion or under-dispersion problems arise in a given situation, the count data do not follow a Poisson distribution.

We focus on the over-dispersion problem since it's more common. Over-dispersion arises when the observed variance in the response data is greater than the theoretical variance $Var(y_i) = \mu_i$. If an over-dispersion problem exists, the parameter estimates based on an assumed Poisson GLM will still be consistent, but the standard errors will be wrong. Thus, the overdispersion problem must be addressed in order for the statistical inference to be valid. If there is an over-dispersion problem, we can introduce a dispersion parameter ϕ such that $Var(y) = \phi E(y)$. This dispersion parameter can be estimated from the data. We will illustrate this in the example below. Where there is an over-dispersion problem, an alternative model of choice is the *negative binomial* GLM, rather than Poisson GLM.

To check the goodness of fit of a Poisson model, we can check the deviance against a χ^2 distribution. For comparing two nested models, we can also use a χ^2 distribution based on the difference of the deviances of the two models. For Poisson models, an alternative goodness of fit measure is the Pearson's χ^2 statistic

$$\chi^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i},$$

which is like scaled (squared) differences between observed counts and estimated counts based on the assumed model, so a large value indicates a poor fit of the assumed model.

The above model diagnostic methods are illustrated in the example below.

Example 2. The dataset contains 1681 householders in Copenhagen who were surveyed on the type of rental accommodation they occupied, the degree of contact they had with other residents, their feeling of influence on apartment management and their level of satisfaction with their housing conditions. These correspond to the following variables: Type, Contact, Influence, and Satisfaction. Satisfaction (Sat) and Influence (Infl) are three-level categorical variables: Low, Medium, High. Contact (Cont) is a 2-level categorical variable: Low and High. Type has 4 categories: Tower, Apartment, Atrium, Terrace. We model Frequencies (counts) using a Poisson GLM and relate the counts to the variables of interest.

```
> library(MASS)
```

```
> attach(housing) # the dataset is available in the R MASS library.
# partial data is shown below
> housing
      Sat    Infl       Type Cont Freq
1     Low     Low      Tower  Low   21
2  Medium     Low      Tower  Low   21
3    High     Low      Tower  Low   28
4     Low  Medium      Tower  Low   34
5  Medium  Medium      Tower  Low   22
...........

# Fit a Poison GLM with 4 categorical covariates
> house.glm0 <- glm(Freq~Infl+Type+Cont+Sat, family=Poisson)
> summary(house.glm0)
.....
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.03545    0.06576  46.160  < 2e-16 ***
InflMedium    0.04978    0.05579   0.892  0.37226
InflHigh     -0.46206    0.06424  -7.193 6.34e-13 ***
TypeApartment 0.64841    0.06170  10.509  < 2e-16 ***
TypeAtrium   -0.51500    0.08176  -6.299 2.99e-10 ***
TypeTerrace  -0.36745    0.07817  -4.701 2.59e-06 ***
ContHigh      0.30575    0.04935   6.195 5.82e-10 ***
Sat.L         0.11592    0.04038   2.871  0.00409 **
Sat.Q         0.26292    0.04515   5.824 5.76e-09 ***
(Dispersion parameter for Poisson family taken to be 1)
    Null deviance: 833.66  on 71  degrees of freedom
Residual deviance: 295.35  on 63  degrees of freedom
AIC: 654.32
```

The above results show that all covariates are highly significant. Note that, for categorical covariates, the baseline category is not shown. The estimate for each category of a categorical variable is relative to the baseline category, which by software default is the first category (this default can be changed by users).

For a Poisson GLM, if the assumed Poisson distribution holds, the mean should be equal to the variance. Let's check to see if this is consistent with the observed data by plotting the estimated variance in the data against the estimated mean.

```
# plot estimated mean versus estimated variance (in log-scale)
> plot(log(fitted(house.glm0)), log((Freq-fitted(house.glm0))^2),
     xlab=expression(hat(mu)), ylab=expression((y-hat(mu))^2),
     xlim=c(0,8),ylim=c(-8,8))
> abline(0,1)
# We can estimate the dispersion parameter by
dp <- sum(residuals(house.glm0,type="pearson")^2)/house.glm0$df.res
> dp
[1] 4.85598
```

Figure 4 shows the plot. From this plot, we see that the estimated variance is much larger

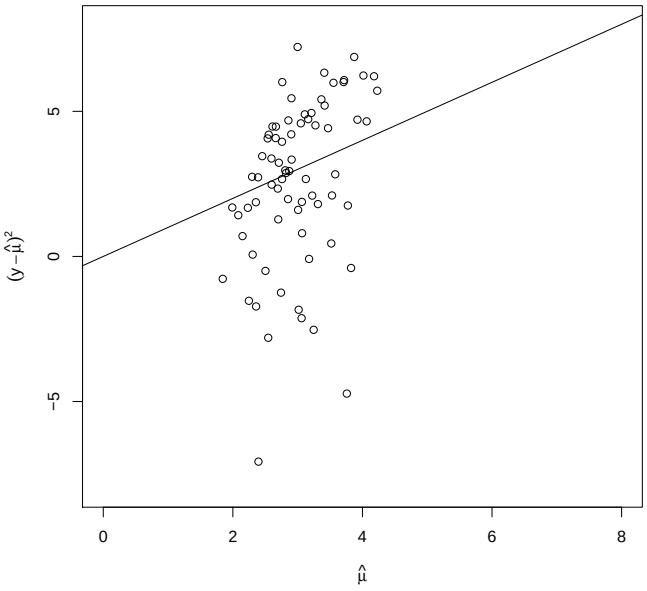

Figure 4: Estimated mean (in log-scale) versus estimated variance (in log-scale). Over-dispersion is obvious.

than the estimated mean (both in log-scale), so there is an over-dispersion. The estimated dispersion parameter is 4.85, which is much larger than 1, so it confirms the over-dispersion problem. Therefore, the assumed Poisson distribution for the Poisson GLM does not hold. This over-dispersion problem must be addressed for reliable inference.

The over-dispersion problem will only affect the standard errors of the parameter estimates in the GLM, so for correct inference of the parameters we should adjust the standard errors. That is, the parameter estimates for the GLM are still consistent even if there is overdispersion, but the standard errors may be incorrect. The following results give more reliable standard errors by incorporating the dispersion parameter.

```
# Introducing a dispersion parameter "dp" to correct standard errors and p-values
> summary(house.glm0, dispersion=dp)
........
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.03545    0.14491  20.947  < 2e-16 ***
InflMedium    0.04978    0.12294   0.405  0.68555
InflHigh     -0.46206    0.14156  -3.264  0.00110 **
TypeApartment 0.64841    0.13597   4.769 1.85e-06 ***
TypeAtrium   -0.51500    0.18016  -2.859  0.00426 **
```

```
TypeTerrace   -0.36745    0.17225  -2.133  0.03291 *
ContHigh       0.30575    0.10875   2.811  0.00493 **
Sat.L          0.11592    0.08898   1.303  0.19266
Sat.Q          0.26292    0.09949   2.643  0.00822 **
(Dispersion parameter for poisson family taken to be 4.85598)
    Null deviance: 833.66  on 71  degrees of freedom
Residual deviance: 295.35  on 63  degrees of freedom
AIC: 654.32
```

The above standard errors, and thus the corresponding z-values and p-values, are more reliable than the original ones since the dispersion problem is addressed by introducing a dispersion problem.

Instead of estimating the dispersion parameter and updating the fitting, we can simply use the *quasi-likelihood* or *quasi-poisson* method:

```
house.glm1 <- glm( Freq ~ Infl + Type + Cont +Sat, family = quasipoisson)
summary ( house.glm1 )
......
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.03545    0.14491  20.947  < 2e-16 ***
InflMedium     0.04978    0.12294   0.405  0.68692
InflHigh      -0.46206    0.14156  -3.264  0.00178 **
TypeApartment  0.64841    0.13597   4.769 1.14e-05 ***
TypeAtrium    -0.51500    0.18016  -2.859  0.00576 **
TypeTerrace   -0.36745    0.17225  -2.133  0.03681 *
ContHigh       0.30575    0.10875   2.811  0.00657 **
Sat.L          0.11592    0.08898   1.303  0.19740
Sat.Q          0.26292    0.09949   2.643  0.01036 *

(Dispersion parameter for quasipoisson family taken to be 4.85598)
```

We obtain the same results. The dispersion parameter is estimated to be 4.85, larger than 1, indicating a possible over-dispersion problem.

For categorical predictors, we can test the significance of each predictor by the F-test as follows, which compares nested models by dropping one predictor at a time.

```
> drop1(house.glm0, test="F")
Single term deletions
       Df Deviance     AIC F value     Pr(F)
<none>      295.35  654.32
Infl    2   373.87  728.84  8.3740 0.0005958 ***
Type    3   671.65 1024.62 26.7556 2.797e-11 ***
Cont    1   334.18  691.15  8.2831 0.0054578 **
Sat     2   340.01  694.98  4.7628 0.0118512 *
```

We can see that all the categorical predictors are significant at 5% level. So all the predictors may be included in the Poisson GLM.

4.10 Extensions

The exponential family contains many other distributions, in addition to the binomial distribution and Poisson distribution. Thus, there are other GLM models. For example, we may consider the negative binomial distribution and the Gamma distribution for some data. While these models may be useful in some cases, they are relatively less commonly used, so we will skip the details here.

The most commonly used GLM is probably the logistic regression model based on the binomial distribution. The model may be extended to allow for a *multinomial distribution* when the response variable has more than two categories. If the categories of the response have orders or if the response variable is an ordinal measurement, we may consider the *proportional odds models*. If the response variable is unordered or is a nominal measurement, we may consider a *multinomial logit model*.

The standard GLM assumes that the response data are *independent*. In practice, the response data may be *clustered* or the response may be repeatedly measured over time (or space), leading to longitudinal data or repeated measurements. In these cases, the standard GLM may be extended to *generalized linear mixed models (GLMMs)* and *generalized estimating equations (GEEs)*. Since longitudinal data and clustered data are very common in practice, GLMMs and GEE models will be discussed in details in the next section.

In a standard GLM, the predictors enter the model in a *linear* fashion (i.e., the linear predictor η_i). In practice, however, the predictors may be associated with the response in complicated ways. The standard GLM may be extended to the following *generalized additive models (GAMs)*:

$$g(E(y_i)) = \beta_0 + q_1(x_1) + q_2(x_2) + \ldots + q_p(x_p),$$

where the functions $q_k(\cdot)$ are *unknown* arbitrary smooth functions and are estimated from the data.

Exercises

1. For a random variable Y following a distribution in the exponential family, prove that

$$E(Y) = \mu = \partial b(\theta)/\partial\theta, \qquad Var(Y) = a(\phi)\partial^2 b(\theta)/\partial\theta^2.$$

2. Show that, for linear regression models, the deviance is simply the familiar residual sum of squares $RSS = \sum_i (y_i - \hat{\mu}_i)^2$.

3. In Example 1, the Cook's distance plot shows that there is one observation which may be influential. Fit a GLM model without this observation and compare the results obtained in Example 1.

4. In biochemistry, the Michaelis-Menten model is the one of the simplest and best-known approaches to enzyme kinetics. The model takes the form of an equation describing the rate of enzymatic reactions by relating reaction rate y to x which is the concentration of a substrate:

$$y = \frac{\alpha x}{\beta + x},$$

where α and β are unknown parameters. What transformation will convert the above model to a linear model?

5. In data analysis, it is desirable to fit different models to the same dataset and compare the results. This is because each model has its own assumptions which may not hold, so we should not rely conclusions on a single model. Moreover, different models allow us to see the problem in different ways, which may give us additional insights. If different models lead to the same or similar conclusions, we are more confident about these conclusions than those based on a single model. In the following, you are asked to analyze a dataset using different models.

The dataset contains crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable "crime rate" depends on the other variables measured in the study. Variable names and definitions for the dataset are as follows (in the order from left to right in the dataset):

1. `R: Crime rate: number of offenses reported to police per million population`
2. `Age: The number of males of age 14-24 per 1000 population`
3. `S: Indicator variable for Southern states (0 = No, 1 = Yes)`
4. `Ed: Mean number of years of schooling x 10 for persons of age 25 or older`
5. `Ex0: 1960 per capita expenditure on police by state and local government`
6. `Ex1: 1959 per capita expenditure on police by state and local government`
7. `LF: Labor force participation rate per 1000 civilian urban males age 14-24`
8. `M: The number of males per 1000 females`
9. `N: State population size in hundred thousands`
10. `NW: The number of non-whites per 1000 population`
11. `U1: Unemployment rate of urban males per 1000 of age 14-24`
12. `U2: Unemployment rate of urban males per 1000 of age 35-39`

13. W: Median value of transferable goods and assets or family income
(in tens of dollars)
14. X: The number of families per 1000 earning below 1/2 the median income

The main objective is to determine which variables are significantly predictive for the crime rate. Please use three different regression models to answer this question.
a) Fit a linear regression model to the crime rate data. What variables are most predictive for the crime rate?
b) A crime rate may be viewed as "high" if it is above 95 and "low" otherwise (a different cutoff value, such as 100, may be used if convergence is a problem). Fit a logistic regression model to the data. What variables are most predictive for a "high" crime rate?
c) Round off crime rate numbers to the nearest integers and then fit a Poisson GLM to the new crime rate data (this is roughly OK here since these crime rates may be viewed as counts). What variables are most predictive for the crime rate?
d) Compare the results from 1) – 3), and comment on what you find. What do you learn from the analysis? What is your final conclusion?

6. Likelihood method is the standard approach for parameter estimation and inference for GLMs. MLEs have nice asymptotic properties such as consistency, efficiency, and asymptotic normality. Standard errors of the MLEs are based on asymptotic formulae. In practice, however, the sample size may not be large enough, which implies that the asymptotic results may not hold exactly and the estimates may not be optimal. In this assignment, you are asked to evaluate the performances of MLEs for logistic regression models via a simulation study.

Consider the following simple logistic regression model

$$\text{logit}\left(\frac{P(y_i=1)}{1-P(y_i=1)}\right)=\beta_0+\beta_1 x_i, \qquad i=1,\cdots,n.$$

MLEs of the parameters β_0 and β_1 are available in most software, such as the "glm" function in R. To evaluate the performances of the MLEs when the sample size n is not large, we can conduct a simulation study as follows: Assume $n=20, \beta_0=1, \beta_1=2$, $x_i \sim N(0,1)$, and y_i follows a Binomial (Bernoulli) distribution. Generate 1000 datasets $\{(x_i,y_i), i=1,2,\cdots,n\}$ from the assumed distributions. For each generated dataset, obtain the MLEs of β_0 and β_1. Then compute the average biases and mean-square-errors (MSE) of the MLEs by comparing them with the true values $\beta_0=1, \beta_1=2$ over the 1000 replications (so, e.g., bias of $\hat{\beta}_1 = \sum_{j=1}^{m}(\hat{\beta}_{1j}-\beta_1)/m$, MSE of $\hat{\beta}_1=\sum_{j=1}^{m}(\hat{\beta}_{1j}-\beta_1)^2/m$, where $\hat{\beta}_{1j}$ is the MLE of β_1 from the j-th

iteration, $m = 1000$). Repeat the above process again with $n = 40$. Summarize your results in tables (or figures), and report your conclusions. Do the performance of MLE improves as n increases?

5 Generalized Linear Mixed Models

5.1 Introduction

Longitudinal studies are very popular in practice. In a longitudinal study, individuals are followed over a period of time and data are collected for each individual at *multiple* time points. These data, which are collected repeatedly over time, are called *longitudinal data*. Longitudinal data are thus closely related to *repeated measures data*, for which repeated or multiple measurements are obtained on each individual in the study but these repeated measurements are not necessarily collected over time (e.g., the repeated measurements can be collected over different locations of a city). In economics and sociology, longitudinal studies are often called *panel studies*, and longitudinal data are thus called *panel data*. Multivariate data, longitudinal data, and repeated measurement data are all examples of *correlated data*.

Examples of longitudinal data include measurements of individuals' weights over time, measurements of individuals' blood pressures over time, and measurements of subjects' depression levels over time. Figure 5 shows an example of longitudinal data. A key characteristic of longitudinal data is that the repeated measurements of a variable on each individual are likely to be *correlated*, since they are data collected over time from the same individuals. Ignoring this correlation in data analysis may lead to inefficient or biased results. Therefore, in the analysis of longitudinal data, a major consideration is to incorporate the correlation of the repeated measurements, as in multivariate analysis. A main advantage of longitudinal studies is that they allow us to study *changes over time* for variables of interests.

Example: depression data

In this longitudinal study, 239 adult subjects with depression were randomly assigned into a treatment group and a control group. Depression scores, along with other variables such as age, gender, and anxiety, were repeatedly measured at day 0 (right before the treatment), and months 4, 12, 24, and 60 during the treatment. The depression scores range from 0 to 5, with a score of 5 being most depressed and a score of 0 being not depressed. Figure 6 shows the longitudinal depression scores. One of the objective is to study the effectiveness of the treatment and how subjects' depression levels change over time.

5.2 Models for Longitudinal Data

In many longitudinal studies, researchers are often interested in understanding the *systematic* variation between individuals. Thus, regression models for longitudinal data are very

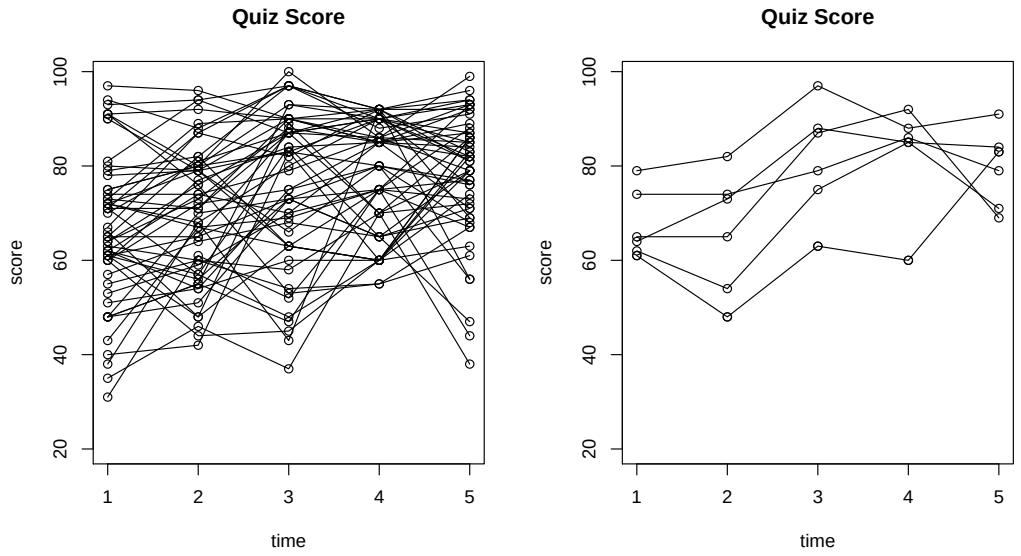

Figure 5: Quiz scores over time. Left: all students. Right: six randomly selected students.

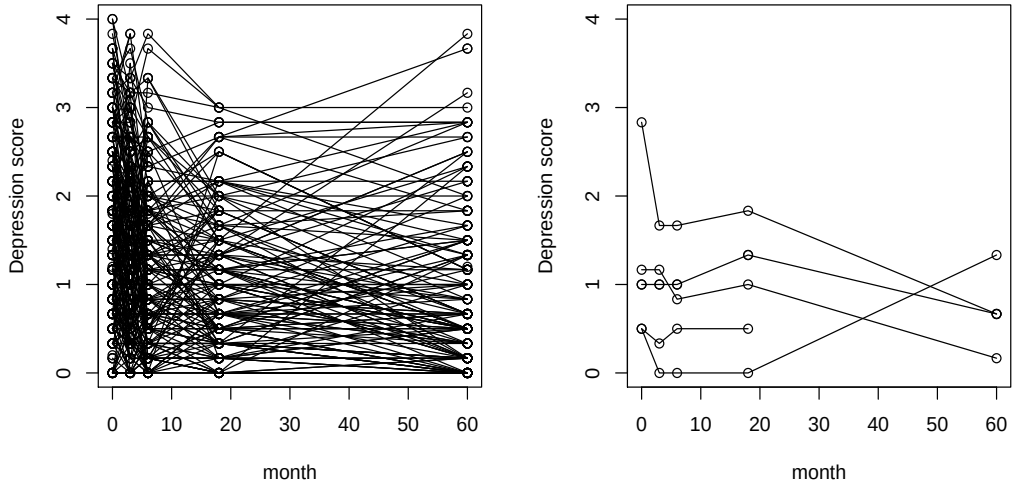

Figure 6: Depression scores for all 239 subjects (left figure) and for 5 randomly selected subjects (right figure).

useful. In regression models, covariates are introduced to partially explain the systematic between-individual variation in the response variable. Note that, in a regression model for longitudinal data, the response data are longitudinal measurements on a variable of interest, but covariates or predictors can either be longitudinal data (time-dependent covariates, including time) or cross-sectional data (time-independent covariates).

A key characteristic of longitudinal data is that observations within the same individual may be *correlated*, although observations between different individuals are assumed to be independent. A major consideration of various statistical methods for longitudinal data analysis is to incorporate the within-individual correlation in different ways.

Three approaches are commonly used in the regression analysis of longitudinal data, and they incorporate the within-individual correlation in different ways:

- The first approach assumes that the repeated measurements within an individual are correlated because these measurements share the same unobserved characteristics of the individual. The unobserved characteristics of an individual can be represented by *random effects* , or *individual effects*, of that individual. Thus, regression models for longitudinal data can be obtained by introducing random effects in the corresponding models for cross-sectional data, leading to *random effects models* or *mixed effects models*.

- The second approach models the longitudinal mean process and the variance-covariance structure separately, based on a set of estimating equations. Such models are often called *marginal models* or *generalized estimating equation (GEE) models*.

- The third approach assumes that the repeated measurements within an individual are correlated because the longitudinal process may be viewed as a Markov process. In other words, the correlation of the within-individual measurements is incorporated through an assumed Markov structure, i.e., the repeated measurements within an individual may be viewed as a transitional process. So such models are called *transitional models*.

Each of the above three approaches has its advantages and limitations. In practice, the choice of the methods for data analysis is often based both on statistical considerations and on scientific considerations.

Mixed effects regression models can be obtained from the corresponding regression models for cross-sectional data by introducing random effects in the models. These models are particularly useful for longitudinal data with large between-individual variation since they

allow model parameters to vary across individuals. They also allow for *individual-specific* inference, as well as for population-average inference as in other regression models.

Marginal GEE models may be viewed as extensions from the corresponding quasi-likelihood models. These models are based on estimating equations similar to likelihood equations but with no distributional assumptions. A main advantage of GEE models is that they only require specifications of the mean and variance-covariance structures of the data, without distributional assumptions. GEE models may be particularly useful for non-normal data, such as binary data or count data, in which over-dispersion problems may arise.

Transitional models may be useful if certain Markov correlation structures are reasonable for the longitudinal processes. A transitional model has a similar form as a classical regression model for cross-sectional data, if previous response observations are viewed as covariates for the current response observation. So model formulations and inference are similar to classical regression models.

For the foregoing three modelling approaches for longitudinal data, mixed effects models may be preferred if the between-individual variation is not small, GEE models may be preferred if distributional assumptions are questionable, and transitional models may be preferred if certain Markov structures are reasonable. In general, if distributional assumptions for the data are reasonable, models with distributional assumptions often produce more efficient estimates than models without distributional assumptions. In the analysis of longitudinal data, it is often desirable to consider different modelling approaches and then compare the results. If the results are similar, conclusions based on these results may be reliable. Otherwise, further investigation of the models and methods may be needed.

5.3 Linear Mixed Effects Models

A linear mixed effects (LME) model can be obtained from a standard linear regression model for cross-sectional data by introducing random effects to the parameters that vary substantially across individuals (i.e., allowing these parameters to be individual-specific).

Let y_{ij} be the response value for individual i at time t_{ij}, $i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i$. Consider the following simple linear regression model for longitudinal data

$$y_{ij} = \beta_0 + \beta_1 t_{ij} + e_{ij}, \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i,$$

where e_{ij} is a random error. If the data show that the intercept β_0 varies substantially for different individuals, as in the depression data in Figure 6, we may introduce a random effect for the intercept β_0, i.e., we may allow the intercepts to be individual-specific. We

thus obtain the following LME model

$$\begin{aligned} y_{ij} &= (\beta_0 + b_{0i}) + \beta_1 t_{ij} + e_{ij} \\ & \qquad i = 1, 2, \cdots, n,\ j = 1, 2, \cdots, n_i, \end{aligned} \tag{19}$$

where b_{0i} is a random effect for individual i, $\beta_{0i} = \beta_0 + b_{0i}$ is the individual-specific intercept for individual i, and the slope β_1 is assumed to be the same (fixed) for all individuals. The parameters β_0 and β_1 are called *fixed effects* or *population parameters*, since they are fixed for all individuals. We often assume that e_{ij} i.i.d. $\sim N(0, \sigma^2)$ and $b_{0i} \sim N(0, d^2)$, and e_{ij} and b_i are independent.

From the above example, we see that we can choose random effects informally based on the *heterogeneous* feature of the data. Formally, we can choose the random effects based on standard tests such as the likelihood ratio test or based on standard model selection methods such as AIC or BIC criteria.

General LME models can be described as follows. Let $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})^T$ be the n_i repeated measurements of the response variable y on individual i, and let $\mathbf{e}_i = (e_{i1}, e_{i2}, \cdots, e_{in_i})^T$ be the corresponding random errors of the repeated measurements, $i = 1, 2, \cdots, n$. A general form of LME models can be written as

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \mathbf{e}_i, \qquad i = 1, 2, \cdots, n, \tag{20}$$

$$\mathbf{b}_i \sim N(0, D), \qquad \mathbf{e}_i \sim N(0, R_i), \tag{21}$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is a $p \times 1$ vector of fixed effects, $\mathbf{b}_i = (b_{i1}, \cdots, b_{iq})^T$ is a $q \times 1$ vector of random effects, matrix X_i $(n_i \times p)$ and matrix Z_i $(n_i \times q)$ are known design matrices which often contain covariates (including times), D is a $q \times q$ covariance matrix of the random effects, and R_i is a $n_i \times n_i$ covariance matrix of the within-individual random errors. The standard assumptions for models (20) and (21) are (i) the individuals are *independent*, (ii) the errors ϵ_i and the random effects $\mathbf{b}_i$ have mean zero, and (iii) the errors ϵ_i and the random effects $\mathbf{b}_i$ are independent and both are normally distributed.

In LME model (20) and (21), the repeated measurements $\{y_{i1}, y_{i2}, \cdots, y_{in_i}\}$ of the response can be taken at *different* time points for different individuals, and the number of measurements n_i may vary across individuals. Thus, a LME model allows *unbalanced data* in the response. In other words, a LME model allows missing data in the response, assuming the missing data are missing at random. Note that LME models differ from linear regression models for cross-sectional data by the term $Z_i\mathbf{b}_i$. In practice, we often assume that $R_i = \sigma^2 I_{n_i}$ in LME model (20) and (21), i.e., the within-individual measurements are assumed to be conditionally independent with constant variance given the random effects.

This assumption greatly reduces the number of parameters in the model and may avoid potential identifiability problems.

Statistical inference for LME models is typically based on the maximum likelihood method. MLE of the fixed parameters $\boldsymbol{\beta}$ can then be obtained using an iterative algorithm such as an *expectation-maximization (EM) algorithm* or a *Newton-Raphson method*.

An R Example (Growth of children).

To study the growth of children, the distance from the pituitary gland to the pterygo-maxillary fissure is measured every two years from 8 years of age until 14 years of age. A sample of 27 children – 16 males and 11 females was obtained by orthodontists from x-rays of the children's skulls. (The dataset is denoted by "Orthodont".) We use this example to illustrate statistical modelling procedures using linear mixed-effects (LME) models.

```
library(nlme)  # get the NLME library for mixed effects models
attach(Orthodont)  # get dataset (it's internal)
# Here is part of the data
Orthodont
    distance age Subject    Sex
1       26.0   8     M01   Male
2       25.0  10     M01   Male
3       29.0  12     M01   Male
4       31.0  14     M01   Male
5       21.5   8     M02   Male
6       22.5  10     M02   Male
.............

# Now let's fit a LME model with random effects on both
# the intercept and the slope
lme.fit1 <- lme(distance~age, data=Orthodont,
                  random = ~age  | Subject, method = "ML")
summary(lme.fit1)
...
Fixed effects: distance ~ age
                Value Std.Error DF   t-value p-value
(Intercept) 16.761111 0.7678975 80 21.827278       0
age          0.660185 0.0705779 80  9.353997       0
...
```

```
# Add Sex and interaction
lme.fit2 <- update(lme.fit1, fixed=distance~Sex*age)
summary(lme.fit2)
.......
Fixed effects: distance ~ Sex + age + Sex:age
                  Value Std.Error DF   t-value p-value
(Intercept)   16.340625 0.9987521 79 16.361042  0.0000
SexFemale      1.032102 1.5647438 25  0.659598  0.5155
age            0.784375 0.0843294 79  9.301322  0.0000
SexFemale:age -0.304830 0.1321188 79 -2.307238  0.0237
......
# There seem significant difference between boys and girls

# Let's do an ANOVA test for comparing the two models
anova(lme.fit1, lme.fit2)
            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lme.fit1     1  6 451.2116 467.3044 -219.6058
lme.fit2     2  8 443.8060 465.2630 -213.9030 1 vs 2 11.40565  0.0033
# Model lme.fit2 fits much better, so we should add Sex and interaction
# in the model.

# Model diagnostics for the LME fit
pdf('lmefig1.pdf')  # Check residual plot
plot(lme.fit2, resid(.,type="p")~fitted(.) | Sex,id=0.05,adj=-0.3)
# residual plot looks OK, but there may be outliers
dev.off()
pdf('lmefig2.pdf')  # Check normality assumption for e
qqnorm(lme.fit2, ~resid(.) | Sex)
# Normality assumption OK for within subject errors
dev.off()
pdf('lmefig3.pdf')  # Check normality assumption for b
qqnorm(lme.fit2, ~ranef(.),id=0.1,cex=0.7)
# Normality assumption OK for random effects
dev.off()
```

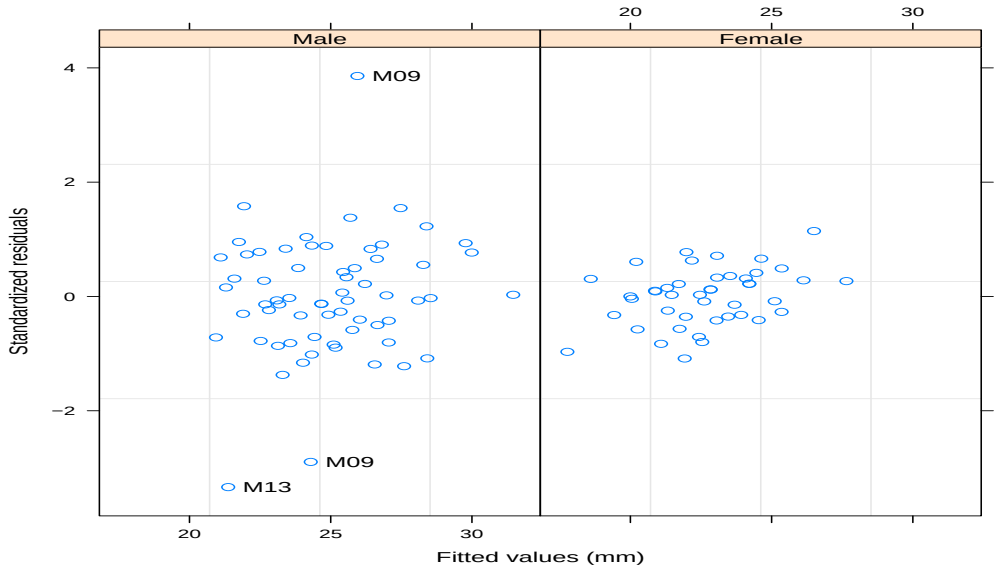

Figure 7: Residual plots.

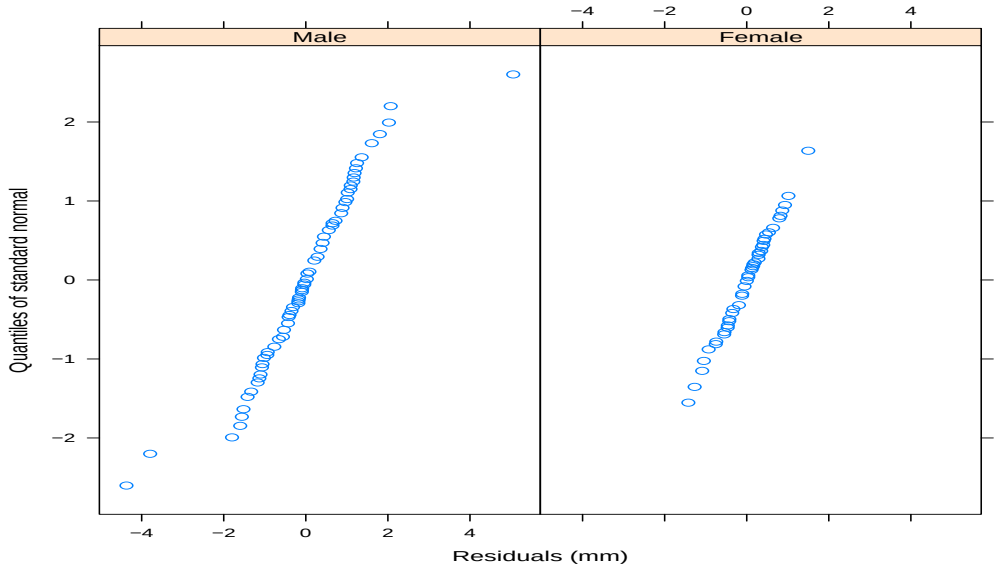

Figure 8: QQ plot for within-individual random errors.

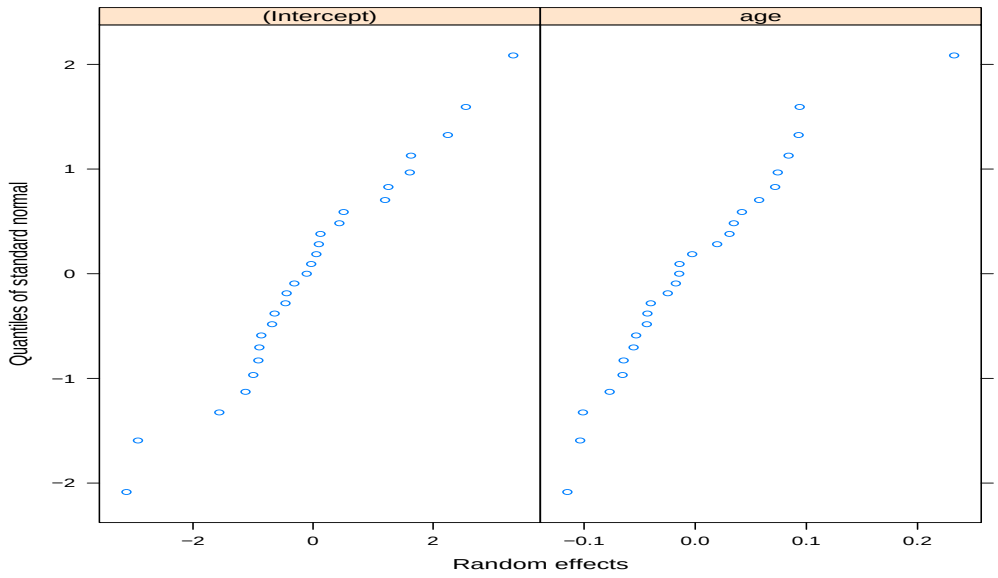

Figure 9: QQ plot for the random effects.

5.4 Generalized Linear Mixed Models

In the previous section, we see that a LME model for longitudinal data can be obtained from the corresponding linear regression model for cross-sectional data by introducing random effects in the model to account for between-individual variation and within-individual correlation. This idea can be extended to other types of regression models. For example, if the longitudinal response is a binary variable, we can extend a logistic regression model for cross-sectional data to a longitudinal regression model by introducing random effects in the logistic model. The resulting longitudinal regression model is an example of generalized linear mixed models (GLMMs). We briefly illustrate the approach as follows.

Consider the depression dataset given earlier. We may wish to study if a subject's mental distress at each measurement time throughout the study is above or below his/her baseline value. Let $y_{ij} = 1$ if the mental distress of subject i at measurement time j is above his/her baseline score and $y_{ij} = 0$ otherwise. Then, the data $\{y_{ij}, i = 1, \cdots, n; j = 1, \cdots, n_i\}$ are longitudinal binary data. Suppose that we are also interested in if the value of y_{ij} is related to the gender (x_i) of subject i. Then, the following *generalized linear mixed model (GLMM)* for longitudinal binary response may be considered:

$$\log\left(\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right) \quad = \quad \beta_{i0} + \beta_1 x_i, \tag{22}$$

$$\beta_{i0} = \beta_0 + b_i, \qquad b_i \sim N(0, d^2), \tag{23}$$

where b_i is a random effect used to incorporate the between-subject variation and within-subject correlation in the longitudinal data, $i = 1, \cdots, n;\ j = 1, \cdots, n_i$. We may also introduce a random effect for parameter β_1 if necessary. Here the responses y_{ij} are assumed to be conditionally independent and follow binomial distributions given the random effects.

More generally, a general GLMM may be written as

$$h(E(\mathbf{y}_i)) = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i, \qquad i = 1, \cdots, n, \tag{24}$$

$$\mathbf{b}_i \sim N(0, D), \tag{25}$$

where $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ are the repeated measurements within individual or cluster i, $h(\cdot)$ is a known monotone link function, X_i and Z_i are known design matrices, $\boldsymbol{\beta}$ contains fixed effects, $\mathbf{b}_i$ contains random effects, and D is a covariance matrix. It is typically assumed that the responses y_{ij} are conditionally independent and follow distributions in the exponential family, given the random effects.

One can see that the difference between a GLM and a GLMM is the random effect term $Z_i\mathbf{b}_i$ in (24). However, this term will cause much of the computational problems since the random effects $\mathbf{b}_i$ are unobservable and are nonlinear in the model. Statistical inference for a GLMM is typically based on the likelihood method. The marginal distribution for $\mathbf{y}_i$ is

$$f(\mathbf{y}_i|\boldsymbol{\beta}, D) = \int \prod_{j=1}^{n_i} \left[f(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}, \phi, \mathbf{b}_i) f(\mathbf{b}_i|D)\right] d\mathbf{b}_i, \tag{26}$$

which usually does not have an analytic or closed-form expression since the model is nonlinear in the random effects $\mathbf{b}_i$. The likelihood for all observed data is given by

$$L(\boldsymbol{\beta}, D|\mathbf{y}) = \prod_{i=1}^{n} \left\{ \int \prod_{j=1}^{n_i} \left[f(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \boldsymbol{\beta}, \phi, \mathbf{b}_i) f(\mathbf{b}_i|D)\right]\ d\mathbf{b}_i \right\}. \tag{27}$$

Likelihood inference for a GLMM can be challenging since the likelihood $L(\boldsymbol{\beta}, D|\mathbf{y})$ involve an intractable multi-dimensional integral with respect to the random effects. The following methods for estimation have been proposed

- Monte Carlo EM algorithms,
- Numerical integration methods such as the Gaussian Hermite quadrature method,
- Approximate methods based on Laplace approximations or Taylor approximations.

A Monte Carlo EM algorithm can be used by treating the random effects $\mathbf{b}_i$ as "missing data", but it sometimes can have convergence problems. Numerical integration methods are often used when the dimension of the integral is low (say 1 or 2). Approximate methods are computationally very efficient, but the accuracy of the approximation sometimes may not be satisfactory.

Example 1. *Logistic regression model with random effects.* Consider a longitudinal binary response y_{ij} taking only two possible values (say, 0 or 1), $i = 1, \cdots, n; j = 1, \cdots, n_i$. A simple logistic regression model with random intercept can be written as

$$\begin{aligned} \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \beta_{0i} + \beta_1 t_{ij} = \beta_0 + b_i + \beta_1 t_{ij}, \\ b_i &\sim N(0, d^2), \end{aligned}$$

where $\mu_{ij} = E(y_{ij}) = P(y_{ij} = 1)$ and $\beta_{0i} = \beta_0 + b_i$. A more general GLMM for binary longitudinal or clustered responses may be written as

$$\begin{aligned} \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) &= \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i, \\ \mathbf{b}_i &\sim N(0, D), \end{aligned}$$

which can be used to model longitudinal or clustered binary data.

Example 2. *Poisson regression model with random effects.* For longitudinal or clustered count responses y_{ij}, we may consider the following Poisson regression models with random effects

$$\begin{aligned} \log(\mu_{ij}) &= \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i, \\ \mathbf{b}_i &\sim N(0, D), \end{aligned}$$

where $\mu_{ij} = E(y_{ij})$. This is another example of a GLMM.

R Example. We consider the depression data described earlier to illustrate R functions to fit GLMMs.

```
library(nlme)
library(MASS)
# LME model fit to the Mental Distress Data
dep.dat <- read.table("dep3.dat",head=T)  # get data
attach(dep.dat)
# part of data
```

```
...
     ID month group gender depression
111 194  0.05     2      2  0.6666667
112 194  0.10     2      2  1.6666667
113 195  0.00     2      2  1.8333333
114 195  0.05     2      2  2.3333333
...
# group data based on clusters
dep1.dat <- groupedData(depression ~ month | ID, data=dep.dat)
# LME model fits
lme1 <- lme(fixed=depression ~ month, random = ~ month,data=dep1.dat)
summary(lme1)
...
Fixed effects: depression ~ month
                 Value  Std.Error  DF   t-value p-value
(Intercept)  1.4015517 0.05442483 728 25.752064       0
month       -0.4076756 0.05886816 728 -6.925231       0
...
# We see that the slope is negative and significant, which indicates
# that depression decreases over time

# Let's add covariate "group" and "gender"
lme2 <- lme(fixed=depression ~ month+group+gender, random = ~ month,data=dep1.dat)
summary(lme2)
......
Fixed effects: depression ~ month + group + gender
                 Value  Std.Error  DF   t-value p-value
(Intercept)  0.8748624 0.22853953 728  3.828057  0.0001
month       -0.4075811 0.05904951 728 -6.902362  0.0000
group       -0.0992603 0.09957314 245 -0.996858  0.3198
gender       0.4004138 0.10351499 245  3.868172  0.0001
# Depression level differs significantly between male and female over time
# (i.e., significant "gender" effect), but the treatment group does
# not seem to have a signficant effect over time.
```

```
# GLMM model fit
# Now, let's convert the depression score into a binary variable (i.e., either
# 1 or 0, depending on whether depression is greater or smaller than average)
dep2 <- as.numeric(dep.dat$depression > mean(dep.dat$depression))
# First, let's try GLM fit, which ignores longitudinal correlation
# and treat all data as i.i.d.
glm1 <- glm(dep2~month, family=binomial, data = dep.dat)
summary(glm1)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.004154   0.078128  -0.053 0.957596
month       -0.723709   0.193432  -3.741 0.000183 ***
# The above results may be biased since the correlation is ignored.

# Next, let's try a GLMM fit which uses random effects to incorporate
# longitudinal correlation over time. Note that the R function glmmPQL
# is based on an approximate method for GLMMs.
# Case 1: random effect on the intercept
glmm1 <- glmmPQL(dep2 ~ month, random = ~ 1 | ID, family = binomial, data = dep.dat)
summary(glmm1)
...
Fixed effects: dep2 ~ month
                Value Std.Error  DF   t-value p-value
(Intercept)  0.0261358 0.1461217 728  0.178863  0.8581
month       -0.9705548 0.2021263 728 -4.801725  0.0000
......
# We see that the estimates differ from GLM fit, when correlation is incorporated.

# Case 2: random effects on both parameters
glmm2 <- glmmPQL(dep2 ~ month, random = ~ 1+month | ID,family = binomial, data = dep.dat
summary(glmm2)
Fixed effects: dep2 ~ month
                Value Std.Error  DF   t-value p-value
```

```
(Intercept)  0.0280688 0.1663374 728  0.168746  0.8660
month       -1.0997344 0.2841096 728 -3.870810  0.0001
```

5.5 Bayesian Generalized Linear Mixed Models

Suppose that the responses $\{y_{i1}, \cdots, y_{in_i}\}$ in the i-th cluster are conditionally independent given the mean parameters $\boldsymbol{\beta}$ and random effects $\mathbf{b}_i$. Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^T$. A full *Bayesian generalized linear mixed model (GLMM)* can be written as

$$\begin{aligned} E(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{b}_i) &= h(X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i), \qquad i = 1, \cdots, n, && (28)\\ \mathbf{b}_i &\sim N(0, D), && (29)\\ \boldsymbol{\beta} &\sim N(\boldsymbol{\beta}_0, \Sigma_0), \qquad D \sim W_q^{-1}(\eta, \Psi), && (30) \end{aligned}$$

where $h(\cdot)$ is a known link function and X_i and Z_i are known design matrices.

Let $\mathbf{b} = (\mathbf{b}_1, \cdots, \mathbf{b}_n)$. We assume that the prior distributions are independent, so

$$f(\boldsymbol{\beta}, D) = f(\boldsymbol{\beta})f(D).$$

Then, the posterior distribution of all parameters can be written as

$$f(\boldsymbol{\beta}, D, \mathbf{b}|\mathbf{y}) \propto \left[\prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \mathbf{b}_i)f(\boldsymbol{\beta})\right]\left[\prod_{i=1}^{n} f(\mathbf{b}_i|D)f(D)\right].$$

For Bayesian inference, note that the full conditionals are given by

$$\begin{aligned} f(\boldsymbol{\beta}|D, \mathbf{b}, \mathbf{y}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \mathbf{b}_i)f(\boldsymbol{\beta}), && (31)\\ f(\mathbf{b}|\boldsymbol{\beta}, D, \mathbf{y}) &\propto \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \mathbf{b}_i)f(\mathbf{b}_i|D), && (32)\\ f(D|\boldsymbol{\beta}, \mathbf{b}, \mathbf{y}) &\propto \prod_{i=1}^{n} f(\mathbf{b}_i|D)f(D), && (33) \end{aligned}$$

where

$$[D|\boldsymbol{\beta}, \mathbf{b}, \mathbf{y}] \sim W_q^{-1}(\eta + n/2,\ \Psi + \sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T/2).$$

Bayesian inference can then be based on the Gibbs sampler along with rejection sampling methods (Zeger and Karim 1991; Gelman et al. 2003). A Gibbs sampler method to generate samples from the posterior distribution $f(\boldsymbol{\beta}, D, \mathbf{b}|\mathbf{y})$ is described as follows. At k-th iteration

- sample $\boldsymbol{\beta}^{(k)}$ from $f(\boldsymbol{\beta}|D^{(k-1)}, \mathbf{b}^{(k-1)}, \mathbf{y})$;
- sample $D^{(k)}$ from $f(D|\boldsymbol{\beta}^{(k)}, \mathbf{b}^{(k-1)}, \mathbf{y})$;
- sample $\mathbf{b}^{(k)}$ from $f(\mathbf{b}|\boldsymbol{\beta}^{(k)}, D^{(k)}, \mathbf{y})$, $k = 1, 2, 3, \cdots$.

Beginning with starting values $(\boldsymbol{\beta}^{(0)}, D^{(0)}, \mathbf{b}^{(0)})$, after a warm-up period we obtain a sample of $(\boldsymbol{\beta}, D, \mathbf{b})$ from the posterior distribution $f(\boldsymbol{\beta}, D, \mathbf{b}|\mathbf{y})$. Once we generate many such samples, the posterior mean and posterior covariance can be approximated by the sample mean and sample covariance based on the simulated samples. As one can imagine, this procedure can be computationally intensive.

5.6 GEE Models

In GEE models the mean structure and variance-covariance structure are specified separately, without any distributional assumptions for the data, and emphasis is placed on the correct specification of the mean structure. Such models are particularly useful for non-normal data, such as binary or count data, when GLM may be considered and over-dispersion can be a potential issue. The choice of the variance-covariance structure only affects the *efficiency* of the estimates: the closer the variance-covariance structure to the true one, the more efficient the resulting GEE estimates. Parameter estimators are obtained by solving a set of GEEs. The resulting GEE estimators have attractive asymptotic properties such as asymptotic consistency and asymptotic normality, which can be used for constructing confidence intervals and hypothesis testing.

Let $\mathbf{y}_i = (y_{i1}, \cdots, y_{in_i})^T$ be the longitudinal repeated measurements of the response variable on individual i, and let $\mathbf{x}_i$ be the corresponding covariates, $i = 1, 2, \cdots, n$. A regression model can be written as

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}) = h(\mathbf{x}_i^T\boldsymbol{\beta}), \qquad i = 1, 2, \cdots, n, \tag{34}$$

where $\boldsymbol{\beta}$ is a vector of regression parameters, and $h(\cdot)$ is a known link function. We separately assume a variance-covariance structure for $\mathbf{y}_i$ as follows

$$Cov(\mathbf{y}_i) = \Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{35}$$

where $\boldsymbol{\alpha}$ contains unknown parameters for the variance-covariance structure of $\mathbf{y}_i$.

The variance-covariance structure of the response vector $\mathbf{y}_i$ is often written in the following form

$$\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = V_i^{1/2}(\boldsymbol{\beta})R_i(\boldsymbol{\alpha})V_i^{1/2}(\boldsymbol{\beta}), \tag{36}$$

where $V_i(\boldsymbol{\beta}) = diag(\sigma_{i1}^2, \cdots, \sigma_{in_i}^2)$, with $\sigma_{ik}^2 = var(y_{ik}|\mathbf{x}_i, \boldsymbol{\beta})$, are the variances of the responses for individual i at different times, and the matrix $R_i(\boldsymbol{\alpha})$ is called a *working correlation matrix*, which measures the correlations of the repeated response measurements on individual i.

The following correlation structures are often chosen as working correlation matrices in GEE models:

- the *independence* working correlation matrices:

$$R_i(\boldsymbol{\alpha}) = I_{n_i},$$

  where I_{n_i} is the $n_i \times n_i$ identity matrix;

- the *equicorrelation* working correlation matrices:

$$(R_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha, \qquad \text{for } j \neq k,$$

  where $corr(y_{ij}, y_{ik})$ is the correlation between y_{ij} and y_{ik};

- the *stationary* working correlation matrices:

$$(R_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha|t_{ij} - t_{ik}|, \qquad \text{for } j \neq k;$$

- the *unstructured* working correlation matrices

$$(R_i(\boldsymbol{\alpha}))_{jk} = corr(y_{ij}, y_{ik}) = \alpha_{jk}, \qquad \text{for } j \neq k,$$

where α_{jk}'s are unknown parameters.

To estimate the unknown parameters in GEE models, we can construct a set of equations which are similar to the familiar likelihood equations but without distributional assumptions. Solving these equations lead to GEE estimates of the parameters. Let

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = E(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\beta}), \qquad \Delta_i(\boldsymbol{\beta}) = \partial\boldsymbol{\mu}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}.$$

The *generalized estimating equation (GEE)* for estimating the mean parameters $\boldsymbol{\beta}$ is given by

$$S_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \left[\Delta_i(\boldsymbol{\beta})\Sigma_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}))\right] = 0. \tag{37}$$

GEE (37) can be solved by an iterative algorithm.

If the mean structure $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ is correctly specified, it can be shown that the GEE estimator $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal, *even if* the covariance matrix $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is misspecified. The choice of the covariance matrix $\Sigma_i(\boldsymbol{\beta}, \boldsymbol{\alpha})$ only affects the *efficiency* of the GEE estimator. This is a main advantage of GEE estimators. It suggests that for GEE models we can focus on correct specification of the mean structure. The idea of the GEE approach can be extended to a wide range of problems.

An Example in R (Children's height data). This dataset contains 26 children's heights measured at 9 different times. We use it to illustrate the GEE methods and mixed effects models.

```
oxboys.dat <- read.table("oxboys.dat",head=T)
attach(oxboys.dat)
library(gee)   # A R package for GEE
library(nlme)  # A R package for mixed effects models
# Part of the data
> oxboys.dat[1:4,]
  id     age height occasion
1  1 -1.0000  140.5        1
2  1 -0.7479  143.4        2
3  1 -0.4630  144.8        3
........
# Fitting marginal GEE models with various assumptions
# of the working correlation structures
fit.gee1 <- gee(height~age,id)   # independent correlation
fit.gee2 <- gee(height~age,id,corstr="AR-M") # AR-1 correlation
fit.gee3 <- gee(height~age,id,corstr="exchangeable")  # exchangeable
fit.gee4 <- gee(height~age,id,corstr="unstructured")  # unstructured
# Let's compare the estimates and SE's
> summary(fit.gee1)$coef
             Estimate Naive S.E.    Naive z Robust S.E. Robust z
(Intercept) 149.371801  0.5285648 282.598864    1.554618 96.08266
age           6.521022  0.8169867   7.981797    0.329252 19.80557
> summary(fit.gee2)$coef
             Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.719096  1.5531285 96.39840   1.5847569 94.47449
age           6.547328  0.3177873 20.60286   0.3042478 21.51972
```

```
> summary(fit.gee3)$coef
              Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.371735  1.5615503 95.65605   1.5546081 96.08321
age           6.523916  0.1476602 44.18196   0.3295115 19.79875
> summary(fit.gee4)$coef
              Estimate Naive S.E.  Naive z Robust S.E. Robust z
(Intercept) 149.494178  1.5615644 95.73360   1.5533605 96.23920
age           6.052624  0.3726325 16.24288   0.3205602 18.88139
# The estimates are similar. However, the independent working correlation
# structure produce different SE estimates. This indicates the importance
# of incorporating the correlation in the longitudinal data. The
# unstructured correlation is most general.

# Next, let's try mixed effects models and compare the results.
# Fit a LME model
oxboys.dat1 <- groupedData(height~age|id,data=oxboys.dat)
fit.lme1 <- lme(fixed=height~age, random=~age, data=oxboys.dat1)
summary(fit.lme1)
...
Fixed effects: height ~ age
                Value Std.Error  DF  t-value p-value
(Intercept) 149.37175 1.5854173 207 94.21605       0
age           6.52547 0.3363003 207 19.40370       0
......
# They seem to give similar estimates.
```

5.7 Missing Data Problems

Missing values are very common in multivariate data and longitudinal data, since in practice it is almost unlikely that all data are available for each individual and for each variable. For example, in sample survey, some people may refuse to answer some questions such as incomes. In class, students may fail to hand in some homework. In a longitudinal study, some people may drop out early. Therefore, in the real world, most multivariate data or longitudinal data contain missing values. It is very important to handle these missing data appropriately because simple or naive methods for missing data can lead to severely biased results. For example, if we discard individuals in a sample survey who did not report income, we discard

an important special group of individuals so the remaining data are not representative of all people.

When there are missing data, statistical software may not work. The simplest approach is to delete the missing data, but this will lead to biased results or loss of information. In order to handle missing data appropriately, we should first check the *missing data mechanisms*, i.e., why or how the data are missing. Possible missing data mechanisms are:

- *missing completely at random (MCAR)*: missingness depends neither on the missing values nor on the observed values
- *missing at random (MAR)*: missingness does not depend on the missing values but may depend on observed values
- *nonignorable missing*: missingness depends on the missing values and observed values.

For example, in a sample survey, if a person did not report his income because he forgot, then the missing data is MCAR. If a person did not report his income because he is too old, then the missing data is MAR. If a person did not report his income because his income is too high or too low, then the missing data is nonignorable. To address missing data appropriately, the missing data mechanism must be incorporated in any statistical methods.

A commonly used method for missing data is called the *multiple imputation*. The idea is as follows

- For each missing value, we impute several possible values (say $m = 5$) based on an imputation model, then we obtain several (m) "complete datasets".
- Each "complete dataset" is analyzed using the usual complete-data methods as if all data were observed.
- The m complete-data analyses are then combined to obtain an overall result.

Multiple imputation methods take the *missing data uncertainty* into account, so they are better than single imputation methods such as the mean-imputation method or the last-value-carried-forward method. Software is available for multiple imputations. There are many methods for creating multiple imputations, good understanding of these methods is important for correct analysis.

Consider a multiple imputation method for missing values in data matrix X. Let X_{mis} be the missing parts of X and X_{obs} be the observed part. Imputations can be generated

from the predictive distribution $f(X_{mis}|X_{obs})$ based on the following Bayesian framework

$$X_{mis} \sim f(X_{mis}|X_{obs}) = \int f(X_{mis}|X_{obs}, \theta) f(\theta|X_{obs}) d\theta$$

Let $\hat{\theta}^{(i)}$, $\text{Var}(\hat{\theta}^{(i)})$ be the parameter estimate and its variance based on the i-th imputation, $i = 1, 2, \cdots, m$. The overall estimate of θ is given by

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}^{(i)},$$

with variance

$$\begin{aligned} Var(\hat{\theta}) &= (1 + \frac{1}{m}) \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}^{(i)} - \hat{\theta}) + \frac{1}{m} \sum_{i=1}^{m} Var(\hat{\theta}^{(i)}) \\ &= \text{between imputation variation + within imputation variance} \end{aligned}$$

Another commonly used method for missing data problems is the EM algorithm for likelihood inference. The basic idea is to write down the likelihood for the observed data and then use the EM algorithm to obtain maximum likelihood estimates. We skip the details here.

6 Bootstrap Methods

Bootstrap methods are widely used general approaches for statistical inference. They are computer-intensive resampling methods, and are very useful for some difficult problems, such as variance estimations for intractable estimators and statistical inference when parametric assumptions are in doubt or when parametric inference is highly complicated. For example, it is usually difficult to compute the variances of a sample median or a sample percentile or a sample correlation coefficient. In these cases, it is straightforward to use the bootstrap method to compute estimates of standard errors and confidence intervals of these estimators. Bootstrap methods are often easy to implement, though may be computationally intensive, and can be applied to a wide variety of problems. Therefore, bootstrap has become a very popular statistical tool in modern statistics.

The idea of a bootstrap method is usually to approximate a distribution by the empirical distribution of the *observed data*, implemented by repeatedly resampling from the observed dataset *with replacement* (with the same sample size as the observed dataset). For example, suppose that $(x_1, x_2, \cdots, x_n)$ is an observed dataset, and suppose that one wishes to estimate the variance of the sample median. A simple bootstrap method proceeds as follows. We can sample from this observed dataset with replacement. The resulting sample, denoted by $(x_1^*, x_2^*, \cdots, x_n^*)$, is called a *bootstrap sample*. Then, we compute the sample median of this bootstrap sample. Repeating this process B times (B is often large, say $B = 1000$), we obtain B median estimates from the B bootstrap samples. We then compute the sample variance of these B median estimates and obtain a bootstrap estimate of the variance of the sample median from the original dataset. The *sampling distribution* of these B estimates is an approximation to the "true" distribution of the sample median from the original dataset.

As another example, we know that the MLE of a parameter is asymptotically normally distributed. In practice, this asymptotic distribution is often used to construct approximate confidence intervals and hypothesis testing where the sample size is in fact finite. Since the sample size is finite in practice, we may want to know how close the distribution of the MLE is to normality, so that we can judge how reliable the approximate confidence intervals and testing results are. We can use a bootstrap method to check this, as illustrated as follow.

Suppose that we fit a mixed effects model, such as an NLME model, to a longitudinal dataset (with sample size n) using the likelihood method, and we wish to check if the resulting MLEs of the parameters are approximately normal. A simple bootstrap method can be performed as follows:

- sample from the original dataset with replacement and obtain a bootstrap sample;

- fit the mixed effects model to the bootstrap sample using the likelihood method and obtain MLEs of the parameters;
- Repeating the procedure B times, one obtains B sets of parameter estimates (MLEs).

The sampling distribution of the B estimates of a parameter is an approximation to the "true" sampling distribution of the MLE of this parameter based on the original dataset. One can then, for example, obtain an approximate confidence interval from the bootstrap samples by taking the α and $1-\alpha$ (say, $\alpha = 0.05$) quantiles of the B estimates. A bootstrap estimate of the standard error of the parameter estimate is the sample standard error of the B estimates.

For a parametric bootstrap method, one would fit a parametric model and obtain bootstrap samples from the fitted parametric model. The estimates are again computed from the bootstrap samples.

For more detailed discussions of Bootstrap methods, see Efron and Tibshirani (1993) and Davison and Hinkley (2006).

7 Appendix: Selected Topics

7.1 Likelihood Methods

Likelihood methods are widely used in statistical inference, due to general applicability of likelihood methods and attractive asymptotic properties of MLEs such as asymptotic most efficiency and asymptotic normality. Moreover, the *likelihood principle* says that likelihood functions contain all of the information in the data about unknown parameters in the assumed models. Maximum likelihood estimation is often viewed as the "gold standard" of estimation procedures. Likelihood functions also play an integral role in Bayesian inference. In the following, we provide a brief overview of likelihood methods.

For a likelihood method, once the likelihood for the *observed data* is specified based on the assumed distributions, the MLEs of unknown parameters in the assumed distributions can be obtained by maximizing the likelihood using standard optimization procedures or the EM algorithms. The resulting MLEs will be asymptotically consistent, most efficient (in the sense of attaining the Cramer-Rao lower bound for the variances of the MLEs), and normally distributed, if some common regularity conditions hold. In other words, when the sample size is large, the MLE is approximately optimal if the assumed distributions and some regularity conditions hold. In many problems, the sample sizes do not have to be very large in order for the MLEs to perform well, and the regularity conditions are often satisfied. Violations of the regularity conditions may arise, for example, when the parameters are on the boundary of the parameter space. Therefore, likelihood methods are conceptually straightforward. In practice, difficulties often lie in computation since the observed-data likelihoods can be highly intractable for some complex problems.

The asymptotic normality of MLEs can be used for (approximate) inference in practice where the sample size is finite. For example, we may use the asymptotic normal distributions of MLEs to construct approximate confidence intervals for the unknown parameters and to perform hypothesis testing such as Wald-type tests, the likelihood ratio test, and the score test (these tests will be described below). Likelihood methods are very general and can be used in almost any situations where probability distributions are assumed. Potential drawbacks of likelihood methods are that MLEs are often sensitive to outliers, the assumed distributions may not hold, and MLEs may be biased for finite samples (but the bias should decrease as the sample size increases). Restricted maximum likelihood estimates (REML) are often used to correct some of the biases in MLEs for variance components.

Let $y_1, y_2, \cdots, y_n$ be a sample of independent and identically distributed (i.i.d.) observa-

tions drawn from a distribution with a probability density function (for continuous variables) or a probability mass function (for discrete variables) $f(y;\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)^T$ are unknown parameters. Note that the results below also apply when the observations are independent but not identically distributed, e.g., in regression settings where the mean and variance of y_i may depend on covariates $\mathbf{x}_i$. Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$. The *likelihood function* for the observed data $\mathbf{y}$ is defined as

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta}),$$

which may be roughly interpreted as the probability of observing the data $\mathbf{y}$ under the assumed distribution for the data.

The *maximum likelihood estimate (MLE)* of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is the value of $\boldsymbol{\theta}$ which maximizes the likelihood $L(\boldsymbol{\theta})$, i.e., the MLE is the value of the parameter which makes the observed data most likely to occur. The corresponding *log-likelihood* is given by

$$l(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i;\boldsymbol{\theta}).$$

Since the log-likelihood $l(\boldsymbol{\theta})$ is a monotone function of the likelihood $L(\boldsymbol{\theta})$, maximization of the likelihood $L(\boldsymbol{\theta})$ is equivalent to maximization of the log-likelihood $l(\boldsymbol{\theta})$, but the log-likelihood is easier to handle since a summation is mathematically more manageable than a product. Thus, likelihood inference is often based on the log-likelihood.

The MLE $\hat{\boldsymbol{\theta}}$ satisfies the following estimating equation (likelihood equation)

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0.$$

Note that the MLE may not be unique or may not even exist, but if the MLE exists it should satisfy the above estimating equation. Note also that, if the likelihood function has multiple modes, a solution to the above estimating equation may be a local maximum/minimum, depending on the choice of starting values. So the choice of starting values is important for complex likelihood functions. In practice, a simple approach is to try different starting values and check if the solutions differ. The vector

$$\mathbf{s}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \sum_{i=1}^{n} \frac{\partial \log f(y_i;\boldsymbol{\theta})}{\partial \theta_p} \right)^T$$

is called the *Fisher efficient score* or the *score*. It can be shown that $E(s(\boldsymbol{\theta})) = 0$.

The *Fisher's information function (matrix)* is defined by

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} \right) = (I(\boldsymbol{\theta})_{jk})_{p\times p}, \quad \text{with} \quad I(\boldsymbol{\theta})_{jk} = -E\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right).$$

The information matrix $I(\boldsymbol{\theta})$ quantifies the expected amount of information in the data about the unknown parameters $\boldsymbol{\theta}$. Note that the second derivatives $\partial^2 l(\boldsymbol{\theta})/\partial\theta_j^2$ describe th e curvature of the likelihood in the neighborhood of θ_j, so the greater the value of $-\partial^2 l(\boldsymbol{\theta})/\partial\theta_j^2$, the sharper is the peak of the likelihood function and thus the greater is the information about θ_j. The Fisher's information matrix can also be expressed as

$$I(\boldsymbol{\theta}) = E\left[\left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right],$$

which only involves the first derivatives so sometimes may be easier to evaluate. The matrix

$$H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \frac{\partial \mathbf{s}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

is called the *Hessian matrix*. The *observed information* is defined as

$$i(\boldsymbol{\theta}) = -H(\boldsymbol{\theta}) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2},$$

which sometimes can be used to approximate $I(\boldsymbol{\theta})$ since $I(\boldsymbol{\theta}) = E(i(\boldsymbol{\theta}))$.

Under some regularity conditions, the MLE is consistent, asymptotically efficient, and asymptotically normally distributed. These regularity conditions can be stated as follows:
R1. The parameter space Θ of $\boldsymbol{\theta}$ is an open subset of the whole space R^p.
R2. The set $A = \{y : f(y;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$.
R3. The function $f(y;\boldsymbol{\theta})$ is three times continuously differentiable with respect to $\boldsymbol{\theta}$ for all y.
R4. The following equations hold

$$E(\partial l(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}) = 0, \qquad Cov(\partial l(y;\boldsymbol{\theta})/\partial\boldsymbol{\theta}) = I(\boldsymbol{\theta}), \quad \text{for all } \boldsymbol{\theta}.$$

R5. The expectations of all the derivatives of $f(y;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ exist and are finite.
The above regularity conditions are satisfied for a wide variety of models and are relativel y easy to verify. Note that there are variations of these conditions, and weaker conditions are available.

Under the regularity conditions R1 – R5, the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ has the following large-sample properties:

- The MLE $\hat{\boldsymbol{\theta}}$ is *consistent*, i.e.,

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}, \qquad \text{as } n \to \infty;$$

- The MLE $\hat{\boldsymbol{\theta}}$ is *asymptotically efficient*, i.e., the asymptotic variance of $\hat{\boldsymbol{\theta}}$ attains the *Cramer-Rao lower bound*, which is $I^{-1}(\boldsymbol{\theta})$;

- The MLE $\hat{\boldsymbol{\theta}}$ is *asymptotically normal*, i.e.,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, I^{-1}(\boldsymbol{\theta})), \quad \text{as } n \to \infty.$$

Thus, the MLE is asymptotically *optimal*. Note that, however, the MLE is not necessary unbiased for finite samples. In some cases, the bias of MLE may be substantial. On the other hand, the MLE is asymptotically unbiased, i.e., its bias tends to zero as the sample size increases. Due to the above attractive asymptotic properties of MLEs, likelihood methods are widely used in statistical inference.

Based on the asymptotic normality of the MLE $\hat{\boldsymbol{\theta}}$, in practice when the sample size is finite, an approximate level $1-\alpha$ confident interval for θ_j, the j-th component of $\boldsymbol{\theta}$, is given by

$$\hat{\theta}_j \pm z_{\alpha/2} \cdot s.e.(\hat{\theta}_j),$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution $N(0,1)$ and $s.e.(\hat{\theta}_j) = I^{-1/2}(\hat{\boldsymbol{\theta}})_{jj}$ is the approximate standard error of the MLE $\hat{\theta}_j$. For hypothesis testing, the following three likelihood-based large-sample tests are widely used: the Wald test, the likelihood ratio test (LRT), and the efficient score test. These three tests are briefly described as follows.

Consider testing the hypotheses

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

The following three tests are based on asymptotic results and widely used in practice:

- *Wald-type test.* The Wald-type test statistic for testing H_0 versus H_1 is given by

$$T_W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \hat{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

  where $\hat{\Sigma} = I(\hat{\boldsymbol{\theta}})^{-1}$ is an estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$. The test statistic $T_W \sim \chi^2_p$ asymptotically under H_0, where p is the dimension of parameter $\boldsymbol{\theta}$. To test an individual component of $\boldsymbol{\theta}$, say $H_{0j} : \theta_j = \theta_{j0}$ versus $H_1 : \theta_j \neq \theta_{j0}$, we may consider individual Wald-type test statistic

$$T_W^{(j)} = \frac{(\hat{\theta}_j - \theta_0)^2}{\widehat{var}(\hat{\theta}_j)}$$

  where $\widehat{var}(\hat{\theta}_j) = (I(\hat{\boldsymbol{\theta}})^{-1})_{jj}$. The test statistic $T_W^{(j)} \sim \chi^2_1$ asymptotically under H_{0j}.

- *Likelihood ratio test (LRT).* Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, and let $L(\boldsymbol{\theta}_0)$ and $L(\hat{\boldsymbol{\theta}})$ be the likelihood functions evaluated at $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ respectively. The LRT test statistic for testing H_0 versus H_1 is given by

$$T_L = -2\log\left(\frac{L(\boldsymbol{\theta}_0)}{L(\hat{\boldsymbol{\theta}})}\right) = 2\log L(\hat{\boldsymbol{\theta}}) - 2\log L(\boldsymbol{\theta}_0).$$

  The test statistic $T_L \sim \chi^2_p$ asymptotically under H_0.

- *Score test.* The score test statistic for testing H_0 versus H_1 is given by

$$T_S = \mathbf{s}(\boldsymbol{\theta}_0)^T I(\boldsymbol{\theta}_0)^{-1}\mathbf{s}(\boldsymbol{\theta}_0),$$

  where $\mathbf{s}(\boldsymbol{\theta}_0)$ is the score function at $\boldsymbol{\theta}_0$. The test statistic $T_S \sim \chi^2_p$ asymptotically under H_0.

The above three tests are asymptotically equivalent, but they may differ with finite samples. The LRT is equivalent to the deviance test which is widely used in GLMs. The Wald test requires the least computational effort. The score test does not require computing the MLE since the test statistic is evaluated under the null hypothesis.

Note that the above asymptotic results do not hold for order-restricted tests or constrained tests, such as the one-sided test $H_0 : \boldsymbol{\theta} = 0$ versus $H_1 : \boldsymbol{\theta} > 0$. In this case, the above three tests can be constructed in a similar way, but their asymptotic distributions are no longer χ^2-distributions but are mixtures of χ^2-distributions.

7.2 Optimization Methods and the Newton-Raphson Algorithm

In estimation problems, especially in likelihood methods, one often needs to find a maxima or minima of a function, say $L(\boldsymbol{\theta})$. This problem is often equivalent to finding a root of the function $g(\boldsymbol{\theta}) \equiv \partial \log L(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \partial l(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, or solving equation

$$g(\boldsymbol{\theta}) = 0.$$

There are different general optimization procedures to accomplish this. The most widely used one is perhaps the *Newton-Raphson method* or the *Newton's method*, which is briefly described below.

The Newton-Raphson method is an iterative algorithm based on a Taylor series expansion. In the univariate case $\boldsymbol{\theta} = \theta$, the Newton-Raphson method iteratively solves the following equation

$$\theta_k = \theta_{k-1} - \frac{g(\theta_{k-1})}{g'(\theta_{k-1})}, \qquad k = 1, 2, \cdots,$$

where $g'(\theta)$ is the derivative of $g(\theta)$. Beginning with an initial value θ_0, the algorithm will usually converge to a possibly local maxima or minima of $L(\theta)$. In the multi-dimensional case, the algorithm can be written as

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{k-1}} \cdot g(\boldsymbol{\theta}_{k-1}), \qquad k = 1, 2, \cdots,$$

where

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1}, \cdots, \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_p}\right) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right)_{p\times p}.$$

Note that the choice of the initial value $\boldsymbol{\theta}_0$ is important. The closer $\boldsymbol{\theta}_0$ to the root of $g(\boldsymbol{\theta})$, the better. However, since we often do not know the root of $g(\boldsymbol{\theta})$, a guessed value $\boldsymbol{\theta}_0$ is usually chosen. If the function $l(\boldsymbol{\theta})$ has multi-mode (so $l(\boldsymbol{\theta})$ has two or more roots), the Newton-Raphson algorithm only converges to a local maxima. To find a global maxima, one should try several different initial values. Note that the Newton-Raphson algorithm may not converge in some cases, such as the initial value being too far from the true zero or the derivatives being zeros or near zeros (so the tangent line is nearly horizontal). In practice, one may want to put an upper limit on the number of iterations. For more detailed discussions of the Newton-Raphson methods and other optimization methods, see, e.g., Deuflhard (2004) and Press et al. (2007).

7.3 Numerical Integration Methods

To evaluate intractable integrals, numerical integration methods are alternatives to Monte Carlo methods. Numerical integration methods approximate an integral by a weighted sum, with suitably chosen points and weights. These methods include the Simpson's rule and quadrature methods. In the following, we briefly describe the popular Gauss-Hermite quadrature method. Evans and Swartz (2000) provided a detailed discussion of various approaches.

Consider the following integral

$$I = \int g(x)f(x)dx,$$

where $g(x)$ is a continuous function and $f(x)$ is a normal density function. We first consider the $N(0,1)$ density and let $f(x) = \exp(-x^2)$. The *Gauss-Hermite quadrature* method approximates the integral by

$$I = \int \exp(-x^2)g(x)dx \approx \sum_{i=1}^{k} w_i(x_i)g(x_i),$$

where the node x_i is the i-th root of the Hermite polynomial $H_k(x)$ with degree of k, $w_i(x_i)$ is the weight which depends on $H_{k-1}(x_i)$:

$$w_i(x_i) = \frac{2^{k-1}k!\sqrt{\pi}}{k^2(H_{k-1}(x_i))^2},$$

and the *Hermite polynomials* are orthogonal polynomials defined by

$$H_k(x) = (-1)^k e^{x^2/2}\frac{d^k\, e^{-x^2/2}}{dx^k}.$$

The above approximation can be arbitrarily accurate when the number k of nodes increases. When $g(x)$ is a polynomial of degree up to $2k-1$, the approximation is exact. Note that the first several Hermite polynomials are:

$$\begin{aligned} H_0(x) &= 1, \qquad H_1(x) = x, \qquad H_2(x) = x^2 - 1, \\ H_3(x) &= x^3 - 3x, \qquad H_4(x) = x^4 - 6x^2 + 3, \qquad \cdots. \end{aligned}$$

If $f(x)$ is the density function of a general normal distribution $N(\mu, \sigma^2)$, we may consider transformation $x = \sqrt{2}\sigma z + \mu$, and we then have

$$I = \int g(x)f(x)dx \approx \sum_{i=1}^{k} w_i^*(x_i)g(\sqrt{2}\sigma z_i + \mu),$$

where $w_i^*(x_i) = \pi^{-1/2}w_i(x_i)$.

If $\mathbf{x} = (x_1, \cdots, x_m)^T$ is a m-dimensional vector, we have

$$I = \int_{R^m} g(\mathbf{x})f(\mathbf{x})d\mathbf{x} \approx \sum_{i_1=1}^{k_1} w_{i_1}^{(1)} \cdots \sum_{i_m=1}^{k_m} w_{i_m}^{(m)} g(x_{i_1}^{(1)}, \cdots, (x_{i_m}^{(m)})$$

where $x_{i_j}^{(j)}$ is the i_j-th root of the Hermite polynomial with degree k_j and $w_{i_j}^{(j)}$ is the corresponding weight. Note that the number of nodes increases exponentially with the number m of dimensions, so the method can be very inefficient for high-dimensional integrals. In practice, one may use the method for integrals with dimensions up to 5 or 6. See Evans and Swartz (2000) and Fahrmeir and Tutz (2001) for more detailed discussions.

7.4 Monte Carlo Methods

In likelihood or Bayesian inference, we often need to evaluate intractable integrals which do not have analytic or closed-form expressions. Monte Carlo methods are widely used to

approximate these integrals. In this section, we briefly describe two popular Monte Carlo methods: rejection sampling methods and importance sampling methods.

We first consider the following integral

$$I = \int g(x)f(x)dx,$$

where $g(\cdot)$ is a continuous function and $f(\cdot)$ is a probability density function. Suppose that I does not have an analytic expression. A Monte Carlo method can be used as follows. If we can generate an i.i.d. sample $x_1, x_2, \cdots, x_m$ from the density $f(x)$, we can then approximate the integral I by

$$I \approx \hat{I} = \frac{1}{m}\sum_{j=1}^{m} g(x_j).$$

The accuracy of this approximation increases as the number m increases. Thus, the problem is how to sample from the distribution $f(x)$. Unfortunately, in many problems the density function $f(x)$ is highly complicated, especially when x is a vector, so it may not be straightforward to generate these samples. The rejection sampling methods and the importance sampling methods are two classes of general and widely used methods to generate samples from intractable distributions.

Rejection Sampling Methods

Suppose that we wish to generate a sample from a complicated density function $f(x)$. Since $f(x)$ is complicated, it may be hard to sample directly from $f(x)$. Suppose, however, that we know how to sample from the distribution with density $h(x)$, and that there is a known constant c such that

$$f(x) \leq c\, h(x), \qquad \text{for all } x.$$

Then, a rejection sampling method may proceed as follows:

- generate a value x^* from the distribution $h(x)$;
- generate a value u from the uniform distribution on (0,1);
- accept x^* if

$$u < \frac{f(x^*)}{c\, h(x^*)},$$

  otherwise reject x^*.

Repeating this procedure and retaining only accepted values of x^*, we obtain a sample $x_1^*, x_2^*, \cdots$, from the target distribution $f(x)$.

The efficiency of the rejection sampling method depends strongly on how well the function $h^*(x) = c\,h(x)$, often called the *envelope function*, approximates the target function $f(x)$. If the ratio $f(x)/h^*(x)$ is small, the probability of acceptance will be small, so the algorithm will spend most of time rejecting x^* values. A multivariate version of the rejection sampling method is similar. See Evans and Swartz (2000) and Robert and Casella (2004) for more details.

It is often not easy to find a good envelope function. The following adaptive rejection sampling method is widely applicable and very popular.

The Adaptive Rejection Sampling Method

The *adaptive rejection sampling method* (Gilks and Wild 1992) is a very useful rejection sampling method when the target density function $f(x)$ is *log-concave*, i.e., function $\log f(x)$ is concave. It is particularly useful in Gibbs sampling where the full conditional distributions may be intractable but are known to be log-concave. In its original version, the adaptive rejection sampling method constructs the envelope functions based on a set of *tangents* to the function $\log f(x)$. In later versions, the method updates the envelope functions to correspond more closely to the target function $\log f(x)$ whenever a point is rejected, and thus improves the efficiency of the method. More flexible algorithms, which relax the log-concavity requirement, have also been proposed (Evans and Swartz 2000).

Importance Sampling Methods

Let $g(x)$ be a continuous function and $f(x)$ be a probability density function. Consider again the integral

$$I = \int g(x)f(x)dx = E(g(x)).$$

Suppose that it is hard to directly sample from $f(x)$, but we know how to sample from a distribution $h(x)$, which does not need to be an envelope function as in a rejection sampling method. Since

$$I = \int g(x)f(x)dx = \int \frac{g(x)f(x)}{h(x)}h(x)dx,$$

if $x_1, x_2, \cdots, x_m$ is a sample generated from the distribution $h(x)$, we then have

$$I \approx \hat{I} = \frac{1}{m}\sum_{i=1}^{m}\frac{g(x_i)f(x_i)}{h(x_i)} = \frac{1}{m}\sum_{i=1}^{m} w_i g(x_i),$$

where $w_i = f(x_i)/h(x_i),\ i = 1, \cdots, m$, are called importance weights and $h(x)$ is called the *importance function*. Note that here we do not require the condition $f(x) \le ch(x)$, and, unlike rejection sampling methods, here we use all the generated x_j's. The efficiency of the importance sampling method depends on the choice of the importance function. The closer

the function $h(x)$ approximates $f(x)$, the better the importance sampling method. Often, the function $h(x)$ is chosen to have larger tails than $f(x)$. Evans and Swartz (2000) and Robert and Casella (2004) provided more detailed discussions.

7.5 EM Algorithm

The expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) is an iterative procedure used to compute maximum likelihood estimates in the presence of missing data or unobservables. It has become extremely popular in the last few decades, and it has also been widely used outside the missing data area. The popularity of the EM algorithm compared with other numerical methods such as Newton-Raphson methods is EM's superior stability properties and easy for implementation. Specifically, the advantages of the EM algorithm include (i) its convergence is stable and each iteration increases the likelihood, (ii) the M-step involves only complete-data maximization so it is often computationally straightforward, and (iii) it is very general and can be used in almost any maximum likelihood estimation problems with unknown quantities. A disadvantage is that the EM algorithm may be slow to converge for many problems. In recent years, there has been extensive research on the EM algorithm itself, such as methods to speed up convergence and extensions of the standard EM algorithm. McLachlan and Krishnan (1997) provided a comprehensive review of these developments.

An EM algorithm *iterates* between an E-step and an M-step as follows:

- *E-step*: computes the conditional expectation of the "complete-data" log-likelihood given the observed data and the current parameter estimates, where the "complete-data" contain both the observed data and the missing data,
- *M-step*: maximizes the conditional expectation in the E-step with respect to the unknown parameters to produce updated estimates of the parameters.

Given some starting values, we iterate between the E-step and the M-step until the parameter estimates converge. It can be shown that, under some reasonable regularity conditions, each EM iteration *increases* the likelihood, so the EM algorithm is guaranteed to converge to a local maxima (Wu 1983). Since the EM algorithm only converges to a local maxima, when the likelihood may have multiple modes, it is important to choose good starting values or try different starting values to make sure that the EM algorithm eventually converges to a global maxima.

In the following, we present some simple examples to illustrate the EM algorithms.

Example: An EM algorithm for normal data

Let $y_1, y_2, \cdots, y_n$ be an i.i.d. sample from normal distribution $N(\mu, \sigma^2)$,

and let $\boldsymbol{\theta} = (\mu, \sigma^2)$ be the unknown parameters. Suppose that $y_1, y_2, \cdots, y_r$ are observed, but $y_{r+1}, y_{r+2}, \cdots, y_n$ are missing, where $r < n$. Assume that the missing data are MAR or MCAR. Let $\mathbf{y}_{obs} = (y_1, y_2, \cdots, y_r)$ be the observed data and let $\mathbf{y}_{mis} = (y_{r+1}, y_{r+2}, \cdots, y_n)$ be the missing data. The "complete data" is then $\mathbf{y}_{com} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}) = (y_1, y_2, \cdots, y_n)$. The observed data log-likelihood is given by

$$l_{obs}(\boldsymbol{\theta}) = -\frac{r}{2}\log(2\pi r\sigma^2) - \frac{1}{2}\sum_{i=1}^{r}\frac{(y_i-\mu)^2}{\sigma^2},$$

and the "complete-data" log-likelihood is given by

$$l_{com}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi n\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^2}.$$

Let $\boldsymbol{\theta}^{(k)}$ be the parameter estimate from the $(k-1)$th EM iteration, $k = 1, 2, 3, \cdots$. At k-th EM iteration, the E-step computes the conditional expectation of the "complete-data" log-likelihood given the current parameter estimates $\boldsymbol{\theta}^{(k)}$ and the observed data $\mathbf{y}_{obs}$, i.e., the E-step computes

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) &= E\left(l_{com}(\boldsymbol{\theta})|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right) \\ &= \left[-\frac{n}{2}\log(2\pi n\sigma^2) + \frac{n\mu^2}{2\sigma^2}\right] - \frac{1}{2\sigma^2}\left[E\left(\sum_{i=1}^{n} y_i^2|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right)\right. \\ &\qquad \left. -2\mu E\left(\sum_{i=1}^{n} y_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right)\right], \end{aligned}$$

where

$$E\left(\sum_{i=1}^{n} y_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right) = \sum_{i=1}^{r} y_i + (n-r)\mu^{(k)}, \tag{38}$$

$$E\left(\sum_{i=1}^{n} y_i^2|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right) = \sum_{i=1}^{r} y_i^2 + (n-r)(\mu^{(k)2} + \sigma^{(k)2}). \tag{39}$$

The M-step of the EM algorithm then updates the parameter estimates by maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$, which leads to

$$\begin{aligned} \mu^{(k+1)} &= \frac{1}{n}E\left(\sum_{i=1}^{n} y_i|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right) = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + (n-r)\mu^{(k)}\right], \\ (\sigma^{(k+1)})^2 &= \frac{1}{n}E\left(\sum_{i=1}^{n} y_i^2|\boldsymbol{\theta}^{(k)}, \mathbf{y}_{obs}\right) - (\mu^{(k+1)})^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{r} y_i^2 + (n-r)(\mu^{(k)2} + \sigma^{(k)2})\right] - (\mu^{(k+1)})^2. \end{aligned}$$

Iterate the E-step and the M-step (for $k = 1, 2, 3, \cdots$) until convergence, we obtain the fo llowing MLEs

$$\hat{\mu} = \frac{1}{r}\sum_{i=1}^{r} y_i, \qquad \hat{\sigma}^2 = \frac{1}{r}\sum_{i=1}^{r} y_i^2 - \hat{\mu}^2.$$

Note that in this example the EM algorithm is in fact not needed. We use it as an illustration of the EM algorithm, since it is simple and contains the essential ideas of the EM algorithm.

7.6 Bayesian Methods

In this section, we describe some general concepts and approaches for Bayesian inference. These general concepts and approaches can be easily extended to more specific models, such as mixed effects models.

Let $\mathbf{y}$ be the data following an assumed parametric distribution with probability density function $f(\mathbf{y}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains unknown parameters. A *Bayesian method* assumes that the unknown parameters $\boldsymbol{\theta}$ are random variables following a distribution with probability density function $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{\theta}_0)$, called a *prior distribution*. The parameters $\boldsymbol{\theta}_0$ in the prior distribution are called *hyper-parameters* and are often assumed to be known, which can be chosen based on similar studies or expert opinion or even non-informative.

Bayesian inference for the unknown parameters $\boldsymbol{\theta}$ is based on the *posterior distribution* $f(\boldsymbol{\theta}|\mathbf{y})$ given the data $\mathbf{y}$. Specifically, given the prior distribution $f(\boldsymbol{\theta})$, the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ can be obtained via the Bayes's theorem:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}). \tag{40}$$

Bayesian inference for $\boldsymbol{\theta}$ is then based on the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$. For example, a Bayesian estimator of $\boldsymbol{\theta}$ is the *posterior mean*:

$$\hat{\boldsymbol{\theta}}_B = E(\boldsymbol{\theta}|\mathbf{y}) = \int \theta f(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta},$$

with its precision being measured by the *posterior variance*:

$$Cov(\hat{\boldsymbol{\theta}}_B) = Cov(\boldsymbol{\theta}|\mathbf{y}) = \int (\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{y}))(\boldsymbol{\theta} - E(\boldsymbol{\theta}|\mathbf{y}))^T f(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}.$$

The posterior mean is an optimal estimator of $\boldsymbol{\theta}$ under the quadratic loss.

When the hyper-parameters are unknown, one approach is to estimate them from the data, and the resulting Bayesian estimates are called *empirical Bayesian estimates*.

The choice of prior distribution $f(\boldsymbol{\theta})$ may affect Bayesian estimation. In other words, Bayesian inference may be influenced by a strong prior. In practice, we can try different prior distributions or different values of the hyper-parameters for *sensitivity analysis*. In the absence of any prior information, we may choose a *non-informative prior*: $f(\boldsymbol{\theta}) \propto 1$. Note that the likelihood $L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta})$, so we have

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta}|\mathbf{y}) f(\boldsymbol{\theta}).$$

Therefore, Bayesian methods and likelihood methods are linked. In particular, if we choose a non-informative prior distribution for $f(\boldsymbol{\theta})$, Bayesian inference is equivalent to likelihood inference.

Although the Bayesian paradigm seems conceptually straightforward, implementation of a Bayesian method is often non-trivial since the integrations involved in Bayesian computation, such as that in (40), are often of high dimensions and intractable, and these integrals usually do not have closed-form or analytic expressions, except in some special cases. Computational challenges are partially due to possible high dimensionality of the parameters $\boldsymbol{\theta}$. The developments of MCMC methods, such as the Gibbs sampler, make such tedious computation feasible.

MCMC methods are often used to generate large samples from the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$, and these samples are then used for Bayesian inference. For example, the widely used Gibbs sampler method breaks down the dimensionality of $\boldsymbol{\theta}$ by iteratively sampling from lower dimensional distributions which are easier to sample. These MCMC methods are often combined with rejection sampling methods or importance sampling methods. Note that, however, although modern computers are increasingly fast, these MCMC methods can still be computationally very intensive and it is not always easy to check the convergence of these iterative algorithms.

To avoid numerical integrations or simulation methods, which may be computationally intensive, alternatively we can consider Bayesian estimation based on *posterior mode* rather than posterior mean (e.g., Santner and Duffy 1989). The idea is to find an estimator $\tilde{\boldsymbol{\theta}}_B$, called *posterior mode estimator*, which maximizes the posterior density $f(\boldsymbol{\theta}|\mathbf{y})$ or maximizes the log posterior likelihood:

$$l_p(\boldsymbol{\theta}|\mathbf{y}) = \log L(\boldsymbol{\theta}|\mathbf{y}) + \log f(\boldsymbol{\theta}).$$

If a non-informative prior is chosen for $\boldsymbol{\theta}$, the posterior mode estimator coincides with the MLE. Note that the posterior mode estimation is also closely related to the Laplace approximation method (Breslow and Clayton 1993).

It can be shown that, under similar regularity conditions as that for asymptotic normality of MLE, the posterior mode estimator $\tilde{\boldsymbol{\theta}}_B$ is asymptotically normal:

$$\tilde{\boldsymbol{\theta}}_B \xrightarrow{d} N(\boldsymbol{\theta}, I_p^{-1}(\boldsymbol{\theta})), \qquad \text{as } n \to \infty,$$

where

$$I_p(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l_p(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right).$$

The posterior mode $\tilde{\boldsymbol{\theta}}_B$ and the curvature $I_p^{-1}(\tilde{\boldsymbol{\theta}}_B)$ can be used to approximate the posterior mean and covariance matrix when they are difficult to compute.

7.7 Prior Distributions

In Bayesian inference, the choice of prior distributions is important since it may affect the final results. In choosing the prior distributions, if there is no inherent reason to prefer one prior distribution over another, a conjugate prior is sometimes chosen for simplicity. A *conjugate prior* is a (parametric) prior distribution for which the resulting posterior distribution also belongs to the same family of distributions. This is important since Bayesian inference is based on the posterior distribution. Specifically, the prior distribution $f(\boldsymbol{\theta})$ is *conjugate* to $f(\mathbf{y}|\boldsymbol{\theta})$ if the posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ is in the same family as the prior distribution $f(\boldsymbol{\theta})$.

For example, the normal distribution (Gaussian family) is conjugate to itself, i.e., if a prior distribution is normal then the posterior distribution is also normal. In fact, all members of the exponential family have conjugate priors (Gelman et al. 2003). In regression models, we typically choose a multivariate normal distribution as a prior distribution for the mean parameters $\boldsymbol{\beta}$, i.e., we typically assume that $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \Sigma_0)$, where $\boldsymbol{\beta}_0$ and Σ_0 are hyper-parameters. For a non-informative prior, we can choose $\Sigma_0^{-1} = 0$ or $\boldsymbol{\beta} \sim$ uniform$(-\infty, \infty)$. For variance-covariance matrices, we typically choose Wishart distributions as prior distributions, which are described as follows.

The Wishart distribution is a generalization of the χ^2 distribution to multiple dimensions or a generalization of the gamma distribution. It is useful for estimation of covariance matrices. Suppose that Z is an $n \times p$ matrix, with i-th row $\mathbf{z}_i \sim N_p(0, V)$ independently, where the $p \times p$ covariance matrix V is positive definite. Then, the probability distribution of

$$W = Z^T Z$$

has a *Wishart distribution* with degrees of freedom n, denoted by $W_p(V, n)$ or $W(V, n)$, and a density function given by

$$f(W) = \frac{|W|^{(n-p-1)/2}}{2^{np/2}|V|^{n/2}\Gamma_p(\frac{n}{2})} \exp\left(-\frac{1}{2}tr(V^{-1}W)\right),$$

where $W > 0$ (positive definite), and $\Gamma_p(\cdot)$ is the multivariate gamma function defined as

$$\Gamma_p(n/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma((n+1-j)/2).$$

The Wishart distribution $W_p(V, n)$ has the mean nV and the mode $(n-p-1)V$ for $n \geq p+1$. When $p = 1$ and $V = 1$, the Wishart distribution $W_p(V, n)$ reduces to a χ^2_n-distribution. Note that the Wishart distribution is the distribution of the MLE for the covariance matrix in a multivariate normal distribution.

In Bayesian inference, a conjugate prior for the covariance matrix of a multivariate normal distribution is the inverse Wishart distribution, defined as follows. If a $p \times p$ random matrix $A \sim W_p(V, n)$, then $B = A^{-1}$ has an *inverse Wishart distribution* (or *inverted Wishart distribution*), denoted by $W_p^{-1}(V^{-1}, n)$ or $W^{-1}(V^{-1}, n)$, with probability density function

$$f(B) = \frac{|V|^{-n/2}|B|^{-(n+p+1)/2} \exp(-tr(V^{-1}B^{-1})/2)}{2^{np/2}\Gamma_p(n/2)}.$$

The mean of $B \sim W_p^{-1}(V^{-1}, n)$ is given by

$$E(B) = V^{-1}/(n-p-1).$$

Let $X = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, with $\mathbf{x}_i \sim N_p(0, \Sigma)$. If we assume a prior distribution $\Sigma \sim W_p^{-1}(\Phi, m)$, then the posterior distribution is given by

$$\Sigma|X \sim W_p^{-1}(XX^T + \Phi,\ m+n).$$

When $p = 1$, the inverse Wishart distribution becomes a inverse gamma distribution.

7.8 MCMC Methods

In likelihood inference of mixed effects models with incomplete data, Monte Carlo EM algorithms are often used in which the E-step requires sampling from multi-dimensional and intractable distributions. Similarly, in Bayesian inference the target posterior distributions are often highly complicated, and approximate Bayesian inference is usually based on large

samples drawn from the target posterior distributions. In both cases, one needs to generate large numbers of samples from highly complicated and multi-dimensional distributions. *Markov chain Monte Carlo (MCMC)* methods are great tools for such tasks and they have become very popular.

MCMC methods are algorithms for generating samples from intractable distributions. The key idea of MCMC methods is to construct *Markov chains* that have the desired distributions as their stationary distributions. After a large number of steps, called a *burn-in period*, the Markov chain will converge to its stationary distribution, and thus the last state of the chain can be used as a sample from the desired distribution. A key characteristic of a Markov chain is that the current state depends on the previous one, so there may be many ways to construct a Markov chain which converges to the same target distribution.

MCMC methods have revolutionized Bayesian inference since they have made highly complicated Bayesian computations feasible. These MCMC methods are also very useful tools in likelihood inference since many likelihood computations encount similar problems as in Bayesian inference. The most useful MCMC method is probably the *Gibbs sampler*, which is briefly described below. Detailed discussions of MCMC methods can be found in Gilks et al. (1996), Gelman et al. (2003), and Robert and Casella (2004).

The Gibbs sampler

Gibbs sampling or the *Gibbs sampler* is an example of MCMC methods, which is perhaps the most widely used MCMC method. It was devised by Geman and Geman (1984). The Gibbs sampler is typically used to obtain random samples from a multi-dimensional probability distribution, which is either intractable or is not known explicitly. The desired samples can be obtained by sequentially sampling from lower-dimensional conditional distributions which are easier to sample from. These samples then comprise a Markov chain, whose stationary distribution is the target distribution. The Gibbs sampler is widely used because it is often easier to sample from the *lower-dimensional* conditional distributions than the original distribution. We describe the details as follows.

Suppose that we wish to generate samples from the probability distribution $f(\mathbf{u}|\boldsymbol{\theta})$, where $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T, \cdots, \mathbf{u}_q^T)^T$ is a random vector, with each component $\mathbf{u}_j$ being possibly also a random vector. Suppose also that $f(\mathbf{u}|\boldsymbol{\theta})$ is highly intractable or even not known explicitly, so it is difficult to generate samples from $f(\mathbf{u}|\boldsymbol{\theta})$ directly. Note that the components $\mathbf{u}_j$'s are typically *unobserved* quantities which may have different dimensions or different types. For example, for missing covariates $\mathbf{x}_i$ in a mixed effects model $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{b}_i, \boldsymbol{\theta})$, with $\mathbf{b}_i$ being the random effects, we may want to generate samples from the intractable distribution $f(\mathbf{x}_{mis,i}, \mathbf{b}_i|\mathbf{x}_{observation,i}, \mathbf{y}_i, \boldsymbol{\theta})$ in the EM algorithm for likelihood estimation. In this case, we

can choose

$$q = 2, \quad \mathbf{u}_1 = \mathbf{x}_{mis,i}, \quad \mathbf{u}_2 = \mathbf{b}_i.$$

As another example, let $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$. Suppose that we want to simulate from the posterior distribution $f(\mu, \Sigma|\mathbf{y})$ in Bayesian inference. In this case, we can choose

$$q = 2, \quad \mathbf{u}_1 = \boldsymbol{\mu}, \quad \mathbf{u}_2 = \Sigma.$$

In the following, we describe the Gibbs sampler method to generate samples from $f(\mathbf{u}|\boldsymbol{\theta})$, assuming $\boldsymbol{\theta}$ is known for simplicity. Let

$$\mathbf{u}_{-j} = (\mathbf{u}_1^T, \cdots, \mathbf{u}_{j-1}^T, \mathbf{u}_{j+1}^T, \cdots, \mathbf{u}_q^T)^T, \qquad j = 1, 2, \cdots, q,$$

be the sub-vector of $\mathbf{u}$ without component $\mathbf{u}_j$. It is often easier to generate samples from the lower-dimensional conditional distributions $f(\mathbf{u}_j|\mathbf{u}_{-j}, \boldsymbol{\theta}), j = 1, 2, \cdots, q$, which are called *full conditionals*. The Gibbs sampler proceeds as follows: beginning with starting values $(\mathbf{u}_1^{(0)}, \cdots, \mathbf{u}_q^{(0)})$, at step k,

- sample $\mathbf{u}_1^{(k)}$ from $f(\mathbf{u}_1|\mathbf{u}_2^{(k-1)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- sample $\mathbf{u}_2^{(k)}$ from $f(\mathbf{u}_2|\mathbf{u}_1^{(k)}, \mathbf{u}_3^{(k-1)}, \cdots, \mathbf{u}_q^{(k-1)}, \boldsymbol{\theta})$;
- $\cdots$;
- sample $\mathbf{u}_q^{(k)}$ from $f(\mathbf{u}_q|\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_{q-1}^{(k)}, \boldsymbol{\theta}), \quad k = 1, 2, \cdots$.

The sequence $\{(\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)}),\ k = 1, 2, 3, \cdots\}$ then comprises a Markov chain with stationary distribution $f(\mathbf{u}|\boldsymbol{\theta})$. Therefore, when k is large enough (after a burn-in period), we can view $\mathbf{u}^{(k)} = (\mathbf{u}_1^{(k)}, \cdots, \mathbf{u}_q^{(k)})^T$ as a sample generated from the target distribution $f(\mathbf{u}|\boldsymbol{\theta})$.

Repeating the above process m times, or taking an independent sample of size m after burn-in, we obtain a sample of size m from the intractable distribution $f(\mathbf{u}|\boldsymbol{\theta})$. When m is large, we can approximate the mean and variance of the distribution $f(\mathbf{u}|\boldsymbol{\theta})$ by the sample mean and sample variance respectively, or we can approximate the density curve $f(\mathbf{u}|\boldsymbol{\theta})$ by the empirical density function based on the simulated samples.

The key idea of the Gibbs sampler is to sequentially sample from lower dimensional conditional distributions in order to generate samples from the original higher dimensional and intractable distribution. Usually it is easier to sample from these lower dimensional conditional distributions (full conditionals) than the original distribution. Sometimes, however, sampling from the lower-dimensional conditional distributions may not be easy either. In

this case, we can use rejection sampling methods to sample from these full conditionals. That is, we can combine the Gibbs sampler with rejection sampling methods or other sampling methods.

Note that the Gibbs sampler or other MCMC methods can only *approximate* the target distributions. The accuracy of the approximation improves as the number of steps (burn-in period) increases. It may not be easy to determine the burn-in period for the Markov chain to converge to the stationary distribution within acceptable random errors. Determining the convergence criteria is an important issue. See Gelman et al. (2003) for a more detailed discussion.

WinBUGS is a statistical software that is widely used to do Gibbs sampling. It is based on the BUGS project (Bayesian inference Using Gibbs Sampling). For details of the WinBUGS software, see webpage:

http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml.

Example. Consider a simple example of generating samples from the bivariate normal distribution $\mathbf{u} = (u_1, u_2)^T \sim N(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and Σ is a 2×2 covariance matrix with diagonal elements being 1 and off-diagonal elements being ρ ($|\rho| < 1$). Assume that $\boldsymbol{\mu}$ and Σ are known. Here it is not difficult to sample from $N(\boldsymbol{\mu}, \Sigma)$, but we consider a Gibbs sampler for illustration purpose.

Consider the Gibbs sampler to sample from the target distribution $N(\boldsymbol{\mu}, \Sigma)$. The full conditionals are

$$\begin{aligned} u_1|u_2, \boldsymbol{\mu}, \rho &\sim N(\mu_1 + \rho(u_2 - \mu_2),\ 1 - \rho^2), \\ u_2|u_1, \boldsymbol{\mu}, \rho &\sim N(\mu_2 + \rho(u_1 - \mu_1),\ 1 - \rho^2). \end{aligned}$$

The Gibbs sampler proceeds as follows: beginning with starting value $(u_1^{(0)}, u_2^{(0)})$, at k-th step,

- generate $u_1^{(k)}$ from $N(\mu_1 + \rho(u_2^{(k-1)} - \mu_2),\ 1 - \rho^2)$;
- generate $u_2^{(k)}$ from $N(\mu_1 + \rho(u_1^{(k)} - \mu_2),\ 1 - \rho^2)$, $\qquad k = 1, 2, \cdots$.

Then, the sequence $\{(u_1^{(k)}, u_2^{(k)}), k = 0, 1, 2, \cdots\}$ forms a Markov chain with stationary distribution $N(\boldsymbol{\mu}, \Sigma)$. Thus, when k is large (say 200), we may consider $(u_1^{(k)}, u_2^{(k)})$ as a sample from $N(\boldsymbol{\mu}, \Sigma)$.

The Metropolis-Hastings Algorithm

The Gibbs sampling algorithm is a special case of the *Metropolis-Hastings algorithm*. That is, the Metropolis-Hastings algorithm is a more general method for creating a Markov chain that can be used to generate samples from an intractable probability distributions, and it is often faster and easier to use than the Gibbs sampler but it is less generally applicable. The algorithm can be used to draw samples from a distribution without knowing the normalization factor, i.e., from $g(\mathbf{u}) \propto f(\mathbf{u})$, where $f(\mathbf{u})$ is the target distribution of interest but the normalization factor may be very difficult to compute (e.g., in Bayesian inference).

Suppose that we wish to simulate a sample from an intractable distribution $f(\mathbf{u})$. Let $p(\mathbf{u}|\mathbf{v})$ be a proposal density. Beginning with a starting value $\mathbf{u}^{(0)}$, at step k we generate a value $\mathbf{u}^*$ from $p(\mathbf{u}|\mathbf{v})$. Then the Metropolis-Hastings algorithm proceeds as follows: simulate a value a from the uniform distribution $U(0,1)$ on $(0,1)$, then accept $\mathbf{u}^*$ as the next value $\mathbf{u}^{(k+1)}$ (i.e., choose $\mathbf{u}^{(k+1)} = \mathbf{u}^*$) *if*

$$a < \frac{f(\mathbf{u}^*)\, p(\mathbf{u}^{(k)}|\mathbf{u}^*)}{f(\mathbf{u}^{(k)})\, p(\mathbf{u}^*|\mathbf{u}^{(k)})}.$$

Otherwise, the proposal is not accepted and the current value is retained (i.e., $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)}$). The sequence $\{\mathbf{u}^{(k)}, k = 0, 1, 2, \cdots\}$ then forms a Markov chain with stationary distribution $f(\mathbf{u})$. After a burn-in period, we can view $\mathbf{u}^{(k)}$ as a sample from $f(\mathbf{u})$.

The algorithm works best if the proposal density $p(\mathbf{u}|\mathbf{v})$ is close to $f(\mathbf{u})$, but often this is difficult to do. Gibbs sampling is a special case of the Metropolis-Hastings sampling where the proposal is always accepted (see, e.g., Gelman et al. 2004, page 328).