Summary of Generalized Linear Models

Jinhong Du

March 9, 2020

Contents

1 Generalized Linear Models **7**

1.1 Introduction . . . 7

1.1.1 Components of GLMs . . . 7

1.1.2 One-Parameter Exponential Families . . . 7

1.1.3 Canonical Link Functions . . . 8

1.2 Estimation . . . 8

1.2.1 Log-Likelihood Functions . . . 8

1.2.2 Likelihood Score Equations . . . 9

1.2.3 Properties . . . 10

1.2.4 Calculation of $\text{Var}(\hat{\boldsymbol{\beta}})$ . . . 10

1.2.5 The Distribution of $h(\hat{\boldsymbol{\beta}})$ . . . 11

1.3 Inference . . . 11

1.3.1 Hypothesis Tests . . . 11

1.3.2 Deviance Analysis . . . 12

1.4 Computation . . . 13

1.4.1 Newton-Raphson Method . . . 13

1.4.2 Fisher Scoring Method . . . 13

1.4.3 Iterative Reweighted Least Squares . . . 14

2 Models For Binary Data **15**

2.1 Binomial GLMs . . . 15

2.1.1 Model Settings . . . 15

2.1.2 Applications . . . 16

2.1.3 Estimation . . . 17

2.1.4 Inference . . . 18

2.1.5 Computation . . . 20

2.2 Beta-Binomial GLMs . . . 21

2.2.1 Model Settings . . . 21

2.2.2 Drawbacks . . . 22

3 Multinomial Response Models 23

3.1 Data Input . . . 23
3.2 Baseline Category Logit Models . . . 23
3.3 Multivariate GLMs . . . 24
3.3.1 Model Settings . . . 24
3.3.2 Estimation . . . 24
3.4 Cumulative Models . . . 25
3.4.1 Model Settings . . . 25
3.4.2 Estimation . . . 25
3.4.3 Comparison with OLS . . . 26

4 Models For Count Data 27

4.1 Poisson GLMs . . . 27
4.1.1 Model Settings . . . 27
4.1.2 Estimation . . . 27
4.1.3 Deviance Analysis . . . 27
4.1.4 Offset . . . 27
4.1.5 Poisson Modeling for Contingency Tables . . . 28
4.1.6 Connections Between Poisson and Multinomial GLMs . . . 30
4.1.7 Overdispersion . . . 31
4.2 Negative Binomial GLMs . . . 31
4.2.1 Model Setting . . . 31
4.2.2 Connections Between Negative Binomial and Poisson GLMs . . . 31
4.3 Zero-Inflated Poisson/Negative Binomial Models . . . 31
4.3.1 Zero-Inflated Counts . . . 31
4.3.2 Model Settings . . . 32

5 Quasi-Likelihood Methods 33

5.1 Quasi-Likelihoods . . . 33
5.2 Estimating Equations . . . 33
5.2.1 Properties . . . 34
5.2.2 Sandwich Covariance Adjustment for Variance Misspecification . . . 34

6 Generalized Linear Mixed Models 35

6.1 Introduction . . . 35
6.1.1 Marginal Models . . . 35

6.1.2 Generalized Linear Mixed Models . . . 35
6.1.3 Relationship between Marginal Models and GLMMs . . . 36
6.2 Binomial GLMMs . . . 36
6.2.1 Model Settings . . . 36
6.2.2 Latent Variable Threshold Model . . . 36
6.2.3 Properties . . . 37
6.2.4 Relationship Between Binomial GLMMs and GLMs . . . 37
6.3 Poisson GLMMs . . . 38
6.4 Normal Linear Mixed Models . . . 38
6.4.1 Random Intercept Models . . . 38
6.4.2 Multilevel Models . . . 39
6.5 Estimation of LMMs . . . 39
6.5.1 MLE of β . . . 39
6.5.2 Best Linear Unbiased Prediction (BLUP) of Random Effects . . . 40
6.5.3 Residual Maximum Likelihood (REML) . . . 41

7 Survival Analysis **43**
7.1 Introduction . . . 43
7.2 Estimating The Survival Functions without Covariates . . . 44
7.2.1 Non-parametric approach . . . 44
7.2.2 Parametric Models . . . 44
7.3 Proportional Hazards Regression Model . . . 45
7.4 Inference . . . 45
7.4.1 Log-Rank Test . . . 45

Chapter 1

Generalized Linear Models

1.1 Introduction

1.1.1 Components of GLMs

Definition 1. *Generalized linear models (GLMs) extend standard linear regression models to encompass non-normal response distributions and possibly nonlinear functions of the mean. They have three components,*

1. *Random component: It consists of a response variable $\boldsymbol{y}$ with independent observations $(y_1, \ldots, y_n)^\top$ having probability density or mass function for a distribution in the exponential family.*
2. *Linear predictor: For a parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ and a $n \times p$ model matrix $\boldsymbol{X}$ that contains values of p explanatory variables for the n observations, the linear predictor is $\boldsymbol{\eta} = \boldsymbol{X\beta}$.*
3. *Link function: This is a monotonic and differentiable function g applied to each component of $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$ that relates it to the linear predictor,*

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta}.$$

1.1.2 One-Parameter Exponential Families

Definition 2. *A one-parameter exponential family is a set of probability densities or mass functions of the form*

$$f(t;\theta) = e^{t\theta - b(\theta)} f_0(t) \tag{1.1}$$

where

- $\theta \in \mathbb{R}$ *: natural / canonical parameter.*
- $t \in \mathbb{R}$*: sufficient statistics.*
- $b(\theta)$*: normalizing or cumulant function such that* $b(\theta) = \displaystyle\int_{\mathbb{R}} e^{t\theta} f_0(t) \mathrm{d}t$.

- f_0: *a non-negative function that only depends on* t.

Properties

Let y be a random variable with densities in exponential family (1.1), then

1. $\mu = \mathbb{E}(y) = b'(\theta)$;
2. $\nu = \mathrm{Var}(y) = b''(\theta) = \frac{\partial \mu}{\partial \theta} > 0$.

Proof. Since

$$e^{b(\theta)} = \int_{\mathbb{R}} e^{b(\theta)} f(y;\theta)\mathrm{d}y = \int_{\mathbb{R}} e^{y\theta} f_0(y)\mathrm{d}y,$$

we have

$$b'(\theta)e^{b(\theta)} = \frac{\partial e^{b(\theta)}}{\partial \theta} = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} e^{y\theta} f_0(y)\mathrm{d}y = \int_{\mathbb{R}} y e^{y\theta} f_0(y)\mathrm{d}y.$$

Then

$$\begin{aligned} b'(\theta) &= \int_{\mathbb{R}} y e^{y\theta - b(\theta)} f_0(y)\mathrm{d}y = \mathbb{E}(y), \\ b''(\theta) &= \int_{\mathbb{R}} \frac{\partial}{\partial \theta} y e^{y\theta - b(\theta)} f_0(y)\mathrm{d}y \\ &= \int_{\mathbb{R}} y(y - b'(\theta)) e^{y\theta - b(\theta)} f_0(y)\mathrm{d}y \\ &= \mathbb{E}(y^2) - [\mathbb{E}(y)]^2 \\ &= \mathrm{Var}(y). \end{aligned}$$

■

When the distribution of y is in the exponential family, the relation between the mean and the variance characterizes the distribution.

1.1.3 Canonical Link Functions

Definition 3. *The link function* g *that transforms* μ *to the natural parameter* θ *is called the canonical link.*

1.2 Estimation

1.2.1 Log-Likelihood Functions

$$\begin{aligned} L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} L_i(\beta_i) \\ &= \sum_{i=1}^{n} \log f(y_i;\theta_i) \end{aligned}$$

$$= \sum_{i=1}^{n} [y_i\theta_i - b(\theta_i) + \log f_0(y_i)]$$

1.2.2 Likelihood Score Equations

For Canonical Link

Since $\theta_i = g(\mu_i) = \eta_i = \boldsymbol{x_i}^\top \boldsymbol{\beta}$, we have

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} [y_i x_{ij} - b'(\boldsymbol{x_i}^\top \boldsymbol{\beta}) x_{ij}] = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij}.$$

or equivalently, in matrix form,

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\mu}).$$

Also,

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} \nu_i x_{ij} x_{ik}$$

and

$$H = \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right)_{ij} = -\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} \preceq 0,$$

where $V = \mathrm{diag}(\nu_1, \ldots, \nu_n)$. So the log-likelihood function for canonical link is a concave function and the minimizer of $L(\boldsymbol{\beta})$ is unique.

For General Link

$$\begin{aligned} 0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{\partial L_i(\beta)}{\partial \beta_j} \\ &= \sum_{i=1}^{n} \frac{\partial L_i(\boldsymbol{\beta})}{\partial \theta_i} \frac{\partial \theta_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \sum_{i=1}^{n} (y_i - b'(\theta_i)) \cdot \frac{1}{b'(\theta_i)} \cdot \frac{1}{g'(\mu_i)} \cdot x_{ij} \\ &= \sum_{i=1}^{n} \frac{y_i - \mu_i}{\nu_i} \cdot \frac{x_{ij}}{g'(\boldsymbol{\mu_i})}. \end{aligned}$$

In matrix form,

$$0 = \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}),$$

where

$$\boldsymbol{V} = \begin{bmatrix} \nu_1 & & \\ & \ddots & \\ & & \nu_n \end{bmatrix}, \quad \boldsymbol{D} = \begin{bmatrix} g'(\mu_1) & & \\ & \ddots & \\ & & g'(\mu_n) \end{bmatrix}^{-1}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

In this case, the log-likelihood function is not necessary a concave function. Although $\boldsymbol{\beta}$ does not appear in these equations, it is there implicitly through $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, since $\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $\nu_i = \frac{\partial \mu_i}{\partial \theta_i}$. The likelihood equations are nonlinear functions of $\boldsymbol{\beta}$ that must be solved iteratively.

1.2.3 Properties

Asymptotic Distribution of GLMs

Let p be fixed, $\boldsymbol{\beta_0}$ be the true parameter and $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$ be the maximum likelihood estimator of $\boldsymbol{\beta_0}$ with sample size n. When $n \to \infty$, we have

1. (Consistency) $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} \to \boldsymbol{\beta_0}$
2. (Asymptotic Normality) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta}_0) \to N(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}_0})$, then $\frac{1}{n}\boldsymbol{V}_{\boldsymbol{\beta}_0}$ is the estimated variance matrix of $\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}$.

1.2.4 Calculation of $\mathrm{Var}(\hat{\boldsymbol{\beta}})$

Notice that $L(\boldsymbol{\beta_0})$ and its derivatives are functions of random vector $\boldsymbol{y}$. Two useful facts:

1. $\mathbb{E}[L'(\boldsymbol{\beta_0})] = 0$.
2. $\mathrm{Var}[L'(\boldsymbol{\beta_0})] = -\mathbb{E}[L''(\boldsymbol{\beta_0})]$.

We use delta methods. From

$$0 = L'(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}) \approx L'(\boldsymbol{\beta_0}) + L''(\boldsymbol{\beta_0})(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta_0}),$$

we have

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta_0} \approx -[L''(\boldsymbol{\beta_0})]^{-1} L'(\boldsymbol{\beta_0}).$$

Thus

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}} - \boldsymbol{\beta_0}) \approx -\left[\frac{L''(\boldsymbol{\beta_0})}{n}\right]^{-1} \left(\sqrt{n} \cdot \frac{L'(\boldsymbol{\beta_0})}{n}\right).$$

By law of large numbers,

$$\frac{1}{n}L''(\beta_0) = \frac{1}{n}L_i''(\beta_0) \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{1}{n}L''(\beta_0)\right].$$

By central limit theorem, we have

$$\sqrt{n}\frac{L'(\beta_0)}{n} \xrightarrow{D} N\left(0, \frac{1}{n}\mathrm{Var}(L'(\beta_0))\right).$$

By Slutsky theorem, we have as $n \to \infty$,

$$\begin{aligned}
\boldsymbol{V}_{\boldsymbol{\beta}_0} = n\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}) &\approx -\left[\mathbb{E}\left(\frac{L''(\boldsymbol{\beta_0})}{n}\right)\right]^{-1} \frac{1}{n}\mathrm{Var}(L'(\boldsymbol{\beta_0})) \left[\mathbb{E}\left(\frac{L''(\boldsymbol{\beta_0})}{n}\right)\right]^{-1} \\
&= n\left[\mathbb{E}\left(L''(\boldsymbol{\beta_0})\right)\right]^{-1} \mathrm{Var}(L'(\boldsymbol{\beta_0})) \left[\mathbb{E}\left(L''(\boldsymbol{\beta_0})\right)\right]^{-1} \\
&= n\left[\mathrm{Var}(L'(\boldsymbol{\beta_0}))\right]^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= n[\mathrm{Var}(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}))]^{-1} \\
&= n(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}\mathrm{Var}(\boldsymbol{y})\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X})^{-1} \\
&= n(\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X})^{-1}
\end{aligned}
$$

and

$$
\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}) = \frac{1}{n} V_{\boldsymbol{\beta_0}} \approx (\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X})^{-1},
$$

where $\boldsymbol{I} = \boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X}$ is called the *fisher information matrix*.

1.2.5 The Distribution of $h(\hat{\boldsymbol{\beta}})$

Let h be a real-value functon continuous at $\boldsymbol{\beta_0}$, then by delta method, we have $\sqrt{n}[h(\hat{\boldsymbol{\beta}}_{\boldsymbol{n}}) - h(\boldsymbol{\beta_0})] \to N\left(0, \left[\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]^\top \boldsymbol{V}_{\boldsymbol{\beta}_0} \left[\frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]\right)$

1.3 Inference

1.3.1 Hypothesis Tests

For the model parameter β, we focus on tests of

$$
H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\alpha_0} \quad \text{v.s.} \quad H_1 : \boldsymbol{A}\boldsymbol{\beta} \neq \boldsymbol{\alpha_0},
$$

for $\boldsymbol{A} \in \mathbb{R}^{d\times p}$ and $\boldsymbol{\alpha_0} \in \mathbb{R}^d$.

Next we introduce three tests concerning this hypothesis. The Wald test is based on the curvature of $L(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_{MLE}$. The likelihood-ratio test statistic is twice the vertical distance between values of $L(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_{MLE}$ and at $\boldsymbol{\beta}$ such that $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\alpha}_0$. The score test uses the slope and curvature of $L(\beta)$ at $\boldsymbol{\beta}_0 = \boldsymbol{\alpha}_0$.

Wald Tests

The Wald test statistics is given by

$$
T = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha_0})^\top S^{-1}_{\boldsymbol{A}\hat{\boldsymbol{\beta}}} (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{\alpha_0}) \overset{H_0}{\sim} \chi^2_d
$$

where $S_{\boldsymbol{A}\hat{\boldsymbol{\beta}}} = \boldsymbol{A}\boldsymbol{V}_{\hat{\boldsymbol{\beta}}}\boldsymbol{A}^\top$ is the *estimated standard error (SE)* of $\hat{\boldsymbol{\beta}}$.

Likelihood Ratio Tests

The likelihood ratio test statistics is given by

$$
-2\log\Lambda = -2[L(\tilde{\boldsymbol{\beta}}) - L(\hat{\boldsymbol{\beta}})],
$$

where $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under the constraint $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\alpha}_0$, and $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ without any constraint. As $n \to \infty$, $-2\log\Lambda \to \chi^2_d$.

Score Tests

For a simple case when $\boldsymbol{A} = \boldsymbol{I_p}$, the score test statistics is given by

$$T = -L'(\boldsymbol{\alpha_0})^\top [L''(\boldsymbol{\alpha_0})]^{-1} L'(\boldsymbol{\alpha_0}).$$

Under H_0 and as $n \to \infty$, $T \to \chi^2_p$.

1.3.2 Deviance Analysis

Definition 4. *The deviance is a function that measure the distance between two distribution in the same family* $\{f(y;\theta) : \theta \in \Theta\}$ *with parameter* θ_1 *and* θ_2*,*

$$D(\theta_1, \theta_2) = 2\mathbb{E}_{\theta_1}\left[\log \frac{f(Y;\theta_1)}{f(Y;\theta_2)}\right],$$

where $\mathbb{E}_{\theta_1}$ *denotes the expectation taken with respect to* Y *with density function* $f(y;\theta_1)$*. For GLMs, since* θ *and* μ *have a one-to-one mapping, we also write* $D(\mu_1, \mu_2)$ *and*

$$D(\mu_1, \mu_2) = 2\mathbb{E}_{\theta_1}[Y(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)] = 2[\mu_1(\theta_1 - \theta_2) - b(\theta_1) + b(\theta_2)].$$

- Generally $D(\theta_1, \theta_2) \neq D(\theta_2, \theta_1)$ and therefore D may not be a metric.
- $\frac{1}{2}D(\mu_1, \mu_2)$ is the KL divergence.
- If f is the normal density with $\sigma = 1$, then $D(\mu_1, \mu_2) = (\mu_1 - \mu_2)^2$.

Definition 5. *A saturate model is a model whose number of parameters is euqal to the sample points.*

In GLMs, we have some statistics and properties that are similar to those of the linear models.

1. The *total deviance/residual deviance* is given by
$$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \sum_i D(y_i, \hat{\mu}_i) = 2\sum_{i=1}^n \log \frac{f(y_i;\theta_{y_i})}{f(y_i;\theta_{\hat{\mu}_i})},$$
the sum of deviance between the saturated model $\boldsymbol{\mu} = \boldsymbol{y}$ and another model with $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ ($\boldsymbol{\theta}$ is therefore determined by $\boldsymbol{\mu}$).
2. The *null deviance* is given by $D_+(\boldsymbol{y}, \overline{\boldsymbol{y}}) = \sum_i D(y_i, \overline{y})$.
3. $R^2 = 1 - \frac{D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{D_+(\boldsymbol{y}, \overline{\boldsymbol{y}})}$
4. *(Additivity Theorem)* $D_+(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}^{(1)}) = D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}^{(1)})$.

Deviance Analysis for Nested Models

In GLMs, consider the canonical link $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \cdots, \boldsymbol{\beta}^{(J)})^\top$ and $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(J)})$ where $\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{p_j}$, $\boldsymbol{x}^{(j)} \in \mathbb{R}^{n \times p_j}$, and $\sum_{j=1}^J p_j = p$.

We call $\mathcal{M}_k$ with $\theta = \sum_{j=1}^k \boldsymbol{x}^{(k)}\boldsymbol{\beta}^{(k)}$ a *nested model* of the full model for $k = 1, \ldots, J$. We can use deviance to select best nested models.

The deviance table is given by

MLE	$2L_{\max}$	Deviance	Test
$\hat{\boldsymbol{\beta}}^{(\mathbf{0})} = \overline{y}$	$2L(\hat{\boldsymbol{\beta}}^{(\mathbf{0})})$		
$\hat{\boldsymbol{\beta}}^{(1)}$	$2L(\hat{\boldsymbol{\beta}}^{(1)})$	$D_+(\hat{\boldsymbol{\mu}}^{(1)}, \hat{\boldsymbol{\mu}}^{(0)})$	$\chi^2_{p_1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\hat{\boldsymbol{\beta}}^{(J)} = \hat{\boldsymbol{\beta}}$	$2L(\hat{\boldsymbol{\beta}}^{(J)})$	$D_+(\hat{\boldsymbol{\mu}}^{(J)}, \hat{\boldsymbol{\mu}}^{(J-1)})$	$\chi^2_{p_J}$

Table 1.1: Deviance table.

1.4 Computation

1.4.1 Newton-Raphson Method

At the t-th step, since

$$L(\boldsymbol{\beta}) \approx L(\boldsymbol{\beta}^{(t)}) + L'(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)})^\top L''(\boldsymbol{\beta}^{(t)})^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}),$$

taking derivative with respect to $\boldsymbol{\beta}$ and setting to zero yields

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = L'(\boldsymbol{\beta}^{(t)}) + L''(\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}) = \mathbf{0}.$$

So the next iteration is given by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - [L''(\boldsymbol{\beta}^{(t)})]^{-1} L'(\boldsymbol{\beta}^{(t)}).$$

The negative of the Hessian matrix, $-\boldsymbol{H}^{(t)} = -L''(\boldsymbol{\beta}^{(t)})$, is called the *observed information*.

- The Newton-Raphson method converges to the global maximum if $L(\boldsymbol{\beta})$ is strongly concave

1.4.2 Fisher Scoring Method

We replace the observed information with the *expected information* $\boldsymbol{J} = -\mathbb{E}[L''(\boldsymbol{\beta})]$. But we don't know the actual value of $\boldsymbol{J}$ since we don't know the true $\boldsymbol{\beta}$. So we use the estimated expected information at the t-th step,

$$\boldsymbol{J}^{(t)} = \mathrm{Var}[L'(\boldsymbol{\beta})]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}} = \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X}$$

where $\boldsymbol{W}^{(t)} = \boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(t)}}$.

Then the Fisher Scoring iteration is given by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + [\boldsymbol{J}^{(t)}]^{-1} L'(\boldsymbol{\beta}^{(t)}).$$

Noted that for canonical links,

$$\boldsymbol{H}^{(t)} = L''(\boldsymbol{\beta}^{(t)}) = -\boldsymbol{X}^\top \boldsymbol{V}^{(t)} \boldsymbol{X} = -\boldsymbol{J}^{(t)},$$

i.e., the Fisher Scoring iteration is equivalent to the Newton-Raphson iteration.

1.4.3 Iterative Reweighted Least Squares

At t-th iteration, since $\mu_i = g^{-1}(\eta_i)$ is a function of η_i, we take the Taylor expension at $\eta_i^{(t)} = \boldsymbol{x}_i\boldsymbol{\beta}^{(t)}$,

$$\mu_i \approx \mu_i^{(t)} + \frac{\partial \mu_i}{\partial \eta_i}\Big|_{\eta_i^{(t)}} (\eta_i - \eta_i^{(t)})$$

Then

$$\boldsymbol{\mu} \approx \boldsymbol{\mu}^{(t)} + \boldsymbol{D}^{(t)}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{(t)}).$$

So

$$\begin{aligned}
L'(\boldsymbol{\beta}) &= \boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \\
&\approx \boldsymbol{X}^\top \boldsymbol{D}^{(t)}(\boldsymbol{V}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \\
&\approx \boldsymbol{X}^\top \boldsymbol{D}^{(t)}(\boldsymbol{V}^{(t)})^{-1}[\boldsymbol{y} - \boldsymbol{\mu}^{(t)} - \boldsymbol{D}^{(t)}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})] \\
&= \boldsymbol{X}^\top \boldsymbol{W}^{(t)}[(\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) - (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}\boldsymbol{\beta}^{(t)})],
\end{aligned}$$

which is linear to $\boldsymbol{\beta}$. Let $\boldsymbol{z}^{(t)} = \boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$, then

$$L'(\boldsymbol{\beta}) \approx \boldsymbol{X}^\top \boldsymbol{W}^{(t)}(\boldsymbol{z}^{(t)} - \boldsymbol{X}\boldsymbol{\beta}).$$

Setting to zero yields the $(t+1)$-th iteration,

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}^{(t)}\boldsymbol{z}^{(t)},$$

which is a weighted least squares estimator with weight $\boldsymbol{W}^{(t)}$.

Noted that

$$\begin{aligned}
\boldsymbol{\beta}^{(t+1)} &= (\boldsymbol{X}^\top \boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{W}^{(t)}[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)})] \\
&= \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}^\top \boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{D}^{(t)}(\boldsymbol{V}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \\
&= \boldsymbol{\beta}^{(t)} + (\boldsymbol{X}^\top \boldsymbol{W}^{(t)}\boldsymbol{X})^{-1}L'(\boldsymbol{\beta}^{(t)}),
\end{aligned}$$

which is the same as the Fisher Scoring iteration.

Chapter 2

Models For Binary Data

2.1 Binomial GLMs

2.1.1 Model Settings

Input Data

In the binary GLM, the observations are $y_i \sim \text{Binomial}(n_i, p_i)$ with probability mass function

$$\mathbb{P}(y_i = y) = \binom{n_i}{y} p_i^y (1-p_i)^{n_i - y} = \binom{n_i}{y} \left(\frac{p_i}{1-p_i}\right)^y (1-p_i)^{n_i} = e^{\log\left(\frac{p_i}{1-p_i}\right) y - n_i \log(1-p_i)} \binom{n_i}{y},$$

which is in the one-parameter exponential family. If $n_i = 1$, then $y_i \sim \text{Bernoulli}(p_i)$.

1. Ungrouped data. Each observation comes from $\text{Bernoulli}(p_i)$.
2. Grouped data. We combine the observations that share the same values of x_i, which give us the grouped data that $y_i \sim \text{Binomial}(n_i, p_i)$.

y_i	1	1	0	1	0
n_i	1	1	1	1	1
x_i	0	0	0	1	1

$\tilde{y}_i$	2	1
$\tilde{n}_i$	3	2
$\tilde{x}_i$	0	1

Table 2.1: Ungrouped and grouped data.

The log-likelihood functions for these two types of data are given by

$$\begin{aligned}
L(\boldsymbol{\beta}) &= \sum_{i=1}^{N} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) - \log(1-p_i) \right] \\
\tilde{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{\tilde{N}} \left[\tilde{y}_i \log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) - \tilde{n}_i \log(1-\tilde{p}_i) + \log\binom{\tilde{n}_i}{\tilde{y}_i} \right] \\
&= \sum_{i=1}^{\tilde{N}} \sum_{j=1}^{\tilde{n}_i} \left[y_{i_j} \log\left(\frac{p_{i_j}}{1-p_{i_j}}\right) - \log(1-p_{i_j}) \right] + \sum_{i=1}^{\tilde{N}} \log\binom{n_i}{y_i}
\end{aligned}$$

$$= L(\boldsymbol{\beta}) + \sum_{i=1}^{\tilde{N}} \log \binom{n_i}{y_i}$$

since $\tilde{p}_i = p_{i_1} = \cdots = p_{i_{\tilde{n}_i}}$ and $\tilde{y}_i = \sum_{j=1}^{\tilde{n}_i} y_{i_j}$. Then the likelihood equations for these two forms of data are the same.

Link Functions

The expectation of each sample is $\mathbb{E}(y_i) = n_i p_i$ where n_i is a known constant. Thus we define the link function as a function of p_i

$$g(p_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

Equivalently, $p_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \in [0,1]$. As g^{-1} is monotone (since g is a one-to-one function and continuous), a natual choice of g^{-1} is to make it as a cdf of some distribution. We denote $F(z) = g^{-1}(z)$. If y_i is binary, let $\epsilon_i \overset{iid}{\sim} F$, then the latent variable is $y_i^* = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ and $y_i = \begin{cases} 1 & , \text{ if } y_i^* \geq 0 \\ 0 & , \text{ if } y_i^* < 0 \end{cases}$, which is called a *latent variable threshold model*. Some choices of the link function are

1. The probit link. $F(z) = \Phi(z)$ and $g(x) = \Phi^{-1}(x)$.
2. The logit link. $F(z) = \frac{e^z}{1+e^z}$ and $g(x) = \log\left(\frac{x}{1-x}\right) \triangleq \text{logit}(x)$. The logit link is the canonical link of the Binomial distribution
3. The log-log link.

2.1.2 Applications

Odds Ratios

When x_i and y_i are both binary data, we can use a 2-by-2 table to represent the data. Assume that (x_i, y_i) are i.i.d., the *odds ratio* for the response variable Y is given by

$$OR = \frac{\mathbb{P}(Y=1|X=1)/\mathbb{P}(Y=1|X=0)}{\mathbb{P}(Y=0|X=1)/\mathbb{P}(Y=0|X=0)}.$$

If we fit the model $\text{logit}(p_i) = \beta_0 + \beta_1 x_i$, then $e^{\beta_1} = OR$. Similarly, when we have more than one covariate, we can define the odds ratio for one covariate conditional on other covariates.

Case-Contral Studies

1. Prospective design. It takes long time and has limited power for rare events. We compare $\mathbb{E}(Y=1|X=1)$ with $\mathbb{E}(Y=1|X=0)$ where X is exposure and Y is disease.
2. Retrospective study. In a retrospective cohort we separate groups by exposure and then look at disease status.

3. Case-control study. In a case control study we separate groups by disease status and then look backwards for exposures. We compare $\mathbb{E}(X = 1|Y = 1)$ with $\mathbb{E}(X = 1|Y = 0)$. If we only care about OR, then

$$OR = \frac{\mathbb{P}(Y = 1|X = 1)/\mathbb{P}(Y = 1|X = 0)}{\mathbb{P}(Y = 0|X = 1)/\mathbb{P}(Y = 0|X = 0)} = \frac{\mathbb{P}(X = 1|Y = 1)/\mathbb{P}(X = 1|Y = 0)}{\mathbb{P}(X = 0|Y = 1)/\mathbb{P}(X = 0|Y = 0)}.$$

Classfication

- A classification table.

y \ $\hat{y}$	0	1
0	True negative	False positive
1	False negative	True positive

Table 2.2: A calssfication table.

- ROC curve. The sensitivity is defined as $\mathbb{P}(\hat{y} = 1|y = 1)$ and the specificity is defined as $\mathbb{P}(\hat{y} = 0|y = 0)$. The sensitivity is the *true positive rate (TPR)*, and $\mathbb{P}(\hat{y} = 1|y = 0) = (1-\text{specificity})$ is the *false positive rate (FPR)*. The *receiver operating characteristic (ROC) curve* is a plot of the true positive rate as a function of the false positive rate as π_0 decreases from 1 to 0.

2.1.3 Estimation

We will mainly focus on the Logistic regression where we use the logit link functon for the binary data.

Score Equations

When $p_i = F(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, the score equations for either grouped or ungrouped data are given by

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) = 0,$$

i.e.,

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{v_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{N} \frac{[y_i - n_i F(\boldsymbol{x}_i^\top \boldsymbol{\beta})]x_{ij}}{F(\boldsymbol{x}_i^\top \boldsymbol{\beta})[1 - F(\boldsymbol{x}_i^\top \boldsymbol{\beta})]} f(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = 0,$$

where f is the derivative of F.

When F is the standard logistic function, we have

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{N} \left(y_i - n_i \frac{\boldsymbol{x}_i^\top \boldsymbol{\beta}}{1 + \boldsymbol{x}_i^\top \boldsymbol{\beta}} \right) x_{ij} = 0,$$

since $f(x) = F(x)[1 - F(x)]$.

Calculation of Var($\hat{\boldsymbol{\beta}}$)

Since the logit link is the canonical link, we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}$$

where $W = V = \mathrm{diag}(n_1 p_1(1-p_1), \ldots, n_N p_N(1-p_N))$. So the estimated variance of $\hat{\boldsymbol{\beta}}$ is $(\boldsymbol{X}^\top \hat{\boldsymbol{W}} \boldsymbol{X})^{-1}$ where $\hat{W} = \hat{V} = \mathrm{diag}(n_1 \hat{p}_1(1-\hat{p}_1), \ldots, n_N \hat{p}_N(1-\hat{p}_N))$.

2.1.4 Inference

Hypothesis Tests

Consider a simple case. Under the null model, the group data is $\sum_{i=1}^{N} y_i \sim \mathrm{Binomial}(N, p)$. We want to test for $H_0 : \beta = \mathrm{logit}(p_0)$ (or equivalently: $H_0 : p = p_0$) where β is the constant coefficient. Define $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, then the MLE is $\hat{p} = \overline{y}$. The test statistics are

- **Wald Test.**

$$Z_1 = \left(\frac{\hat{\beta} - \mathrm{logit}(p_0)}{\mathrm{SE}(\hat{\beta})}\right)^2 = [\mathrm{logit}(y) - \mathrm{logit}(p_0)]^2 N y(1-y),$$

  or

$$Z_2 = \left(\frac{\hat{p} - p_0}{\mathrm{SE}(\hat{p})}\right)^2 = \frac{(y-p_0)^2}{\frac{1}{N} y(1-y)}.$$

- **Likelihood Ratio Test.**

$$L = -2(L_0 - L_1) = -2\log\left[\frac{p_0^{Ny}(1-p_0)^{N(1-y)}}{y^{Ny}(1-y)^{N(1-y)}}\right].$$

- **Score Test.**

$$T = -L'(\beta_0)^\top [L''(\beta_0)]^{-1} L'(\beta_0) \frac{(y-p_0)^2}{\frac{1}{N} p_0(1-p_0)}.$$

Example 1. *For the following two data sets,*

$$A: \quad N = 25, \sum_{i=1}^{N} y_i = 24, p_0 = \frac{1}{2},$$

$$B: \quad N = 25, \sum_{i=1}^{N} y_i = 23, p_0 = \frac{1}{2},$$

we have

We can see that Wald test is not ideal for logistic regression since Z_1 is too conservatibe and Z_2 is too liberal. It is less stable when y is close to 0 or 1, and it depends on scale.

Statistics	Value for A	Value for B
Z_1	9.7	11
Z_2	137.8	60
L	26.3	20.7
T	21.2	17.6

Table 2.3: Comparison of different statistics.

Deviance Analysis

The total/residual deviance for a binomial GLM is of the form,

$$\begin{aligned} D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) &= \sum_{i=1}^{N} D(y_i, n_i\hat{\pi}_i) = 2\sum_{i=1}^{N} \log\left[\frac{f\left(y_i; \frac{y_i}{n_i}\right)}{f(y_i; \hat{\pi}_i)}\right] \\ &= 2\sum_{i=1}^{N} \log\left[\frac{\binom{n_i}{y_i}\left(\frac{y_i}{n_i}\right)^{y_i}\left(1-\frac{y_i}{n_i}\right)^{n_i-y_i}}{\binom{n_i}{y_i}\hat{\pi}_i^{y_i}\,(1-\hat{\pi}_i)^{n_i-y_i}}\right] \\ &= 2\sum_{i=1}^{N} y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + 2\sum_{i=1}^{N}(n_i-y_i)\log\left(\frac{n_i-y_i}{n_i-n_i\hat{\pi}_i}\right). \end{aligned}$$

Noted that the estimated linear coefficients for the same binomial GLM with ungrouped and grouped data are the same because of the same likelihood equation.

For ungrouped data, the total deviance is given by

$$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{N} y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + 2\sum_{i=1}^{N}(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\pi}_i}\right).$$

For grouped data, the deviance for the same model is given by

$$\tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\tilde{\boldsymbol{\mu}}}) = 2\sum_{i=1}^{\tilde{N}}\left[\tilde{y}_i \log\left(\frac{\tilde{y}_i}{\tilde{n}_i\hat{\tilde{\pi}}_i}\right) + (\tilde{n}_i-\tilde{y}_i)\log\left(\frac{\tilde{n}_i-\tilde{y}_i}{\tilde{n}_i-\tilde{n}_i\hat{\tilde{\pi}}_i}\right)\right].$$

where $\hat{\tilde{\mu}}_i$ and $\hat{\mu}_j$ are the MLEs of $\tilde{\mu}_i$ and μ_j of the same binary GLM, respectively. Here $\frac{1}{n_i}\hat{\tilde{\mu}}_i = \hat{\mu}_{i_1} = \cdots = \hat{\mu}_{i_{n_i}}$ and $\hat{\tilde{\pi}}_i = \hat{\pi}_{i_1} = \cdots = \hat{\pi}_{i_{n_i}}$ for indexes $i_1, \ldots, i_{n_i} \in \{1, \ldots, N\}$ that are in group i. Then

$$\tilde{y}_i \log\left(\frac{\tilde{y}_i}{\tilde{n}_i\hat{\pi}_i}\right) + (\tilde{n}_i-\tilde{y}_i)\log\left(\frac{\tilde{n}_i-\tilde{y}_i}{\tilde{n}_i-\tilde{n}_i\hat{\pi}_i}\right) = \sum_{j=1}^{n_i}\left[y_{i_j}\log\left(\frac{\tilde{y}_i}{\tilde{n}_i\hat{\pi}_i}\right) + (1-y_{i_j})\log\left(\frac{\tilde{n}_i-\tilde{y}_i}{\tilde{n}_i-\tilde{n}_i\hat{\pi}_i}\right)\right].$$

So

$$\tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\tilde{\boldsymbol{\mu}}}) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{\tilde{N}}\sum_{j=1}^{n_i}\left[y_{i_j}\log\left(\frac{\tilde{y}_i}{\tilde{n}_i y_{i_j}}\right) + (1-y_{i_j})\log\left(\frac{\tilde{n}_i-\tilde{y}_i}{\tilde{n}_i(1-y_{i_j})}\right)\right].$$

Each term in the above summation does not necessarily equal to zero in general, i.e., the estimated mean $\frac{1}{\tilde{n}_i}\tilde{y}_i$ may not equal to the estimated mean $y_{i_1}, \ldots, y_{i_4}$ for the two saturated models. However, the different of total deviance between two different models are the same for ungrouped and grouped data, i.e.,

$$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_1) - D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}_2) = \tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\tilde{\boldsymbol{\mu}}}_1) - \tilde{D}_+(\tilde{\boldsymbol{y}}, \hat{\tilde{\boldsymbol{\mu}}}_2).$$

2.1.5 Computation

Methods

We can use IRLS to compute binary GLMs. The tth iteration is given by

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)},$$

where

$$\begin{aligned} \boldsymbol{z}^{(t)} &= \boldsymbol{X}\boldsymbol{\beta}^{(t)} + (\boldsymbol{D}^{(t)})^{-1}(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}) \\ z_i^{(t)} &= \log\left(\frac{p_i^{(t)}}{1 - p_i^{(t)}}\right) + \frac{y_i - n_i p_i^{(t)}}{n_i p_i^{(t)}(1 - p_i^{(t)})} \\ W_{ii}^{(t)} &= V_{ii}^{(t)} = n_i p_i^{(t)}(1 - p_i^{(t)}). \end{aligned}$$

Since the logit link is canonical, the Newton-Raphson method, Fisher Scoring method and the iterative reweighted least squares method are essentially the same for the Logistic regression.

Infinite Parameter Estimates

- **Complete/Perfect Separation.** If there exists $\boldsymbol{\beta}$ such that if $\boldsymbol{x}_i^\top \boldsymbol{\beta} > 0$ then $y_i = 1$ otherwise $y_i = 0$, then

$$\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \to \begin{cases} 1 & , \text{ if } \boldsymbol{x}_i^\top \boldsymbol{\beta} > 0 \\ 0 & , \text{ otherwise} \end{cases},$$

  as some $\beta_j \to \infty$. Thus, the solution of the score equation is infinite and not unique.

- **Quasi-Complete Separation.** If there exists $\boldsymbol{\beta}$ such that if $\boldsymbol{x}_i^\top \boldsymbol{\beta} > 0$ then $y_i = 1$, if $\boldsymbol{x}_i^\top \boldsymbol{\beta} < 0$ then $y_i = 0$, and if $\boldsymbol{x}_i^\top \boldsymbol{\beta} > 0$ then y_i can be either 0 or 1 (allow data points on the separation hyperplane with both outcomes), then

$$\frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \to \begin{cases} 1 & , \text{ if } \boldsymbol{x}_i^\top \boldsymbol{\beta} > 0 \\ 0 & , \text{ if } \boldsymbol{x}_i^\top \boldsymbol{\beta} < 0 \\ \frac{1}{2} & , \text{ if } \boldsymbol{x}_i^\top \boldsymbol{\beta} = 0 \end{cases},$$

  as some $\beta_j \to \infty$. Thus, the solution of the score equation is still infinite and not unique.

- **Control of Complete/Quasi-Complete Separation.** With an infinite estimate, we can still compute likelihood-ratio tests and make statistical inference. With quasi-complete separation, some parameter estimates and SE values may be unaffected, and even Wald inference methods are available with them.
  Alternatively, we can make some adjustment so that all estimates are finite. Some approaches smooth the data, thus producing finite estimates, such as Bayesian approaches. Some approaches maximize a penalized likelihood function.

2.2 Beta-Binomial GLMs

2.2.1 Model Settings

For grouped binary data, the real $\{y_i\}$ may exhibit more variability than the binomial allows. This can happen in two common ways.

1. Heterogeneity of π_i. Observations at a particular setting of explanatory variables have success probabilities that vary according to values of unobserved variables.
2. Positively correlated Bernoulli trials. The Bernoulli trials at each i are positively correlated.

To deal with the overdispersion problem, we can use the *beta-binomial distribution*, which results from a beta distribution mixture of binomials. Suppose that $y|\pi \sim \text{Binomial}(n, \pi)$, and $\pi \sim \text{Beta}(\alpha_1, \alpha_2)$. The density function of π is given by

$$f(t; \alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} t^{\alpha_1 - 1} t^{\alpha_2 - 1}, \quad 0 \le t \le 1,$$

with parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. The beta distribution for π has mean and variance

$$\mathbb{E}(\pi) = \mu, \qquad \text{Var}(\pi) = \mu(1-\mu)\frac{\theta}{1+\theta},$$

where $\mu = \frac{\alpha_1}{\alpha_1+\alpha_2}$ and $\theta = \frac{1}{\alpha_1+\alpha_2}$. For the beta-binomial random variable y,

$$\mathbb{E}(y) = n\mu, \qquad \text{Var}(y) = n\mu(1-\mu)\left[1 + (n-1)\frac{\theta}{1+\theta}\right].$$

In fact, $\rho = \frac{\theta}{1+\theta}$ is the correlation between each pair of the individual Bernoulli random variables that sum to y.

Assume $y_i \sim \text{Beta-Binomial}(n_i, \mu_i, \theta)$ and $y_1, \ldots, y_N$ are independent. Models using the beta-binomial distribution usually let θ_i be the same unknown constant for all observations. Models can use any link function for binary data, but the logit is most common. Then the beta-binomial GLM is given by

$$\text{logit}(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

When $n_i = 1$ $(i = 1, \ldots, N)$, then it is not possible to have overdispersion or underdispersion, and this model is still valid.

2.2.2 Drawbacks

The beta-binomial distribution is not in the exponential dispersion family, even for known θ. When the linear predictor is correct, the beta-binomial ML estimator $\boldsymbol{\beta}$ is not consistent if the actual distribution is not beta-binomial.

Chapter 3

Multinomial Response Models

3.1 Data Input

Assume that we have c categories and the ith sample $\boldsymbol{y}_i = \begin{bmatrix} y_{i_1} & \cdots & y_{i_c} \end{bmatrix}^\top$ follows Multinomial$(n_i, \boldsymbol{p}_i)$ where $\boldsymbol{p}_i = \begin{bmatrix} p_{i_1} & \cdots & p_{i_c} \end{bmatrix}^\top$. The probability mass function for $\boldsymbol{y}_i$ is given by

$$\mathbb{P}(y_{i_1} = t_1, \ldots, y_{i_c} = t_c) = \binom{n_i}{t_1 \cdots t_c} p_1^{t_1} \cdots p_c^{t_c}$$

for $t_1, \ldots, t_c \in \mathbb{N}$ and $\sum_{k=1}^{c} t_k = n_i$.

The responses can either be grouped or ungrouped as in the binary GLMs.

Also, there are two types of them,

1. Nomial: categories have no order. We use vector form $\boldsymbol{y}_i \in \mathbb{R}^c$.
2. Ordinal: categories have an order. We use $y_i \in \mathbb{R}$, denoting which category the ith sample belongs to.

For the last case, we will have different models to deal with the orders. Noted that even though the form of repsonse is different, we use the same multinomial distributions when calculating the likelihood (with different model assumptions). So y_{i_k} and p_{i_k} will appear in the their likelihoods.

3.2 Baseline Category Logit Models

Since we have talk about binary GLM before, we can extend the idea to use pairwise binary GLM models. First assume that class c is the baseline category, and for $i = 1, \ldots, c-1$, we have

$$\frac{p_{i_k}}{p_{i_k} + p_{i_c}} = F(\boldsymbol{x}_i^\top \boldsymbol{\beta}_k)$$

for some cumulative distribution function F. A desired property of F is that the model does not depend on which category we choose as the baseline. Then, we need

1. For each k, there exists some $\tilde{\boldsymbol{\beta}}_k$ such that $\frac{p_{ic}}{p_{i_k}+p_{i_c}} = F(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_k)$.
2. For any $k_1, k_2 \neq c$, there exists some $\tilde{\boldsymbol{\beta}}_{k_1 k_2}$ such that $\frac{p_{i_{k_1}}}{p_{i_{k_1}}+p_{i_{k_2}}} = F(\boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}}_{k_1 k_2})$.

If F corresponds to the logit link, then the two requirements are satisfied as $\frac{p_{ik}}{p_{ic}} = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_k}$. This is called the *baseline-category logit model*. If there is a natural baseline category in some applications (categories not "exchangeable"), other links can still be used. The Baseline-category logit model is closely related to the discrete-choice model in economics.

Under the baseline-category logit model, we have

$$p_{ic} = \frac{1}{1+\sum_{l=1}^{c-1} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l}}, \qquad p_{ik} = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_k}}{1+\sum_{l=1}^{c-1} e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_l}}, \qquad k = 1, \ldots, c-1.$$

Treating each pair as a logistic regression, we can get the asymptotic distribution of each $\hat{\boldsymbol{\beta}}_k$. However, there are problems with this model:

1. $\hat{\boldsymbol{\beta}}_k$ are correlated across k, but we cannot make inference about $(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_c)$ jointly.
2. We don't know the distribution of $h(\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_c)$.
3. It may be not efficient to ignore other categories when calculate $\hat{\boldsymbol{\beta}}_k$.

3.3 Multivariate GLMs

3.3.1 Model Settings

To overcome the above problems, we can generalize the univariate GLM to a multivariate GLM. The distribution of the multivariate exponential family is of the form

$$f(\boldsymbol{y}_i; \boldsymbol{\theta}_i) = e^{\boldsymbol{y}_i^\top \boldsymbol{\theta}_i - b(\boldsymbol{\theta}_i)} f_0(\boldsymbol{y}_i).$$

We know that the multinomial distribution is in the multivariate exponential family. So the multivariate GLM is given by

$$g(\boldsymbol{\mu}_i) = \tilde{\boldsymbol{x}}_i^\top \tilde{\boldsymbol{\beta}}$$

with $g_k(\boldsymbol{p}_i) = \log\left(\frac{p_{i_k}}{p_{i_c}}\right)$ for $k = 1, \ldots, c-1$ and

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i_1} \\ \vdots \\ y_{i_c} \end{bmatrix} \qquad \boldsymbol{\mu}_i = \boldsymbol{p}_i = \begin{bmatrix} p_{i_1} \\ \vdots \\ p_{i_c} \end{bmatrix} \qquad g(\boldsymbol{\mu}_i) = \begin{bmatrix} g_1(\boldsymbol{\mu}_i) \\ \vdots \\ g_c(\boldsymbol{\mu}_i) \end{bmatrix} \qquad \tilde{\boldsymbol{x}}_i = \begin{bmatrix} \boldsymbol{x}_i & & \\ & \ddots & \\ & & \boldsymbol{x}_i \end{bmatrix} \qquad \tilde{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_c \end{bmatrix}.$$

3.3.2 Estimation

We consider the simple case with the ungrouped data. The log-likelihood function is

$$L(\tilde{\boldsymbol{\beta}}; \boldsymbol{y}) = \log\left[\prod_{i=1}^{N}\left(\prod_{k=1}^{c} p_{i_k}^{y_{i_k}}\right)\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{N}\left[\left(\sum_{k=1}^{c-1} y_{i_k}\log p_{i_k}\right) + \left(1-\sum_{k=1}^{c-1} y_{i_k}\right)\log p_{i_c}\right] \\
&= \sum_{i=1}^{N}\left[\left(\sum_{k=1}^{c-1} y_{i_k}\log \frac{p_{i_k}}{p_{i_c}}\right) + \log p_{i_c}\right] \\
&= \sum_{i=1}^{N}\left[\left(\sum_{k=1}^{c-1} y_{i_k}\log \boldsymbol{x}_i^\top\boldsymbol{\beta}_k\right) - \log\left(1+\sum_{l=1}^{c-1} e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_l}\right)\right].
\end{aligned}
$$

The score functions are

$$
\frac{\partial L(\tilde{\boldsymbol{\beta}};\boldsymbol{y})}{\partial \beta_{k_j}} = \sum_{i=1}^{N}\left(y_{i_k}x_{ij} - \frac{e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_k}x_{ij}}{1+\sum_{l=1}^{c-1} e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}_l}}\right) = \sum_{i=1}^{N}(y_{i_j}-p_{i_j})x_{ij} = 0,
$$

which have the same forms as we saw before for canonical link. For computation, we can find that Fisher-scoring is the same as Newton's method.

3.4 Cumulative Models

3.4.1 Model Settings

For ordial data, we don't need different $\boldsymbol{\beta}_k$ for different categories. Instead, we are assuming they share the same $\boldsymbol{\beta}$. Also, the responses are the samples' orders.

A naive solution is to ignore the categorical nature of y_i, choose scores for different categories and fit a linear model. However, usually there is no clear-cut choice for the scores.

Another approach is to use a latent variable $y_i^* = -\boldsymbol{x}_i^\top\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\epsilon}_i \overset{iid}{\sim} F$. Suppose that there are some cutpoints $-\infty = \alpha_0 \le \cdots \le \alpha_c = \infty$ such that we observe $y_i = k$ if $\alpha_{k-1} < y_i^* \le \alpha_k$. Then, we have

$$
\mathbb{P}(y_i \le k) = \mathbb{P}(y_i^* \le \alpha_k) = F(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta}).
$$

When we take F as the cumulative distribution functon of standard logistic (or standard Gaussian) distribution, we get the *cumulative logit models* (or *cumulative probit models*).

We want the model to learn both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_c)^\top$. Therefore, $\boldsymbol{x}_i$ here does not include the intercept term for identifiability.

Another equivalent way to define the cumulative logit model is

$$
\text{logit}[\mathbb{P}(y_i \le k)] = \log\frac{p_{i_1}+\cdots+p_{i_k}}{p_{i_{k+1}}+\cdots+p_{i_c}} = \alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta}.
$$

3.4.2 Estimation

We consider the simple case with the ungrouped data. The log-likelihood function is given by

$$
L(\boldsymbol{\alpha},\boldsymbol{\beta}) = \log\left[\prod_{i=1}^{N}\prod_{j=1}^{c} p_{i_j}^{y_{i_j}}\right] = \sum_{i=1}^{N}\sum_{k=1}^{c} y_{i_k}\log[F(\alpha_k+\boldsymbol{x}_i^\top\boldsymbol{\beta}) - F(\alpha_{k-1}+\boldsymbol{x}_i^\top\boldsymbol{\beta})].
$$

The score functions are

$$\frac{\partial L(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N}\sum_{k=1}^{c} y_{i_k} x_{i_j} \frac{f(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta}) - f(\alpha_{k-1} + \boldsymbol{x}_i^\top\boldsymbol{\beta})}{F(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta}) - F(\alpha_{k-1} + \boldsymbol{x}_i^\top\boldsymbol{\beta})} = 0$$

$$\frac{\partial L(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \alpha_k} = \sum_{i=1}^{N}\left[\frac{y_{i_k} f(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta})}{F(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta}) - F(\alpha_{k-1} + \boldsymbol{x}_i^\top\boldsymbol{\beta})} - \frac{y_{i_{k+1}} f(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta})}{F(\alpha_{k+1} + \boldsymbol{x}_i^\top\boldsymbol{\beta}) - F(\alpha_k + \boldsymbol{x}_i^\top\boldsymbol{\beta})}\right] = 0$$

The computation of MLEs is complicated, but we can still use Fisher-scoring/Newton's method to solve it and we can still calculate the asymptotic variances of $\boldsymbol{\beta}$ and each α_k.

3.4.3 Comparison with OLS

Limitation of the cumulative link models:

- Settings are stochastically ordered. If $\boldsymbol{x}_i^\top\boldsymbol{\beta} \geq \boldsymbol{x}_j^\top\boldsymbol{\beta}$ for $i < j$ (or $i > j$) then we have $\mathbb{P}(y_i \leq k) \geq \mathbb{P}(y_j \leq k)$ for all k.

Disadvantages of modeling ordered categories using a lienar model:

- Usually no clear cut for the numerical scores.
- Linear model does not allow for the measurment error is discretization.
- From the linear model you can not get estimated probabilities of each category for a particular sample.
- Linear model ignores that the variability in each category can be different.

Chapter 4

Models For Count Data

4.1 Poisson GLMs

4.1.1 Model Settings

The density function of Poisson distribution is given by

$$f(y) = \frac{e^{-\mu}\mu^y}{y!} = e^{y\log\mu-\mu}\frac{1}{y!}, \qquad y = 0, 1, 2, \ldots,$$

which is in the one-parameter exponential family. And the canonical link is $\log(\mu_i) = \eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, which is called the *Poisson loglinear model*.

For a Poisson loglinear model, the mean satisfies the exponential relation

$$\mu_i = (e^{\beta_1})^{x_{i1}} \cdots (e^{\beta_p})^{x_{ip}}.$$

A 1-unit increase in x_{ij} has a multiplicative impact of e^{β_j}.

4.1.2 Estimation

For Poisson distribution, the variance ν_i is equal to the mean μ_i, i.e., $\nu_i = \mu_i = \mathrm{Var}(y_i)$.

The estimated variance of $\hat{\boldsymbol{\beta}}$ is given by $\mathrm{Var}(\hat{\beta}) = (\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X})$, where $\hat{W}_{ii} = \hat{\nu}_i = \hat{\mu}_i$.

4.1.3 Deviance Analysis

$$D_+(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left[y_i\log\left(\frac{y_i}{\mu_i}\right) - y_i + \mu_i\right]$$

4.1.4 Offset

Suppose that the count response y_i is proportional to an index t_i. One can model the rate $\frac{y_i}{t_i}$ by

$$\log\left(\frac{\mu_i}{t_i}\right) = \boldsymbol{x}_i^\top\boldsymbol{\beta},$$

i.e.,

$$\log(\mu_i) = \log(t_i) + \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where $\log(t_i)$ is called the *offset*. The fit corresponds to using $\log t_i$ as an explanatory variable in the linear predictor for $\log(\mu_i)$ and forcing its coefficient to equal 1.

4.1.5 Poisson Modeling for Contingency Tables

The One-Way Layout for Poisson Counts

For one-way layouts, we only have one count response, which is denoted by A. Suppose that $\{y_{ij}\}$ are independent counts having Poisson distributions with means $\{\mu_{ij}\}$ that satisfy $\mu_{ij} = \mu_i$.

A	1	2	$\cdots$	c
observations	$y_{1_1}, \ldots, y_{1_{n_1}}$	$y_{2_1}, \ldots, y_{2_{n_1}}$	$\cdots$	$y_{c_1}, \ldots, y_{c_{n_c}}$
number of observations	n_1	n_2	$\cdots$	n_c

Table 4.1: The one-way layout for Poisson counts.

The Poisson model is given by

$$\log(\mu_i) = \beta_0 + \beta_i.$$

Since this model is non-identifiable, we need extra constraint that $\beta_0 = 0$. Then the model can be written in vector form,

$$\log(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta},$$

where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \mathbf{1}_{n_1} \\ \vdots \\ \mu_c \mathbf{1}_{n_c} \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} \mathbf{1}_{n_1} & & \\ & \ddots & \\ & & \mathbf{1}_{n_c} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \mathbf{1}_{n_1} \\ \vdots \\ \beta_c \mathbf{1}_{n_c} \end{bmatrix}.$$

The Two-Way Layout for Poisson Counts

For two-way layouts, we consider $r \times c$ tables that cross-classify two categorical response variables, which we denote by A and B.

- Independence. Suppose that $\{y_{ij}\}$ are independent counts having Poisson distributions with means $\{\mu_{ij}\}$ that satisfy $\mu_{ij} = \mu\phi_i\psi_j$, where μ is a constant, $\{\phi_i\}$ and $\{\psi_j\}$ are positive constants satisfying $\sum_{i=1}^r \phi_i = \sum_{j=1}^l \psi_j = 1$. The GLM is given by

  $$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B.$$

  Identifiability requires a constraint on $\{\beta_i^A\}$ and on $\{\beta_j^B\}$.
  The MLEs are given by $\hat{\mu} = y_{++}$, $\hat{\phi}_i = \frac{y_{i+}}{y_{++}}$ and $\hat{\psi}_j = \frac{y_{+j}}{y_{++}}$, where $y_{++} = \sum_{i=1}^r \sum_{j=1}^l y_{ij}$, $y_{i+} = \sum_{j=1}^l y_{ij}$ and $y_{+j} = \sum_{i=1}^r y_{ij}$.

A \ B	1	$\cdots$	l
1	y_{00}	$\cdots$	y_{0l}
$\vdots$	$\vdots$	$\ddots$	$\vdots$
r	y_{r0}	$\cdots$	y_{rl}

Table 4.2: The two-way layout for Poisson counts.

- Dependence. When A and B is dependent, we add a two-factor interaction term to loglinear model

$$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B + \gamma_{ij}^{AB}.$$

We need identifiability conditions such as $\gamma_{i1} = \gamma_{1j}^A B = 0$ for all i, j. The association parameters γ_{ij}^{AB} are related to odds ratios.

Three-way Contingency Tables

We illustrate for $r \times c \times l$ cross-classifications of three categorical response variables, which we denote by A, B, and C. The models apply to Poisson sampling with independent cell counts $\{y_{ijk}\}$ having means $\{\mu_{ijk}\}$. They also apply to a multinomial distribution with cell probabilities $\{\pi_{ijk}\}$ having $\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\pi_{ijk}=1$.

- Mutual independence.

$$\mathbb{P}(A = i, B = j, C = k) = \mathbb{P}(A = i)\mathbb{P}(B = j)\mathbb{P}(C = k).$$

Equivalently, the loglinear form is

$$\log(\mu_{ijk}) = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C.$$

- Joint independence.

$$\mathbb{P}(A = i, B = j, C = k) = \mathbb{P}(A = i)\mathbb{P}(B = j, C = k).$$

Equivalently, the loglinear form is

$$\log(\mu_{ijk}) = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}.$$

- Conditional independence.

$$\mathbb{P}(A = i, B = j|C = k) = \mathbb{P}(A = i|C = k)\mathbb{P}(B = j|C = k).$$

Equivalently, the loglinear form is

$$\log(\mu_{ijk}) = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ij}^{AB}.$$

This model permits all three pairs of variables to be conditionally dependent.

- Saturated model.

$$\log(\mu_{ijk}) = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ij}^{AB} + \gamma_{ijk}^{ABC}.$$

4.1.6 Connections Between Poisson and Multinomial GLMs

Connections Between Poisson and Multinomial Distributions

For independent Poisson random variables $(y_1, \ldots, y_c)$ with means $(\mu_1, \ldots, \mu_c)$, the conditional probability on $\sum_{i=1}^c y_i = k$ is

$$\begin{aligned}
\mathbb{P}\left(y_1 = k_1, \cdots, y_c = k_c \middle| \sum_{i=1}^c y_i = k\right) &= \frac{\mathbb{P}(y_1 = k_1, \cdots, y_c = k_c)}{\mathbb{P}\left(\sum_{i=1}^c y_i = k\right)} \\
&= \frac{\prod_{i=1}^c \mathbb{P}(y_i = k_i)}{\mathbb{P}\left(\sum_{i=1}^c y_i = k\right)} \\
&= \frac{\prod_{i=1}^c e^{-\mu_i} \mu_i^{k_i} \frac{1}{k_i!}}{e^{-\sum_{j=1}^c \mu_j} \left(\sum_{j=1}^c \mu_j\right)^k \frac{1}{k!}} \\
&= \binom{k}{k_1 \cdots k_c} \prod_{i=1}^c \left(\frac{\mu_i}{\sum_{j=1}^c \mu_j}\right)^{k_i},
\end{aligned}$$

which follows a multinomial distribution with parameters k and $\boldsymbol{p}$, where $p_i = \frac{\mu_i}{\sum_{j=1}^c \mu_j}$.

Connections Between Logistic and Loglinear Models

Loglinear models for contingency tables treat all categorical classifications symmetrically and regard the cell count as the response. They are useful for modeling the joint distribution of categorical variables. By contrast, logistic models distinguish between response and explanatory classifications. Although different in purpose, the two types of models are connected.

We illustrate with the homogeneous association loglinear model. Suppose we treat A as a response variable and B and C as explanatory, conditioning on $\{n_{+jk}\}$. For the binary case $r = 2$, we are then modeling cl binomial distributions on A. When we construct the logit for each binomial distribution of A, we obtain

$$\begin{aligned}
\log \frac{\mathbb{P}(A = 1|B = j, C = k)}{\mathbb{P}(A = 2|B = j, C = k)} &= \log \frac{\mathbb{P}(A = 1, B = j, C = k)}{\mathbb{P}(A = 2, B = j, C = k)} \\
&= \log \mu_{1jk} - \log \mu_{2jk} \\
&= (\beta_1^A - \beta_2^A) + (\gamma_{1j}^{AB} - \gamma_{2j}^{AB}) + (\gamma_{1k}^{AC} - \gamma_{2k}^{AC})
\end{aligned}$$

which is equivalent to a binomial logistic model

$$\text{logit}[\mathbb{P}(A = 1|B = j, C = k)] = \lambda + \delta_j^B + \delta_k^C.$$

4.1.7 Overdispersion

For the Poisson distribution, the variance equals the mean. In practice, count observations often exhibit variability exceeding that predicted by the Poisson, which is called *overdispersion*.

If we view μ_i to be random and $y_i|\mu_i \sim \text{Poisson}(\mu_i)$, then $\mathbb{E}(y_i) = \mathbb{E}[\mathbb{E}(y_i|\mu_i)] = \mathbb{E}(\mu_i)$ and

$$\text{Var}(y_i) = \mathbb{E}[\text{Var}(y_i|\mu_i)] + \text{Var}[\mathbb{E}(y_i|\mu_i)] = \mathbb{E}(\mu_i) + \text{Var}(\mu_i) \geq \mathbb{E}(y_i),$$

which explains why we will have larger variance than mean in general when the count data follows some mixture of Poisson distributions.

4.2 Negative Binomial GLMs

4.2.1 Model Setting

To deal with overdispersion data, we can introduce the negative binomial distributions. Let $Y \sim \text{NB}(\mu, k)$, the density function for this negative binomial variable is given by

$$f(y; \mu, k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \qquad y = 0, 1, 2, \ldots,$$

where $\mu = \mathbb{E}(Y)$ and $r = \frac{1}{k}$ is the dispersion parameter. Also, we have $\text{Var}(y) = \mu + r\mu^2 > \mathbb{E}(Y)$.

With k fixed, this is a member of an exponential dispersion family appropriate for discrete variables, with natural parameter $\log\left(\frac{\mu}{\mu+k}\right)$. However, negative binomial GLMs commonly use the log link, as in Poisson loglinear models, rather than the canonical link. Also, k will be the same across all samples.

4.2.2 Connections Between Negative Binomial and Poisson GLMs

The negative binomial distribution can be viewed as a gamma mixture of the Poisson distributions, if we let $y|\lambda \sim \text{Poisson}(\lambda)$ and $\lambda \sim \Gamma(k, \mu)$ where μ is the mean and k is the shape parameter.

4.3 Zero-Inflated Poisson/Negative Binomial Models

4.3.1 Zero-Inflated Counts

For $y \sim \text{Poisson}(\mu)$, $\mathbb{P}(y = 0) = e^{-\mu}$. For $y \sim \text{NB}(\mu, k)$, $\mathbb{P}(y = 0) = \left(\frac{k}{\mu+k}\right)^k$. However, there may be more 0 counts than what these distributions can allow in practice. For example, what we observe is 0 for some probability and Poisson counts otherwise. This type of data is called *zero-inflated data*.

4.3.2 Model Settings

The zero-inflated Poisson (ZIP) model is given by

$$y_i \sim \begin{cases} 0 & , \text{ with probability } 1-\phi_i \\ \text{Poisson}(\mu_i) & , \text{ with probability } \phi_i \end{cases},$$

which can be decomposed as two parts,

$$\text{logit}(\phi_i) = \boldsymbol{x}_{1i}^\top \boldsymbol{\beta}_1, \qquad \log(\mu_i) = \boldsymbol{x}_{2i}^\top \boldsymbol{\beta}_1.$$

Here we allow the features to be differnt for two parts of the model.

By introducing a latent variable $z_i \sim \text{Binomial}(\phi_i)$, we have $\mathbb{E}(y_i) = \mathbb{E}[\mathbb{E}(y_i|z_i)] = \phi_i\mu_i$ and $\text{Var}(y_i) = \mathbb{E}[\text{Var}(y_i|z_i)] + \text{Var}[\mathbb{E}(y_i|z_i)] = \phi_i\mu_i[1-(1-\phi_i)\mu_i] > \mathbb{E}(y_i)$. So over-dispersion still occurs related to a Poisson model.

Similarly, the zero-inflated negative binomila (ZINB) model is given by

$$y_i \sim \begin{cases} 0 & , \text{ with probability } 1-\phi_i \\ \text{NB}(\mu_i, k) & , \text{ with probability } \phi_i \end{cases}.$$

Chapter 5

Quasi-Likelihood Methods

5.1 Quasi-Likelihoods

Overdispersion occurs for both Binomial and Poisson data. To deal with this problem, we can use Beta-Binomial and Negative Binomial distributions respectively. Otherwise, we can use quasi-likelihoods which aims to modify the theorectical vairance (under specific distribution assumptions) to adapt to the empirical variance.

Consider the score functions of the exponential family distributed data,

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\nu_i} \frac{1}{g'(\mu_i)} = 0, \quad j = 1, \ldots, p$$

which only involve μ_i and ν_i related to the distribution. Suppose that we are satisfied with the link functions, and thus satisfied with the forms of μ_i and $g'(\mu_i)$ as a function of $\boldsymbol{\beta}$. However, the form of ν_i does not fit the data as we see phenomenons like over-dispersion. For quasi-likelihood, we replace ν_i by some other mean-variance relationship $a(\mu_i, \phi)$ that typically involves another unknown dispersion parameter ϕ. Noted that we are only required to know the mean and variance of y_i when using the quasi-likelihoods.

Some common forms of $a(\mu_i, \phi)$:

- Proportional variance. $a(\mu_i, \phi) = \phi\nu(\mu_i)$.
- Grouped Binomial variance. $a(\mu_i, \phi) = \frac{1}{n_i}\phi\mu_i(n_i - \mu_i)$.
- Poisson variance. $a(\mu_i, \phi) = \mu_i + \phi\mu_i^2$.
- Correlated Bernoulli variance. $a(\mu_i, \phi) = \frac{1}{n_i}\mu_i(n_i - \mu_i)[1 + (n_i - 1)\phi]$.

5.2 Estimating Equations

The estimating equations are the quasi-score equations,

$$u_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_j} \cdot \frac{y_i - \mu_i}{\nu(\mu_i)} = 0, \quad j = 1, \ldots, n,$$

which reduces to be the score equations of GLMs when $\eta_i = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Notice that the quasi-score equations are more general than the score equations. Usually, we need extra assumptions to ensure that the solutions of the quasi-score equations have good properties.

5.2.1 Properties

If $\mathbb{E}[\boldsymbol{u}(\hat{\boldsymbol{\beta}}_0)] \to 0$ as $n \to \infty$ and $\boldsymbol{u}(\boldsymbol{\beta}) = 0$ has a unique solution, then

- (**Consistency**) For any ν, $\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_0$ as $n \to \infty$, as long as the specification is correct for the link function and linear predictor.
- (**Asymptotic Normality**) Notice that $\boldsymbol{u}(\boldsymbol{\beta_0})$ and its derivatives are functions of random vector $\boldsymbol{y}$.

  $$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$$

  where

  $$\boldsymbol{V} = \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^\top \frac{1}{\nu(\mu_i)} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right]^{-1}$$

  is the model-based covariance matrices.

5.2.2 Sandwich Covariance Adjustment for Variance Misspecification

Since

$$\boldsymbol{0} = \boldsymbol{u}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{u}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

we have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \boldsymbol{u}(\boldsymbol{\beta}),$$

and

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx \left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1} \mathrm{Var}[\boldsymbol{u}(\boldsymbol{\beta})] \left(\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^{-1}.$$

Since

$$\mathrm{Var}[\boldsymbol{u}(\boldsymbol{\beta})] = \mathrm{Var}\left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^\top \frac{y_i - \mu_i}{\nu(\mu_i)}\right] = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^\top \frac{\mathrm{Var}(y_i)}{[\nu(\mu_i)]^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)$$

and $-\mathbb{E}\left[\frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = \mathbb{E}\left[\frac{\partial^2 \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] \xrightarrow{\mathbb{P}} \boldsymbol{V}$ when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, we have

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{V}\left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^\top \frac{\mathrm{Var}(y_i)}{[\nu(\mu_i)]^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)\right] \boldsymbol{V},$$

which simplifies to $\boldsymbol{V}$ if $\mathrm{Var}(y_i) = \nu(\mu_i)$. It is called a *sandwich estimator*, because the empirical evidence is sandwiched between the model-based covariance matrices.

In practice, not knowing the true variance function, we obtain a robust estimator of this actual asymptotic variance by replacing $\mathrm{Var}(y_i)$ by $(y_i - \overline{y}_n)^2$.

Chapter 6

Generalized Linear Mixed Models

6.1 Introduction

For a d-dimensional response variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)$. Each subject has a cluster of d observations. Often d varies by cluster, such as when some subjects drop out of the study and are missing some observations. For multivariate data, observations within a cluster are typically correlated, and models need to account for that correlation. There are two primary types of models for multivariate responses.

1. A marginal model, which simultaneously models only each marginal distribution, but takes into account the correlation structure in finding valid standard errors.
2. Generalized linear mixed model (GLMM), which includes random effects in addition to the usual fixed effects.

6.1.1 Marginal Models

A marginal model for $\boldsymbol{y}_i$, with link function g, has the form

$$g[\mathbb{E}(y_{ij})] = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta},$$

where $y_{ij} \in \mathbb{R}$, $\boldsymbol{x}_{ij}, \boldsymbol{\beta} \in \mathbb{R}^d$. A marginal model has the usual GLM structure for each component in the multivariate vector. To complete the model, we assume a parametric joint distribution for $\boldsymbol{y}_i$. Then, with independent observations for $i = 1, \ldots, n$, we can fit the model by maximum likelihood.

6.1.2 Generalized Linear Mixed Models

A generalized linear mixed model (GLMM) for $\boldsymbol{y}_i$ has the form

$$g[\mathbb{E}(y_{ij}|\boldsymbol{u}_i)] = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i,$$

for $i = 1, \ldots, n$, $j = 1, \ldots, d_i$, where $y_{ij} \in \mathbb{R}$, $\boldsymbol{x}_{ij}, \boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{z}_{ij}, \boldsymbol{u}_i \in \mathbb{R}^q$. Here $\boldsymbol{z}_{ij}$'s are the known explanatory variables. The random effects $\{\boldsymbol{u}_i\}$ are usually assumed to be independent from a

$\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$ distribution specified by unknown variance and correlation parameters. The adjective mixed in generalized linear mixed model refers to the presence of both fixed effects $\boldsymbol{\beta}$ and random effects $\boldsymbol{u}_i$ in the linear predictor.

6.1.3 Relationship between Marginal Models and GLMMs

For the GLMM, by inverting the link function,

$$\mathbb{E}(y_{ij}|\boldsymbol{u}_i) = g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i).$$

Marginally, averaging over the random effects, the mean is

$$\mu_{ij} = \int_{\mathbb{R}^q} g^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i) f(\boldsymbol{u}_i; \boldsymbol{\Sigma}_u) \mathrm{d}\boldsymbol{u}_i.$$

For the identity link function, this reduces to $\mu_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}$. So GLMMs imply marginal models when g is the identity link.

The GLMM approach is preferable when we want to estimate cluster-specific effects, estimate their variability, specify a mechanism for generating nonnegative association among clustered observations, or model the joint distribution. When between-cluster effects are the main focus, it can be simpler to model them directly using marginal models.

6.2 Binomial GLMMs

6.2.1 Model Settings

Similar to binomial GLMs, we have mainly two types of binomial GLMMs:

1. Logistic-Normal binomial GLMMs.

$$\text{logit}(\mathbb{P}(y_{ij}|\boldsymbol{u}_i)) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i,$$

   where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

2. Probit-Normal binomial GLMMs.

$$\Phi^{-1}(\mathbb{P}(y_{ij}|\boldsymbol{u}_i)) = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i,$$

   where $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u)$.

6.2.2 Latent Variable Threshold Model

From the latent variable threshold modeling prospective, we assume there is a latent y_{ij}^* where

$$y_{ij}^* = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i - \epsilon_{ij},$$

where $\epsilon_{ij} \overset{iid}{\sim} F$ and F is a cumulative distribution function (of normal, Logistic, etc). We have

$$y_{ij} = \begin{cases} 1 & , \ y_{ij}^* \geq 0 \\ 0 & , \ y_{ij}^* < 0 \end{cases}.$$

6.2.3 Properties

Conditional Independence

$$\mathbb{P}(y_{i1} = t_1, \ldots, y_{id_i} = t_{d_i} | \boldsymbol{u}_i) = \mathbb{P}(y_{i1} = t_1) \cdots \mathbb{P}(y_{id_i} = t_{d_i} | \boldsymbol{u}_i)$$

Marginal Correlation

$$\begin{aligned} \text{Cor}(y_{ij}, y_{ik}) &= \mathbb{E}[\text{Cov}(y_{ij}, y_{ik} | \boldsymbol{u}_i)] + \text{Cov}[\mathbb{E}(y_{ij} | \boldsymbol{u}_i)\mathbb{E}(y_{ik} | \boldsymbol{u}_i)] \\ &= \text{Cov}[F(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i), F(\boldsymbol{x}_{ik}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ik}^\top \boldsymbol{u}_i)] \end{aligned}$$

If $\boldsymbol{z}_{ik} = 1$, then $\text{Cor}(y_{ij}, y_{ik}) > 0$.

6.2.4 Relationship Between Binomial GLMMs and GLMs

Marginally, $\mathbb{P}(y_{ij} = 1) \neq F(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})$. However, we have the following relationship. Consider the case when $\boldsymbol{z}_{ij} = 1$. For probit-normal binomial GLMMs, marginally we have

$$\begin{aligned} \mathbb{P}(y_{is} = 1) &= \mathbb{E}[\mathbb{E}(\mathbb{1}_{\{\epsilon_{ij} - u_i \leq \boldsymbol{x}_i^\top \boldsymbol{\beta}\}} | u_i)] \\ &= \mathbb{P}(\epsilon_{ij} - u_i \leq \boldsymbol{x}_i^\top \boldsymbol{\beta}) \\ &= \Phi\left(\frac{\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}}{\sqrt{1 + \sigma_u^2}}\right) \end{aligned}$$

since $\epsilon_{ij} + u_i \sim \mathcal{N}(0, 1 + \sigma_u^2)$. Also notice that

$$\mathbb{P}(y_{ij} = 1 | u_i) = \Phi(\boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i)$$

and u_i has mean zero. So the coefficient of the probit-link binomial GLM will be $\tilde{\boldsymbol{\beta}} = \frac{\boldsymbol{\beta}}{\sqrt{1+\sigma_u^2}}$, which has smaller absolute magnitude than the coefficient $\boldsymbol{\beta}$ of probit-normal binomial GLMM.

Similar results hold for Logistic-normal binomial GLMMs. Let $F(x) = \frac{e^x}{1+e^x}$ be the Logistic function. Notice that $F(x) \approx 1.7\Phi(x) = \Phi\left(\frac{x}{\sqrt{1.7}}\right)$, for the Logistic-normal binomial GLMM,

$$\begin{aligned} \mathbb{P}(y_{is} = 1) &= \mathbb{E}[\mathbb{E}(\mathbb{1}_{\{\epsilon_{is} - u_i \leq \boldsymbol{x}_{is}^\top \boldsymbol{\beta}\}} | u_i)] \\ &= \mathbb{P}(\epsilon_{is} - u_i \leq \boldsymbol{x}_{is}^\top \boldsymbol{\beta}) \\ &\approx \Phi\left(\frac{\boldsymbol{x}_{is}^\top \boldsymbol{\beta}}{\sqrt{1.7 + \sigma_u^2}}\right) \end{aligned}$$

$$\mathbb{P}(y_{is} = 1|u_i) = F(\boldsymbol{x}_{is}^\top \boldsymbol{\beta} + u_i) \approx \Phi\left(\frac{\boldsymbol{x}_{is}^\top \boldsymbol{\beta}}{\sqrt{1.7}}\right).$$

While the first one represents the ordinary logistic GLM, the coefficients in GLM should be $\frac{\sqrt{1.7}\boldsymbol{\beta}}{\sqrt{1.7+\sigma_u^2}}$. So the GLM shrinks the coefficients compared to the GLMM with extra random effects.

6.3 Poisson GLMMs

The log-link Poisson GLMM is given by

$$\log[\mathbb{E}(y_{ij}|\boldsymbol{u}_i)] = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i.$$

6.4 Normal Linear Mixed Models

A special case of GLMMs is the normal liner mixed models, where g is the identity link. The model is given by

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i + \epsilon_{ij}, \quad i =, \ldots, b, \quad j = 1, \ldots, d,$$

where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\boldsymbol{u}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_u^2)$, and $\{\boldsymbol{u}_i\}$ and $\{\epsilon_{ij}\}$ are independent of each other. The matrix form of the model is given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{Z}_i \boldsymbol{u}_i + \boldsymbol{\epsilon}_i,$$

where

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{id} \end{bmatrix}, \qquad \boldsymbol{X}_i = \begin{bmatrix} \boldsymbol{x}_{i1}^\top \\ \vdots \\ \boldsymbol{x}_{id}^\top \end{bmatrix}, \qquad \boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{z}_{i1}^\top \\ \vdots \\ \boldsymbol{z}_{id}^\top \end{bmatrix}, \qquad \boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_d \end{bmatrix}.$$

When ϵ_{ij}'s are independent across $j = 1, \ldots, d$, then $\boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_d)$. More generally to allow correlated errors, we have $\boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$.

6.4.1 Random Intercept Models

An important special case of a linear mixed model has $\boldsymbol{u}_i = u_i \mathbf{1}$, $\boldsymbol{Z}_i = \frac{1}{q}\mathbf{1}\mathbf{1}^\top$, and $\mathrm{Var}(\boldsymbol{u}_i) = \sigma_u^2$, that is,

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + u_i \mathbf{1} + \boldsymbol{\epsilon}_i.$$

For this random-intercept model, marginally

$$\mathrm{Var}(\boldsymbol{y}_i) = \sigma_u^2 \mathbf{1}\mathbf{1}^\top + \sigma_\epsilon^2 \boldsymbol{I}.$$

The two variances in this expression are referred to as variance components.

This model has the exchangeable correlation structure, for $j \neq k$,

$$\text{Corr}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2},$$

called *compound symmetry*. Greater within-cluster correlation occurs as σ_u^2 increases.

6.4.2 Multilevel Models

Let y_{ijk} denote the kth response among samples with i at the level 1 and value j at the level 2. A multilevel model with fixed effects $\boldsymbol{\beta}$ for explanatory variables and random effects $\{u_i \in \mathbb{R}\}$ for level 1 and $\{v_{ij} \in \mathbb{R}\}$ for level 2 has the form

$$y_{ijk} = \boldsymbol{x}_{ijk}^\top \boldsymbol{\beta} + u_i + v_{ij} + \epsilon_{ijk}.$$

We assume that the random effects u_i and v_{ij} and the errors ϵ_{ijk} are independent with distributions $\mathcal{N}(0, \sigma_u^2)$, $\mathcal{N}(0, \sigma_v^2)$, and $\mathcal{N}(0, \sigma_\epsilon^2)$ having unknown variances.

The intraclass correlation for different samples at the same levels is given by

$$\text{Corr}(y_{ijk_1}, y_{ijk_2}) = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\epsilon^2}$$

for any $k_1 \neq k_2$.

The intraclass correlation at the same level 1 is given by

$$\text{Corr}(y_{ij_1k_1}, y_{ij_2k_2}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\epsilon^2}$$

for all $j_1 \neq j_2$, while k_1 and k_2 can be the same.

6.5 Estimation of LMMs

6.5.1 MLE of β

We express the linear mixed model simultaneously for all n observations as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon},$$

where

$$\boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \in \mathbb{R}^{nd},\ \boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{bmatrix} \in \mathbb{R}^{nd \times p},\ \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & \\ & \ddots & \\ & & \boldsymbol{Z}_n \end{bmatrix} \in \mathbb{R}^{nd \times nq},\ \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_n \end{bmatrix} \in \mathbb{R}^{nq},\ \boldsymbol{\epsilon} = \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{bmatrix} \in \mathbb{R}^{nd},$$

and $\boldsymbol{\beta} \in \mathbb{R}^p$. For the random components, we have $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_u)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_\epsilon)$ where

$$\tilde{\boldsymbol{\Sigma}}_u = \begin{bmatrix} \boldsymbol{\Sigma}_u & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_u \end{bmatrix}, \qquad \tilde{\boldsymbol{\Sigma}}_\epsilon = \begin{bmatrix} \boldsymbol{\Sigma}_\epsilon & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}_\epsilon \end{bmatrix}.$$

Marginally, $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top + \tilde{\boldsymbol{\Sigma}}_\epsilon)$. Let $\boldsymbol{V} = \boldsymbol{Z}\tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top + \tilde{\boldsymbol{\Sigma}}_\epsilon$, then the log-likelihood function for the model is

$$l(\boldsymbol{\beta}, \boldsymbol{V}) = -\frac{1}{2}|V| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

If V is known,then maximize $l(\boldsymbol{\beta}, \boldsymbol{V})$ with respect to $\boldsymbol{\beta}$ yields the generalize least squares solution

$$\hat{\boldsymbol{\beta}}(\boldsymbol{V}) = (\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{y} = \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \sim \mathcal{N}\left(\boldsymbol{\beta}, \left(\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1}\right),$$

since $\boldsymbol{V}$ has block-diagonal form with $d \times d$ blocks $\boldsymbol{V}_i = \boldsymbol{Z}_i\tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}_i + \tilde{\boldsymbol{\Sigma}}_\epsilon$.

However, if $\boldsymbol{V}$ is unknown, we can estimate it by the method given in the section to obtain $\hat{\boldsymbol{V}}$ and then estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{V}})$. Under regularity conditions, its asymptotic distribution is the same as the normal distribution that applies when $\boldsymbol{V}$ is known. Inference about fixed effects can use the usual methods, such as likelihood-ratio tests and Wald confidence intervals.

6.5.2 Best Linear Unbiased Prediction (BLUP) of Random Effects

After we have obtained $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{V}}$, we can predict values of the random effects $\{\boldsymbol{u}_i\}$. The *best linear unbiased prediction (BLUP)* of random effects $\boldsymbol{u}_i$ is a random vector $\hat{\boldsymbol{u}}_i$ such that $\mathbb{E}(\hat{\boldsymbol{u}}_i) = \mathbb{E}(\boldsymbol{u}_i) = 0$ and for any linear combination $\boldsymbol{a}^\top \boldsymbol{u}_i$, $\mathbb{E}[(\boldsymbol{a}^\top \hat{\boldsymbol{u}}_i - \boldsymbol{a}^\top \boldsymbol{u}_i)^2]$ is minimized among all such linear unbiased predictors.

The joint distribution of $(\boldsymbol{y}, \boldsymbol{u})^\top$ is given by

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{Z}\tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top + \tilde{\boldsymbol{\Sigma}}_\epsilon & \boldsymbol{Z}\tilde{\boldsymbol{\Sigma}}_u \\ \tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top & \tilde{\Sigma}_u \end{bmatrix}\right).$$

We differentiate the log-density with respect to $\boldsymbol{\beta}$ and $\boldsymbol{u}$ to obtain *Henderson's mixed-model equations*,

$$\begin{bmatrix} \boldsymbol{X}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{X} & \boldsymbol{X}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{X} & \tilde{\boldsymbol{\Sigma}}_u^{-1} + \boldsymbol{Z}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{Z} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{y} \\ \boldsymbol{Z}^\top \tilde{\boldsymbol{\Sigma}}_\epsilon^{-1} \boldsymbol{y} \end{bmatrix}.$$

The solution $\hat{\boldsymbol{\beta}}$ is identical to the generalized least squares solution. The solution $\hat{\boldsymbol{u}}$ is the BLUP of $\mathbb{E}(\boldsymbol{u}|\boldsymbol{y})$. Since from the conditional normal distribution $\mathbb{E}(\boldsymbol{u}|\boldsymbol{y}) = \tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$, the estimate of $\boldsymbol{u}$ is

$$\hat{\boldsymbol{u}} = \tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \tilde{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^\top \boldsymbol{V}^{-1}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}]\boldsymbol{y}$$

when $\boldsymbol{V}$ is known. The prediction $\hat{\boldsymbol{u}}$ is a weighted combination of $\mathbf{0}$ and the generalized least squares estimate based on treating $\boldsymbol{u}$ as a fixed effect. In practice, we use estimated variances of $\tilde{\Sigma}_{\boldsymbol{\epsilon}}$ and $\tilde{\Sigma}_{\boldsymbol{u}}$, which gives an *empirical BLUP*.

6.5.3 Residual Maximum Likelihood (REML)

To estimate the covariance matrices and variance components of random effects, we use *residual maximum likelihood (REML)*. Let $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$, then $\boldsymbol{LX} = \mathbf{0}$ and $\boldsymbol{Ly} = \boldsymbol{L}(\boldsymbol{Zu} + \boldsymbol{\epsilon}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{LVL}^\top)$. Since the likelihood of $\boldsymbol{Ly}$ does not involve $\boldsymbol{\beta}$ and we can maximize this likelihood to find the estimate of $\boldsymbol{V}$.

Chapter 7

Survival Analysis

7.1 Introduction

Elements	Continuous Time	Discrete Time
survival time	$T \in \mathbb{R}^+$ with density $f(t)$	$T \in \mathbb{N}$ with mass function $f_t = \mathbb{P}(T = t)$
survival function/curve	$S(t) = \int_t^\infty f(t)\mathrm{d}t$	$S_t = \sum_{j \geq t} f_j$
hazard rate/function	$h(t) = \frac{f(t)}{S(t)}$	$h_t = \frac{f_t}{S_t}$
accumulative hazard rate/function	$H(t) = \int_0^t h(t)\mathrm{d}t$	$H_t = \sum_{j \leq t} h_t$

Table 7.1: Basic concepts of survival analysis.

An important fact is that knowing one of the three functions of $H(t)$, $h(t)$ and $S(t)$ will enable inferring the other two functions:

$$h(t) = -\frac{\partial \log S(t)}{\partial t}, \qquad H(t) = -\log[S(t)], \qquad S(t) = e^{-H(t)},$$
$$S_t = \prod_{j=1}^{t-1}(1 - h_j)$$

Similar results apply to the case of discrete time.

For n samples, denote their survival time as $T_1, T_2, \ldots, T_n$. However, we may not be able to observe every T_i's. Let C_i denote the censoring time for the ith subject. Then the event indicator $d_i = \begin{cases} 1 & , \text{ if } T_i \leq C_i \\ 0 & , \text{ if } T_i > C_i \end{cases}$ indicates whether T_i is observed or not. Finally, the observed response is $Y_i = \min\{T_i, C_i\}$. When each sample also has its covariate, what we observe can be denoted as $(Y_i, \boldsymbol{X}_i, d_i)$ for $i = 1, 2, \ldots, n$.

Here, we only consider *non-informative censoring*, i.e. $T_i \perp C_i | \boldsymbol{X}_i$, which means that the censoring time is not associated with the survival time, at least conditioning on other known covariates $\boldsymbol{X}_i$.

7.2 Estimating The Survival Functions without Covariates

7.2.1 Non-parametric approach

Empirical Estimates When No Censoring

We consider an intercept model where there is no $\boldsymbol{X}_i$.

When there is no censoring, we can estimate $S(t)$ by the empirical distribution function $\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{T_i \geq t\}}$. Note that $\mathbb{1}_{\{T_i \geq t\}} \sim \text{Bernoulli}(S(t))$.

Properties of $\hat{S}(t)$:

(1) (Consistency) $\hat{S}(t) \xrightarrow{\mathbb{P}} S(t)$.

(2) $\sqrt{n}(\hat{S}(t) - S(t)) \xrightarrow{D} \mathcal{N}(0, S(t)[1 - S(t)])$.

However, when there is censoring this method does not work.

Kaplan-Meier Estimator

For the discrete survival time, we can divide the survival time into bins. For each bin k, assume we observe n_k samples that are still alive at the beginning of this time bin, y_k death during this time bin and l_k drop-outs at the end of this time bin. Then, as the n_k samples are i.i.d. at this time point, we have $y_k \sim \text{Bernoulli}(n_k, h_k)$. Then an unbiased estimator of h_i is

$$\hat{h}_t = \frac{y_t}{n_t}.$$

Therefore,

$$\hat{S}_t = \prod_{j \leq t-1} (1 - \hat{h}_t) = \prod_{j \leq t-1} \frac{n_j - y_j}{n_j}.$$

For continuous survival time, the bin can be smaller and smaller, and we get the Kaplan-Meier estimator as

$$\hat{S}(t) = \prod_{j:\tau_j \leq t} \frac{n_j - y_j}{n_j},$$

where $\{\tau_1, \tau_2, \ldots, \tau_K\}$ is the set of K observed death times(i.e. the sorted distinct values of $\{Y_i : d_i = 1, i = 1, \ldots, n\}$), y_j is the number of death during τ_{j-1} and τ_j, and n_j is the total number of people who are at risk right before τ_j.

7.2.2 Parametric Models

We can assume T follows some parametric distribution, e.g.

- Exponential distribution $f(t) = \lambda e^{-\lambda t}$ for $\lambda > 0$. Then the survival function is $S(t) = e^{-\lambda t}$ and the hazard rate is $h(t) = \lambda$.
- Weibull distribution $f(t) = \kappa\lambda t^{\kappa-1} e^{-\lambda t^\kappa}$ for $\lambda, \kappa > 0$. Then the survival function is $S(t) = e^{-\lambda t^\kappa}$ and the hazard rate is $h(t) = \kappa\lambda t^{\kappa-1}$. When $\kappa = 1$, the Weibull distribution reduces to be the exponential distribution. When $\kappa > 1$ the hazard rate increases as t increases. When $\kappa < 1$ the hazard rate decreases as t increases.

To decide which distribution to use, we can use likelihood ratio tests or visualization of the Kaplan-Meier curve.

Maximum Likelihood Estimates

To constructing the likelihood with censoring,

- If $d_i = 1$, then $T_i = y_i$, the likelihood for the i-th sample is $L_i = f(y_i) = S(y_i)h(y_i)$.
- If $d_i = 0$, then $T_i \geq y_i$, the likelihood for the i-th sample is $L_i = S(y_i)$.

Thus the total likelihood is

$$L = \prod_{i=1}^{n} L_i = \prod_{i=1}^{n} S(y_i)h(y_i)^{d_i}.$$

Specifically

- If $T_1, \ldots, T_n \overset{iid}{\sim} \text{Exp}(\lambda)$, then $L(\lambda) = \prod_{i=1}^n e^{-\lambda y_i} \lambda^{d_i}$ and $l(\lambda) = -\lambda \sum_{i=1}^n y_i + \log\lambda \sum_{i=1}^n d_i$. The MLE is $\hat{\lambda} = \dfrac{\sum_{i=1}^n d_i}{\sum_{i=1}^n y_i}$.
- If $T_1, \ldots, T_n \overset{iid}{\sim} \text{Weibull}(\lambda, \kappa)$, then $L(\lambda, \kappa) = \prod_{i=1}^n e^{-\lambda y_i^\kappa} (\kappa\lambda y_i^{\kappa-1})^{d_i}$ and $l(\lambda) = -\lambda \sum_{i=1}^n y_i^\kappa + (\log\kappa + \log\lambda) \sum_{i=1}^n d_i + (\kappa - 1) \sum_{i=1}^n d_i \log y_i$.

7.3 Proportional Hazards Regression Model

7.4 Inference

7.4.1 Log-Rank Test

To compare two survival distributions, the hypotheses are $H_0 : S_{1,t} = S_{2,t}$ and $H_1 : S_{1,t} \neq S_{2,t}$. We can use a nonparametric *log-rank test*. First we divide survival time into bins. For bin i, we will have data like Table 7.2.

If the margins of this table are considered fixed, then $y_{i1} \overset{H_0}{\sim} \text{Hypergeometric}(n_i, n_{i1}, y_i)$, with probability mass function

$$\mathbb{P}(y_{i1} = k) = \frac{\binom{n_{i1}}{k}\binom{n_{i2}}{y_i - k}}{\binom{n_i}{y_i}},$$

Group	**Death**	**Alive**	**Total at Risk**
1	y_{i1}	$n_{i1} - y_{i1}$	n_{i1}
2	y_{i2}	$n_{i2} - y_{i2}$	n_{i2}
Total	y_i	$n_i - y_i$	n_i

Table 7.2: Summary table in bin i.

and mean and variance

$$\mathbb{E}(y_{i1}) = y_i \frac{n_{i1}}{n_i}, \qquad \text{Var}(y_{i1}) = \frac{n_{i1} n_{i2} y_i (n_i - y_i)}{n_i^2 (n_i - 1)}.$$

The Cochran-Mantel-Haenszel log-rank test statistics is

$$X^2_{CMH} = \frac{\left[\sum_{i=1}^{n} \left(y_{i1} - \frac{n_{i1}}{n_i} y_i\right)\right]^2}{\sum_{i=1}^{n} \frac{n_{i1} n_{i2} y_i (n_i - y_i)}{n_i^2 (n_i - 1)}} \overset{H_0}{\sim} \chi^2_1.$$

For continuous survival time, we want to test $H_0 : S_1(t) = S_2(t)$ and $H_1 : S_1(t) \neq S_2(t)$. Similarly, by dividing survival time into bins, the Cochran-MantelHaenszel log-rank test statistics is

$$X^2_{CMH} = \frac{\left[\sum_{i=1}^{K} \left(y_{i1} - \frac{n_{i1}}{n_i} y_i\right)\right]^2}{\sum_{i=1}^{K} \frac{n_{i1} n_{i2} y_i (n_i - y_i)}{n_i^2 (n_i - 1)}} \overset{H_0}{\sim} \chi^2_1,$$

where the bins are separated by $\{\tau_1, \tau_2, \ldots, \tau_K\}$ the set of K observed death times.